Utrecht
University

MASTER THESIS

Machine Learning Many-Body Interactions for Charged Colloidal Suspensions

Author :

Floris VAN DEN BOSCH

Supervisors :

Prof. Dr. René VAN ROIJ
Institute for Theoretical Physics

Prof. Dr. Ir. Marjolein DIJKSTRA
Debye Institute for Nanomaterials Science

Daily Supervisors :

Dr. Gerardo CAMPOS-VILLALOBOS
Debye Institute for Nanomaterials Science

Dr. Giuliana GIUNTA
Debye Institute for Nanomaterials Science

August 5, 2022

Abstract

Suspensions of charged colloidal particles are complex systems of mesoscopic colloidal particles in an electrolyte solution composed of microscopic water molecules and ions. An accurate description of these systems is vital for predicting their properties and the stability of the suspension. Direct computer simulations can provide us with accurate predictions, but are too computationally expensive for large systems due to the huge amount of microscopic water molecules and ions. Direct simulation of the entire system is also inefficient, as our interest lies primarily in the behavior of the mesoscopic colloids. We employ machine learning methods to learn effective many-body interactions for a colloid-only system from direct simulations of the full system. The accuracy of this machine learning approach is compared to established effective interactions such as Poisson-Boltzmann and DLVO theory, and its efficiency against the direct simulation of the full colloidal suspension.

Acknowledgements

Before presenting my thesis to you, I would like to thank all the people that helped me reach this point.

First of all, I would like to thank my supervisors Prof. René van Roij and Prof. Marjolein Dijkstra for all the guidance, helpful discussion and support during my thesis. Additionally, I would also like to thank Dr. Gerardo Campos Villalobos and Dr. Giuliana Giunta for the hands-on advice and feedback on my project. In extension, I want to thank Sebastian Bielfeldt and Tim Veenstra for joining the weekly supervisor meetings, the occasional collaboration and the general company you provided.

Beyond these people that were closely involved in my thesis, I want to express my gratitude to all my family and friends that supported me during this fun but, at times, challenging endeavor!

Contents

Chapter 1

Introduction

Colloidal suspensions are systems of mesoscopic colloidal particles suspended in a solvent. The typical size of these colloids is of the order of 10-1000 nm, meaning that they are orders of magnitude larger than the solvent molecules, but still small enough to display Brownian motion [1, 2] due to thermal fluctuation of the solvent. The magnitude of these thermal forces at these particle scales is large enough to overcome gravitational effects on individual colloids, which would result in sedimentation. Ordinarily, the free colloids in a solvent tend to aggregate over time due to attractive interactions, such as the Van der Waals interaction [3, 4]. These systems are called unstable as the aggregation increases the effective particle sizes that shift the balance between the thermal forces and sedimentation forces towards the latter. This causes the colloids to eventually precipitate out of the solvent. To stabilize a colloidal suspension, one thus needs a longer-range repulsive interaction between the colloids to overcome the attractive interactions. Examples of such repulsive interactions are polymer chain interactions and electric double-layer forces corresponding to so-called sterically [5] and charge [6] stabilized colloidal suspensions, respectively.

The charge of the colloids originates from ionizable chemical groups on the surface of the colloids [7]. When submerged in a polar liquid, these groups release ions into the solvent, while the colloid itself gains an equal and opposite surface charge. The ions released by the surface groups of the colloids are called counterions, as they have an opposite charge to the colloids. All salt ions that also have an opposite charge to the colloids are called counterions as well. Any ions with the same polarity as the colloids are called coions. The balance between the thermal motion, the dispersion of the counterions, and the Coulomb attraction to the colloids results in a diffuse layer of counterions around the colloids, which we call an electric double-layer. These double-layers give rise to effective repulsive interactions between the colloids that can counteract attractive effects, such as the Van der Waals interaction. This allows these colloidal particles to remain in suspension where they would otherwise tend to coagulate and precipitate out of solution.

Already more than 70 years ago Derjaguin, Landau, Verwey, and Overbeek [8, 9] showed that one can treat many charge-stabilized colloidal suspensions using a screened Coulomb or otherwise known as a Yukawa pair potential with an approximate range equal to the Debye screening length κ^{-1}. This is the well-known DLVO theory that is one of the cornerstones of colloid science. Such a system is often denoted as a Yukawa system and its electrostatic interactions are purely repulsive. This approach works well for systems with low colloid density and high salt concentration, corresponding to a shorter screening length (as we will see in subsection 2.2.1). The application of this model can be extended to higher densities and concentrations by renormalizing the charges of the colloids and the screening length of the medium, as effective colloid charges of dense systems become lower [10].

At the end of the last century, however, there were several pieces of experimental evidence that the DLVO theory does not hold up well in various cases. For high-density systems, the measured phase diagram does not match up to the predicted phase diagram of a corresponding Yukawa system [11, 12]. Furthermore, there have been additional experimental observations of clustering [13], gas-liquid phase transition [14, 15] which a Yukawa system does not show, and also computer simulations [16, 17] implying the existence of attractive electrostatic forces. These findings all contradict the DLVO theory, whose electrostatic force is purely repulsive.

Efforts to correct or extend the DLVO theory mainly fall into two categories: (i) incorporating ion-ion correlations into the pair potential, and (ii) considering many-body interactions between the colloids. Many-body terms can be accounted for by including an additional volume term [18–21] that is density-dependent and takes into account the relative proximity of the particles, or by explicitly calculating higher-order corrections to the Yukawa pair potential [22]. Many-body interactions between charged colloids have also been experimentally measured [23].

Many-body interactions become relevant when inter-particle distances become comparable to or smaller than the screening length of the solution. In this regime, we expect overlaps between multiple double-layers to become relevant, which results in higher-order corrections to the Yukawa pair interaction. Such overlaps are, for example, also important in colloid-polymer mixtures, where the excluded volume of the colloids results in an entropic attractive force between the colloids. This excluded volume is dependent on the overlaps of the exclusion zones of all the colloidal particles. Recently, machine learning methods have been used to accurately determine the many-body terms for the excluded volume [24]. Such a machine learning approach may also be applied to a charged colloidal suspension to determine the many-body electrostatic interaction.

Simulations of charged colloidal suspensions usually make use of the primitive model, where the solvent is treated as a continuous medium and colloids and ions are treated as hard-sphere particles with long-range Coulomb interactions. The challenge in the simulation of these systems is twofold: (i) the number of particles increases rapidly with the charge asymmetry between colloid and ion, and (ii) the different components of the system have very different relevant time- and length scales. An increase in particle number due to charge asymmetry increases the computation time due to the sheer increase of particles to loop over. Due to the nature of the electric double-layers, one also expects an increase in counterion buildup near the colloids which also increases the number of interactions between particles to consider, which has a large impact on computational times. As we are mostly interested in the properties of the colloidal particles, it is preferable to run simulations on time and length scale appropriate to the mesoscopic colloids. Explicit simulation of the microscopic ions, however, requires that we run the simulation in their, much smaller, time and length scales. This discrepancy requires much longer simulation times to obtain results relevant for the colloidal particles.

In this thesis, we aim to speed up simulations of charged colloidal suspensions by integrating out the ions in the solvent, which would drastically reduce the total number of particles in the system and allow us to simulate on time scales more relevant to the colloidal particles. In contrast to the theoretical approaches based on DLVO theory, we will make use of empirical machine learning methods. These methods have already been proven to work effectively in colloid-polymer mixtures [24] and we will now test their use on the description of charged colloidal suspensions.

The remainder of this thesis is structured as follows. In chapter 2 we will give a brief overview of the theoretical background on charged colloidal suspensions and the theories to describe them. In chapter 3 we will discuss the methods used to construct effective interactions for charged colloidal suspension using Machine Learning. In chapter 4 we will discuss the resulting models and compare their efficacy to conventional models. In chapter 5 we will summarize our results and its contributions to the field in addition to outlining future extensions of the methods presented in this thesis.

Chapter 2

Theory

2.1 Charged Colloidal Suspensions

Suspensions of charged colloidal particles are complex systems of large charged particles in a electrolyte. The charge of these large colloidal particles comes from ionic groups on the surface of the particle. These charged groups form when ions dissociate from an initially charge-neutral surface of the colloid, resulting in a charged ion in the solvent and an oppositely charged group on the surface of the colloid. The ions that are released from the colloid's surface are called counterions, as they balance the charge of the colloids, leaving the system charge neutral. This solvent may or may not also include ions from a dissolved salt.

The entire system of a charged colloidal suspension now consists of charged colloidal particles, counterions with opposite charges originating from the colloid's surface and counterions are coions originating from a dissolved salt. Of course, all of these charged particles are surrounded by a sea of solvent particles.

Treating all particles in this system explicitly in all their degrees of freedom would be unfeasible and in addition highly unnecessary to give an accurate description of the physics of the system. For this reason, we usually make some reasonable assumptions to simplify the system and greatly decrease the number of degrees of freedom.

2.1.1 Primitive Models

In order to effectively describe the behavior of ions in a solution, one often introduces some simplification to make the result easier to compute and in a sense more general. A common approximation leads to the Primitive Model (PM).

The PM greatly reduces the complexity of the system by reducing all the individual solvent molecules into a continuous medium. The ions in the system are treated as charged hard-sphere particles that interact via a Coulomb pair potential. The effect of the solvent is then taken into account via its dielectric constant ϵ, that scales the Coulomb interaction strength between the particles. Often though a more intuitive quantity for the strength of the Coulomb interaction is the Bjerrum-length

$$\lambda_B = \frac{e^2}{4\pi\epsilon k_B T},$$
(2.1)

where e denotes the proton charge, ϵ the dielectric constant of the medium and k_B is the Boltzmann constant and T the absolute temperature.

The Bjerrum Length represents the distance at which the Coulomb interaction energy of oppositely charged monovalent ions is equal to the thermal energy $k_B T$ of the system. For water at room temperature the Bjerrum length is approximately $\lambda_B \approx 0.71$ nm.

A given PM of N ion species with hard-sphere radii a_i and valencies z_i, hence charges $q_i = z_i e$, for all species $i = 1, \ldots, N$ in a solvent with Bjerrum length λ_B is then described by the pair potentials between species i and j with distance r between the two particle centers

$$\beta V_{ij}^{\mathrm{PM}}(r) = \begin{cases} \infty & \text{if } r < (a_i + a_j); \\ \lambda_B \dfrac{z_i z_j}{r} & \text{if } r \geq (a_i + a_j). \end{cases}$$
(2.2)

The PM is used to describe very general systems of charged particles in a solution. It admits size asymmetric and charge asymmetric systems with many distinct species of ion. However, a common implementation of the PM is for simple salt solutions. Common salt solutions like table salt in water are charge symmetric meaning they have equal but opposite valency $z_+ = -z_-$, a so-called 1:1 electrolyte, and their radii are comparable $a_+ \approx a_-$.

Restricted Primitive Model

So for describing these salt solutions one often uses an even more simplified version of the PM, namely the Restricted Primitive Model (RPM). In the RPM, we only consider two ion species (+/-) of equal radius $a \equiv a_+ = a_-$, but with opposite valency $z_\pm = \pm 1$. The RPM describes a 1:1 electrolyte, ensuring that the system is charge neutral.

To describe this binary system, we have the simplified pair potential between species i and j with distance r between the two particle centers

$$\beta V_{ij}^{\mathrm{RPM}}(r) = \begin{cases} \infty & \text{if } r < a; \\ \lambda_B \dfrac{z_i z_j}{r} & \text{if } r \geq a. \end{cases}$$
(2.3)

In Figure 2.1 we schematically visualize the full atomic model, the primitive model and the restrictive primitive model of a salt solution.

Colloidal Primitive Model

For charged colloidal suspensions, the RPM is unfortunately too restrictive, but the simplifications made in this model are still applicable to all ions in a charged colloidal suspension. This version of the PM is called the Colloidal Primitive Model (CPM). Charged colloidal suspensions consist of 4 types of particles: Colloids and their counterions plus the two ion species (+/-) from the salt. The main differences between the RPM and the charged colloidal suspensions are the asymmetries in the system between colloids and ions. The colloids are larger and carry more charge. The size discrepancy between the ions and the colloids is accounted for by allowing particles of two distinct sizes: macroions, the colloids, with size a_M and microions, counterions and salt ions, with size $a_\mu < a_M$. We assume these macroions to have negative charge $q_M = -Ze$ and microions with unit charges $q_\mu = \pm e$. The corresponding macroion and microions valencies are then Z and ± 1, respectively. The charged macroions supply the system with counterions that counter the macroions charge, such that the system is charge neutral. Let N_M be the number of macroions in the system, their negative charges are then countered by $N_c = ZN_M$ counterions, microions with positive charge e. In addition to these ions there are N_s salt pairs present in the solution giving a total of $N_+ = N_s + ZN_M$ positive ions and $N_- = N_s$ negative ions.

The interaction in the CPM can be described by the pair potential between species i and j with distance r between the two particle centers

$$\beta V_{ij}^{\text{CPM}}(r) = \begin{cases} \infty & \text{if } r < a_{ij}; \\ \lambda_B \dfrac{q_i q_j}{r} & \text{if } r \geq a_{ij}. \end{cases} \tag{2.4}$$

In this potential q_i denotes the charge of particle i in units of e and a_{ij} is the notation for the arithmetic mean of the radii of particles i and j as given by $a_{ij} = \frac{1}{2}(a_i + a_j)$ in which a_i, a_j denotes the individual radii of particles i and j respectively.
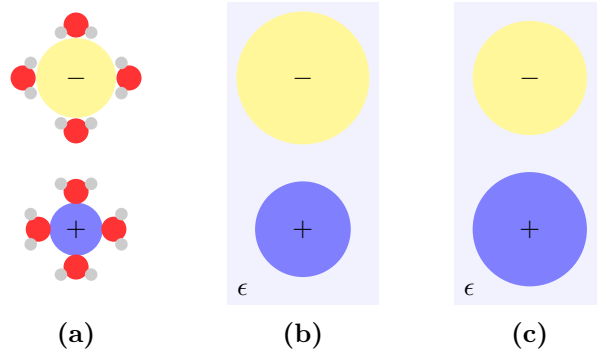

Figure 2.1: Schematic representation of electrolyte models for NaCl solution in water (not to scale). Here we show **(a)** the full atomic model of hydrated ions in bulk water, **(b)** the primitive model representing the ions with an effective hydrated hard-sphere radius and the water as a medium with dielectric constant ϵ, and **(c)** the restricted primitive model, where additionally the salt ion radii are equal.

2.2 Effective interactions

As we recall from statistical mechanics, the equilibrium properties of a thermodynamic system are largely determined by the free energy of the system. In the case of a canonical ensemble consisting of N identical particles in a volume V at temperature T the Helmholtz free energy F equals

$$F(N, V, T) = -k_B T \ln Z(N, V, T), \tag{2.5}$$

with the canonical partition function Z in terms of the Hamiltonian H and all N particle positions $\vec{r}^N$ and particle momenta $\vec{p}^N$ given by

$$Z(N, V, T) = \frac{1}{N! \, h^{3N}} \int d\vec{r}^N d\vec{p}^N e^{-\beta H} \equiv \mathrm{tr} \left[e^{-\beta H} \right]. \tag{2.6}$$

In Eq. (2.6) we defined $\beta \equiv 1/k_B T$ to be the inverse temperature, h is the Planck constant, and tr[.] denotes the classical canonical trace defined by the phase-space integral over the particle degrees of freedom in the same equation.

The Hamiltonian determines the dynamics of a particular configuration of the system and is split up into a kinetic energy part K and potential energy U as

$$H(\vec{r}^N, \vec{p}^N) = K(\vec{p}^N) + U(\vec{r}^N). \tag{2.7}$$

For all classical systems the kinetic energy will be given by

$$K(\vec{p}^N) = \sum_i^N \frac{p_i^2}{2m}, \tag{2.8}$$

where m is the particle mass of all N identical particles. The potential energy, however, is dependent on the system in question and can consists of pair interactions, such as Lennard-Jones or Coulomb interaction and/or external interactions, such as an electric field, to name a few examples.

The description above holds for one-component systems, but charged colloidal suspensions are systems with multiple distinct components, namely macroions N_M and microions N_μ. The partition function of such a system then becomes

$$Z(N_M, N_\mu, V, T) = \mathrm{tr}_M \big(\big[\mathrm{tr}_\mu \left[e^{-\beta H} \right] \big] \big). \tag{2.9}$$

In Eq. (2.9) we have taken the canonical trace over the macroions and microions separately. The key step to defining an effective interaction and reducing the multi-component system back into a one-component system of only macroions is to perform the canonical trace over the microions and integrate out the microion degrees of freedom. This results in

$$Z(N_M, N_\mu, V, T) = \mathrm{tr}_M \left[e^{-\beta H_{\mathrm{eff}}} \right], \tag{2.10}$$

where we have defined an effective Hamiltonian H_{eff} in which the microions have been integrated out into their effective contributions.

In the context of the CPM defined in subsection 2.1.1, we know the Hamiltonian H as given by the kinetic energy in Eq. (2.8) and the potential energy as the sum of pair potentials in Eq. (2.4) over all particle pairs. We can derive what the effective Hamiltonian H_{eff} of the one-component system of only colloidal particles would equate to. The CPM is formulated in terms of pair interactions V_{ij} between species i and j given by Eq. (2.4). This give us the option to split up the total Hamiltonian into multiple parts: Macroion-macroion H_M, macroion-microion $H_{M\mu}$ and microion-microion H_μ parts as

$$H = H_M + H_{M\mu} + H_\mu; \tag{2.11}$$

$$H_M = K_M + \sum_{i,j|i<j}^{N_M} V_{MM}(|R_i - R_j|); \tag{2.12}$$

$$H_{M\mu} = \sum_{i}^{N_M} \sum_{j}^{N_\mu} V_{M\mu}(|\vec{R}_i - \vec{r}_j|); \tag{2.13}$$

$$H_\mu = K_\mu + \sum_{i,j|i<j}^{N_\mu} V_{\mu\mu}(|r_i - r_j|). \tag{2.14}$$

This split-up Hamiltonian gives us the opportunity to more explicitly define the effective Hamiltonian $H_{\text{eff}} \equiv H_M + F_\mu$ with

$$F_\mu \equiv -k_B T \ln \text{tr}_\mu \left[e^{-\beta(H_\mu + H_{M\mu})} \right]. \tag{2.15}$$

The effective Hamiltonian consists of the macroion energy H_M and the effective contribution from the microions F_μ. Notice that this definition of F_μ describes the free energy of the microions in the external field defined by the macroions.

