



**FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA**

Physikalisch-Astronomische Fakultät

# **Computer aided design of nanoparticulate polymeric drug carriers**

DISSERTATION

*by*

M.Sc. Mingzhe Chi

born on February 28, 1992 in Xuzhou, P. R. China

*Submitted to*

Faculty of Physics and Astronomy

Friedrich Schiller University Jena

To fulfill the requirements for the degree of  
*doctor rerum naturalium (Dr. rer. nat.)*

15 December 2023

## **Reviewers**

1. Prof. Dr. Marek Sierka, Friedrich Schiller University Jena
2. Prof. Dr. Adam Kubas, Polish Academy of Sciences
3. Prof. Dr. Stefanie Gräfe, Friedrich Schiller University Jena

**Submitted on: 31.01.2024**

**Disputation date: 04.06.2024**

# Acknowledgements

---

I am deeply grateful to my family for their continuous love, encouragement, and emotional support. Their unwavering belief in me has been a constant source of motivation during this challenging PhD journey.

I sincerely thank my supervisor, Prof. Dr. Marek Sierka, for his continuous guidance, expertise, and support throughout my research. His valuable insights and advice have played a pivotal role in shaping this thesis and the related research papers. When I needed help with my research, he always made time to help in any way he could. Additionally, I am grateful for his assistance with administrative tasks and the provision of a comfortable work environment that fostered creativity and productivity. I also greatly appreciate his willingness to share his research and work experience, which was invaluable information for a foreign student in Germany.

I would like to express my appreciation to Mr. Lutz Neumann for his technical assistance and troubleshooting support. Additionally, I want to acknowledge the secretaries of the Otto Schott Institute of Materials Research, Ms. Langer, Ms. Schweter and Ms. Partschefeld, for their assistance with administrative matters like conference applications and poster printing.

I am thankful to my amazing colleagues - Carolin Müller, Martin Becker, Ghada Belhadj-Hassine, Tim Schrader, Ya-Fan Chen, Felix Arendt, Jamoliddin Khanifaev, Antonia Weber, and Kira Klebesz. Their support and insights during the group seminars have been invaluable, making this journey both intellectually stimulating and enjoyable.

I would like to give a special thanks to Tim Schrader, Rihab Gargouri, Yingfeng Teng, Chenjie Zhu, Antonia Weber and Ya-Fan Chen. Tim, Rihab, Yingfeng, and Chenjie helped me in collecting simulation data and early machine learning research

(Chapter I). They did a great job and completed their thesis based on it. Antonia helped me simulate some complex polymers, which was challenging, but she succeeded in the end. Ya-Fan and I worked together to build a polymer database for machine learning, which will be very helpful not only for me, but for future projects.

I want to acknowledge Professor Dr. Ulrich S. Schubert and his group member Dr. Christine Weber, Dr. Antje Vollrath and Dr. Mira Behnke. Part of my work (Chapter II) would have been nearly impossible without their expertise in experiments and the data they provided. Moreover, I would like to thank Dr. Irina Muljajew and Ms. Natalie Göppert for their excellent work and valuable polymer data. They were members of Prof. Schubert's group, and it was an honor to work with them.

Additionally, I would like to thank Dr. Wanling Foo from the faculty of medicine, University Jena. Our collaboration on polymer micelles for encapsulating siRNA has been successful (Chapter III), and I have been impressed by the many interesting discoveries that have come out of the work.

I appreciate the reviewers for their time and efforts in assessing my thesis and providing constructive feedback.

Finally, I acknowledge the financial support from DFG SFB 1287 PolyTarget. This support has also enabled me to participate in many conferences and scientific events, which helped me to enhance the quality of my research and connect with peers as well as the leading experts in the field.

I sincerely thank everyone mentioned above for their invaluable contributions and unwavering support throughout my PhD journey.

# Table of contents

---

<b>Acknowledgements .....</b>	<b>i</b>
<b>Abstract.....</b>	<b>1</b>
<b>Zusammenfassung .....</b>	<b>3</b>
<b>Introduction.....</b>	<b>6</b>
<b>Theory and methodology.....</b>	<b>12</b>
1 Flory-Huggins parameter.....	12
2 Hildebrand solubility parameter .....	14
3 Computer simulation for material science.....	17
3.1 Quantum mechanics simulation.....	17
3.2 Molecular dynamics simulation.....	18
3.3 Dissipative particles dynamic simulation.....	20
3.4 Machine learning .....	23
<b>MD calculation and ML prediction of Hildebrand solubility parameters .....</b>	<b>26</b>
1 Methods and materials.....	26
1.1 Programs.....	26
1.2 Machine learning process .....	27
1.3 Hildebrand SP: MD simulation .....	28
1.4 Molecular descriptors .....	29
1.5 Data cleaning and descriptors selection.....	34
1.6 Molecules: polymer and small molecule .....	36
2 Results and discussion .....	39
2.1 Descriptors selection.....	39
2.2 ML algorithm and Hyperparameters optimization .....	41
2.3 ML performance for dataset 1 .....	44
2.4 ML performance for dataset 2 .....	48
3 Conclusion .....	52

<b>Prediction of polymer-drug miscibility based on the Flory-Huggins theory .....</b>	<b>55</b>
1 Materials and methods.....	55
1.1 Polymer candidates and reference drugs .....	55
1.2 Simulations and calculations .....	58
2 Results and discussion .....	59
2.1 Simulations for PEA-Indomethacin mixture .....	59
2.2 Simulations for New PEAs-BRP drugs .....	61
3 Conclusion .....	65
<b>DPD simulations of modified block copolymer micelles with specific therapeutic agents .....</b>	<b>67</b>
1 Materials and methods.....	67
1.1 Polymers and small interfering RNA .....	67
1.2 Dissipative particle dynamics simulation .....	69
2 Results and discussion .....	71
2.1 DPD simulations for charged siRNA .....	71
2.2 DPD simulations for charged polyplex micelle EN15 .....	75
2.3 DPD simulations for neutral polyplex micelles.....	77
3 Conclusion .....	81
<b>Conclusion and outlook.....</b>	<b>82</b>
<b>Appendix.....</b>	<b>85</b>
Chapter I .....	85
1 MD simulation procedure.....	85
2 Polymer REs and SMILES .....	87
3 Datasets and descriptors .....	87
4 RFE selection of descriptors.....	87
5 Computing time of ML models .....	90
Chapter II.....	92
1 Details of polymers.....	92
2 Intermolecular RDF plots of H bonds (PEA+IMC) .....	92

Chapter III.....	93
1 Atomistic structures of all beads .....	93
2 Repulsive parameters $a_{ij}$ for all beads.....	94
3 MD simulation procedure for siRNA .....	94
<b>List of Abbreviations .....</b>	<b>96</b>
<b>Reference .....</b>	<b>99</b>
<b>List of Publications .....</b>	<b>111</b>
<b>Declaration of Authorship .....</b>	<b>113</b>
<b>Ehrenwörtliche Erklärung.....</b>	<b>114</b>
<b>Curriculum Vitae.....</b>	<b>115</b>

## Abstract

Polymer drug delivery systems have developed tremendously over the past two decades, with a variety of polymeric carrier formulations and structures being created in the laboratory. The compatibility between drugs (pharmaceutical agents) and polymers is always the key factor, regardless of the differences in the shapes or detailed properties of polymeric carriers. Compatibility can be studied in terms of polymer-drug mixtures and Gibbs free energy change of mixing  $\Delta G_{\text{mix}}$  is a reliable tool for assessing polymer-drug pair compatibility. The Flory-Huggins interaction parameter (FH)  $\chi$  is a critical quantity in the calculation of  $\Delta G_{\text{mix}}$ , which itself is also a common indicator for analyzing the miscibility of polymer blends. Hildebrand solubility parameter (SP)  $\delta$  is again the key quantity for calculating  $\chi$  and itself can also provide good indication of solubility, particularly for nonpolar compounds.  $\delta$  can be determined experimentally or from simulations, but for polymers, it is almost impossible to measure directly by experiments. Computer simulation techniques have shown the power in this field and can provide detailed information from the atomistic scale to the mesoscale and even the macroscale.

In this study, a “ $\delta - \chi - \Delta G_{\text{mix}}$ ” hierarchical approach based on *in silico* methods was established for the prediction of compatibility between tailor-made polymers and drugs. Molecular dynamics (MD) simulation was applied to obtain  $\delta$  based on the atomistic structures of pure compounds and mixtures. Then  $\chi$  and  $\Delta G_{\text{mix}}$  were calculated with the simulation results and the predictions were in good agreement with experimental observations for specific polymer-drug systems. MD simulations can provide results with high accuracy, but require also high computational cost, especially for polymers with complex structure or long chains. Machine learning (ML) methods were then applied on the basis of a database that included our historical simulation results and encoded structural information (molecular descriptors). In current work,



various of ML models were built to predict  $\delta$  of polymers and the predictions had good accuracy compared to the simulation results but were faster and less computationally expensive.

In this study, not only atomistic scale simulations but also mesoscale simulations were applied to study the morphology of polymer-drug clusters. Dissipative particle dynamics (DPD) simulation, as a popular mesoscale method, were used to qualitatively study the encapsulation of small interfering RNA (siRNA) with PEG-*b*-PAGE block copolymer polyplex micelle. The FH parameter was essential for describing the interparticle interactions in DPD simulations and was a key element in bridging the two scales of simulation. In current work, the simulations processed both the charged and neutral models, and the simulation results were in good agreement with the experimental observations. The simulation results also revealed that the weight ratio of the PEG blocks was a key factor in the encapsulation ability of the PEG-*b*-PAGE micelles, which provided valuable theoretical support to the experimental team.