Note that the same procedure can be used when the microions are treated grand canonically by replacing the partition function of Eq.(2.9) with its semi grand canonical equivalent and substituting the canonical trace over the microion degrees of freedom $\text{tr}_\mu[.]$ by the grand canonical trace $\text{Tr}_\mu[.]$ defined by the grand canonical partition function $\mathcal{Z}$. A semi grand canonical system with N_M identical macroions in a volume V at temperature T in contact with a microion reservoir at chemical potential μ is described by grand canonical partition function

$$\mathcal{Z}(N_M, \mu, V, T) = \frac{1}{N_M! h^{3N_M}} \int d\vec{r}^{N_M} d\vec{p}^{N_M} \sum_{N_s=0}^{\infty} \frac{e^{\beta\mu N_\mu}}{N_\mu! h^{3N_\mu}} \int d\vec{r}^{N_\mu} d\vec{p}^{N_\mu} e^{-\beta H} \tag{2.16}$$

$$\equiv \text{tr}_M \left[\text{Tr}_\mu \left[e^{-\beta H} \right] \right], \tag{2.17}$$

where N_s is the number of salt pairs in the system such that the total number of microions N_μ is given by $N_\mu = 2N_s + ZN_M$. For such a system we would then define an effective Hamiltonian as $H_{\text{eff}} \equiv H_M + \Omega_\mu$ with

$$\Omega_\mu \equiv -k_B T \ln \text{Tr}_\mu \left[e^{-\beta(H_\mu + H_{M\mu})} \right]. \tag{2.18}$$

This equation now describes the grand potential of the microions in the external field defined by the macroions and $\text{Tr}_\mu[.]$ is the grand canonical trace over the microion degrees of freedom.

2.2.1 Poisson Boltzmann Theory

At the start of the 20th century, Gouy and Chapman [25, 26] both independently formulated a description of the ion concentration around an electrode as a diffuse layer of ions. For this description, they used the Poisson equation as an expression for the electrostatic potential ψ dependent on the monovalent ion concentrations $\rho_\pm$ of ions with charges $\pm e$ in a medium of dielectric constant ϵ

$$\nabla^2 \psi(\vec{r}) = -\frac{4\pi e}{\epsilon} \left(\rho_+(\vec{r}) - \rho_-(\vec{r}) \right). \tag{2.19}$$

The Poisson equation describes not only the electrostatic potential in ionic solutions, but in all electrostatic systems. In this context it is equivalent to Coulomb's Law and by extension Gauss' law. Furthermore, they argued that in a mean-field approximation the average ion concentration would be distributed like a Boltzmann distribution as

$$\rho_\pm(\vec{r}) = \rho_s e^{\mp \beta e \psi(\vec{r})}, \tag{2.20}$$

where ρ_s is the bulk ion concentration of a connected salt reservoir and $\beta \equiv 1/k_B T$ is the inverse thermal energy. Eqs. (2.19) and (2.20) combined give the Poisson-Boltzmann (PB) equation as a mean-field description of ion concentrations in electrostatic liquids:

$$\nabla^2 \psi = \frac{4\pi e \rho_s}{\epsilon} \sinh(\beta e \psi). \tag{2.21}$$

Most often the PB equation is stated in terms of dimensionless potential $\phi = \beta e \psi$ and inverse screening length

$$\kappa = \sqrt{8\pi \lambda_B \rho_s} \tag{2.22}$$

$$\nabla^2 \phi = \kappa^2 \sinh(\phi). \tag{2.23}$$

The PB equation is a second-order non-linear differential equation, for which analytical solutions do not always exist. For this reason, the PB equation is often linearized by assuming $\sinh(\phi) \approx \phi$, which is a reasonable approximation when ϕ is small. The linearized PB equation reads

$$\nabla^2 \phi = \kappa^2 \phi. \tag{2.24}$$

The particular solution of the linear PB equation is of course dependent on the boundary conditions of the problem in question. Therefore, when solving this second-order differential equation, we will need two additional boundary conditions (actually six if we count all three dimensions separately).

One common boundary condition is to set the gradient of the potential at charged surfaces proportional to surface charge density σ as $\hat{n} \cdot \vec{\nabla} \phi|_{\text{surface}} = 4\pi \lambda_B \sigma$ with $\hat{n}$ being the surface normal. The other remaining boundary condition is usually dependent on the boundary of the system itself. For example, in infinite systems with localized charges one often requires the potential to vanish far from any charges in the system $\phi(\infty) = 0$. This implies that the ion concentrations far away returns to equilibrium, $\rho_\pm(\infty) = \rho_s$.

2.2.2 Density Functional Theory

Density Functional Theory (DFT) is, exactly as the name implies, a method in which one describes quantities as a functional of particle density. A functional $\mathcal{F}$ is a function that takes another function as its input. In the case of a DFT for a one-component system this input function is a density function $\rho(\vec{r})$. We denote a functional and its input as $\mathcal{F}[\rho(\vec{r})]$. The advantage that functionals give in this situation is the ability to take functional derivatives to determine certain equilibrium conditions. In particular, we know that an equilibrium density function ρ^{eq} exists that minimizes the free energy functional and the equilibrium free energy $\mathcal{F}^{\text{eq}}$ equals the free energy of the equilibrium density function:

$$\left.\frac{\delta\mathcal{F}[\rho]}{\delta\rho}\right|_{\rho=\rho^{\text{eq}}} = 0, \qquad \mathcal{F}[\rho^{\text{eq}}] = \mathcal{F}^{\text{eq}}. \tag{2.25}$$

In the case of a charged colloidal suspension, the relevant density distributions are those of the microions $\rho_\pm(\vec{r})$, as the effective interactions between macroions are dependent on these unknown functions. Other useful derived descriptors in this system are the total microion charge density function $q_\mu(\vec{r}) = (\rho_+(\vec{r}) - \rho_-(\vec{r}))$ in units of e and total microion number density function $n_\mu(\vec{r}) = (\rho_+(\vec{r}) + \rho_-(\vec{r}))$.

Using the quantities $\rho_\pm(\vec{r})$, $q_\mu(\vec{r})$, $n_\mu(\vec{r})$ we can express the components of the grand potential $\Omega_\mu = U - TS - \mu_+ N_+ - \mu_- N_-$. This is the grand potential of the microions in an external field described by the macroions and in contact with a salt reservoir at salt density ρ_s just like in Eq. (2.18). This can be described as a functional of the ion density functions $\Omega_\mu[\rho_\pm]$ by its components

$$U[\rho_\pm] = U_\mu[\rho_\pm] + U_{M\mu}[\rho_\pm] \tag{2.26}$$

$$= \frac{e^2}{2\epsilon}\int d\vec{r}\,d\vec{r}'\,\frac{q_\mu(\vec{r})q_\mu(\vec{r}')}{|\vec{r}-\vec{r}'|} + \int d\vec{r}\,[q_\mu(\vec{r})V(\vec{r}) + n_\mu(\vec{r})W(\vec{r})], \tag{2.27}$$

where the first term $U_\mu[\rho_\pm]$ represents Coulombic microion-microion interactions and the second term $U_{M\mu}[\rho_\pm]$ the macroion-microion interactions, where $V(\vec{r})$ and $W(\vec{r})$ represent the multi-centered Coulomb and hard-sphere interactions of the macroions respectively. Notice that we left out any microion-microion correlations, as we assume them to be point-like for the rest of this derivation. The next component of the microions grand potential is the entropic term of an ideal ionic gas with thermal wavelength $\Lambda = h/\sqrt{2\pi m k_B T}$ given by

$$TS[\rho_\pm] = -k_B T \sum_\pm \int d\vec{r}\,\rho_\pm(\vec{r})\left[\ln\left(\rho_\pm(\vec{r})\Lambda^3\right) - 1\right]. \tag{2.28}$$

Lastly, we have the chemical potential contributions of both species

$$\mu_\pm N_\pm[\rho_\pm] = \mu_\pm \int d\vec{r}\,\rho_\pm(\vec{r}), \tag{2.29}$$

where chemical potentials of these species may be defined as $\mu_\pm \equiv k_B T \ln(\rho_s \Lambda^3)$ with a given reservoir microion concentration ρ_s. We can combine Eqs. (2.28) and (2.29) using the definition of the chemical potentials to find the ideal contribution Ω_{id} to the grand potential

$$\beta\Omega_{\text{id}}[\rho_\pm] = \sum_\pm \int d\vec{r}\,\rho_\pm(\vec{r})\left[\ln\frac{\rho_\pm(\vec{r})}{\rho_s} - 1\right]. \tag{2.30}$$

This gives us a total grand potential

$$\beta\Omega_\mu[\rho_\pm] = \beta\Omega_{\mathrm{id}}[\rho_\pm] + 2\pi\lambda_B \int d\vec{r}\,d\vec{r}'\,\frac{q_\mu(\vec{r})q_\mu(\vec{r}')}{|\vec{r}-\vec{r}'|} + \int d\vec{r}\,[q_\mu(\vec{r})\,\beta V(\vec{r}) + n_\mu(\vec{r})\,\beta W(\vec{r})]\,. \quad (2.31)$$

We may find an expression for the equilibrium profiles of $\rho_\pm(\vec{r})$ by considering the functional derivative of the grand potential $\beta\Omega_\mu[\rho_\pm]$ with respect to these functions. In equilibrium we would have that these ion profiles minimize the grand potential and their functional derivatives thus vanish as

$$\frac{\delta\Omega_\mu[\rho_\pm(\vec{r})]}{\delta\rho_\pm(\vec{r})} = 0. \quad (2.32)$$

Substituting Eq. (2.31) into this functional derivative leads to

$$\ln\frac{\rho_\pm(\vec{r})}{\rho_s} \pm \phi(\vec{r}) + \beta W(\vec{r}) = 0, \quad (2.33)$$

where we define the dimensionless potential

$$\phi(\vec{r}) \equiv \lambda_B \int d\vec{r}'\,\frac{q_\mu(\vec{r}')}{|\vec{r}-\vec{r}'|} + \beta V(\vec{r}). \quad (2.34)$$

In the region of the system outside the hard-sphere of the macroions, we have that $\beta W(\vec{r}) \equiv 0$. In this case, Eq. (2.33) reduces to exactly the Boltzmann Eq. (2.20). The corresponding version of the Poisson Eq. (2.19) is then

$$\nabla^2\phi(\vec{r}) = \lambda_B \int d\vec{r}'\nabla^2\frac{q_\mu(\vec{r}')}{|\vec{r}-\vec{r}'|} + \beta\nabla^2 V(\vec{r}) \quad (2.35)$$

$$= -4\pi\lambda_B \int d\vec{r}'\delta(|\vec{r}-\vec{r}'|)q_\mu(\vec{r}') + \sum_i^{N_M}\nabla^2 v(\vec{R}_i - \vec{r}) \quad (2.36)$$

$$= -4\pi\lambda_B q_\mu(\vec{r}) - \frac{Z\lambda_B}{a^2}\sum_i^{N_M}\delta(|\vec{R}_i - \vec{r}| - a). \quad (2.37)$$

Here we used the multi-centered nature of the electrostatic potential $V(\vec{r}) = \sum_i v(\vec{r} - \vec{R}_i)$. The multi-centered electrostatic macroion potential $v(\vec{r})$ is defined by

$$\beta v(\vec{r}) = \begin{cases} \dfrac{-Z\lambda_B}{a} & \text{if } |\vec{r}| < a; \\[2ex] \dfrac{-Z\lambda_B}{|\vec{r}|} & \text{if } |\vec{r}| > a. \end{cases} \quad (2.38)$$

The corresponding charge distribution according to Gauss' Law is

$$\beta\nabla^2 v(\vec{R}_i - \vec{r}) = \frac{-Z\lambda_B}{a^2}\delta(|\vec{R}_i - \vec{r}| - a). \quad (2.39)$$

Now combining this Poisson Eq. (2.37) with the Boltzmann Eq. (2.33) gives us the PB from this DFT:

$$\nabla^2 \phi(\vec{r}) = \kappa^2 \sinh \phi(\vec{r}) - \frac{Z\lambda_B}{a^2} \sum_i^{N_M} \delta(|\vec{R}_i - \vec{r}| - a). \tag{2.40}$$

Unfortunately, Eq. (2.40) has no analytic solutions. We must make some approximations to come to a solution that approximates the equation. Working out the approximate solutions to Eq. (2.40) will be too involved for the scope of the thesis. We will give an overview of the procedure one would follow to obtain an effective macroion potential from DFT.

First, one would make approximation to the initial density functional by expanding the ideal-gas contributions $\beta\Omega_{\mathrm{id}}[\rho_\pm]$ about the mean ion densities $\bar{\rho}_\pm = \int d\vec{r}\rho_\pm(\vec{r})/V$ up to quadratic order. Then the equilibrium conditions according to Eq. (2.32) will become linear in $\rho_\pm - \bar{\rho}_\pm$.

Secondly, the hard-sphere exclusion of the colloidal particles makes the problem difficult to solve. Replacing with a pseudo-hard-sphere potential with finite strength βw_0 instead of ∞ for $r < a$. We can later choose w_0 such that $\rho_\pm(r) = 0$ for $r < a$. These assumptions make it possible to solve a version of PB Eq. (2.40). [19, 21, 27]

After finding our equilibrium profiles $\rho_\pm^{\mathrm{eq}}(\vec{r})$ as solution of the PB equation, we can express the equilibrium density functional $\beta\Omega^{\mathrm{eq}}(\rho_\pm)$ by entering said equilibrium profiles $\rho_\pm^{\mathrm{eq}}(\vec{r})$. Now $\beta\Omega^{\mathrm{eq}}(\rho_\pm^{\mathrm{eq}})$ defines the microion induced effective potential as defined in Eq. (2.18).

The method described above gives an effective pair potential, namely the DLVO potential, as a result. The DLVO pair potential is given by the equation

$$\beta V_{\mathrm{DLVO}}(r) = Z^2 \lambda_B \left(\frac{e^{\kappa a}}{1 + \kappa a} \right)^2 \frac{e^{-\kappa r}}{r}, \tag{2.41}$$

where Z is the macroion charge in units of e, λ_B is the Bjerrum length, κ is the inverse Debye screening length and a is the macroion radius. The DLVO pair potential is a special case of a general class of pair interactions, Yukawa potentials

$$\beta V_{\mathrm{Yuk}}(r) = C \frac{e^{-\kappa r}}{r}. \tag{2.42}$$

The Yukawa potential only has two parameters: C the energy scale and κ the inverse screening length.

2.3 Computer Simulations

Computer simulations provide a useful tool for exploring model systems. They allow us to set parameters in the model very precisely and measure quantities with very high accuracy. In this way computer simulations allow us to test our models against experiments or test theoretical simplifications against their full description counterparts. In this thesis we will use computer simulation in the opposite way, by simulating a full description system we aim to find a simplified effective description using the measurement from these simulations.

Computer simulations can be run at many scales, ranging from the atomic scale to molecule scale to macro-molecules and beyond. When simulating large molecular systems, it is often inefficient and unnecessary to describe the system in terms of individual atoms. This would lead to insurmountable complexity and thus computation times. A common approach is to coarse-grain the system: reduce certain elements to simplified effective descriptions. For charged colloidal suspensions, we employ coarse-graining to reduce the medium to a dielectric and the particles to spherical approximations, just like in the PM described in subsection 2.1.1.

For a more in depth exposition of the techniques and theory behind computer simulations, the book by Frenkel and Smit [28] is strongly recommended as additional reading beyond the summary below.

2.3.1 Molecular Dynamics Simulation

In Molecular Dynamics (MD) simulation one uses Newton's second Law, $\vec{F} = m\vec{a}$, force $\vec{F}$ equals mass m times acceleration $\vec{a}$, to propagate a system in time and in this way simulate the physical behavior of the system in question as it would behave in an experiment. A basic MD program goes through a cycle of computing the forces on all particles in the current configuration and then performing time integration to generate the next configuration.

Force computation

Calculating the forces of a given configuration is in principle relatively straightforward. Let us assume all the forces on the particles can be approximated by pair interactions, then we loop over all the pairs and sum the corresponding forces, giving us the net force on each particle. This does work, but is not very efficient and more importantly does not scale well with system size. The amount of times the pair interactions would have to be evaluated is of quadratic order $O(N^2)$ in the number of particles in the system N.

One way to improve upon this is not to consider all particle pairs, but only the necessary or relevant pairs. As many types of interactions, such as Van der Waals or Lennard-Jones interaction, are only relevant on short ranges, one can choose a suitable cutoff range r_c. Particle pairs further apart than this cutoff range are approximated to not interact, and their interactions do not have to be directly evaluated. This reduces the amount of interactions to be computed to all particles within a fixed volume of around each particle. This scales only with the amount of particles in the system, thus $O(N)$.

Of course, we have now only shifted the problem to determining which atoms are within interaction range of each other, for which we still have to compute $O(N^2)$ distances. Another trick to avoid this is using neighbor lists. These are lists that keep track of what particles are near one another. An example of such a method would be using cell lists, which divides up the total system into boxes, or cells. These boxes have side lengths l larger than the cutoff range r_c. We can easily determine to which box each particle belongs using $O(N)$ operations. Because the boxes are such that $l > r_c$ we know for sure that only particles in neighboring boxes can interact, otherwise there would be at least one box in between making their separation at least as large as l and thus beyond the interaction range r_c. In total, this would make the evaluation of all interactions with a limited range scale as $O(N)$.