In this dissertation, the completed work will be presented in three chapters: Chapter I describes the ML procedure for predicting SP based on MD results; Chapter II focuses on the prediction of miscibility of specific polymer-drug mixtures using MD simulations; and Chapter III is a DPD simulation to qualitatively study the encapsulation ability of PEG-*b*-PAGE block copolymer polyplex micelles for encapsulating siRNA.

## Zusammenfassung

Polymere Systeme zur Verabreichung von Arzneimitteln haben sich in den letzten zwei Jahrzehnten enorm weiterentwickelt, wobei im Labor eine Vielzahl von polymeren Trägerformulierungen und -strukturen entwickelt wurden. Die Kompatibilität zwischen Arzneimitteln (pharmazeutischen Wirkstoffen) und Polymeren ist immer der Schlüsselfaktor, unabhängig von den Unterschieden in den Formen oder detaillierten Eigenschaften der polymeren Träger. Die Kompatibilität kann anhand von Polymer-Wirkstoff-Mischungen untersucht werden, und die Änderung der freien Gibbs-Energie beim Mischen  $\Delta G_{\text{mix}}$  ist ein zuverlässiges Tool zur Bewertung der Kompatibilität von Polymer-Wirkstoff-Mischungen. Der Flory-Huggins-Wechselwirkungsparameter (FH)  $\chi$  ist eine Schlüsselgröße für die Berechnung von  $\Delta G_{\text{mix}}$ , die ihrerseits auch ein gängiger Indikator für die Analyse der Mischbarkeit von Polymermischungen ist. Der Hildebrand-Löslichkeitsparameter (SP)  $\delta$  ist ebenfalls die Schlüsselgröße für die Berechnung von  $\chi$  und kann ebenfalls ein guter Indikator für die Löslichkeit sein, insbesondere für unpolare Verbindungen.  $\delta$  kann experimentell oder durch Simulationen bestimmt werden, aber für Polymere ist es fast unmöglich, ihn direkt durch Experimente zu messen. Computersimulationstechniken haben ihre Effizienz auf diesem Gebiet bewiesen und können detaillierte Informationen von der Atomskala bis zur Mesoskala und sogar bis zur Makroskala bieten.

In dieser Untersuchung wurde ein hierarchischer " $\delta - \chi - \Delta G_{\text{mix}}$ " Ansatz auf der Grundlage von *in silico*-Methoden für die Vorhersage der Kompatibilität zwischen maßgeschneiderten Polymeren und Arzneimitteln entwickelt. Die Simulation der Molekulardynamik (MD) wurde angewandt, um SP auf der Grundlage der atomistischen Strukturen von reinen Verbindungen und Mischungen zu erhalten. Dann wurden  $\chi$  und  $\Delta G_{\text{mix}}$  mit den Simulationsergebnissen berechnet und die Vorhersagen stimmten gut mit den experimentellen Beobachtungen für bestimmte Polymer-Wirkstoff-Systeme überein. MD-Simulationen können Ergebnisse mit hoher

Genauigkeit liefern, erfordern aber auch hohe Rechenkosten, insbesondere für Polymere mit komplexer Struktur oder langen Ketten. Anschließend wurden Methoden des maschinellen Lernens (ML) auf der Grundlage einer Datenbank angewendet, die unsere historischen Simulationsergebnisse und codierte Strukturinformationen (molekulare Deskriptoren) enthielt. In der aktuellen Arbeit wurden verschiedene ML-Modelle zur Vorhersage von  $\delta$  von Polymeren erstellt, und die Vorhersagen hatten eine gute Genauigkeit im Vergleich zu den Simulationsergebnissen, waren aber schneller und weniger rechenintensiv.

In dieser Arbeit wurden nicht nur atomistische, sondern auch mesoskalige Simulationen angewandt, um die Morphologie von Polymer-Wirkstoff-Clustern zu untersuchen. Die Simulation der dissipativen Partikeldynamik (DPD), eine populäre mesoskalige Methode, wurde zur qualitativen Untersuchung der Einkapselung von small interfering RNA (siRNA) mit Polyplexmizellen aus PEG-b-PAGE-Blockcopolymeren verwendet. Der FH Parameter  $\chi$  war für die Beschreibungen der Interpartikel-Interaktionen in DPD-Simulationen von wesentlicher Bedeutung und stellte ein Schlüsselement für die Überbrückung der beiden Simulationsskalen dar. In der aktuellen Arbeit wurden in den Simulationen sowohl geladene als auch neutrale Modelle verarbeitet, und die Simulationsergebnisse waren in guter Übereinstimmung mit den experimentellen Beobachtungen. Die Simulationsergebnisse zeigten auch, dass das Massenverhältnis der PEG-Blöcke ein Schlüsselfaktor für die Verkapselungsfähigkeit der PEG-b-PAGE-Micellen war, was eine wertvolle theoretische Unterstützung für das Versuchsteam war.

In dieser Dissertation wird die abgeschlossene Arbeit in drei Kapiteln vorgestellt: Kapitel I beschreibt das ML-Verfahren zur Vorhersage von SP auf der Grundlage von MD-Ergebnissen; Kapitel II konzentriert sich auf die Vorhersage der Mischbarkeit spezifischer Polymer-Wirkstoff-Mischungen mittels MD-Simulationen; und Kapitel III ist eine DPD-Simulation zur qualitativen Untersuchung der Verkapselungsfähigkeit von PEG-b-PAGE-Blockcopolymer-Polyplexmizellen zur Verkapselung von siRNA.

---

The following software, programs, or tools were utilized in the creation of this document:

1. Microsoft Word for making the Thesis document.
2. Materials studio for MD and DPD simulations.
3. Materials Visualizer for viewing the structures of all compounds
4. Turbomole (GUI: Tmolex) for DFT simulations
5. Padel descriptor for generating molecular descriptors
6. Anaconda, Spyder and Python for ML
7. Microsoft PowerPoint for preparing figures.
8. Microsoft Excel for making tables

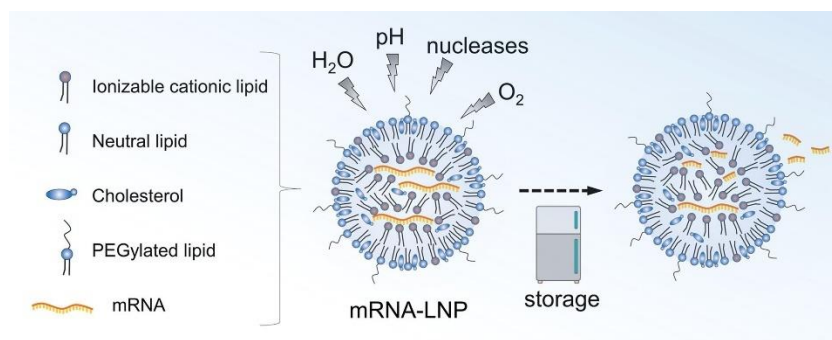
## Introduction

---

How to take the drugs? This question has always accompanied the development of human civilization. From muddy, bitter broths and powders to tiny pills, capsules and injections, the way humans take medicine has become more efficient and safer as technology advances. After the end of the tragic wars worldwide, and the general improvement of health care in many countries, the main diseases causing human deaths were changed. According to the data of World Health Organization (WHO), cardiovascular diseases caused by aging, such as ischemic heart disease and stroke, have become the leading cause of death worldwide.<sup>[1]</sup> Unfortunately, for a number of reasons conventional medication is limited in its ability to control these diseases. One reason for this is the toxic side effects associated with the drug, which can attack healthy tissues or organs while it is working, which can cause great pain to patients, especially the elderly and young children.<sup>[2]</sup> At the same time, there are still many diseases that cannot be completely cured, such as acquired immunodeficiency syndrome (AIDS). Drug treatment can only temporarily stop the progression of the disease and must be taken regularly and consistently. These issues have served as the motivation for scientists to develop novel drug delivery systems.<sup>[3]</sup> These systems will have the ability to deliver drugs more precisely to reduce side effects, or to act as carriers of vaccines that effectively gain immunity to specific viruses.

Polymers are playing an important role in the development of novel drug carriers and have already made significant progress in application.<sup>[4]</sup> The latest achievements of polymer drug carrier is the mRNA vaccine against COVID-19 virus, which was the first large-scale application of mRNA vaccines in reality.<sup>[5]</sup> The mRNA carrier is a lipid-

based nanoparticle and polyethylene glycol (PEG) is applied to prolong the residence time of the nanoparticles in the human body and improve the stability of the nanoparticles (figure 1).



**Figure 1.** The structure of mRNA-lipid nanoparticle COVID-19 vaccines (Reprinted from [6] open access: CC BY 4.0 DEED)

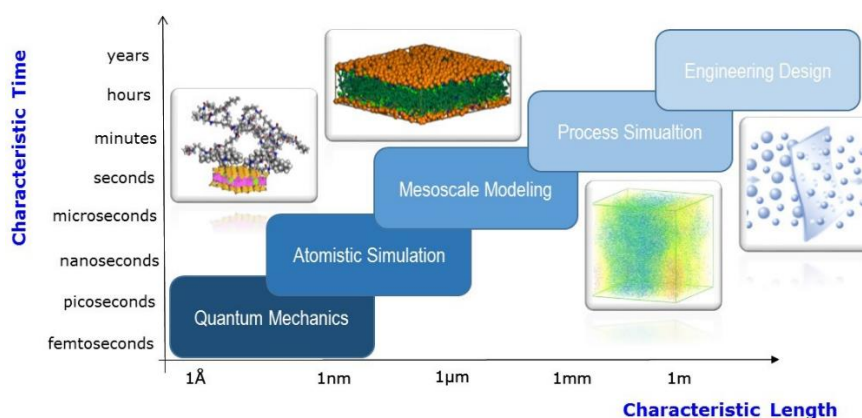
Although the main body of the aforementioned nanoparticles consists of biomolecules (lipids), polymers exhibit an important role in improving the performance of drug carriers and increasing the efficiency of drug delivery.<sup>[4, 7]</sup> Polymer-based nanocarriers have attracted the attention of the scientists because the properties of polymers may be able to overcome the disadvantages of lipid-based carriers, such as the lack of targeted delivery and harsh storage requirements.<sup>[7]</sup>

Research on the use of polymers for drug release has been conducted for decades, and essentially several well-established technical lines have been developed.<sup>[7, 8]</sup> Polymer can be simply used as a matrix to contain the drug and the release will be controlled by diffusion. In such systems, polymer performs just as a container and plays no role in the release process. Another way is using dehydrated hydrophilic polymers to encapsulate drug. The polymer matrix will swell when exposed to an aqueous environment and the drug will move out of the matrix based on diffusion or difference in solubility.<sup>[8]</sup> In such systems, polymer becomes not only a container but also a controller of drug release when the environment changes. Not only the above-mentioned approaches, but biodegradable polymers also have a place. Biodegradable polymer can be shaped into different nano structures such as micelles or nanoparticles, which can maintain quite long-term stability until they were triggered by change of pH,

temperature or redox potential.<sup>[9]</sup> The ideal degradable systems will degrade into small and nontoxic compounds, which can be cleaned by natural clearance mechanism. This dissertation focuses mostly on the degradable polymer systems, which is also one of the core topics of PolyTarget project.