There are also particle interactions, such as Coulombic interactions, that are relevant on long ranges, and cutting them off at a certain radius could significantly change the physics of the system. Luckily, there still are some methods to improve upon a brute force $O(N^2)$ computation of pair interactions. One improvement comes in the form of long-range solvers. Long-range solvers often translate the system to reciprocal space using a Fourier transform and compute the forces on all particles in reciprocal space. This translation makes the range of the interaction essentially infinite. An example of such a long-range solver is the Particle-Particle Particle-Mesh (PPPM or P^3M) algorithm [29] often used for Coulombic interactions. This algorithm improves on the direct computation of all interactions by reducing the scaling to $O(N \log N)$.

Time integration

Now, knowing the forces on all the particles, it is time to use those to propagate the system in time. Given the current position of a particle $\vec{r}(t)$ and the force $\vec{f}(t)$ on that particle at time t we would like to know the position at some later time $t + \Delta t$. There are many integration schemes to propagate the system in time, and one simple yet accurate example is Verlet integration [30].

For Verlet integration we do require an additional position of the particle at the previous time step $t - \Delta t$. The next position according to Verlet integration is then given by

$$\vec{r}(t + \Delta t) = 2\vec{r}(t) - \vec{r}(t - \Delta t) + \frac{\vec{f}(t)}{m}\Delta t^2. \tag{2.43}$$

To run an actual simulation using this integration scheme, we need two starting positions $\vec{r}(t_0)$, $\vec{r}(t_0 + dt)$ instead of one. The second position is usually not explicitly given but instead, an initial velocity $\vec{v}(t_0)$ is given to compute $\vec{r}(t_0 + \Delta t)$ as

$$\vec{r}(t_0 + \Delta t) = \vec{r}(t_0) + \vec{v}(t_0)\Delta t + \frac{\vec{f}(t_0)}{2m}\Delta t^2. \tag{2.44}$$

A derived integration scheme is Velocity Verlet integration [31]. This integration scheme does not require two previous positions, but instead keeps track of particle velocity. This integration scheme uses the following equations to compute the next particle position and velocity:

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{\vec{f}(t)}{2m}\Delta t^2, \tag{2.45}$$

$$\vec{v}(t + \Delta t) = \vec{v}(t) + \frac{\vec{f}(t) + \vec{f}(t + \Delta t)}{2m}\Delta t. \tag{2.46}$$

Just like the regular Verlet algorithm, it has an error of order $O(\Delta t^4)$ and uses similar amounts of computational resources. In Figure 2.2 we show a visualization of a time step in this integration scheme.

Velocity Verlet integration is often used in MD as it provides good accuracy with an error on the order of $O(\Delta t^4)$. Furthermore, it is numerically stable, respecting the preservation of energy, and, lastly, it is computationally inexpensive.
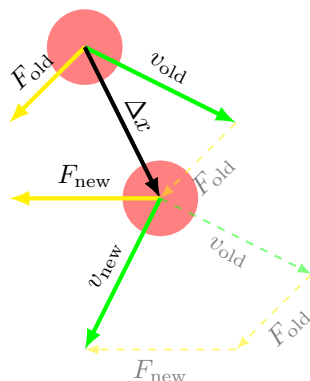

Figure 2.2: Schematic visualization of a single time step in the velocity Verlet algorithm given by Eqs. (2.45) and (2.46). In this figure, velocities v and forces F are represented by the integrated displacement equivalents $v\Delta t$ and $\frac{1}{2m}F\Delta t^2$, respectively.

Thermostat scheme

The force calculation and time integration are in principle sufficient to perform MD simulation. However, due to the nature of Newton's equations, simulations using only time integration are energy preserving. This corresponds to a simulation of a microcanonical ensemble (NVE), while temperature conserving ensembles like the canonical ensemble (NVT) are more common in physical systems.

In order to simulate a canonical ensemble, we need a thermostat that ensures that the velocities of the atoms in the system are distributed according to the Maxwell-Boltzmann velocity distribution for that given temperature. One simple approach is to simulate actual collisions with a heat bath in which we randomly select particles each time step to undergo a collision with the heat bath, setting their velocity randomly according to the Maxwell-Boltzmann distribution. This approach is the so-called Andersen thermostat [32]. This thermostat, unfortunately, gives unphysical simulation results. It affects the dynamics due to sudden changes in velocity, which influences the velocity auto-correlation and thus diffusion of the particles. Another thermostat that does not rely on stochastic methods is the Nosé-Hoover thermostat [33]. This thermostat modifies the equations of motion by introducing a thermodynamic friction coefficient ξ and its effective 'mass' Q. Considering a system of N particles with positions $\vec{r}_i$, momenta $\vec{p}_i$ and masses m_i for $i = 1, \ldots, N$ these equations of motions for a canonical ensemble become

$$\dot{\vec{r}}_i = \vec{p}_i/m_i, \tag{2.47}$$

$$\dot{\vec{p}}_i = \vec{f}_i - \xi\vec{p}_i, \tag{2.48}$$

$$\dot{\xi} = \frac{1}{Q}\left(\sum_i \frac{\dot{\vec{p}}^2}{m_i} - 3Nk_BT\right). \tag{2.49}$$

Where $\vec{f}_i$ denotes the force on particle i.

The Nosé-Hoover thermostat generates a canonical distribution much like the Andersen thermostat in most cases without the disturbances in the particle's velocities that the Andersen thermostat brings along. Yet the Nosé-Hoover thermostat also has shortcomings. It fails to produce canonical distributions, and instead more alike microcanonical distributions, when there are multiple conserved quantities in the system. An extension to the Nosé-Hoover thermostat is a Nosé-Hoover Chain [34] in which we have a chain of M friction coefficients $\xi_1, \ldots \xi_M$. Only the first of these coefficients, ξ_1, is coupled to the momenta of the particles and the rest are coupled to each other in a chain-like fashion. This added chain expands the number of conservation laws a given system can have for which the Nosé-Hoover thermostat can reproduce a canonical distribution.

Molecular dynamics code

For an efficient and fast implementation of all these algorithms used for MD simulation, we will use the open-source LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) molecular dynamics code [35]. LAMMPS provides various other optimizations to the algorithms described before, such as system partitioning and force lookup tables, along with excellent parallel performance. LAMMPS also includes many built-in options to simulate common systems and allows for simulations on various scales: From atomic scale to continuum scale. In addition to any built-in options, LAMMPS comes with many prepackaged extensions for Monte Carlo methods, Brownian Dynamics, Long-range solvers and GPU acceleration just to name a few.

2.3.2 Grand Canonical Monte Carlo simulation

A limitation of MD simulations is the inability to make discrete changes in your simulation. For example, there is no continuous parameter for the number of particles to change over time, so it is not straightforward to use only MD to simulate a Grand Canonical ensemble. While there are methods similar to the Nosé-Hoover thermostat for dynamically changing the particle number [36], these are involved methods and hard to implement.

This is where Monte Carlo (MC) methods enter the picture. MC methods are stochastic algorithms that use randomness to sample an entire configuration space according to a certain distribution. In the context of simulating a thermodynamic ensemble, MC simulations perform random walks in the parameter space of the system, randomly displacing particles, incrementing/decrementing system volume, and/or inserting/removing particles. Each step of the random walk is either accepted or rejected according to a criterion that ensures that all the samples generated by the random walk are distributed according to the ensemble distribution.

For Grand Canonical Monte Carlo (GCMC) simulation one performs particle exchange moves, these moves either attempt to insert or remove a particle at random. For an insertion move, a particle will be added to the system in a random location, and for a removal move, a random particle will be removed from the system. These moves are accepted with probabilities P_{insert} and P_{insert} according to the Metropolis method [37]

$$P_{\text{insert}}(N \to N+1) = \min\left[1, \frac{V}{\Lambda^3(N+1)} \, e^{\beta(\mu - U(N+1) + U(N))}\right], \tag{2.50}$$

$$P_{\text{remove}}(N \to N-1) = \min\left[1, \frac{\Lambda^3 N}{V} \, e^{-\beta(\mu + U(N-1) - U(N))}\right]. \tag{2.51}$$

In these equations N is the number of particles, V is the system volume, Λ is the thermal wavelength of the particle, β the inverse temperature, μ the chemical potential of a virtual reservoir and U is the system energy as function of the particle configuration.

In the metropolis method, one always accepts a move that reduces the energy of the system and accepts moves that increase the energy of the system with a probability that decreases exponentially with said increase in energy. The metropolis method ensures that the underlying distribution of states generated by the MC simulation is equal to the distribution of the thermodynamic system.

The chemical potential μ of a given system is kept fixed for GCMC simulations and determines the equilibrium particle number of the system.

Chapter 3

Method

3.1 Machine Learning Model

Finding an effective interaction for our system of charged colloids using a Machine Learning (ML) model requires us to fit a spatial description (e.g., positions, orientations, etc.) of the system to some descriptor of the physics of our system (e.g., energy or forces).

The full physical descriptions of our system consist of colloid or macroion positions $\vec{R}^{N_M} = \{\vec{R}_i\}$ for $i = 1, \ldots, N_M$ together with all microion positions $\vec{r}^{N_\mu} = \{\vec{r}_j\}$ for $j = 1, \ldots, N_\mu = 2N_s + ZN_M$.

The CPM, detailed in subsection 2.1.1, can then be used to determine the physics by using the pair potentials to compute a total system energy or the forces on each particle. The CPM can be seen as functions U and F that map all $3N_{\text{tot}}$ particle coordinates to the system energy U and $3N_{\text{tot}}$ forces $\vec{F}^{N_{\text{tot}}} = \{\vec{F}_i\}$ for $i = 1, \ldots, N_{\text{tot}}$, where $N_{\text{tot}} = N_M + N_\mu$ is the total number of particles in the system, hence:

$$\textbf{Primitive Model}: \vec{R}^{N_M}, \vec{r}^{N_\mu} \rightarrow U, \vec{F}^{N_{\text{tot}}}.$$

The goal now is to replace this computationally expensive function with an easy-to-compute function for an effective interaction, as defined in section 2.2, that only relies on the macroions positions $\vec{R}^{N_M}$ as an input while still accurately describing the physics of the macroions in the full system. This is where a machine learning model is used to find functions for estimated system energy $\hat{U}$ and the corresponding macroion forces $\vec{F}^{N_M}$:

$$\textbf{Machine Learning Model}: \vec{R}^{N_M} \rightarrow \hat{U}, \vec{F}^{N_M}.$$

The machine learning model that we will use will focus on determining the forces on the macroions depending on the macroion positions. To get the model to 'learn' the patterns and relations between the positions and their corresponding forces, we need training data. This training data consists of input and output pairs $\{\vec{R}^{N_M}, \vec{F}^{N_M}\}$ that are the result of explicit calculation of the CPM.

We could then use a machine learning algorithm to find a model that maps bare particle position to particle forces. However, this approach is not ideal as the input to the model is dependent on every individual particle position and thus dependent on the global properties of the system. This means that the approach is not suitable for scaling up the system in terms of size nor particle number. In addition, from a theoretical standpoint, we expect the interaction to be of a limited range due to screening effects [9]. A better approach would then be to use a local description of each particle's environment and use this description as input for the machine learning model. A commonly used implementation of these kinds of local descriptors are Symmetry Functions (SFs) [38] as will be discussed in subsection 3.1.1.

As a machine learning method we use linear regression subsection 3.1.2 to iteratively select from a large pool of SFs to find the ones that best describe the forces in the training data.
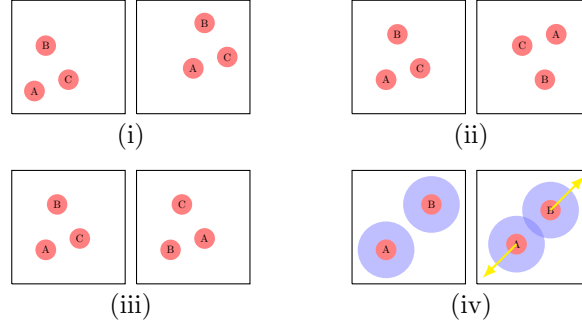
3.1.1 Symmetry Functions

Recall that our machine learning model is essentially a function that transforms macroion input configurations $\vec{R}^{N_M}$ into system energy U and macroion forces $\vec{F}^{N_M}$.

Training a ML model directly on the input configurations is not practical as training on these raw configurations would provide only very specific information and thus require many distinct configurations to be able to interpolate between configuration. If we instead utilize the internal symmetries of the system we can use our data much more efficiently.

Additionally, training a model on a global description of the system, where the input is every particle position in space $\vec{R}^{N_M}$, would not allow us to easily change the size of the system in terms of the number of particles and the volume/density as the machine learning model would need to have been trained on these systems. Instead, we want to train on a local description of the structure around the particles.

In order for a local description to be as useful as possible, it must obey the symmetries that are inherent to the system. In the case of a suspension of identical spherical colloids, this implies that the local description must be symmetric under (i) translations of all particles, (ii) rotations of the system and (iii) permutations of identical particles. Furthermore, we require that (iv) this descriptor is local, meaning that a particle's local description is only affected by other particles within a certain range. These four properties are visualized in the figure below. In this figure, the symmetries (i) through (iii) are presented by two system equal under the corresponding transformation. The locality (iv) is visualized by spheres of influence around particles that determine which particles affect each other.

Symmetry Functions (SFs) provide such a way to transform our input data, which is a global description of all the particle positions $\vec{R}^{N_M}$, into a more appropriate set of local descriptors. For SFs this set $\{G_k^{N_M}\}$ for $k = 1, \ldots, D$ would consist of D values of distinct symmetry functions computed for all N_M macroions individually.

The role of SFs within the entire ML approach is displayed in the figure below, where SFs are used to translate the global description $\vec{R}^{N_M}$ to a set of local descriptors $\{G_k^{N_M}\}$, which is consequently used for linear regression to fit the particle forces $\vec{F}^{N_M}$.
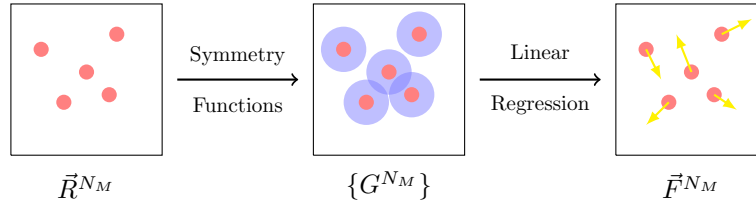
For our machine learning model we will use two families of SFs, namely radial SFs and angular SFs as they were described in great detail by Behler [38]. The radial SFs, as the name implies, describe spherically symmetric relations between particles and only depend on inter-particle distances. The angular SFs do not only depend on the distances between pairs of particles, but also on the angles between triplets.

SFs are scalar functions meant to describe interaction energies, meaning that they are not directly suitable to fit to particle forces, as these are vectors. However, the gradients with respect to particle position are vectors that can be fitted using linear regression, such that the force on a particle i equals the gradient of a linear combination of SFs. We can approximate the system energy U of the CPM with an estimate $\hat{U}$ given a weighted sum of SFs as

$$\hat{U} = \sum_{k=1}^{D} w_k \left[\sum_{j=1}^{N_M} G_k(j) \right]. \tag{3.1}$$

Here w_k is the weight of the kth SF and $G_k(j)$ is the value of the kth SF for particle j in a given configuration. From this energy the corresponding force $\langle \vec{F_i} \rangle$ on particle i is easily obtained by taking the gradient $\vec{\nabla}_i$ with respect to the position $\vec{R}_i$ of particle i as

$$\langle \vec{F}_i \rangle \equiv -\vec{\nabla}_i \hat{U} \tag{3.2}$$

$$= -\sum_{k=1}^{D} w_k \vec{\nabla}_i \left[\sum_{j=1}^{N_M} G_k(j) \right]$$

$$= \sum_{k=1}^{D} w_k \left[-\vec{\nabla}_i G_k(i) - \sum_{j \neq i}^{N_M} \vec{\nabla}_i G_k(j) \right].$$

Here we expressed the forces in terms of SF gradients by substituting Eq. (3.1) for $\hat{U}$. We separated the SF value for i itself from the other particles, as the expression for this gradient may differ. This immediately displays the benefit of using a linear machine learning method, as the system energy can still be recovered by undoing the gradient, which is just equal to the linear combination of the original SFs.

Cutoff function

To ensure that the SFs are local descriptions of a particle's surroundings, they make use of a cut-off function. This cut-off function smoothly reduces the effect of particles at larger distances from each other and eliminates any effect on the SFs by particles beyond the cut-off radius r_{cut}. The chosen cut-off function is in the form of

$$f_c(r) = \begin{cases} \tanh^3(1 - r/r_{\text{cut}}) & \text{if } r \leq r_{\text{cut}}, \\ 0 & \text{if } r > r_{\text{cut}}. \end{cases} \tag{3.3}$$

as used before by Campos-Villalobos et al. [24], Behler [38], and Singraber, Behler, and Dellago [39].

Radial symmetry function

The family of radial SFs $G^2(i)$ is a parametrized function to describe the local environment of particle i in terms of inter-particle distances. This type of SF is in essence a Gaussian shifted from the center of the particle as given by

$$G^2(i) = \sum_j e^{-\eta(R_{ij} - R_s)^2} f_c(R_{ij}). \tag{3.4}$$

Here $R_{ij} = |\vec{R}_i - \vec{R}_j|$ is the distance between particles i and j. The parameters of these SFs are η, determining the radial width of the Gaussian, and R_s which sets the distance between the particle center and the peak of the Gaussian. In Figure 3.1 we display several normalized radial SFs with a range of parameter values for η and R_s.