It is difficult to find suitable polymeric carriers for specific drugs using only experimental methods. Selecting candidates, analyzing properties, encapsulating drugs, release test, etc. Each step requires extensive laboratory work, and scientists have to repeat the process endlessly to obtain an acceptable result. Some good results were obtained from such trial-error process<sup>[10]</sup> at the beginning but not sustainable due to the experimental cost.

With the development of computational technology scientists are able to build fine mathematical models for single molecule or a cluster of molecules based on the theory from theoretical physics and chemistry.<sup>[11]</sup> With different simulation methods, various material properties and chemical or physical processes can be studied at different time scales and size scales (figure 2).



**Figure 2.** Simulation approaches at different time and length scales (picture from: <https://www.molbni.it/node/116>)

Since the first simulation of liquid carried out by Metropolis et al.<sup>[12]</sup>, computer simulation i.e., *in silico* method has become a powerful tool for selecting ideal candidate from the ocean of materials or for designing molecules with specific properties, even those that have not yet been synthesized.

Simulation can also provide strong support for the study and design of polymer drug carrier from different dimensions, and good progress has been made on many topics. At atomic or molecular scale, Density Functional Theory (DFT) can be applied to investigate the polymer-drug interaction and to explain some features of specific polymer carrier.<sup>[13]</sup> By using empirical descriptions of force as well as classical mechanics, Molecular Dynamic (MD) simulation can be used to study the properties of a much larger system containing thousands of atoms or molecules at longer time scale and the properties can be related to macroscopic characteristics of polymer-drug conjugates.<sup>[14]</sup> Complex atomic structures can be further simplified to combinations of larger beads (i.e., coarse graining), and simulations using such coarse grained models are often referred to as mesoscale. Dissipative particle dynamics (DPD) method is one of the representatives of mesoscale simulations. With coarse grained models and a modified description of forces, DPD has been applied to study more complex molecules or clusters such as proteins, nucleic acids and polymer nanoparticles / micelles.<sup>[15]</sup> In addition, DPD is able to simulate the system over a longer period of time, which is comparable to the real process, for example the self-assembly of nano structures or the interactions of nanoparticles with bio-membrane.<sup>[16]</sup> A multi-scale approach combining atomic to mesoscale and even macroscale simulations is considered to provide a complete view of the design of polymeric drug carriers, but in practice there are a number of challenges.<sup>[17, 18]</sup>

Machine learning (ML), a field of artificial intelligence, aims to improve the performance of computers on specific tasks by using structured and labeled data. Today all areas of society are gradually becoming digital, and the internet has become pervasive in all aspects of human life. More than ever before in history, humans need efficient ways to process and use data, and ML is one of the most talked about methods today. Back to the topic of drug delivery, ML is not only a fresh idea in this scientific field but has shown potential to support the design and analysis of drug carriers.<sup>[19]</sup> As mentioned above, selecting the appropriate candidate polymeric carrier for a particular



drug often means extensive work in the laboratory. Simulations can help speed up the process, but still require sufficient computational resources. Bringing in ML may significantly reduce the time and resources required for the above work, which is one of the current hot topics.<sup>[15, 20]</sup>

Before ML became so popular, scientists were already trying to predict various physical and chemical properties through simple mathematical and statistical methods. Quantitative Structure - Activity Relationship (QSAR) and Quantitative Structure - Property relationship (QSPR) are computational modeling approaches for revealing the relationship between compound's structure and various activities or properties.<sup>[21]</sup> The goal of QSAR/QSPR is to build a mathematical function of a specific property by using structural and/or physiochemical information as inputs:

$$P = f \left( \text{physiochemical} \begin{array}{c} \text{and} \\ \text{or} \end{array} \text{structural information} \right) + \text{error}.$$

The advent of ML has brought more mathematical tools and possibilities to QSAR/QSPR method and has expanded the boundaries of the data that traditional QSAR/QSPR can use as input.<sup>[22]</sup> ML can be used not only for predicting compound properties, but also for many other applications in the field of polymer drug delivery, such as drug dose optimization, polymer carrier design, drug release time prediction, etc.<sup>[18]</sup> Also, for a long-running scientific project, ML can be an effective solution to the problem of handling old data.<sup>[23]</sup>

Back to the polymeric drug delivery, the compatibility of polymers and drug molecules is a key factor that affects the encapsulation of polymeric drug carriers. Gibbs free energy change of mixing  $\Delta G_{\text{mix}}$  is a widely used thermodynamic index to determine the miscibility of two compounds and can be described as the combination of entropy and enthalpy

$$\Delta G_{\text{mix}} = \Delta H_{\text{mix}} - T\Delta S_{\text{mix}} , \quad (1)$$

where  $\Delta H_{\text{mix}}$  is the enthalpy change of mixing,  $\Delta S_{\text{mix}}$  is the entropy change of mixing and  $T$  is the temperature.

The negative  $\Delta G_{\text{mix}}$  indicates a good miscibility of the two compounds, and a

positive value indicates the opposite.<sup>[24]</sup> The equation can be further derived into different expressions and in the current work derivation based on Flory-Huggins mean-field theory was applied.<sup>[25]</sup>

$$\Delta G_{\text{mix}} = RT(n_1 \ln(\varphi_1) + n_2 \ln(\varphi_2) + n_1 \varphi_2 \chi_{1-2}), \quad (2)$$

where  $R$  is the gas constant,  $T$  is the temperature,  $n_1$  and  $n_2$  are number of moles of component 1 and 2,  $\varphi_1$  and  $\varphi_2$  are volume fraction of 1 and 2, respectively.  $\chi_{1-2}$  is the Flory-Huggins interaction parameter (FH) and is the key quantity of the calculation.

FH theory is a lattice model for the thermodynamics of polymer solutions and can be used to describe the properties of not only polymer-solvent systems but also polymer blends.<sup>[26]</sup> The FH parameter  $\chi$  is an important quantity in the field of polymer thermodynamics and can itself be used as an indicator of miscibility. Take semi-dilute polymer solution as an example, if  $\chi$  in the range of 0 to 0.5, the solution is in well mixed state and  $\chi = 0.5$  becomes the critical point of phase separation of semi-dilute polymer solution.<sup>[27]</sup> In reality, however, the behavior of polymer solutions and mixtures is often complex. The FH parameter is difficult to use directly to determine the miscibility of a particular system, such as a polymer-drug mixture<sup>[28]</sup> but is still useful. Like  $\Delta G_{\text{mix}}$ , FH parameter can also be derived in different from and in current work the derivation based on Hildebrand solubility parameter  $\delta$  was used.

In this dissertation, the completed work aims to establish a framework for predicting properties of polymeric nanocarriers by coupling advanced data analysis with physicochemical concepts, which could become a platform for direct property prediction of polymeric nanocarriers or as part of efficient large-scale simulation protocols. The thesis is presented in three chapters: Chapter I describes the ML procedure for predicting SP based on the results of MD simulations; Chapter II focuses on predicting the miscibility of specific polymer-drug mixtures using MD simulations and FH theory; and Chapter III describes DPD simulations investigating the encapsulation mechanism of PEG-b-PAGE block copolymer polyplex micelles for encapsulating siRNA.

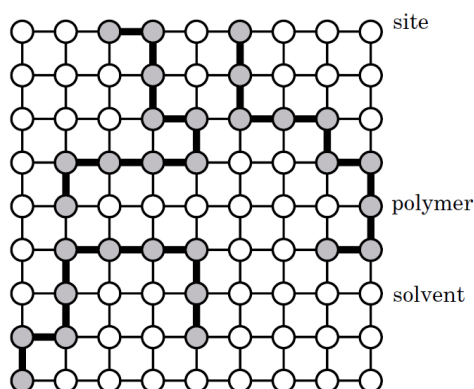
# Theory and methodology

---

## 1 Flory-Huggins parameter

The basic idea to characterize the mixing ability of polymer mixtures using  $\Delta G_{\text{mix}}$  in this study has been briefly introduced in the previous sections. Depending on the purpose of the task and the required accuracy, the FH parameter  $\chi$  can be derived into different expressions.

FH mean-field theory uses the lattice model to represent polymer solution, as shown in figure 3. Taking the polymer-solvent system as an example, the lattice model assumes that one polymer repeating unit or one solvent molecule occupies one lattice site, and all sites are of exactly the same size.<sup>[27]</sup> And there can be no empty spaces in the lattice model, which means that the polymer and solvent molecules must fill the entire lattice.