The gradients of the radial SF, that can be used to describe forces, can be straightforwardly calculated using repeated use of the chain rule for derivation. The gradient does take a different form if the gradient is taken with respect to the particle for which the SF is computed $\vec{\nabla}_i G^2(i)$ or with respect to a different particle $\vec{\nabla}_i G^2(j)$ where $j \neq i$. These are given by

$$\vec{\nabla}_i G^2(i) = \sum_{j|j \neq i} e^{-\eta(R_{ij} - R_s)^2} [f_c'(R_{ij}) - 2\eta(R_{ij} - R_s)f_c(R_{ij})] \frac{\vec{R}_{ij}}{R_{ij}}, \tag{3.5}$$

$$\vec{\nabla}_i G^2(j) = e^{-\eta(R_{ij}-R_s)^2} \left[f'_c(R_{ij}) - 2\eta(R_{ij}-R_s) f_c(R_{ij}) \right] \frac{\vec{R}_{ij}}{R_{ij}}. \tag{3.6}$$

Here we use the convention that $\vec{R}_{ij} = \vec{R}_i - \vec{R}_j$, $R_{ij} = |\vec{R}_{ij}|$ and f'_c is shorthand for the radial derivative $\frac{df_c}{dr}$ of th cutoff function f_c.
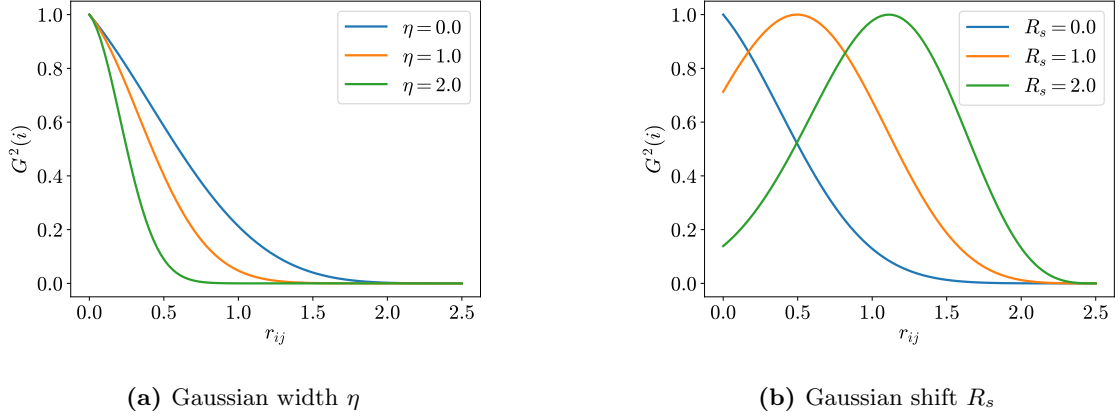

(a) Gaussian width η **(b)** Gaussian shift R_s

Figure 3.1: Normalized versions of radial symmetry functions with cutoff radius $R_{\text{cut}} = 2.5$ for a particle i with one neighbour j at distance R_{ij} . **(a)** Various values for Gaussian width η at $R_s = 0$. **(b)** Various values for Gaussian shift R_s at $\eta = 1.0$.

Angular symmetry functions

The family of angular SFs $G^3(i)$ used both inter-particle distances and angles to give a description of a particle's environment. This type of SF is similar to the radial SF, but with an additional angular term as given by

$$G^3(i) = 2^{1-\xi} \sum_{j,k\neq i} (1 + \lambda\cos\theta_{ijk})^\xi e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}). \tag{3.7}$$

Here $R_{ij} = |\vec{R}_i - \vec{R}_j|$ denotes the distance between particles i and j, θ_{ijk} is the angle at particle i between $\vec{R}_{ij}$ and $\vec{R}_{ij}$ such that $\cos\theta_{ijk} = \frac{\vec{R}_{ij}\cdot\vec{R}_{ik}}{|\vec{R}_{ij}||\vec{R}_{ij}|}$. The Angular SF is parameterized by λ, ξ, η. The parameter λ equals ± 1 and determines the parity of the angle term, this affects the position of the peak in the angular term of the SF. For $\lambda = 1$ the angular is maximal at $\theta_{ijk} = 0°$, and for $\lambda = 0$ the angular is maximal at $\theta_{ijk} = 180°$. Next, ξ determines the peak width of the angular term, while η determines the width of the Gaussian term just like for the radial SF. In Figure 3.2 we visualize the shapes of several parameter choices for the family of angular SFs.

Computation of the gradient is still a straightforward application of the chain rule, although more involved than for the radial SF. The gradient with respect to the particle concerning the SF $\vec{\nabla}_i G^3(i)$ is given by Eq. (3.8) and the other gradients $\vec{\nabla}_j G^3(i)$ such that $i \neq j$ are given by Eq. (3.9).

$$\vec{\nabla}_i G^3(i) = 2^{1-\xi} \sum_{j,k|j\neq i,k\neq i,j<k} \omega_{ijk} \left[(\phi_{ijk} - 2\eta + \chi_{ij})\vec{R}_{ij} + (\phi_{ikj} - 2\eta + \chi_{ik})\vec{R}_{ik} \right] \qquad (3.8)$$

$$\vec{\nabla}_i G^3(j) = 2^{1-\xi} \sum_{k|k\neq i,k\neq j} \omega_{jik} \left[(\phi_{jik} - 2\eta + \chi_{ij})\vec{R}_{ij} + (\psi_{jik} - 2\eta + \chi_{ik})\vec{R}_{ik} \right] \qquad (3.9)$$

Here we used the following four equations as abbreviations:

$$\omega_{ijk} = (1 + \lambda \cos\theta_{ijk})^{\xi} e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}), \qquad (3.10)$$

$$\psi_{ijk} = -\frac{1}{R_{ij}R_{ik}} \frac{\lambda\xi}{1 + \lambda\cos(\theta_{ijk})}, \qquad (3.11)$$

$$\phi_{ijk} = \frac{1}{R_{ij}} \left[\frac{1}{R_{ik}} - \frac{1}{R_{ij}} \cos(\theta_{ijk}) \right] \frac{\lambda\xi}{1 + \lambda\cos(\theta_{ijk})}, \qquad (3.12)$$

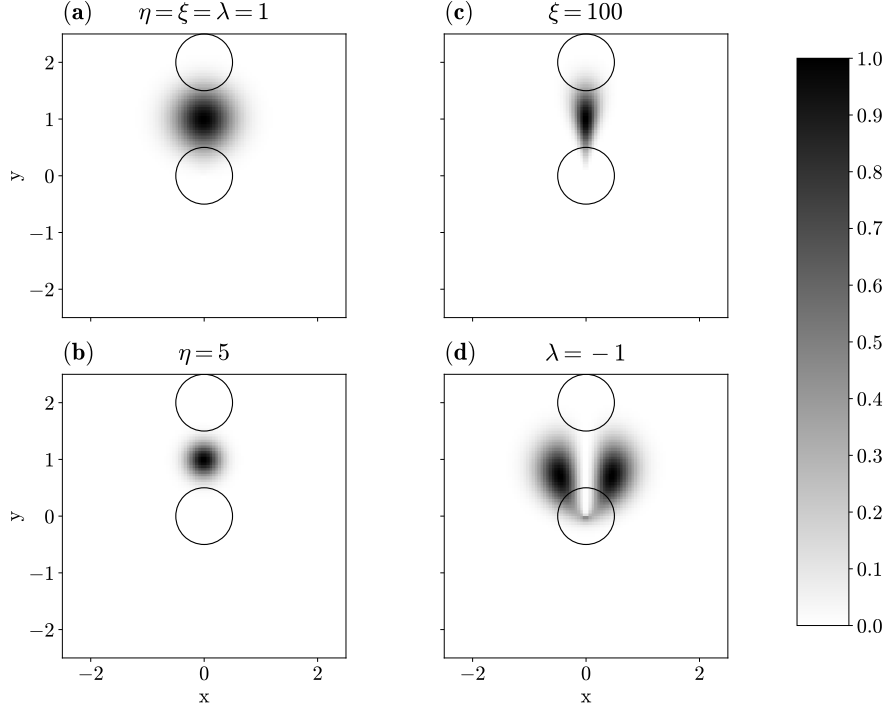$$\chi_{ij} = \frac{1}{R_{ij}f_c(R_{ij})} f_c'(R_{ij}). \qquad (3.13)$$

Figure 3.2: Normalized versions of angular symmetry functions with cutoff radius $R_{\mathrm{cut}} = 2.5$ for particle i with two neighbours j and k. Particle i is fixed at origin and the first neighbour j is fixed at $(0, 2)$. The value of $G^3(i)$ at (x, y) corresponds to the configuration where particle k is fixed at (x, y). **(a)** Standard parameter values. **(b)** Greater η with $\xi = \lambda = 1$. **(c)** Greater ξ with $\eta = \lambda = 1$. **(d)** Opposite λ with $\eta = \xi = 1$.

3.1.2 Linear Regression

The final step in obtaining our machine learning model is fitting the model to the training data. For this fitting procedure, we use linear regression to construct a predictive model for the microion-averaged forces $\langle \vec{F} \rangle^{N_M}$ as a linear combination of SFs gradients $\vec{\nabla}_i G_k(i)$ and $\vec{\nabla}_i G_k(j)$ with weights w_k as

$$\langle \vec{F_i} \rangle = \sum_{k=1}^{D} w_k \left[-\vec{\nabla}_i G_k(i) - \sum_{j \neq i}^{N_M} \vec{\nabla}_i G_k(j) \right] .$$

Combining the radial and angular SFs we have a pool of N_{SF} SFs to build our model. From this pool we will progressively find optimal sets of $D < N_{\mathrm{SF}}$ SFs in such a way that each added SF adds the most information to the previous set.

This pool of SFs consist of various parameter combinations for both radial and angular SFs described in subsection 3.1.1. For each parameter p we choose a set of N_p values $\{p_i\}$ for $i = 1, \ldots, N_p$. The total number of parameter combinations for the radial SFs would then be $N_\eta \cdot N_{R_s}$ and the amount of angular SFs equals $N_\eta \cdot N_\xi \cdot N_\lambda$ for a total of $N_{\mathrm{SF}} = N_\eta \cdot N_{R_s} + N_\eta \cdot N_\xi \cdot N_\lambda$ distinct SFs. The cutoff radius is set to the same value R_{cut} for all SFs.

To ensure that the next chosen SF adds the most information, we compute the Root Mean Squared Error (RMSE) of the model, that quantifies its absolute error. A lower RMSE corresponds to a better model and a RMSE = 0 would indicate a perfect model. Given a certain subset of all SFs, we then compute the RMSE for each remaining SF that is still in the pool. In principle, all SFs should add some information and thus decrease the RMSE, but only the SF that decreases RMSE the most is added to the set of selected SFs.

To determine the RMSE we first need to fit the set of SFs to the forces using the normal equation for linear regression. The normal equation finds weights w for the SFs by projecting the target forces y onto the space spanned all the SFs. This normal equation reads

$$w = (X^T X)^{-1} X^T y. \tag{3.14}$$

Here y is the column vector of all $N = 3N_M \cdot N_C$ force components, X is a $N \times D$ matrix with the N gradient components of D SFs as its columns. The normal equation ensures that w is the column vector of D weights w_k for the SFs G_k with $k = 1, \ldots, D$ that minimizes the Sum of Square Residuals (SSR),

$$\mathrm{SSR} = |y - Xw|^2. \tag{3.15}$$

The RMSE in turn is related to the SSR as

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SSR}}{N}} \tag{3.16}$$

$$= \sqrt{\frac{|y - Xw|^2}{N}}. \tag{3.17}$$

Here $N = 3N_M \cdot N_C$ is the total number of force components and y, X and w are as defined before.

Notice that the version of the normal equation in Eq. (3.14) does not exactly describe conventional linear regression, as we omitted a column of only ones in X. This column represents an added constant term, so without it we only fit scaling parameters for the SFs. This ensures that the forces that we will fit using this method do not acquire a constant shift and are zero in the large separation limit.

3.2 Computer Simulation

3.2.1 System Parameters

The specific system from which we will collect training data for our machine learning models is based on a system described in a paper by Royall et al. [40]. This paper describes experimental systems, ensuring that the parameters used in our simulations are in the realm of physical systems.

The computer simulation that we use to simulate these systems are MD simulation, described in subsection 2.3.1, as they are fast and resource efficient. Unfortunately, MD simulate cannot actually simulate a CPM, as this model includes a Hard-Sphere (HS) potential. The problem here is that MD simulations require forces to be finite to perform the time integration and the HS potential is either zero or infinite, which is not allowed. The solution is to replace the HS potential by a softer short-range repulsive force.

The chosen potential to replace the HS potentials is the Weeks Chandler Andersen (WCA) potential [41], which is essentially a Lennard-Jones potential cut-off and shifted to 0 at its minimum as defined by

$$V_{\text{WCA}}(r) = \begin{cases} 4\,\epsilon_{\text{WCA}} \left[\left(\dfrac{\sigma_{ij}}{r} \right)^{12} - \left(\dfrac{\sigma_{ij}}{r} \right)^{6} \right] + \epsilon_{\text{WCA}} & \text{if } r < 2^{1/6}\sigma_{ij}; \\ 0 & \text{if } r \geq 2^{1/6}\sigma_{ij}. \end{cases} \tag{3.18}$$

The WCA potential has two parameters: ϵ_{WCA} as the strength of the interaction and σ_{ij} as the interaction distance between particles i and j. We set this distance as the sum of the particle radii a_i and a_j, thus $\sigma_{ij} = a_i + a_j$, just like for a HS potential.

The Coulomb forces on all particle in the CPM are calculated by using a combination of exact evaluation up to a cutoff radius $r_{\text{cut}} = 5a$, with a the macroion radius, and a long-range solver as described in subsection 2.3.1. The particular long-range solver used is the particle-particle particle-mesh (PPPM) solver [29]. These two computations of the forces are combined to compute the Coulombic forces.

In Table 3.1 we detail the system parameters that will be used for all simulation throughout this thesis. As seen in this table, the ratio between Bjerrum length λ_B and the ion radius a_μ is roughly 1:5 which indicates that the coupling between microions of opposite charge is relatively weak, and thus we should not expect to see significant ion pair formation.

Macroion radius	a	1.0
Microion radius	a_μ/a	0.05
Macroion valency	Z	90
Microion valency	z	± 1
Bjerrum Length	λ_B/a	0.0098
Soft sphere strength	$\beta\epsilon_{\mathrm{WCA}}$	10

Table 3.1: Parameters used in all MD simulations of the CPM in terms of dimensionless quantities.

3.2.2 Single Density Canonical Ensemble

One set of training data that we use to train a machine learning model on, is obtained from MD simulations of a charged colloidal suspension in a canonical (NVT) ensemble with fixed particle numbers, fixed system volume/density and fixed temperature.

The system density of this system is set by the macroion packing fraction equal to $\eta = \frac{4}{3}\pi a^3 N_M/V = 0.187$. The microion concentration for this system is determined by the relation between ion concentration and reservoir inverse screening length $\kappa a = \sqrt{8\pi (\lambda_B/a)(\rho_s a^3)}$. This relation is defined in Eq. (2.22) in which the salt density $\rho_s = N_s/V$ only depends on the number of ion pairs from an external reservoir N_s. For this system we set the inverse screening length as $\kappa a = 0.83$.

For this system we use MD simulations to obtain $N_C = 200$ configurations of $N_M = 32$ macroions each with $N_s = 2026$ cation-anion pairs added for a total of $N_\mu = 6932$ microions in the system. For each macroion configuration, $N_S = 10000$ samples of microion configuration were generated also using MD simulations for the purposes of determining the mean macroion forces for each configuration, as will be described in subsection 3.2.4.

3.2.3 Multi Density Hybrid Ensemble

A variation on the before mentioned ensemble using a fixed microion concentration would be to treat the microions grand canonically as if the system was connected to a salt reservoir by means of a semipermeable membrane. In simulations, this is done using Grand Canonical Monte Carlo (GCMC) methods, detailed in subsection 2.3.2, where ion pairs are exchanged between the simulated system and a virtual salt reservoir.

While these GCMC simulations better mimic the real physical systems, that are often connected to a salt bath, they also have their downsides. GCMC simulations are much more computationally intensive than MD simulations and on top of that these simulations are mostly sequential and normally cannot benefit from parallel computing as is the case for MD simulations.

One method to improve the efficiency of simulations of a Grand Canonical system is to combine MD simulations and GCMC simulations by utilizing the MD algorithm for performing time steps and employing a GCMC algorithm only for insertion and deletion of particles in the system. Typically, particle exchange moves are performed between a larger number of MD time steps, effectively alternating between MD and GCMC simulations.

In addition to being more physically accurate, the use of Grand Canonical methods to determine the microion concentration in the system allows us to extend our system to various densities much better. The resulting microions concentration in a given system is dependent on the packing fraction of the colloidal particles, the screening effect of the ions on the colloid charge, and the interaction between the microions from the reservoir and from the macroions. Predicting the microion concentration in a given system ab initio is complex. Measuring the microion concentration of a system simulation using GCMC methods, however, is trivial.

In hybrid MD and GCMC form the simulation times of grand canonical simulation can be improved, yet still these simulations are orders of magnitude slower than MD simulations by themselves. This makes GCMC method still unfeasible for simulating the hundreds of configurations for our training data. In order to bring the simulation times down but still use grand canonical methods to determine the microion concentration of our system, we split up the data generation into a grand canonical part for determining average microion concentrations and a canonical part for generating the training data given these microion concentrations. For the first part we use hybrid GCMC methods, and for the latter we can restrict ourselves to more efficient MD simulations.

In the following subsections, we will go through the individual steps that will eventually lead to the training data for a system across multiple densities.

$$\text{Screening: } \kappa a \;\rightarrow\; \text{Chemical potential: } \mu \;\rightarrow\; \text{Salt concentrations: } \rho_s \;\rightarrow\; \text{Forces: } \langle \vec{F} \rangle^{N_M}$$

The first step in the process would be to establish the relation between microion concentration and chemical potential in a salt reservoir. The next step is to determine the average ion concentrations for colloidal suspension at varying densities and screening lengths. In subsection subsection 3.2.4 we will cover the calculation of the mean forces of different colloid configurations similarly to the single density case.