**Figure 3.** Two-dimensional lattice model of a polymer-solvent mixture [27].

Based on the above assumptions, the FH parameter  $\chi$  can be defined by the interactions among sites

$$\chi = \frac{Z \left[ \varepsilon_{PS} - \frac{\varepsilon_{PP} + \varepsilon_{SS}}{2} \right]}{k_B T}, \quad (3)$$

where  $Z$  is the number of contacts of the sites (lattice coordinates).  $\varepsilon_{ij}$  is the interaction energy of polymer-solvent, polymer-polymer and solvent-solvent, respectively.  $k_B$  is Boltzmann constant and  $T$  is the temperature. If  $T$  is kept constant, it can be clearly seen that the polymer-solvent interaction plays a dominant role in the positive and negative variation of  $\chi$ : A negative  $\chi$  indicates that the energy after mixing is lower than the energy before mixing, implying that polymer-solvent contacts are preferred, i.e., good miscibility.

Then the enthalpy of mixing can be described by  $\chi$

$$\Delta H_{\text{mix}} = \chi \varphi (1 - \varphi) * n_{\text{site}} RT, \quad (4)$$

where  $\varphi$  is volume fraction of one component,  $n_{\text{site}}$  is the total number of lattice sites and  $R$  is gas constant. When  $\chi = 0$ , the solution is in athermal state and  $\Delta H_{\text{mix}} = 0$  regardless of volume fraction. However, for real polymer solutions, this athermal state is only an ideal situation when the polymer concentration is very low.

Another key threshold of  $\chi$  is  $\chi = 0.5$  and in the FH mean-field theory, a solvent can solve the specific polymer at  $\chi = 0.5$  is the theta solvent and in such solution the polymer chain acts like ideal chain.<sup>[27]</sup> If a solvent with  $\chi < 0.5$ , it can be considered as a good solvent and if  $\chi > 0.5$ , the solvent can be a poor solvent or even nonsolvent for a specific polymer.

In practice, if the energy of mixing  $\Delta E_{\text{mix}}$  is determined ( $\Delta E_{\text{mix}} \approx \Delta H_{\text{mix}}$ ),  $\chi$  can be calculated by re-writing equation 4

$$\chi = \frac{V_m}{RT} \Delta E_{\text{mix}}, \quad (5)$$

where  $V_m$  is the molar volume of one site. For calculating  $\Delta E_{\text{mix}}$ , another popular thermodynamic quantity will be introduced: Hildebrand solubility parameter  $\delta$ <sup>[29]</sup>

$$\delta = \sqrt{\frac{E_{\text{coh}}}{V_m}}, \quad (6)$$

where  $E_{\text{coh}}$  is cohesive energy and  $V_m$  is the molar volume. According to FH theory,  $E_{\text{coh}}$  is equivalent to the total interaction energy of a system, like  $\varepsilon_{ij}$  in equation 3. In

practice, term  $\frac{E_{\text{coh}}}{V_m}$  can also be named as cohesive energy density *CED*.

After introducing  $\delta$ , equation 5 can be re-written as

$$\chi = \frac{V_m}{RT} (\delta_p - \delta_s)^2, \quad (7)$$

where  $\delta_i$  is the Hildebrand solubility parameter of polymer and solvent, respectively.

Of course, equation can also be applied to polymer blends or other mixtures. Expanding the squared term in the equation, one finds that the interaction of the two components of the mixture is approximated as the product of  $\delta$  from pure components

$$\chi = \frac{V_m}{RT} (\delta_p^2 + \delta_s^2 - 2\delta_p\delta_s). \quad (8)$$

This means that there is no need to obtain the interaction energy between the two components when calculating  $\chi$  using equation 8, thus reducing the computational cost of obtaining the data. The method has proven to be reliable and easy to use over many years of scientific practice, however, this simplification inevitably introduces a bias in the accurate description of the state of the mixture.<sup>[27]</sup> In current study, another method to calculate  $\chi$  with the consideration of exact interaction energy of the mixture system is applied<sup>[30]</sup>

$$\chi = \frac{V_m}{RT} (\varphi_p CED_p + \varphi_s CED_s - CED_{p-s}), \quad (9)$$

where  $\varphi$  is volume fraction of one component ( $\varphi_p$  for polymer and  $\varphi_s$  for small molecule, such as solvent and drug),  $CED_p$  is the cohesive energy density of the pure component and  $CED_{p-s}$  is the cohesive energy density of the mixture. In actual scientific work,  $\delta$  can also be used independently as an indicator to characterize solubility, as will be described in detail below.

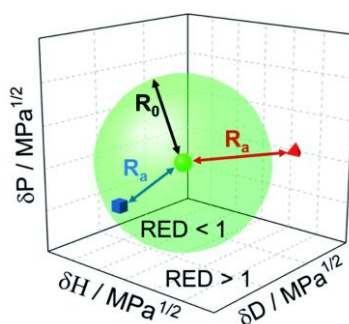
## 2 Hildebrand solubility parameter

In the 1930s Joel Hildebrand introduced solubility parameter  $\delta$  to explain the concept of regular solution using equation 6.<sup>[29]</sup> Later it became popular because this single

parameter was very easy to obtain and apply in experiments. Regardless of the properties of the solvent, once the evaporation energy  $E_{\text{coh}}$  (or cohesive energy) is determined,  $\delta$  can be easily calculated. And  $\delta$  can provide simple prediction of solubility by comparing the  $\delta$  value of two components, based on “like dissolves like” rule: when the  $\delta$  of two compounds (solvent and solute) are close enough and “close” normally described by the difference of solubility parameter  $\Delta\delta$  in the range  $\pm 2 \text{ MPa}^{1/2}$ , such two compounds have good miscibility.<sup>[29]</sup>  $\delta$  is easy to use, but of course has its limitation: it cannot provide reliable prediction on polar compounds or the system with hydrogen bonding.<sup>[31]</sup> To overcome this limitation, in the 1960s Charles Hansen divided  $\delta$  into three terms<sup>[29]</sup>

$$\delta^2 = \delta_{\text{D}}^2 + \delta_{\text{P}}^2 + \delta_{\text{H}}^2, \quad (10)$$

where  $\delta_{\text{D}}$ ,  $\delta_{\text{P}}$ , and  $\delta_{\text{H}}$  are the contribution of dispersive interaction (van der Waals), polar interaction and Hydrogen bonds, respectively. These three parameters are also known as Hansen solubility parameters and are still widely used today. The Hansen solubility parameters improve the solubility prediction of polar compounds by analyzing intermolecular interactions separately, allowing a more differentiated comparison of the solubility of two compounds. Furthermore, by using Hansen parameters, the solubility can be visualized by 3D graphics as in Figure 4.



**Figure 4.** 3-Dimensional Hansen space with axes representing three energy contributions  $\delta_{\text{D}}$ ,  $\delta_{\text{P}}$ , and  $\delta_{\text{H}}$ . The solute is located at the center of the sphere of radius  $R_0$ , with poor liquid (red tetrahedron) outside the sphere and good liquid (blue cube) inside the sphere. ([32] open access: CC BY-NC 4.0 DEED)

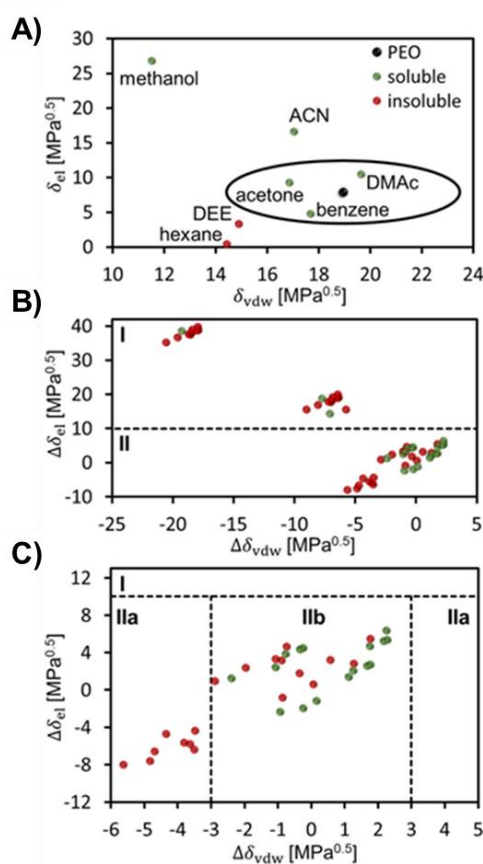
A three-dimensional coordinate system can be constructed by using Hansen parameters (Hansen space) and the  $\delta$  of a compound is visualized as a point in Hansen

space. The difference of  $\delta$  between two compounds is exactly the distance between two points. If the solubility limit is known, then a sphere can be built with radius  $R_0$  (Hansen sphere). If the point of a compound is inside the sphere, it is theoretically considered as miscible and outside the sphere is non-miscible. Compared to Hildebrand  $\delta$ , the Hansen parameters provide better predictions for polar systems and are still very easy to use but also have limitations. One of the limitations is that the measurement of the parameters is often difficult, especially for the contribution of hydrogen bonding  $\delta_H$ .<sup>[31]</sup> Therefore, a further modified form of  $\delta$  with only two terms is introduced<sup>[26]</sup>

$$\delta^2 = \delta_{\text{vdw}}^2 + \delta_{\text{el}}^2, \quad (11)$$

where  $\delta_{\text{vdw}}$  is the contribution of van der Waals (non-polar) interactions and  $\delta_{\text{el}}$  is the contribution of electrostatic (polar) interactions. This form is easier to apply in both experimental and theoretical analysis (figure 5).<sup>[33]</sup>

Using  $\delta$  or  $CED$  to predict solubility of polymers still has one serious problem: the “polymer” itself. Polymers cannot be vaporized, and it means that evaporation energy  $E_{\text{cov}}$  cannot be obtained directly from experimental measurements. Although some indirect measurement methods have been used for polymers,<sup>[34]</sup> they only work for a few polymer species and are often of doubtful accuracy. The rapid development of computer technology has brought new solutions to this field. Computer simulations can overcome such difficulties in experiments and also provide a more microscopic view.<sup>[26, 35]</sup>



**Figure 5.** Solubility predictions employing electrostatic  $\delta_{ei}$  and van der Waals  $\delta_{vdw}$  contributions to SPs for PEG (PEO) and various solvents: **A)** solvents enclosed by the circle (black line) are predicted to dissolve PEG for arbitrary solution compositions; **B,C)** differences  $\delta_{ei}$  and  $\delta_{vdw}$  for various polymer–solvent combinations; dashed lines indicate arbitrary separation of  $\delta_{ei}$  into areas **B)** I and II as well as **C)** further separation of II into IIa and IIb for  $\delta_{vdw}$ . Green and red colors indicate experimental solubility data for PEG. (Reprinted from [33] open access: CC BY-NC 4.0 DEED, Copyright 2023 the Authors.)