Screening Length and Chemical Potential

The relation between the screening length and the chemical potential of the salt reservoir is important for determining the parameters we need to set for simulations of the full colloidal system. Both the inverse screening length κ and the chemical potential of the microions μ_s are related to the mean salt concentration ρ_s of the system. The inverse screening length by its definition $\kappa^2 \equiv 8\pi \lambda_B \rho_s$ with ρ_s being the ion concentration of a connected salt reservoir. In the ideal gas limit, one can determine the relation between chemical potential and concentration to be $\mu_s = k_B T \ln(\rho_s \Lambda^3)$, where λ is the thermal wavelength of the particles. This gives us the relation for κ as a function of μ:

$$\kappa a = \sqrt{8\pi \frac{\lambda_B}{a} \left(\frac{a}{\Lambda}\right)^3 e^{\beta \mu_s/2}}. \tag{3.19}$$

In the CPM, however, we describe microions as hard-sphere particles with electrostatic interaction, this means they do not behave in an ideal manner. For this reason, we run GCMC simulations of a salt reservoir at various chemical potentials to verify the validity of this relation.

The GCMC simulation uses the same ion and solvent system parameters of the CPM as the full system with colloids, as in subsection 3.2.1. The system size is set to a cube of volume $(2a)^3$. The microions in a GCMC simulation have to be added and deleted as anion-cation pairs in order to keep the overall system charge neutral. Additionally, these pair must form a fixed structure with fixed distance between the two microions for the GCMC implementation used in this thesis. This fixed distance d between the microions does affect the equilibrium microion concentration as in the presence of macroions microion pairs with large inter-particle distance d are hard to fit in. For this reason, the inter-particle distance is set to $d = 0.2a$ as it is small enough to not interfere with placement and does not place the microions within the range of the WCA potential between them.

The chemical potential of the reservoir is varied from $\beta\mu \in \{1/12, 1/6, \ldots, 4\}$ in steps of $\Delta\mu = 1/12$ defined in units of $k_B T$ per particle, where we have set $\Lambda/a = 1$ without loss of generality as this parameter effectively only shift μ_s by a constant value. The mean salt concentration was averaged over 500 measurements taken with 100 insertion/removal attempts between each measurement. These were then used to compute the corresponding inverse screening lengths κ according to Eq. (2.22) This relation between inverse screening length κ and chemical potential μ is presented in Figure 3.3

For all averaged salt concentrations ρ_s the corresponding inverse screening length κ was computed. These inverse screening lengths were plotted against their input chemical potentials and a generalized form of Eq. (3.19), namely $\kappa(\mu) = A \exp(B\mu)$, was fit to the data with fit parameters A and B. The measurements and the resulting fit are in excellent agreement with the ideal gas approximation for most values of the chemical potential. At low chemical potential and thus lower density, we see some deviation. This is most likely due to limited system size in combination with the limitation of the particular GCMC method which requires at least one salt pair to be present in the system.

For the next steps, all we have to do is invert the relation we found such that we have a function $\mu(\kappa)$ for the chemical potential given a particular inverse screening length.
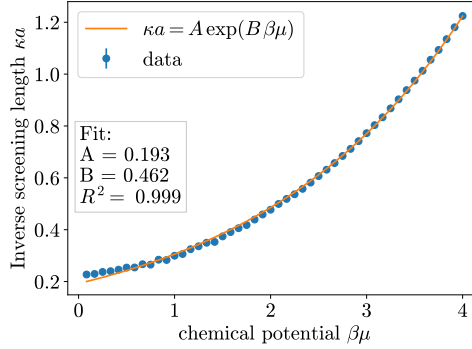
Figure 3.3: The inverse screening length $\kappa = \sqrt{8\pi\lambda_B\rho_s}$, computed from density ρ_s, as a function of salt reservoir chemical potential μ_s. Measurement from GCMC simulation of this relation are presented as dots. The solid lines represent fits of the form $\kappa(\mu)a = A\exp(B\beta\mu)$ was performed to the measurements.

Ion Concentration in Charged Colloidal Suspensions

Now that we can determine the chemical potential of a given reservoir for it to be at a certain inverse screening length κ or salt concentration ρ_s, we are able to set the right parameters and compute average salt concentrations in charged colloidal suspensions. To do this we need to run very costly GCMC simulations. For this reason, we introduce some initialization steps before starting the actual GCMC run from which we measure the salt concentration.

The first step is to generate a valid configuration of only colloids and their associated counterions, no additional salt ions included yet. This system is then brought to an equilibrated state during a canonical NVT MD simulation. After this equilibration step of the colloid configuration, we introduce the salt ions to the system by first performing 15000 insertion/removal attempts at once to get the salt concentration in the system to approximately the equilibrium level. After this second initialization step, we run a GCMC/MD hybrid simulation which periodically performs 100 insertion/removal attempts and lets the system equilibrate between these sets of insertion/removal attempts. The salt concentration is averaged over 100 of these cycles.

A naive model for the average ion concentration in the system would be to dismiss any electrostatic effects and assume the number of ions to be proportional to the free volume in the system. This would be the volume not occupied by the colloids $V_f = V(1-\eta)$ and then the average number of salt ions would be equal to $\overline{N_s} = \rho_s V(1-\eta)$, where η is the colloid packing fraction, V is the system volume and ρ_s is the mean salt concentration of the reservoir. A slight correction would be to consider that not only the colloids exclude volume, but also their accompanying counterions. A general approach to consider discrepancies between the effective occupied volume of the colloidal particles and their packing fraction would be to include a scalar correction to the packing fraction. The model for the average number of ions would then be of the form $\overline{N_s}(\eta) = A(1-B\eta)$, where A and B are fit parameters.

As we can see in figure Figure 3.4, the naive model does not accurately describe the measurements. From this we conclude that the electrostatic effects that we ignored in the naive model are important. Instead of using such a model for determining the average salt concentration in CPM, we use the measurements of these average themselves.

For this system we perform MD simulations of three distinct macroion packing fractions $\eta \in \{0.187, 0.374, 0.561\}$. For each packing fraction we generate 100 configurations for a total of $N_C = 300$ macroion configuration of $N_M = 32$ macroions each. The number of cation-anion pairs added to each configuration is determined by the nearest integer value of the measured average number of salt pairs reported in Table 3.2. For each configuration the mean forces on each macroion was averaged over $N_S = 5000$ microion configurations. More details of this process are discussed in subsection 3.2.4. This entire process was done for two distinct inverse screening lengths $\kappa a \in 0.25, 0.5$ resulting is two separate sets of training data for the training of two separate models.

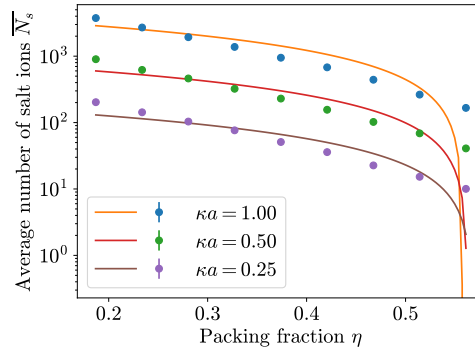| **Packing Fraction**	**Inverse Screening Length**		
η	$\kappa a = 0.25$	$\kappa a = 0.5$	$\kappa a = 1.0$
0.187	101.39 ± 1.2	450.1 ± 2.2	1875 ± 3
0.374	25.5 ± 0.5	115.0 ± 1.2	473.9 ± 2.1
0.561	5.02 ± 0.22	20.4 ± 0.5	83.2 ± 1.0

Table 3.2: The average number of salt pairs $\overline{N}_s$ at different macroion packing fractions η and chemical potential of salt reservoir μ_s corresponding to inverse screening lengths κa measured via GCMC simulations.

Figure 3.4: The average number of salt pairs $\overline{N}_s$ at different macroion packing fractions η and chemical potential of salt reservoir μ_s corresponding to inverse screening lengths κa measured via GCMC simulations. A fit was performed of the form $\overline{N}_s(\eta) = A(1 - B\eta)$ for each value of κa.

3.2.4 Training Data

The set of training data for fitting the machine learning model consists of N_C configurations of N_M colloidal particles, resulting in a total of $N_M \cdot N_C$ position and force vector pairs or $3N_M \cdot N_C$ distinct scalar pairs.

The positions of the colloids $\vec{R}^{N_M}$ in these N_C configurations were extracted from direct simulations of the CPM. Each configuration is in equilibrium. Obtaining accurate values for the forces on each particle $\vec{F}^{N_M}$ for each of the N_C configurations requires an additional step, as the forces are sensitive to specific microion arrangements $\vec{r}^{N_\mu}$ and tend to fluctuate a lot. To remedy this, the forces $\vec{F}^{N_M}$ were averaged over N_S distinct microion arrangements $\vec{r}_i^{N_\mu}$ for $i = 1, \ldots, N_S$ resulting is a microion-averaged force per particle. This average was taken by performing additional MD simulations of each of the N_C configurations in which only the microion dynamics are computed while the colloids are kept fixed in configuration $\vec{R}^{N_M}$ as

$$\langle \vec{F} \rangle^{N_M} = \frac{1}{N_S} \sum_{i=1}^{N_S} \vec{F}_n^{N_M}. \tag{3.20}$$

Here $\langle \vec{F} \rangle^{N_M}$ represents the microion-averaged forces on the colloids in positional configuration $\vec{R}^{N_M}$. Furthermore, $\vec{F}_n^{N_M}$ are the individually computed forces per particle in configuration of macroion positions $\vec{R}^{N_M}$ and microion positions $\vec{r}_n^{N_\mu}$ for $n = 1, \ldots, N_S$.

The resulting set of positions and microion-averaged force vectors for each particle $\{\vec{R}^{N_M}, \langle \vec{F} \rangle^{N_M}\}^{N_C}$ will be split into a 80%-20% training-validation setup.

Force auto-correlation

When computing the mean forces in the CPM for a given configuration we have to make sure we sample independent and thus uncorrelated microion configurations.

To quantify the correlation of these configurations, we instead look at the resulting forces in these configurations. Essentially, we only need these forces to be uncorrelated for the sampling to be efficient. The exact configurations of the microions do not matter. Computing the force autocorrelation $R_{FF}(\Delta t)$ of the forces across a range of time step separations Δt can tell us about the correlation between the sampled configuration. The force autocorrelation is defined as

$$R_{FF}(\Delta t) = \frac{1}{N_S - \Delta t} \sum_{t=t_0}^{t_{N_S} - \Delta t} \frac{\vec{F}(t) \cdot \vec{F}(t + \Delta t)}{|\vec{F}(t)|^2}, \tag{3.21}$$

where Δt is the time separation in number of time steps, N_S is the total number of samples, t_0 is the time of the first sample and t_{N_S} corresponds to the last sample, $\vec{F}(t)$ is the force at time step t.

In Figure 3.5 and Figure 3.6 we show the autocorrelation functions for all systems treated in this thesis. In these figures, we see that almost all samples separated by at least 100 time steps are practically uncorrelated. For the single density system we do see that subsequent sample that are separated by 50 time steps are somewhat correlated and thus not independent.

We also see that all autocorrelations do not tend to 0 as the time step separations increases. This is to be expected as the forces mostly have non-zero mean and fluctuate around this mean. The mean shift present in all forces is of course correlated to itself, and thus all samples are expected to have some correlation greater than 0.
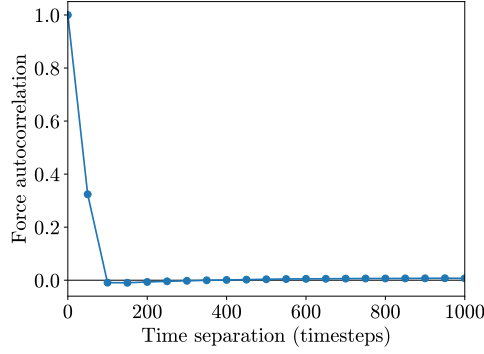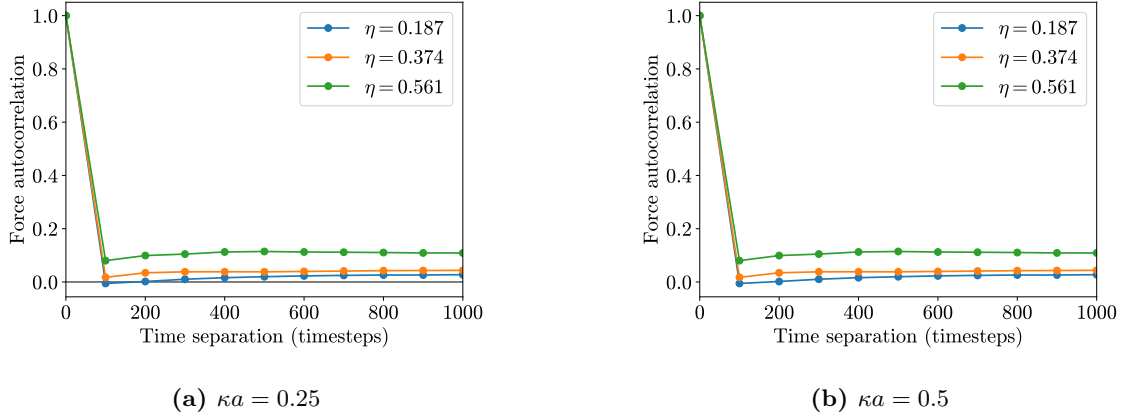

Figure 3.5: Force autocorrelation averaged over all N_C configurations and N_M macroions as function of time step separation. System at packing fraction $\eta = 0.187$ and with inverse screening length $\kappa a = 0.83$.

(a) $\kappa a = 0.25$

(b) $\kappa a = 0.5$

Figure 3.6: Force autocorrelation averaged over all N_C configurations and N_M macroions as function of time step separation for each packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ separately for two different inverse screening lengths κa.

Fluctuation of the forces

Not only the mean of the forces is relevant, it is also relevant to know how much the individual forces fluctuate around this mean. The amount of fluctuation is captured by the Standard Deviation (SD) of the forces over the samples. The SD over the sample is calculated according to

$$SD = \frac{1}{N_S - 1} \sum_{i=0}^{N_S} (F_i - \langle F_i \rangle)^2 \tag{3.22}$$

Where N_S is the sample size, F_i is one component (x, y or z) of the force on a specific macroion and $\langle F_i \rangle$ is the average of this force as computed by Eq. (3.20).

In Table 3.3 we report the average SD of the forces. We can take the averages over the 3 components of the forces, as all three directional components have near identical SD for each macroion in all configuration. The SDs in this table define an effective deviation sphere of the forces with the reported SD as its radius.

Screening Length	Macroion packing fraction		
κa	$\eta = 0.187$	$\eta = 0.374$	$\eta = 0.561$
$\kappa a = 0.83$	55.9	-	-
$\kappa a = 0.5$	42.03	66.5	104.2
$\kappa a = 0.25$	38.47	65.0	102.9

Table 3.3: The average standard deviations of the forces in units of $k_B T / 2a$ averaged of all N_C configuration, N_M macroions and 3 components.

Chapter 4

Results

4.1 Yukawa pair interactions

4.1.1 Two-body interactions

From our MD simulations we can measure the effective interactions between macroions. Before we get to complex many-body systems, it is always good practice to start with a simplified setup and check that we find agreement with established results.

The simplest non-trivial macroion interaction in a charged colloidal suspension is an interaction between two macroions. We can measure the interaction of these two macroions by measuring two forces between them at fixed distances.

We fix two macroions with radius a at a distance $r \in [2a, 5a]$ then we use MD to simulate the microions around these macroions en measure the resulting forces on the two macroions. The microions concentration corresponded to an inverse screening length $\kappa a = 0.83$. These forces are measures for and averaged over $N_S = 5000$ distinct microions configurations. Because of the symmetry of the system, we know that the true mean force vectors of the interaction have to be aligned with their separation vector. For this reason we only consider the projected forces along the line going through the macroion centers given by

$$\langle F \rangle = \frac{1}{N_S} \sum_{i=1}^{N_S} \frac{1}{2} (\vec{F}_n^1 - \vec{F}_n^2) \cdot \vec{R}_{12}, \tag{4.1}$$

where $\vec{F}_n^1$ and $\vec{F}_n^2$ are the forces on the two particles for a microion configuration $\vec{r}_n^{N_\mu}$ for $n = 1, \ldots, N_S$ and $\vec{R}_{12} = \vec{R}_1 - \vec{R}_2$ is the separation vector between the two particles. We also have to take into account the macroion-macroion WCA interaction between the two macroions at short distances by subtracting it from the mean force measurements of the CPM.

The relation between distance and the mean force between the two macroions is shown in Figure 4.1. In this figure, we see that fitting forces according to a Yukawa potential from Eq. (2.42) yields good agreement with the MD measurements. Especially beyond the range of the WCA potential $r/a > 2^{7/6}$ the prediction is very good. At shorter distances we observe an additional attractive interaction that deviates from the Yukawa interaction. This attraction might be due to a depletion force caused by the reduced exclusion of microions when the macroions are close [42]. However, relative to the WCA force itself this attraction is orders of magnitude smaller, which makes the macroions still purely repulsive at this range.
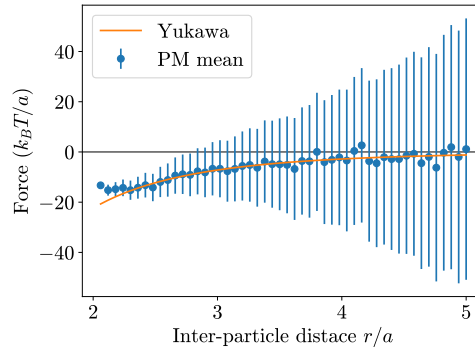
Figure 4.1: The force magnitude between two macroions measured as function of inter-particle distance. Forces were fit using Yukawa pair interaction defined in Eq. (2.42).