## 3 Computer simulation for material science

### 3.1 Quantum mechanics simulation

Since the foundation and development of quantum mechanics in the 20<sup>th</sup> century, theoretical calculations have become an important tool in materials science.<sup>[36]</sup> Lots of information about molecules, such as structure, conformation, dipole moment, etc., can be obtained by calculations, and the results are often in agreement with experimental observations.<sup>[37]</sup> In the mid-20<sup>th</sup> century, the success of digital electronic computers made it possible to calculate on larger spatial and temporal scales,<sup>[37]</sup> and computer



simulation began to play an important role in science.

Comparing with experimental approaches, computer simulations (or *in silico* methods) have many attractive features. The simulations do not require sample preparation and therefore may be much less expensive and safer, especially for studies of radioactive compounds or explosives.<sup>[37]</sup> Also, simulation can provide information for changes or processes that occur very quickly. In addition, the computational results can help scientists gain insight into their research subjects from a more microscopic perspective, information that is often difficult to obtain from experiments. These features remain important even today and are the reason why *in silico* methods is indispensable in scientific research.

Nowadays, with the continuous advances in computer hardware and software, the speed and accuracy of quantum mechanical simulations have increased significantly. At the same time, the development of quantum theory has made it possible to simulate larger sizes or more complex systems. Many programs have been developed specifically for quantum mechanical calculations, which have also led to a gradual lowering of the threshold for simulations. However, the size and complexity of the systems that can be studied are still constrained by limitations stemming from quantum mechanics itself, especially for polymers (e.g., rubber and plastics) and biological macromolecules (e.g., proteins). Since the 1960s, scientists have been searching for non-quantum simulation methods to overcome such limitations and have so far obtained fruitful results, one of which is the emergence and development of molecular dynamics.

### **3.2 Molecular dynamics simulation**

Molecular dynamics (MD) simulation is currently one of the most popular non-quantum methods for solving complex or large-scale systems. MD was born exactly with the birth of electronic computers, so the development of MD is the result of the interaction between computer technology and material science.<sup>[38]</sup> Based on Born-

Oppenheimer-approximation, the motion of electrons is neglected in MD, and the energy of the system is a function of the coordinates of the nucleus,<sup>[38]</sup> which is different from quantum mechanics simulation.

As a "dynamic" approach, MD describes the motion of atoms and molecules using Newton's equations of motion and can provide various information on the evolution of the system over time. Newton's equations of motion describe the motion of an object over a certain period of time and in the presence of forces. In MD, the trajectories of the particles in a given time are described by the potential energy surface (PES) of the system, and the force between the particles is the gradient of PES.

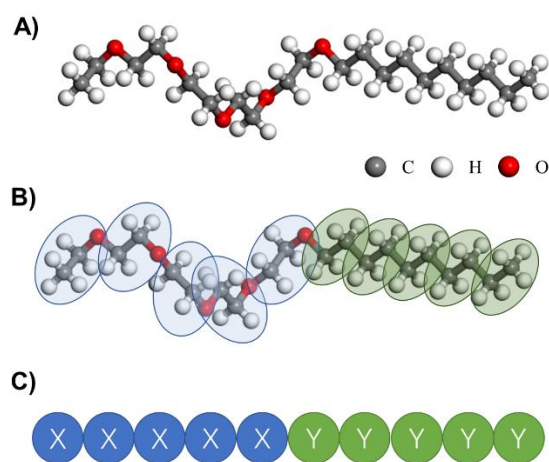
One widely used method of representing PES is the force field, which is usually a series of functions that estimate the inter- and intra-molecular interactions of a system. The parameters of the force field can be obtained through quantum mechanical calculations or experimental measurements, and because of this there are various force fields available for MD.<sup>[38]</sup> Since the 1970s, force fields applicable to biomolecules, polymers, metals, etc. have been developed that can provide accurate thermodynamic or spectroscopic information for many systems. In some scenarios, MD is comparable in accuracy to quantum mechanical calculations, but uses far fewer computational resources and can provide not only thermodynamic information, but also information about the motion of the system.<sup>[39]</sup>

Many open source or commercial MD programs are already available for use on the average PC, and they are now essential tools in laboratories, hospitals or industry. MD shows advantages in simulating the time evolution of complex systems, however only on a limited time scale. For longer physical or chemical processes, such as protein folding, scientists are still working to improve the MD's computational power and have achieved some success.<sup>[38]</sup>

Another way to increase the time scale of dynamic simulations is to simplify the molecular structure, which leads to the so-called mesoscale simulations.

### 3.3 Dissipative particles dynamic simulation

When a system is beyond the scale limit of MD simulation, one possible way to continue the study is to reduce the degrees of freedom in the system and coarse graining is one of the popular approaches to achieve this goal. During the coarse graining, a particular set of atoms or molecules will be considered as one particle (also called bead), and their internal interactions inside the bead are ignored. After coarse graining, the atomic model is reformulated as a coarse-grained model with similar morphology but fewer particles.



**Figure 6.** A demonstration of coarse graining from A) atomistic model to C) beads model.

Coarse-grained models have been widely used in the study of complex systems, such as proteins, fluids, and nucleic acids, and have proven to be a powerful tool for studying their dynamic processes.<sup>[40]</sup> The dynamic simulation of coarse-grained models is closely related to MD simulation, however the MD force field cannot be used directly for coarse-grained models, since the interactions between the beads are different from those of atomic models.<sup>[17]</sup>

One popular dynamic simulation approach for coarse-grained model is dissipative particles dynamic simulation (DPD), which was first developed by Hoogerbrugge and Koelman in 1992 and later modified by Groot et al.<sup>[17]</sup> In the DPD simulation, the

molecules are described as a string of beads connected by a spring-like "soft" force, and of course the positions of all interacting beads evolve over time in accordance with Newton's second law

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i, m_i \frac{d\mathbf{v}_i}{dt} = \mathbf{F}_i, \quad (12)$$

where  $\mathbf{r}_i$ ,  $\mathbf{v}_i$  and  $m_i$  denote the position vector, velocity and mass of bead  $i$ , respectively.  $\mathbf{F}_i$  represents the total force acting on bead  $i$ .  $\mathbf{F}_i$  includes three non-bonded parts from its neighbors: a conservative force  $\mathbf{F}_{ij}^C$ , a dissipative force  $\mathbf{F}_{ij}^D$ , and a random force  $\mathbf{F}_{ij}^R$

$$\mathbf{F}_i = \sum_{i \neq j} (\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R), \quad (13)$$

where all three forces act only within a certain range, a distance known as the cut-off radius  $r_c$ . The conservative force  $\mathbf{F}_{ij}^C$  is a soft force acting on the central linkage of particles, which in the classical DPD approach can be expressed as

$$\mathbf{F}_{ij}^C = \begin{cases} a_{ij}(1 - r_{ij})\hat{\mathbf{r}}_{ij} & (r_{ij} < r_c) \\ 0 & (r_{ij} \geq r_c) \end{cases}, \quad (14)$$

where  $a_{ij}$  denotes the strength of repulsion between particle  $i$  and particle  $j$ . In the above equation,  $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ ,  $r_{ij} = |\mathbf{r}_{ij}|$  and  $\hat{\mathbf{r}}_{ij} = \frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|}$ .

The dissipative force  $\mathbf{F}_{ij}^D$  acts to impede the interaction of the particles, which is the origin of the name "DPD".

$$\mathbf{F}_{ij}^D = -\gamma\omega^D(r_{ij})(\hat{\mathbf{r}}_{ij} \cdot \mathbf{v}_{ij})\hat{\mathbf{r}}_{ij}, \quad (15)$$

where  $\omega^D$  is the weight function, which describes the decay of the dissipative force as the distance between particles  $r_{ij}$  increases and  $\omega^D = 0$  when  $r_{ij} \geq r_c$ .  $\mathbf{v}_{ij}$  is the relative velocity between particles  $i$  and  $j$ ,  $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ .  $\gamma$  is the friction coefficient, which represents the dissipation strength.

In DPD simulation, the kinetic energy of the system is reduced due to the blocking effect of the dissipative forces and this reduction in kinetic energy is compensated by the random force  $\mathbf{F}_{ij}^R$

$$\mathbf{F}_{ij}^R = \sigma\omega^R(r_{ij})\theta_{ij}\hat{\mathbf{r}}_{ij}, \quad (16)$$

where  $\sigma$  represents the noise strength,  $\omega^R(r_{ij})$  is the weight function related to  $r_{ij}$  and  $\theta_{ij}$  denotes a randomly fluctuating Gaussian statistical variable with zero mean and unit variance that ensures that the total momentum is conserved.  $\sigma$ ,  $\omega^R(r_{ij})$  and  $\omega^D(r_{ij})$  can be obtained through following equations

$$\sigma = \sqrt{2\gamma k_B T}, \quad (17)$$

$$\omega^R(r_{ij}) = 1 - \frac{r_{ij}}{r_c}, \quad (18)$$

$$\omega^D(r_{ij}) = [\omega^R(r_{ij})]^2 = \left(1 - \frac{r_{ij}}{r_c}\right)^2 \quad (r_{ij} \leq r_c), \quad (19)$$

where  $T$  is the system temperature and  $k_B$  is the Boltzmann constant.