4.1.2 Many-body interactions

Single Density

The question now is if such a pair interaction extrapolates to systems with more particles. Let us consider a system of $N_M = 32$ macroions at a packing fraction of $\eta = 0.187$ and ion concentration corresponding to an inverse screening length of $\kappa a = 0.83$ as detailed in subsection 3.2.2. We can repeat the same process, by computing the mean forces on the macroions in fixed configurations. There is no single parameter to exhaust all possible configurations of 32 macroions. So instead, we use random sample of macroions configurations to represent the full possibility space. These macroion configurations were picked randomly according to the equilibrium distribution of system (macroions and microion) configurations.

The mean forces were computed from $N_S = 5000$ distinct and uncorrelated microion configurations as described in subsection 3.2.4. The resulting forces can then be fit with a Yukawa pair interaction using the method of non-linear least squares. In Figure 4.2 we show the actual mean forces in the CPM compared to the prediction made by the fitted Yukawa pair interaction. The computed macroion-macroion WCA potential was also subtracted from the mean force measurements of the CPM for these many-body systems. Even for these larger systems the predictions of the Yukawa model are good. Some of the larger magnitude forces get underestimated, but most forces are well approximated, except for some outliers that mostly get overestimated.
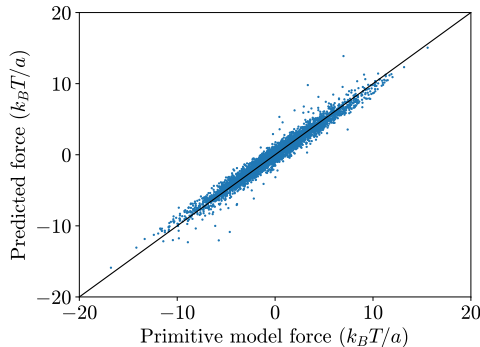
Figure 4.2: The individual force components of each particle in each configuration as measured in the CPM and as predicted by a Yukawa pair interaction plot against each other. System at packing fraction $\eta = 0.187$ and with inverse screening length $\kappa a = 0.83$.

Multi Density

The same method applied in the previous section on single density configurations may be applied to configurations at multiple densities $\eta \in \{0.187, 0.374, 0.561\}$ and for weaker screening $\kappa a \in \{0.25, 0.5\}$ as described in subsection 3.2.3 then we notice a different result. In Figure 4.3 we compare the predictions of a Yukawa interaction with the mean forces measured in the CPM. The Yukawa pair interaction still performs well when the forces are small, but is not able to reproduce the larger forces measured in the CPM. Note that the accuracy of the Yukawa interaction does not depend strongly on the individual densities of the system.

A Yukawa pair interaction describes the two-body and many-body coulomb interactions screened by microion. Given strong microion screening two-body forces are reproduced very well and at low densities many-body interactions are also predicted with fair accuracy. Attempts to extend this description using a Yukawa pair interaction to a system across a range of densities and with weaker microion screening fail. We see that there is no particular density at which it performs as well as for a single density, so it is likely that the Yukawa pair potential does not scale well to multiple densities at once. Another possibility is that the weaker screening is not well-captured by the Yukawa interaction. However, as the performance between the two inverse screening lengths presented in Figure 4.3 is approximately identical, this seems unlikely.
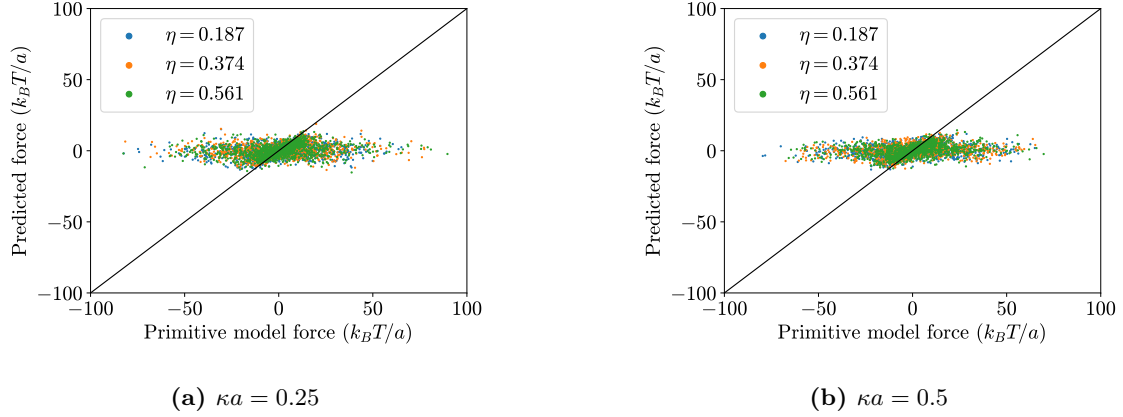
(a) $\kappa a = 0.25$ **(b)** $\kappa a = 0.5$

Figure 4.3: The individual force components of each particle in each configuration as measured in the CPM and as predicted by a Yukawa pair interaction plot against each other. Systems are shown at packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ for two different inverse screening lengths κa.

4.2 Machine Learning Potentials

To address the shortcoming of a Yukawa interaction based description for many-body systems across a density range, we employ a Machine Learning (ML) model based on linear regression of Symmetry Functions (SFs) described in section 3.1.

First we will check if a ML approach is viable for predicting many-body interactions at a single macroion density as was done with the Yukawa model in section 4.1.2. If that succeeds, we will also see if a ML potential does extend to multiple densities, also as in section 4.1.2.

The pool of SFs used to fit was constructed from all parameter combination of $\eta \in \{1/512, 1/256, ..., 16, 32\}$ separated by powers of two, $R_s \in \{0.1, 0.2, ..., 0.8, 0.9\}$ in steps of 0.1, $\xi \in \{1, 2, 4, 8\}$ and $\lambda \in \{1, -1\}$ for a total of $N_{SF} = 288$ SFs in the pool. The cutoff radius for all SFs was set to $R_{\text{cut}} = 5a$. The fitting procedure and SF selection scheme is detailed in section 3.1.
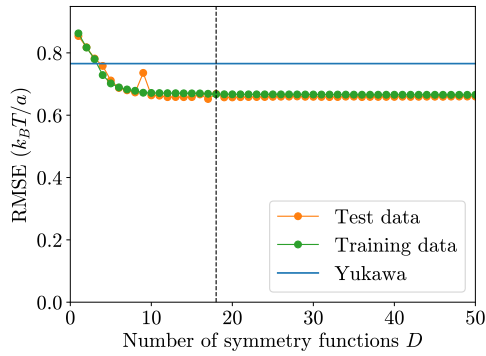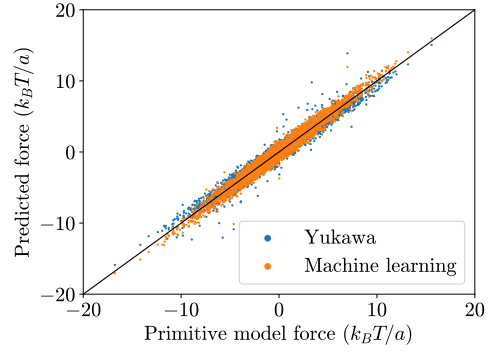
4.2.1 Performance

Single Density

The mean forces that we attempt to predict are measured from MD simulation of the CPM as described in chapter 3 at a fixed macroion packing fraction $\eta = 0.178$ and with inverse screening length of $\kappa a = 0.83$ as detailed in subsection 3.2.2.

In Figure 4.4a we report the performance of the ML potential for the first 50 SFs that are selected. We observe that both the Yukawa potential and the ML potential perform similarly, but the ML potential outperforms the Yukawa potential by a small but significant margin. Only a handful of SF $D = 4$ are needed for the ML to become more accurate than the Yukawa predictions, and beyond $D = 18$ SF we see no significant performance improvement of the ML model.

In Figure 4.4b we see the predictions made by the ML potential at optimal number of SFs $D = 18$. These predictions are compared with the predictions from the Yukawa model. Comparing the two, one sees that the ML model produces much fewer and slightly less severe outliers. A subtle difference that one might spot is that the Yukawa model tends to underestimate large forces while the ML model tends to overestimate larger forces.

(a) The Performance of the ML model as dependent on the number of SFs used to describe the interaction. The optimal number of SFs was determined to be $D = 18$.

(b) The individual force components of each particle in each configuration as measured in the CPM plot against predicted forces by a Yukawa pair interaction and a ML model with $D = 18$ SFs each other.

Figure 4.4: **(a)** Performance and **(b)** predictions of a ML model compared to the Yukawa model for a system at packing fraction $\eta = 0.187$ and with inverse screening length $\kappa a = 0.83$.

Multi Density

We have confirmed that ML potentials, trained on a single macroion density, do have predictive power for charged colloidal suspensions and provide better predictions of forces. Now we move on to apply the same technique to learn the interactions for multiple distinct macroion densities at once.

Again we use the mean forces measured from MD simulation of a CPM at multiple densities $\eta \in \{0.187, 0.374, 0.561\}$ and reservoir inverse screening lengths $\kappa a \in \{0.25, 0.5\}$ as detailed in subsection 3.2.3.

In Figure 4.5 we show the performance progression of the ML potential for both inverse lengths. We observe that the ML potentials both perform much better than the Yukawa model. Just a single SF is already enough to provide a better prediction for the mean forces of the CPM. The performance of the ML potentials at the two inverse screening lengths is near identical. The optimal number of SFs is determined by the last significant decrease in model error. The optimal number of SFs are $D = 21$ and $D = 20$ for $\kappa a = 0.25$ and $\kappa a = 0.5$ respectively.

In Figure 4.6 we set out predictions made with the optimal number of SFs for each inverse screening length against the true mean forces of the CPM. As expected also these predictions by the two ML potentials perform similar between the inverse screening lengths and compared to the predictions Yukawa predictions the ML potentials provide good predictions for both small magnitude forces and large magnitude forces. Still though, it seems that the ML models have the tendency to overestimate forces that are larger in the CPM.

(a) $\kappa a = 0.25$, optimal $D = 21$ **(b)** $\kappa a = 0.5$, optimal $D = 20$
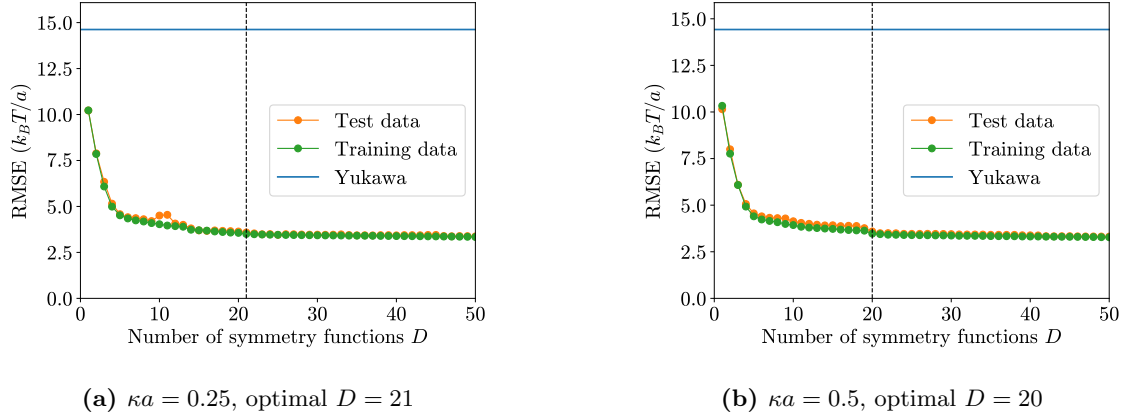
Figure 4.5: The performance of the ML model as dependent on the number of SFs used to describe the interaction for a system at packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ for two different inverse screening lengths κa. The optimal number of SFs D was determined for each inverse screening length.

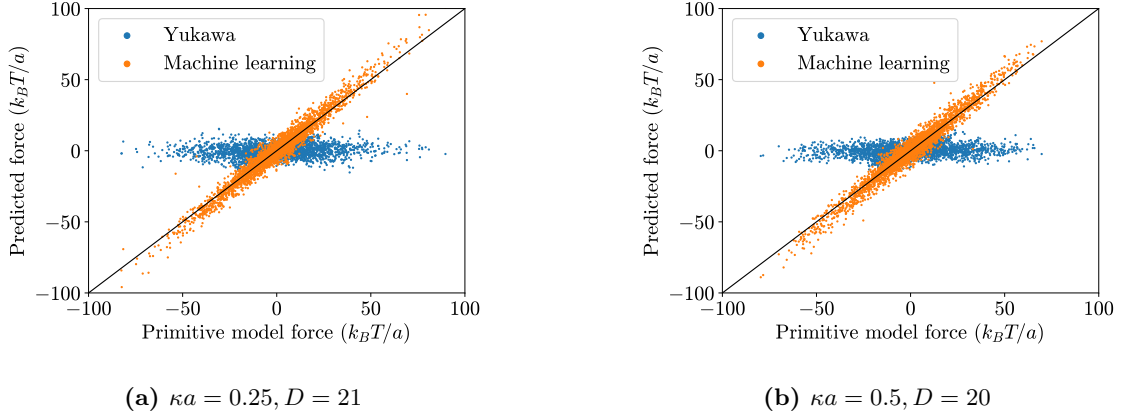**(a)** $\kappa a = 0.25, D = 21$ **(b)** $\kappa a = 0.5, D = 20$

Figure 4.6: The individual force components of each particle in each configuration as measured in the CPM plot against predicted forces by a Yukawa pair interaction and a ML model for a system at packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ for two different inverse screening lengths κa. The optimal number of SFs D was used for each inverse screening length.

4.2.2 Magnitude and alignment comparison

For further insights in the performance of the ML model we shall compare the predicted force vectors. Previously in subsection 4.2.1 we compared individual coordinate components of the forces. While this gives a good feeling for the overall accuracy of the model, it is not coordinate independent and so does not use the symmetries of the system.

We will compare the ML models to the true mean forces of the CPM by using the force magnitude ratio and angle deviation with the CPM force. The magnitude ratio $|\vec{F}|/|\vec{F}_{CPM}|$ is a measure of the accuracy of the scale of the forces, ideally the forces of the model are equal in magnitude to the CPM force and the ratio thus equal to 1. The force alignment $\vec{F} \cdot \vec{F}_{CPM}/|\vec{F}||\vec{F}_{CPM}|$ determines the alignment between the model force and the mean force of the CPM, when the forces are perfectly aligned this value equals 1, when perpendicular 0 and when perfectly anti-aligned -1.

In Figure 4.7, Figure 4.8 and Figure 4.9 we present the distributions of magnitude ratios and force alignments for the single density $\eta = 0.187$ with $\kappa a = 0.83$ and the multi density $\eta \in \{0.187, 0.374, 0.561\}$ at $\kappa a \in \{0.25, 0.5\}$ configurations respectively.

From Figure 4.7a we see that the ML potential produces forces that are more skewed towards the large magnitude end, while the Yukawa model does the opposite. Comparing the two, the Yukawa model predicts many more forces that are much smaller than the true force, while prediction of the ML model have a higher fraction of overestimated forces compared to the Yukawa model. Neither of them has an advantage when it comes to the variance of the magnitude ratios.

Figure 4.7b shows us that there is a slight but no significant difference in the distribution of force alignment between the Yukawa model and the ML model.

In Figure 4.8 we see that for the multi density configuration the skew of the magnitude ratios for the Yukawa model is even larger. A significant portion of the forces predicted by this model are an order of magnitude too small. Again, we also see that the ML model is skewed towards forces that are too large.

Figure 4.9 tells a different story than for the single density configurations. The force alignment of the ML model is much better than the alignment of the Yukawa model. The forces of the Yukawa model are much more randomly oriented for multi density configuration than for single density configurations, while for the ML model this is the other way around. A higher portion of forces is aligned to the CPM force for the multi density configuration.

The predicting of the Yukawa model, when dealing with multiple densities, are worse both is terms of magnitude of the forces and direction of the forces. The ML model on the other hand performs comparably in terms of magnitude prediction for multi density configuration and even becomes more accurate in predicting the direction of the forces in that case.
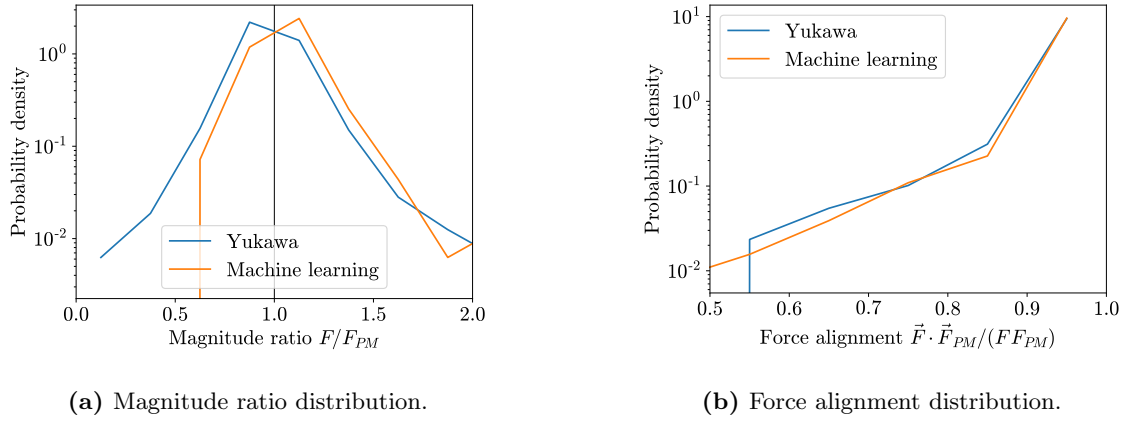
(a) Magnitude ratio distribution.   **(b)** Force alignment distribution.

Figure 4.7: The distributions of **(a)** magnitude ratio and **(b)** force alignment of predicted forces from the Yukawa model and ML model relative to the forces in the CPM. This system has macroion packing fraction $\eta = 0.187$ and inverse screening length $\kappa a = 0.83$.
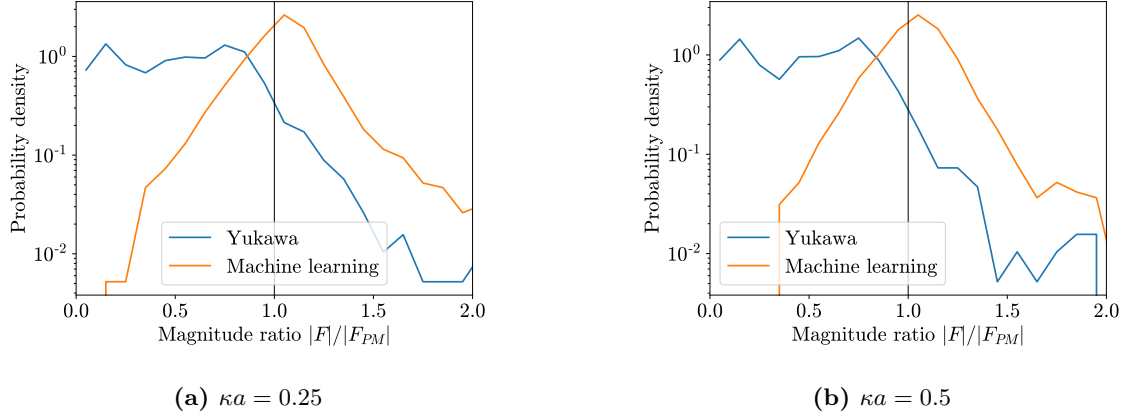
(a) $\kappa a = 0.25$ **(b)** $\kappa a = 0.5$

Figure 4.8: The distributions of magnitude ratio of predicted forces from the Yukawa model and ML model relative to the forces in the CPM. This system has macroion packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ and for two inverse screening lengths κa.
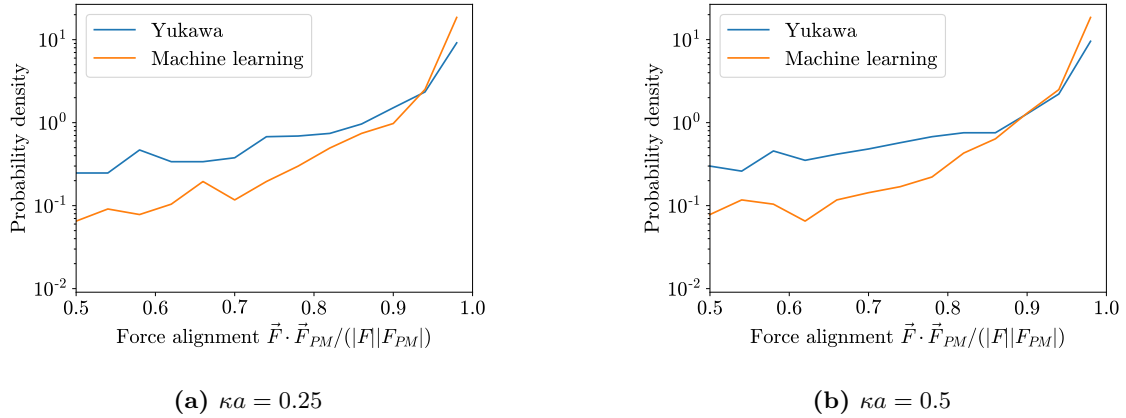

(a) $\kappa a = 0.25$ **(b)** $\kappa a = 0.5$

Figure 4.9: The distributions of angle alignment of predicted forces from the Yukawa model and ML model relative to the forces in the CPM. This system has macroion packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ and for two inverse screening lengths κa.

4.2.3 Performance Dependence

When inspecting the performance progression in Figure 4.4a and Figure 4.5 one observes that beyond a certain number of SFs used to describe the potential the performance of the ML model hardly improves. It seems that there is a set limit to the performance of the ML model. In this section, we discuss some parameters that may influence the performance ceiling.

Averaging Sample Size

One of those parameters would be the sample size N_S of the microion configurations over the mean force is calculated according to subsection 3.2.4. As the saying goes: garbage in, garbage out. So we should not expect to learn the mean forces of the CPM any better than we have approximated them. The Standard Error of the Mean (SEM) of the forces is a measure of the error of a computed mean and the underlying true mean of the sample. It is related to the standard deviation of the forces over all configurations. These Standard Deviations (SD) of the forces are reported in Table 3.3 and the SEM is related to the SD and the sample size N_s by

$$\text{SEM} = \frac{SD}{\sqrt{N_s}}. \tag{4.2}$$

This means we can reduce this error by increasing the sample size and potentially get better performance.

In Figure 4.10a we show the dependence of the performance on the sample size N_s and we see that the performance increases as a function of N_s, however as it only decreases with the square root of the sample the performance gains become smaller and smaller. In Figure 4.10b we instead show the performance as a function of the SEM. From this figure we see that the performance is indeed strongly related to the SEM, and this dependence becomes stronger as the SEM decreases.

Strictly speaking, Eq. (4.2) is incorrect for the single density data that we use for this discussion as these configurations are not entirely uncorrelated, meaning that there is some duplicate information. This equation holds under the assumption that all sample are uncorrelated, when they are correlated the SEM decreases even slower as a function of N_s.
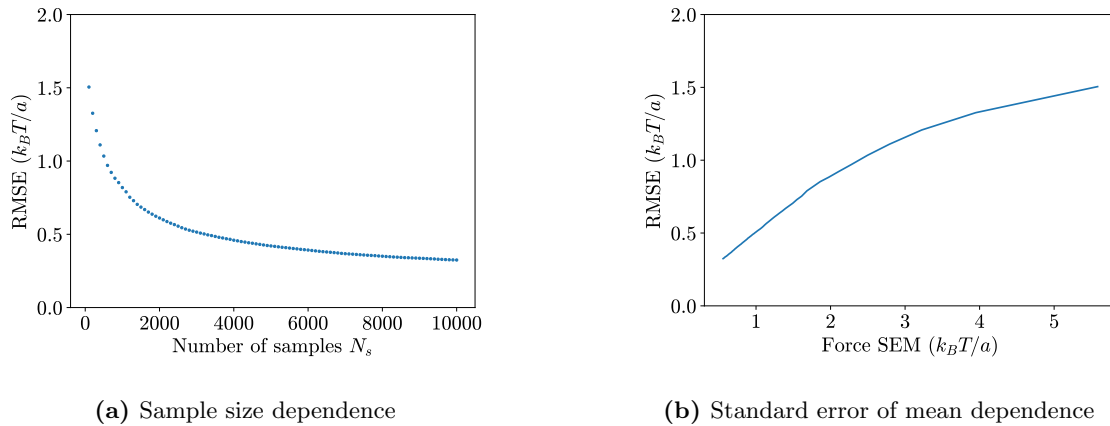

(a) Sample size dependence

(b) Standard error of mean dependence

Figure 4.10: Root-mean-square error of the ML model as function of **(a)** the force averaging sample size N_s of microion configurations and **(b)** standard error of the force mean. The ML model uses $D = 8$ SFs.

Cutoff radius

There are other factors that could influence the performance of the ML model besides the sample size, such as the cutoff radius R_{cut} of the SFs. For the cutoff radius we do not expect the performance to increase indefinitely as it increases. As the effective interactions between the macroions should be of finite range, or at least they should decay very rapidly with distance, it should only be a matter of picking a large enough cutoff radius R_{cut}.

Symmetry function pool

Another factor concerning the SFs is the set of parameters that determines the pool from which the SFs are selected. Logically, one should expect better performance when the SFs in the pool better represent the interaction. Although this probably only holds for a small number D of selected SFs. As more SFs are selected, difference between SFs may also be used to represent the data. With only a couple of SFs the possible combination become numerous very quickly and even initially unrepresentative SFs may become valuable for describing the interaction. When picking parameters for the pool of SFs we should ensure that at least the right length scales of the parameters are present amidst the parameter values.

4.3 Result Validation

4.3.1 Radial distribution functions

We have used testing data to verify the accuracy of our trained ML potentials. Good performance of the ML model on the testing data does not ensure however that the resulting ML potential accurately describes the original system.

We use the ML potentials with the optimal number of SFs to perform Monte Carlo (MC) simulations of the one-component macroion system. From these MC simulations we compute the macroion-macroion Radial Distribution Functions (RDFs) and compare those with the RDFs computed from MD simulation of the full CPM. All RDFs are computed from 1000 configuration of 358 macroions at packing fraction $\eta = 0.187$. As the ML potentials do not learn the direct WCA interactions between macroions, we have the possibility to substitute the WCA potential with another repulsive potential such as a Hard-Sphere (HS) potential.

Single Density

First, we will check the validity of the ML potential trained on configuration of a single density as described in subsection 3.2.2. This ML model was trained of configurations with packing fraction $\eta = 0.187$ at inverse screening lengths $\kappa a = 0.83$. The optimal number of SFs was determined in subsection 4.2.1 and the corresponding set of SFs was used to compute the RDFs of the same system at density $\eta = 0.187$ and $\kappa a = 0.3$. The MC simulation was performed with a WCA potential as short range repulsion.

In Figure 4.11 we show the RDF computed by the CPM and with the ML model. It is clear to see that the ML model is not able to reproduce the result from the full CPM. Although the ML potential has many repulsive components in the set of SFs there is a strong attraction between the macroions. This attraction causes clustering of the particles, which in turn slows down the equilibration of the system.

This unstable behavior is likely to be caused by a lack of data. As the ML model was only trained on a single low macroion density, the model was only fit on inter-particle distances that are relatively large. The ML potential had no examples of forces for particles close together, and thus has to extrapolate to these distances. Clearly this extrapolation was not successful and resulted in a strong short range attraction.
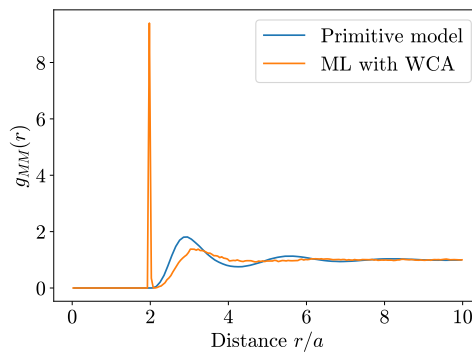
Figure 4.11: Macroion-macroion radial distribution function $G_{MM}(r)$ of a system at packing fraction $\eta = 0.187$ and inverse screening length $\kappa a = 0.83$ computed with the CPM and using a ML model described in section 4.2.1.

Multi Density

Next we will check the validity of the ML potential trained on configuration of multiple densities as described in subsection 3.2.3. Recall that the ML model was trained on multi density configurations at macroion packing fraction $\eta \in \{0.187, 0.374, 0.561\}$ and inverse screening lengths $\kappa a \in \{0.25, 0.5\}$. The optimal sets of SFs defined in subsection 4.2.1 were used to compute the RDFs of a system at density $\eta = 0.187$ for both inverse screening lengths. In these MC simulations we use both the WCA and HS potential as short-range repulsion.

In Figure 4.12 we compare RDFs from different simulations and potentials. The RDFs for the ML potentials are close to the true RDF of the full CPM, although the peak height are less pronounced and the peak positions are slightly shifted. The first peak is shifted toward the center of the macroion, while all other peaks are shifted away from the macroion.

For the RDF corresponding to the ML potentials with a HS repulsion we observe an additional small peak near the macroion surface. This peak might be due to a depletion effect, similar to the attraction found in subsection 4.1.1. Although in this case it might also be a lack of training data with particle very close together, meaning that the ML potential is not trained on this short range in the interaction. As we see, using the WCA potential this peak disappears, which implies that the effect is weaker than the WCA interaction itself.

(a) $\kappa a = 0.25$

(b) $\kappa a = 0.5$
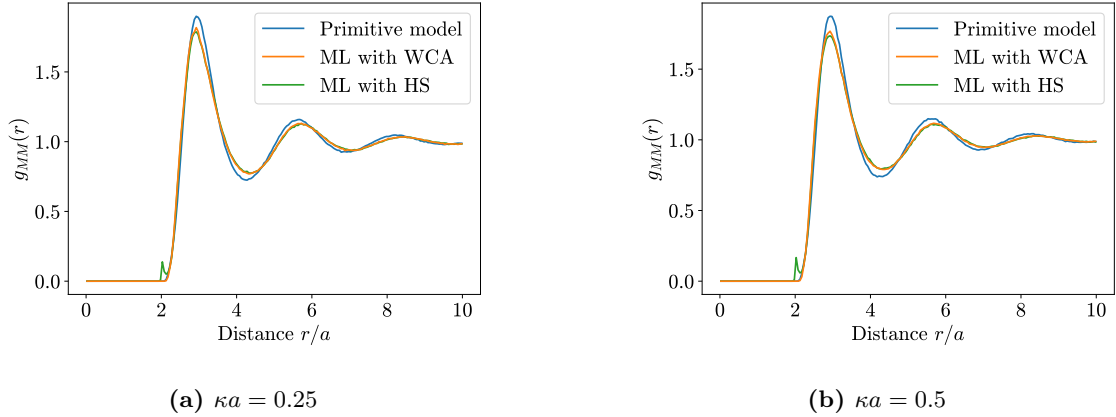
Figure 4.12: Macroion-macroion radial distribution functions $g_{MM}(r)$ for system at packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ for two distinct inverse screening lengths. Computed from configuration from a MD simulation of the CPM and MC simulations of the ML models described in section 4.2.1.

4.3.2 Computation times

The goal of finding an effective macroion interaction to describe charged colloidal suspension is not only for them to be accurate but also computationally efficient.

To compare the computational costs of the full CPM as a baseline versus the ML models from section 4.2 we compute the total amount of core-time (time × number of cores) each model would take to run to generate the 1000 configurations for the RDFs from subsection 4.3.1. This means we are comparing the computing times of a MD simulation of the CPM and a MC simulation of the ML model. The total computation times of each model are reported in Table 4.1. We see that all ML models were faster than the MD methods, but the speedup varies.

For the single density simulation the MC simulation were significantly faster than the MD simulations, however we should add some nuance here. As we saw in subsection 4.3.1 the ML potential obtained from the single density configuration did not reproduce a physical result. This ML potential resulted in heavy clustering, which greatly reduces the possible moves the MC could undertake each step. For an accurate RDF of this system we would then need to perform more MC moves, which would increase the computational times. Furthermore, the MD simulation of the CPM in this case was performed on different hardware than the MD simulation for the multi density configurations. This hardware is less efficient in parallel computing, on the more efficient hardware we could expect a $2 - 3\times$ speedup.

47

For the multi density simulation, we see that the MC methods are still significantly faster than the MD simulations. We also see that the scaling with screening is not as pronounced for the ML model as for the CPM. Stronger screening in the CPM model means more microions which increases computational times, while for the ML model it resulted only in one extra SF added in the optimal set. We expect the ML model to become even more efficient for a system with stronger screening. For weak screening the lower limit would be $\kappa a = 0$, meaning not additional microions are present beyond the counterions balancing the macroion charges. This scenario is already very close to $\kappa a = 0.25$ in terms of particles number and would never result in a speedup of more than 7.6 times for the molecular dynamics simulation of the CPM. So even if we assume that the ML model does not become faster as well for $\kappa a = 0$, the ML model will be faster than the CPM simulation. Extending this reasoning, we can state that a ML model will always be faster computationally than a full CPM simulation for any inverse screening length at macroion packing fraction $\eta = 0.187$.

Nonetheless, there might be situations in which ML models might not outperform a MD simulation of the CPM. The computing times of the ML models depend not only on the number of SFs in the model, but also on the cutoff radius, which is set to $R_{\mathrm{cut}} = 2.5$ for these models. Increasing this cutoff radius quickly increases the number of interactions that need to be taken into account, and thus increases run times. The same hold for higher densities, which also increases the number of interactions to consider.

Another note to consider is that the matchup of MC simulations and MD simulation is inherently unfair to the MC simulations. MD are often more efficient and can utilize multiple core. MC simulations are not as efficient, because they require tuning of the step parameters and do not lend themselves well for parallel computing. This means that if we would implement the ML model into a MD simulation, this would result in even faster computation. For the MD engine LAMMPS used in this thesis it would be possible to implement these ML model by converting them into tabulated pair and three-body interactions using the `table` and `threebody/table` Pair Styles, respectively. The tabulation would come with the benefit that the computation times become independent of the number of SFs as the values in the table are all precomputed.

system	ML Monte Carlo	CPM Molecular dynamics	speedup
single: $\kappa a = 0.83$	24.3 ch	7058 ch	290 $\times$
multi: $\kappa a = 0.5$	66.5 ch	689.9 ch	10.4 $\times$
multi: $\kappa a = 0.25$	60.0 ch	457.9 ch	7.6 $\times$

Table 4.1: Computation times for RDFs of a system at packing fraction $\eta = 0.187$ for several screening lengths in core-hours (ch).

Chapter 5

Conclusion and Outlook

5.1 Summary and Conclusion

The goal of this thesis was to use Machine Learning (ML) to find an effective interaction to describe the many-body interactions between charged colloidal particles in an electrolyte. Such a description of a one-component colloidal system could allow for much faster yet accurate simulations of these charged colloidal suspensions.

Charged colloidal suspensions are systems of large charged colloidal particles in a salt solution. We can describe this system using the Primitive Model (PM) in which we treat all ions as charged hard spheres interacting via a Coulomb interaction in a dielectric medium. Poisson-Boltzmann theory predicts the formation of electric double-layers around the colloidal particles in a charged colloidal suspension. The interactions between these double-layers determine the phase properties of the system. The DLVO theory describes the effective interactions due to the electric double-layers between two colloidal particles as a screened Coulomb or Yukawa potential. This theory breaks down when overlaps between more than two electric double-layers become relevant. This is the case for high colloid densities, high colloid charge and low salt concentrations in the electrolyte, as they result in double-layer sizes larger than the distances between the colloids.