Moreover, a spring force  $\mathbf{F}_{ij}^S$  is introduced to describe the constraint between the bonded beads in molecules

$$\mathbf{F}_{ij}^S = \sum_j C r_{ij}, \quad (20)$$

where  $C$  is the spring constant, and the sum runs over all particles to which particle  $i$  is connected.  $C$  is often set to be 4.0 according to the study by Groot and Warren.<sup>[17]</sup>

The repulsive parameter  $a_{ij}$  in equation 14 lies on the underlying atomistic interactions between bead  $i$  and  $j$  which is linearly related to the Flory-Huggins parameters  $\chi$ <sup>[41]</sup>

$$a_{ij} = a_{ii} + 3.27\chi_{ij} \quad (21)$$

where  $a_{ii}$  is equal to 78<sup>[17, 42]</sup> and  $\chi_{ij}$  can be estimate with Hildebrand solubility parameter  $\delta$  of bead  $i$  and  $j$  by equation 7. The coefficient 3.27 is derived from a linear estimation between  $\chi_{ij}$  and the excess repulsion  $(a_{ij} - a_{ii})$  in a series of linear chain models proposed by Groot and Warren.<sup>[17]</sup>

To this point, a potential multi-scale approach combining MD and DPD simulations, using the Hildebrand solubility parameter as a bridge, could be established. As computational power continues to increase, multiscale simulation has become an increasingly important topic in materials science, and this approach is considered to provide a significant amount of information about a system or a chemical/physical progress over a wider range of spatial or temporal scales than conventional simulation approach.<sup>[43]</sup> Multiscale simulations already have some achievements,<sup>[44]</sup> but still faces

some challenges and there are still many questions that need to be addressed, for example a general validation metric for the simulation results and error propagation within the multiscale models.<sup>[45]</sup>

### **3.4 Machine learning**

With the massive advances in computer hardware and algorithms, the spatial and temporal scale limitations of today's simulations have been greatly improved compared to earlier times. Molecular dynamics, for example, can already simulate systems of more than 10 billion atoms at a scale that overlaps even some mesoscale simulations. But many challenges remain: computing resources are still strained, and the larger and more complex the system being simulated, the more computing power is naturally required. Also, the time scale of the simulation is still insufficient compared to reality, e.g. a particular biochemical process. In addition, the use of data and simulation results is always inefficient because simulated systems tend to have specific environments that are defined according to the requirements of the research project.<sup>[46]</sup> When the project is terminated or the data is unsatisfactory, these simulation results become "hot potatoes" on the hard disk.

Nowadays, Machine learning (ML) and other branches of Artificial intelligence (AI) have made great strides, and their applications have reached into the field of scientific research. With the proliferation of high-performance computer hardware and programming knowledge, as well as a thriving Internet community, ML-related information is being rapidly shared and learned around the world. Thanks to open-source programming packages such as Scikit-learn<sup>[47]</sup> and Pytorch,<sup>[48]</sup> the barriers to learning and using ML have been greatly lowered, allowing many scientists with no programming experience to get up to speed quickly.

Machine learning is essentially a method of data processing and analysis, with the entire process highly automated. The collected and processed data will be fed into a computer which will then generate a specific mathematical model based on the

algorithms and input data which will be applied for different purposes such as prediction, evaluation, decision making, classification etc.

ML provides an alternative way of dealing with experimental, computational and simulation data in the field of materials science, and it is also possible to find links or correlations between data from different sources or methods.<sup>[49]</sup> Many advances have been made in the application of ML in the field of materials science, such as predicting material properties, screening specific molecules, and designing molecular structures.<sup>[49]</sup> The addition of ML has greatly improved the efficiency of related scientific work and its accuracy has been recognized by professional researchers.<sup>[18, 19]</sup>

The first step in ML is always to collect enough data and build a dataset, and in the field of materials science, one of the important prerequisites is to efficiently digitize information about the various properties and structures of molecules. Prior to the large-scale application of ML in materials science, analytical methods such as quantitative structure-activity relationship (QSPR) and quantitative structure-activity relationship (QSAR) have been widely used to predict the properties or chemical/biological activities of molecules based on the similarity of their structures. In QSPR/QSAR, molecular descriptors (descriptors for short) were introduced to encode the structural information or properties of molecules into values for statistics analysis or calculation.<sup>[50]</sup> One of the well-known definitions of molecular descriptors is “The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment”,<sup>[51]</sup> and this suggests that descriptors can be obtained from reliable experiments or simulations.

In practice, simulated data are relatively more complete and tend to have smaller deviations than experimental data, but experimental data are more trustworthy if some specific simulated data are far from experimental observations (low chemical accuracy). In the current work, descriptors are mainly collected from different simulation methods because polymer properties from experimental measurements are often incomplete, and

the methods of measuring polymer properties are limited. Descriptors collected from simulation or calculation are also known as theoretical descriptors, and there are various programs that can calculate all descriptors in batch. The calculation of descriptors always needs the molecular structure as the input and Simplified Molecular Input Line Entry Specification (SMILES) is the most popular tool to encode the molecular structures into ASCII strings. To date, thousands of theoretical descriptors have been made available for use in ML methods, but not all of them are necessary for a particular task, so there is always a need for data cleanup after the database has been created.<sup>[52]</sup>



## Research, results and discussion

### Chapter I

# **MD calculation and ML prediction of Hildebrand solubility parameters**

---

In this chapter, a complete one-step ML procedure was established for predicting SP with MD simulation results. Based on the knowledge and experience from our earlier work,<sup>[53]</sup> the structural information of the polymer chains was newly extracted and encoded into molecular descriptors. Two datasets were constructed from a self-built polymer database and then processed through three methods to select appropriate descriptors. Four popular ML algorithms were trained using processed datasets to generate multiple models. Some of these models showed good performance and comparisons of these models have been made in this Chapter.

## **1 Methods and materials**

### **1.1 Programs**

In the current work, all molecular dynamics simulations were performed in **Materials Studio**,<sup>[54]</sup> and calculations based on density functional theory were performed with program **Turbomole**.<sup>[55]</sup> The machine learning approach was done using **python 3.8** and the **Scikit-learn** package as well as other packages.<sup>[47]</sup> The **Padel Descriptor** program is used to calculate a portion of the molecular descriptors.<sup>[56]</sup>

## 1.2 Machine learning process

ML consists of three components: the computational algorithms that are at the core of decision making, the relevant underlying knowledge with known answers that train the system to learn, and the features that make up the decisions.<sup>[49]</sup> The features in current study were provided by molecular descriptors and will be explained in detail later. First, the training data with a known answer is fed to the model. Then, the algorithm is run for training, the training results are evaluated, and the parameters are adjusted to reduce the difference between the known answer and the output of the algorithm. The algorithm repeats this evaluation and optimization process until an accuracy threshold is met, and finally the trained model is used to make predictions or decisions.

In this study, the prediction of Hildebrand solubility parameters  $\delta$  using ML is considered as a supervised learning-regression process, which means that the training data is processed, building a function (or model) that maps new data on expected output values. The workflow of ML in this study is as follows (figure 7). Collecting data and building a dataset is always the first step in ML, and in the current work the dataset consists of two parts (figure 7): the target quantity-Hildebrand solubility parameter (SP)  $\delta$  and the molecular descriptors used to generate the regression model. The calculations of the dataset components will be explained later.

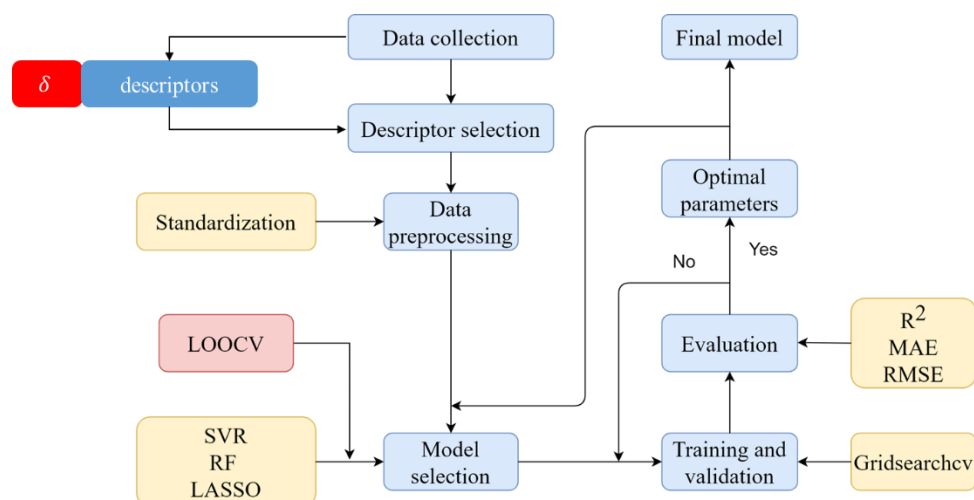


Figure 7. Workflow of ML procedure

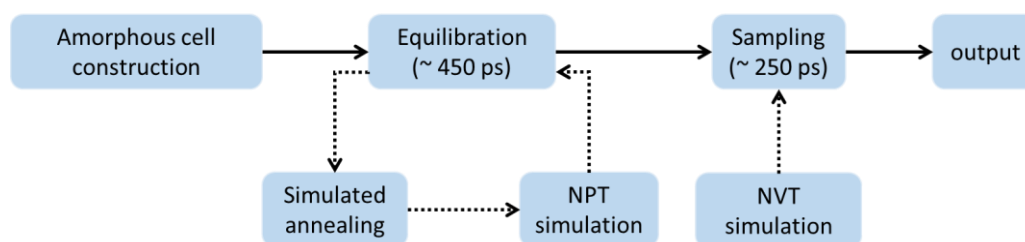
After the dataset is constructed, descriptors in the dataset that are not relevant to the goal and descriptors that are all zeros need to be manually removed. After cleaning, the descriptors will be standardized to avoid large differences between descriptors and overweighting of some descriptors. The standardized dataset will then be analyzed by different algorithms to select the most relevant combination of descriptors to the target quantity, which will then be fed into the computer to generate the corresponding ML model (preprocessing). Four ML algorithms are applied to generate corresponding ML models: Support vector machine (SVR),<sup>[57]</sup> Random forest (RF),<sup>[58]</sup> Least absolute shrinkage and selection operator (Lasso)<sup>[59]</sup> and Elastic net regression (ENR).<sup>[60]</sup> Due to the limited size of dataset, leave-one-out cross validation is used in order to fully utilize each sample in the dataset.

The performance of the ML model is evaluated by three numerical metrics: the coefficient of determination ( $R^2$ ), the mean absolute error (MAE) and the root mean square error (RMSE). During the training process, the hyperparameters of the ML model will be optimized through **Gridsearch** algorithm until the metrics reach the desired criteria. Finally, the whole machine learning process ends with machine learning models containing optimized hyperparameters.

### 1.3 Hildebrand SP: MD simulation

The target quantity, Hildebrand solubility parameter (SP)  $\delta$  was calculated for all polymers using the same molecular dynamics simulation procedure (figure 8). To begin with, the atomic models of all molecules were built, and fed into a simulation cell with periodic boundary condition based on Monte Carlo method.<sup>[61]</sup> Subsequently, during the equilibration process, selected cells would be treated with simulated annealing to reduce the energy, then equilibrated by isothermal-isobaric (NPT) ensemble using Berendsen barostat<sup>[62]</sup> and Parrinello-Rahman barostat<sup>[63]</sup> for the refined structural model. Finally, the cells were scaled to the average cell parameters obtained from the

NPT simulations, and the average values of properties such as cohesive energy density (CED) would be sampled from the canonical (NVT) simulations and  $\delta$  was calculated and output based on CED. The details of MD procedure are described in Appendix Ch.I.1.



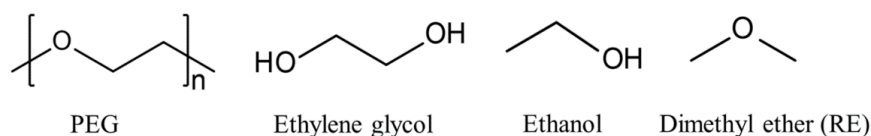
**Figure 8.** Brief illustration of the MD simulation procedure. Time unit: picosecond (ps)

## 1.4 Molecular descriptors

As mentioned in the previous paragraph, the training data for ML are all derived from simulated data. Molecular descriptors are a major component of the training data, and the predictive power of ML models depends heavily on descriptors covering a wide range of structures and compositions.<sup>[23, 49, 50, 52, 64]</sup> In the current study, descriptors were mainly collected from atomistic structure of the polymer or density-functional theory (DFT) simulations because it is difficult to obtain enough relevant experimental data as descriptors. A popular approach to collect descriptors for polymer is to use the features of its monomer,<sup>[50]</sup> but the properties of the polymer may be fundamentally different from those of the monomer. Therefore, organic small molecules with a similar structure to the polymer's repeating unit may be a better choice, and in the current work, selected small molecules are considered as repeating units (REs) of the polymer.<sup>[53]</sup>

For a given polymer, different molecules can be identified to represent the RE, such as polyethylene glycol (PEG) shown in Figure 9. Dimethyl ether was chosen to be the RE because ethylene glycol and ethanol, as polar molecules forming strong hydrogen bonds and are poor choices for deriving molecular descriptors. Therefore, if multiple potential molecules appear as REs, it is important to select molecules without strong

polarity or hydrogen bonding whenever possible.<sup>[53]</sup> An optimal descriptor should have structural interpretation and also be kept as simple as possible. Therefore, in the current work, descriptors were extracted only from the simplest molecule that was similar to a single repeating unit.



**Figure 9.** Polyethylene glycol (PEG) and different choices of small molecules with a structural motif of the repeating element of PEG (Reprinted from [53] open access: CC BY-NC 4.0 DEED, Copyright 2023 the Authors.)

Depending on the number of dimensions of the molecular structure required for the descriptor calculations, molecular descriptors can be classified as zero-dimensional (0D), one-dimensional (1D) and two-dimensional (2D).

### 0D descriptor

All the molecular descriptors for which no information about molecular structure and atom connectivity is needed belong to the class of 0D descriptors. Atom counts as well as sums or averages of atomic attributes are typical features of such descriptors. These descriptors can be always easily calculated, are naturally interpreted, do not require optimization of the molecular structure, and are independent of any conformational problem.<sup>[51, 56]</sup>

### 1D descriptor

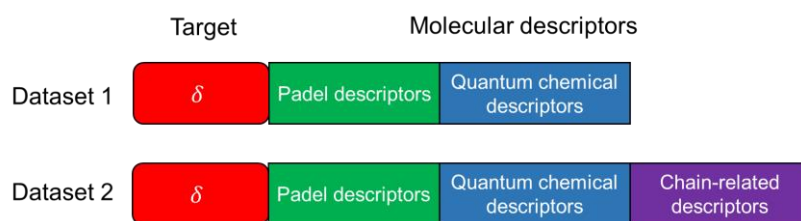
Usually represent molecular information regarding the size, shape, and electronic distribution in the molecule, such as the number of benzene rings, the number of hydrogen bond donors, etc.<sup>[51]</sup>

### 2D descriptor

2D descriptor represents topological features and how the atoms are connected (e.g., adjacency, connectivity). They are usually the descriptors derived from the molecular graph representations of nodes and edges such as size, degree of branching, the

neighborhood of atoms in terms of electronic & steric effects, flexibility, overall shape, etc.<sup>[51]</sup>

In current study, the collected descriptors can be divided into three parts: Padel descriptors contains 0D, 1D and 2D descriptors of polymer RE, quantum chemical descriptors were calculated by program **Turbomole** and chain-related descriptors contains information of polymer chain. Then two datasets were constructed with different data structures (figure 10): dataset 1 contained padel descriptors and quantum chemical descriptors, while dataset 2 further included chain-related descriptors. The two datasets will be discussed in section 2.3 and 2.4.



**Figure 10.** Schematic diagram of the dataset structures

### 1.4.1 Padel descriptors

The program Padel Descriptor<sup>[56]</sup> was used to calculate molecular descriptors using SMILES of polymer REs (Appendix Ch.I.2). In this work, 1444 descriptors were obtained by Padel, which occupy the vast majority of the datasets, including all types of descriptors such as geometrical, informational, charge, functional group counts, molecular properties, etc. The full specification of the descriptors and their interpretations are given in Appendix Ch.I.3.

### 1.4.2 Quantum chemical descriptors

The collection of quantum chemical descriptors is based on density functional theory calculations performed with the **Turbomole** program.<sup>[55]</sup> SMILES of polymer REs were also used as the input for the calculation. The Becke 3-parameter Lee–Yang–Parr (B3-LYP)<sup>[65]</sup> exchange–correlation functional was employed, along with triple zeta valence

plus polarization (def2-TZVP)<sup>[66]</sup> basis sets and Grimme dispersion correction (DFT-D3).<sup>[67]</sup> The geometry convergence criteria for DFT calculations were  $10^{-6}$  hartree for the energy change and  $10^{-3}$  hartree/bohr for the gradient norm.

In the current work, two quantum chemical descriptors were calculated: atomization energy ( $AE$ ), quadrupole moment ( $QM$ ). The atomization energy is the extra energy needed to break up a molecule into separate atoms

$$AE = \sum E_{\text{atom},i} - E_{\text{molecule}}, \quad (23)$$

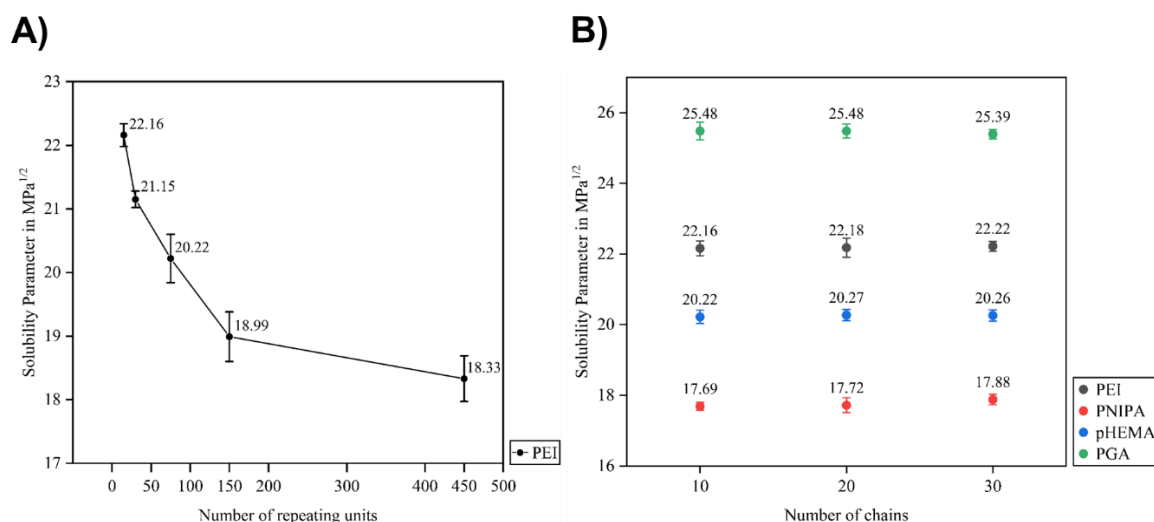
where  $E_{\text{atom},i}$  is the energy of single atom and  $E_{\text{molecule}}$  is the energy of the corresponding molecule. There are different definitions for the quadrupole moment.<sup>[68, 69]</sup> In the present work, the quadrupole moment was defined as the second moment of charge,<sup>[68]</sup> and  $QM$  was taken as

$$QM = \frac{1}{3}(Q_{xx} + Q_{yy} + Q_{zz}), \quad (24)$$

where  $Q_{xx}$ ,  $Q_{yy}$  and  $Q_{zz}$  are diagonal elements of the second moment of the charge tensor.