In this thesis, we present a ML approach to learn the effective many-body interactions between colloidal particles that the DLVO theory fails to describe. This ML method relies on linear regression using Symmetry Functions (SFs). These SFs are families of functions that can be used to model the interaction between spherical particles. The ML model was trained on explicit Molecular Dynamics (MD) simulations of the PM by fitting SFs to the measured mean forces on the colloids using linear regression.

The simulated system has a charge ratio of 90:1 and size ratio 20:1 between colloids and salt ions. The medium was modeled with a Bjerrum length to colloid size ratio $\lambda_B/a = 0.0098$. The hard-sphere interaction of the model were approximated using a WCA potential with a strength of $\beta\epsilon_{\mathrm{WCA}} = 10$. This system was simulation at colloids packing fractions $\eta \in \{0.187, 0.374, 0.561\}$ and salt concentrations corresponding to inverse screening lengths $\kappa a \in \{0.25, 0.5, 0.83\}$.

Our results confirm that a conventional approach based on the DLVO theory describes charged colloidal suspensions at low densities and strong screening. However, a single Yukawa pair potential is not able to describe the interaction of the system at various colloid densities.

Our ML approach was able to find a model that described the effective interactions using a single potential. When trained on configurations across a range of colloid densities, we were able to validate the correctness of these ML potentials by reproducing radial distribution functions of the PM with a Monte Carlo (MC) implementation of the ML model. ML models trained on configurations at only a single density were not able to extrapolate and did not give physical simulation results. The question remains if these ML models can extrapolate and interpolate beyond the colloid densities they were trained on.

Furthermore, it has been shown that the ML models are computationally more favorable compared to the full PM simulations of charged colloidal suspensions. The computation advantage of the ML model over the PM may be further increased by utilizing a MD implementation of these ML potentials instead of the MC implementation presented in this thesis.

In conclusion, we have presented a procedure for obtaining an effective one-component description of colloidal particles in a charged colloidal suspension using machine learning. These machine learning models enable us to accurately and cheaply reproduce simulations of a many-component PM description of the charged colloidal suspension.

5.2 Discussion and Outlook

Before we present an overview of suggested research directions, we first reflect on the methods used in this thesis and point out any known shortcomings and possible improvements.

In section 3.2.4 we alluded to the fact that the fluctuations of the forces are important in addition to the mean of the forces, as they are used to quantify the error or uncertainty in the mean forces. In Figure 4.10b we also saw that this error has a pronounced effect on the performance of a ML model. Table 3.3 shows that the fluctuations in turn depend on the macroion packing fraction η, meaning that ML models trained on multiple densities have input forces with varying uncertainty, so-called heteroscedasticity. In this thesis, we have not dealt with this issue. One possible solution would be to equalize the uncertainties by varying the force sample size N_S for different densities. Another would be to account for this during Linear Regression by implementing Weighted Linear Regression. This last approach is considerably easier and more general.

In addition to implementing Weighted Linear Regression in our ML method, we may also employ k-fold cross-validation to get a more complete picture of the performance of a ML model on the input data. k-fold cross-validation uses all input data as both training and testing data in k batches, as opposed to the 80%-20% data split. This utilizes the input data more efficiently at the cost of a longer computing time for performing the ML fit.

In subsection 4.1.1 and subsection 4.3.1 we mentioned the possibility of a significant depletion force on the macroions caused by the excluded volume of the macroions in the presence of the finite-sized microions. In order to check if this is indeed the case, additional research is required into the depletion effect for this specific implementation of the WCA potential. In which one would determine the range and the magnitude of the depletion effect and check if it fits our results. In physical systems such a depletion effect would be much weaker as the size ratio between macroions and microions is at least an order of magnitude smaller for the simulations in this thesis compared to real charged colloidal suspensions

In subsection 4.3.2 we already alluded to the possibility of a MD implementation of the ML models presented in this thesis. In the same section, we briefly discuss the method of implementation. It is difficult to predict beforehand how much of a speed differential this would imply for simulating with the ML potentials, but it is certainly worth exploring.

In section 3.2 we describe a method for generating training data used for training a ML model. In Figure 4.10b we have also found that the performance of this model is strongly dependent on the error of the mean forces. Using the method described in section 3.2 of this thesis, it becomes unpractical to reduce this error, as this would require a quadratic increase in force sample size N_S. Another method that does not suffer from this dependence of a sample size would be to compute the forces using numerical DFT methods. These methods are very computationally expensive but give very accurate results.

In subsection 4.1.2 we present the prediction of a Yukawa model for both configuration at a single density and at multiple densities at once. Unfortunately, due to the altering of 2 parameters at once, namely macroion packing fractions η and inverse screening length κa, between these systems we cannot determine what causes the Yukawa description to fail for multiple densities. In the same section, we do discuss that the introduction of multiple densities is the likely cause. One could check this by performing fits of a Yukawa interaction on the separate densities of from the multi density configurations. If the multiple densities do not turn out to be the cause, then we would investigate the effect of the inverse screening length κa further.

In section 4.3 we present RDFs computed using the ML models we obtained from training data of CPM simulations. This RDF was computed for macroion packing fraction $\eta = 0.187$ that was also used as training data for the model. In other words the ML model was explicitly trained on similar configurations and to further validate the ML models we would like to check if it can interpolate and extrapolate to packing fractions η it was not trained on.

In subsection 3.2.3 we pointed out that in the particular implementation of the GCMC simulation cation-anion pairs had to be inserted with fixed distance d between them. This is not an ideal implementation as it introduced another parameter d that influences the final microion concentrations. Preferably, one should try a method that allows for microion pairs to be inserted randomly. For example, using a custom MC simulation, although that would come ate the cost of even longer computation times.

The parameters set in section 4.2 are an attempt to create a diverse set of SFs that cover the possible interaction that arise in charged colloidal suspension. Of course there is always room for improvement and finding a parameter set for η, R_s, ξ and λ that is small but still performs well when fitting the data is no easy task. Employing more parameters also does not imply better performance, experimentation of different parameter sets proved the contrary. It would be beneficial for these kinds of ML models to investigate the performance dependent on the input parameters and come up with some method for constructing high-performing parameters sets.

Besides experimenting with those parameters, it would be useful to investigate the performance dependence on the cutoff radius R_{cut}. As this parameter also has a large impact on computation times as discussed in subsection 4.3.2, finding an efficient yet performant cutoff radius for these ML models would certainly increase their value.

Beyond these possible improvements on and additions to the methods used in this thesis, we will also give an outlook on extensions and other application of this ML method for finding effective interactions.

This thesis functions somewhat like a proof of concept for using ML methods to find effective interaction for charged colloidal suspensions. The systems we used to show this were still on the conservative side. To explore the boundaries of this method, we would have to consider systems with higher macroion valence Z with respect to the Bjerrum length λ_B/a and also a wider range of inverse screening lengths κa.

As we have shown, spherical charged colloids can be described using ML potentials. We could repeat the same procedure with some alterations for aspherical charged colloids, such as rod-like, ellipsoid or disc particles. The current families of SF used in this thesis are not suitable to deal with aspherical particles, as they were designed for spherically symmetric particles. These break some symmetries of the system of spherical particles, namely rotation invariant and particle exchange invariant. Different families of SF that respect the intrinsic symmetries of the systems should be used to find a ML model for aspherical charged colloids

Bibliography

[1]   Robert Brown. "XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies". In: *The Philosophical Magazine* 4.21 (Sept. 1828). Publisher: Taylor & Francis, pp. 161–173. ISSN: 1941-5850. DOI: 10.1080/14786442808674769.

[2]   A. Einstein. "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen". en. In: *Annalen der Physik* 322.8 (1905), pp. 549–560. ISSN: 1521-3889. DOI: 10.1002/andp.19053220806.

[3]   F. London. "The general theory of molecular forces". en. In: *Transactions of the Faraday Society* 33 (1937), 8b. ISSN: 0014-7672. DOI: 10.1039/tf937330008b.

[4]   H. C. Hamaker. "The London—van der Waals attraction between spherical particles". en. In: *Physica* 4.10 (Oct. 1937), pp. 1058–1072. ISSN: 0031-8914. DOI: 10.1016/S0031-8914(37)80203-7.

[5]   D. H Napper. "Steric stabilization". en. In: *Journal of Colloid and Interface Science* 58.2 (Feb. 1977), pp. 390–407. ISSN: 0021-9797. DOI: 10.1016/0021-9797(77)90150-3.

[6]   H. C. Hamaker. "A general theory of lyophobic colloïds. I". en. In: *Recueil des Travaux Chimiques des Pays-Bas* 55.12 (1936). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/recl.19360551206, pp. 1015–1026. ISSN: 0165-0513. DOI: 10.1002/recl.19360551206.

[7]   Thomas W. Healy and Lee R. White. "Ionizable surface group models of aqueous interfaces". en. In: *Advances in Colloid and Interface Science* 9.4 (June 1978), pp. 303–345. ISSN: 0001-8686. DOI: 10.1016/0001-8686(78)85002-7.

[8]   B Derjaguin and L Landau. "Theory of the stability of strongly charged lyophobic sols and of the adhesion of strongly charged particles in solutions of electrolytes". en. In: *Progress in Surface Science* 43.1 (May 1993), pp. 30–59. ISSN: 0079-6816. DOI: 10.1016/0079-6816(93)90013-L.

[9]   E. J. W. Verwey. "Theory of the Stability of Lyophobic Colloids." In: *The Journal of Physical and Colloid Chemistry* 51.3 (Mar. 1947), pp. 631–636. ISSN: 0092-7023. DOI: 10.1021/j150453a001.

[10]   S. Alexander et al. "Charge renormalization, osmotic pressure, and bulk modulus of colloidal crystals: Theory". In: *The Journal of Chemical Physics* 80.11 (June 1984). Publisher: American Institute of Physics, pp. 5776–5781. ISSN: 0021-9606. DOI: 10.1063/1.446600.

[11]   Mark. O. Robbins, Kurt Kremer, and Gary S. Grest. "Phase diagram and dynamics of Yukawa systems". In: *The Journal of Chemical Physics* 88.5 (Mar. 1988). Publisher: American Institute of Physics, pp. 3286–3312. ISSN: 0021-9606. DOI: 10.1063/1.453924.

[12]   E. B. Sirota et al. "Complete phase diagram of a charged colloidal system: A synchro- tron x-ray scattering study". In: *Physical Review Letters* 62.13 (1989). Publisher: American Physical Society, pp. 1524–1527. DOI: 10.1103/PhysRevLett.62.1524.

[13]   B. V. R. Tata et al. "Amorphous Clustering in Highly Charged Dilute Poly(chlorostyrene-styrene sulfonate) Colloids". In: *Physical Review Letters* 78.13 (Mar. 1997). Publisher: American Physical Society, pp. 2660–2663. DOI: 10.1103/PhysRevLett.78.2660.

[14]   B. V. R. Tata, M. Rajalakshmi, and Akhilesh K. Arora. "Vapor-liquid condensation in charged colloidal suspensions". In: *Physical Review Letters* 69.26 (Dec. 1992). Publisher: American Physical Society, pp. 3778–3781. DOI: 10.1103/PhysRevLett.69.3778.

[15]   B. V. R. Tata, M. Rajalakshmi, and Akhilesh K. Arora. "Vapor-Liquid Condensation in Charged Colloidal Suspensions". In: *Physical Review Letters* 70.18 (May 1993). Publisher: American Physical Society, pp. 2823–2823. DOI: 10.1103/PhysRevLett.70.2823.

[16]   Per Linse and Vladimir Lobaskin. "Electrostatic Attraction and Phase Separation in Solutions of Like-Charged Colloidal Particles". In: *Physical Review Letters* 83.20 (Nov. 1999), pp. 4208–4211. DOI: 10.1103/PhysRevLett.83.4208.

[17]   Per Linse and Vladimir Lobaskin. "Electrostatic attraction and phase separation in solutions of like-charged colloidal particles". In: *The Journal of Chemical Physics* 112.8 (Feb. 2000). Publisher: American Institute of Physics, pp. 3917–3927. ISSN: 0021-9606. DOI: 10.1063/1.480943.

[18]   Bryan Beresford-Smith, Derek Y. C Chan, and D. John Mitchell. "The electrostatic interaction in colloidal systems with low added electrolyte". en. In: *Journal of Colloid and Interface Science* 105.1 (May 1985), pp. 216–234. ISSN: 0021-9797. DOI: 10.1016/0021-9797(85)90362-5.

[19]   René van Roij and Jean-Pierre Hansen. "Van der Waals–Like Instability in Suspensions of Mutually Repelling Charged Colloids". In: *Physical Review Letters* 79.16 (Oct. 1997). Publisher: American Physical Society, pp. 3082–3085. DOI: 10.1103/PhysRevLett.79.3082.

[20]   René van Roij, Marjolein Dijkstra, and Jean-Pierre Hansen. "Phase diagram of charge-stabilized colloidal suspensions: van der Waals instability without attractive forces". en. In: *Physical Review E* 59.2 (Feb. 1999), pp. 2010–2025. ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.59.2010.

[21]   René van Roij and Robert Evans. "Phase equilibria in a model of low-salt suspensions of charged colloids". en. In: *Journal of Physics: Condensed Matter* 11.50 (Dec. 1999). Publisher: IOP Publishing, pp. 10047–10060. ISSN: 0953-8984. DOI: 10.1088/0953-8984/11/50/301.

[22]   C. Russ et al. "Three-body forces between charged colloidal particles". In: *Physical Review E* 66.1 (July 2002). Publisher: American Physical Society, p. 011402. DOI: 10.1103/PhysRevE.66.011402.

[23]   Matthias Brunner et al. "Direct Measurement of Three-Body Interactions amongst Charged Colloids". In: *Physical Review Letters* 92.7 (Feb. 2004). Publisher: American Physical Society, p. 078301. DOI: 10.1103/PhysRevLett.92.078301.

[24] Gerardo Campos-Villalobos et al. "Machine learning many-body potentials for colloidal systems". In: *The Journal of Chemical Physics* 155.17 (Nov. 2021), p. 174902. ISSN: 0021-9606. DOI: 10.1063/5.0063377.

[25] M. Gouy. "Sur la constitution de la charge électrique à la surface d'un électrolyte". fr. In: *Journal de Physique Théorique et Appliquée* 9.1 (1910), pp. 457–468. ISSN: 0368-3893, 2507-6485. DOI: 10.1051/jphystap:019100090045700.

[26] David Leonard Chapman. "LI. A contribution to the theory of electrocapillarity". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 25.148 (Apr. 1913), pp. 475–481. ISSN: 1941-5982. DOI: 10.1080/14786440408634187.

[27] Bas Zoetekouw and René van Roij. "Volume terms for charged colloids: A grand-canonical treatment". In: *Physical Review E* 73.2 (Feb. 2006). Publisher: American Physical Society, p. 021403. DOI: 10.1103/PhysRevE.73.021403.

[28] Daan Frenkel and Berend Smit. *Understanding molecular simulation : from algorithms to applications. 2nd ed.* Vol. 50. Journal Abbreviation: Physics Today Publication Title: Physics Today. Jan. 1996. DOI: 10.1063/1.881812.

[29] R. W. Hockney and J. W. Eastwood. "The Particle-Mesh Force Calculation". In: *Computer Simulation Using Particles*. CRC Press, 1988. ISBN: 9780367806934.

[30] Loup Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Physical Review* 159.1 (July 1967). Publisher: American Physical Society, pp. 98–103. DOI: 10.1103/PhysRev.159.98.

[31] W. C. Swope et al. "A computer simulation method for the calculation of equilibrium constants for the formation of physi". In: 1981. DOI: 10.1063/1.442716.

[32] Hans C. Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of Chemical Physics* 72.4 (Feb. 1980). Publisher: American Institute of Physics, pp. 2384–2393. ISSN: 0021-9606. DOI: 10.1063/1.439486.

[33] William G. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". In: *Physical Review A* 31.3 (Mar. 1985). Publisher: American Physical Society, pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.

[34] G. Martyna, M. Klein, and M. Tuckerman. "Nosé-Hoover chains : the canonical ensemble via continuous dynamics". In: (1992). DOI: 10.1063/1.463940.

[35] Aidan P. Thompson et al. "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales". en. In: *Computer Physics Communications* 271 (Feb. 2022), p. 108171. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2021.108171.

[36] Hossein Eslami and Florian Müller-Plathe. "Molecular dynamics simulation in the grand canonical ensemble". en. In: *Journal of Computational Chemistry* 28.10 (2007), pp. 1763–1773. ISSN: 1096-987X. DOI: 10.1002/jcc.20689.

[37] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (June 1953). Publisher: American Institute of Physics, pp. 1087–1092. ISSN: 0021-9606. DOI: 10.1063/1.1699114.

[38] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". In: *The Journal of Chemical Physics* 134.7 (Feb. 2011). Publisher: American Institute of Physics, p. 074106. ISSN: 0021-9606. DOI: 10.1063/1.3553717.

[39]   Andreas Singraber, Jörg Behler, and Christoph Dellago. "Library-Based LAMMPS Implementation of High-Dimensional Neural Network Potentials". In: *Journal of Chemical Theory and Computation* 15.3 (Mar. 2019). Publisher: American Chemical Society, pp. 1827–1840. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.8b00770.

[40]   C. Patrick Royall et al. "Re-entrant melting and freezing in a model system of charged colloids". en. In: *The Journal of Chemical Physics* 124.24 (June 2006), p. 244706. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2189850.

[41]   John D. Weeks, David Chandler, and Hans C. Andersen. "Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids". In: *The Journal of Chemical Physics* 54.12 (June 1971), pp. 5237–5247. ISSN: 0021-9606. DOI: 10.1063/1.1674820.

[42]   Xinyi Shen and Ian C. Bourg. "Molecular dynamics simulations of the colloidal interaction between smectite clay nanoparticles in liquid water". en. In: *Journal of Colloid and Interface Science* 584 (Feb. 2021), pp. 610–621. ISSN: 00219797. DOI: 10.1016/j.jcis.2020.10.029.