### 1.4.3 Chain-related descriptors

Based on the results of MD simulations in the current work and observations from other publications, the chain length of the polymer (i.e., the degree of polymerization, DP) has a strong influence on many properties of the polymer.<sup>[70]</sup> Figure 11A shows an example of polyethylenimine (PEI) and it can be seen that the Hildebrand solubility parameter  $\delta$  decreases with the increasing polymer chain length.



**Figure 11.** MD simulation results of **A)** Hildebrand solubility parameter of polyethylenimine (PEI) with different chain length, **B)** Hildebrand solubility parameters of polyethylenimine (PEI), poly(N-isopropylacrylamide) (PNIPA), poly(2-hydroxyethyl methacrylate) (pHEMA) and polyglycolic acid (PGA) with different number of polymer chains in the simulation cell.

At the same time, according to previous studies on polymers, the number of polymer chains in the simulation cell has a minor effect on the simulation results (figure 11B). Based on the above observations, information about polymer chain was considered as an important component in the dataset and in current study three chain-related descriptors were collected: molecular weight of the polymer, the polymer chain length and the connectivity indices.

The zero- and first-order connectivity indices were described in the work of Bicerano.<sup>[71]</sup> Firstly, two atomic indices ( $\sigma$  and  $\sigma_v$ ) are defined, describing the bonding and electronic environment of each non-hydrogen atom. The first one is a simple connectivity index,  $\sigma$ , which is equal to the number of non-hydrogen atoms to which an atom is bound. The latter atomic index,  $\sigma_v$ , contains information about the electron configuration of the atom and is given by

$$\sigma_v = \frac{z^v - N_H}{z - z^v - 1}, \quad (25)$$

where  $z^v$  is the number of valence electrons of the atom,  $N_H$  is the number of hydrogens bound to it, and  $z$  is its atomic number.<sup>[72]</sup> The zeroth-order (atomic) connectivity indices  ${}^0\chi$  and  ${}^0\chi^1$  for the polymer molecule are defined in terms of the summations:



$${}^0\chi = \Sigma \frac{1}{\sqrt{\sigma}} , \quad (26)$$

and

$${}^0\chi^1 = \Sigma \frac{1}{\sqrt{\sigma^v}} , \quad (27)$$

## 1.5 Data cleaning and descriptors selection

After the dataset is constructed, nulls or outliers in the dataset are removed, and descriptors in the dataset that have only zeros are cleared. After manual cleaning of the data, the remaining descriptors in the dataset will be cleaned by different methods or strategies, which is also known as feature selection in the field of machine learning. Depending on the different feature selection method strategies, the whole dataset or part of the dataset will be targeted for cleaning, which will be explained in section 2.3.

### 1.5.1 Pearson correlation coefficient

In statistics, the Pearson correlation coefficient (PCC)<sup>[73]</sup> is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations, given by

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} , \quad (28)$$

where  $r_{xy}$  is the sample Pearson correlation coefficient,  $n$  is sample size,  $x_i$  and  $y_i$  are the individual sample points indexed with  $i$ ,  $\bar{x}$  and  $\bar{y}$  are the mean values of  $x_i$  and  $y_i$ , respectively. In the current work, the sample means the dataset.

PCC is essentially a normalized measurement of the covariance, such that the result always has a value between  $-1$  and  $1$ .<sup>[74]</sup> Correlations of  $-1$  or  $1$  imply an exact linear relationship. Positive correlations imply that as  $x$  increases, so does  $y$  and the situation is reversed when the correlation is negative. In this work, PCC was performed by **Python** with the **stats.pearsonr** function in the **SciPy** package. Padel descriptors

were the target of PCC, since they were the majority in the datasets. After the PCC, the remaining padel descriptors would then be combined with other descriptors to construct a new dataset for the next step.

### 1.5.2 Principal component analysis

PCA (Principal component analysis)<sup>[75]</sup> is a common data analysis method and transforms the original data into a set of linearly independent representations of each dimension by linear transformation, which can be used to extract the main feature components of the data and is often used for dimensionality reduction of high-dimensional data.<sup>[76]</sup> According to the mechanism of PCA, not only the Padel descriptors but the whole dataset was analyzed and performed via the **decomposition.PCA** function in the **Scikit-learn** package.

### 1.5.3 Recursive feature elimination

Recursive feature elimination (RFE)<sup>[77]</sup> is a feature selection method that allows the importance of each descriptor to be obtained by using the relevant attributes of the model returned by the estimator (An estimator is an object that fits a model based on some training data and is capable of inferring some properties on new data.).<sup>[47]</sup> Then, the least important descriptors are removed from the current set of descriptors. This recursive step is repeated over the descriptor set until the desired number of descriptors is finally reached.<sup>[78]</sup>

RFE requires that a specified number of descriptors be retained, however, it is often not known in advance how many descriptors are valid. To find the optimal number of descriptors, the cross-validation method<sup>[79]</sup> is used in RFE (RFECV) to score different subsets of descriptors and select the best scored set of descriptors as the sub-dataset for further training. During RFECV no descriptors are removed, if it would cause a performance loss. This method is relatively good for selecting single model descriptors but has two drawbacks: (1) it is computationally intensive, (2) as the estimator changes,

the optimal combination of descriptors will also change, which in some cases can have a detrimental effect.<sup>[80]</sup>

The **Yellowbrick** visualization tool for python was used to plot the number of descriptors in the model along with their cross-validation test scores, variability and also the number of selected descriptors.<sup>[81]</sup>

After three different feature selection processes, three sub-datasets with different combinations of descriptors were generated on the basis of each original dataset.

## **1.6 Molecules: polymer and small molecule**

For predicting  $\delta$  in current work, up to 82 polymers were selected as the training materials for ML. Table 1 shows all polymers in the datasets. As explained above polymer chain length usually has a strong influence on some of the properties of the polymers, so polymers with the same repeat elements but different chain lengths were treated separately. Therefore, there were in total 167 samples in the dataset 2. The REs of corresponding polymers are summarized in Appendix Ch.I.2.

**Table 1.** The summary of polymers in the datasets

Polymer	Chain length	Polymer	Chain length	Polymer	Chain length
Polyethylene glycol	40, 50, 60	Polyvinyl butyrate	40, 50, 95	Polymethyl cyanoacrylate	35, 50, 85
Polylactic acid	20, 35, 50	Polyethylene terephthalate	20, 30, 50	Polyvinyl ether	25, 40, 50
Polystyrene	30, 45, 50	Polycaprolactone	30, 50, 60	Polynorbornene	10, 30, 50
Polyvinyl ethyl ether	40, 50, 60	Polyethylene	20, 50, 90	Polyethylenimine	15, 30, 75
Polyacrylamide	40, 50, 60	Cis-1,4-polyisoprene	20, 40, 50	Poly(2-hydroxyethyl methacrylate)	15, 30
Polyacrylic acid	40, 50, 60	Polyvinyl alcohol	15,30,50,75	Polyglycolic acid	15, 40, 80
Polybutylene	40, 50, 60	Polyvinyl butyral	15, 35, 50	Poly(N-isopropylacrylamide)	15, 30, 75
Polyacrylonitrile	20, 35, 50	Polyvinyl chloride	20, 35, 50	Polyamide 6	10
Polyallyl cyanide	50, 55, 65	Polychloral	10, 40, 70	Polyamide 66	10
Polyallyl acetate	50, 55, 65	Polyoxymethylene	25, 40, 50	Polybutylene succinate	10
Polypropylene	15, 30, 50, 70, 75, 80	Polyvinyl fluoride	10, 50, 90	Polycarbonate(iso)	10, 20, 30
Polyvinyl acetate	40, 50, 60	Polychloroprene	10, 30, 45	Polyether ether ketone	10
Polymethacrylonitrile	45, 50, 80	Polyacetylene	25, 40, 50	Polyurethane	10
Poly(4-vinylphenol)	50, 55, 65	Polyepichlorohydrin	40, 50, 60	Poly {succinic acid-alt-[bis(2-oxazoline)]}	10

Polyphenylene oxide	10	Polyptalamide	10, 15, 20	Poly {adipic acid-alt-[bis(2-oxazoline)]}	10
Poly(2-ethyl-2-oxazolin)	10	Polyethylene naphthalate	10	Poly {dodecanedioic acid-alt-[bis(2-oxazoline)]}	10
Poly(2-methyl-2-oxazolin)	10	Poly(3-hydroxypropionate)	10	Poly {dimethylglutaric acid-alt-[bis(3-oxazoline)]}	10
Poly(2-propyl-2-oxazolin)	10	Poly(3-hydroxybutyrate)	10	Poly {diethylglutaric acid-alt-[bis(2-oxazoline)]}	10
Polypeptoid alpha	10	Poly(3-hydroxyvalerate)	10	Poly {diethylmalonic acid-alt-[bis(3-oxazoline)]}	10
Polypeptoid beta	10	Poly(3-hydroxyhexanoate)	10	Poly {cyclopentanediacetic acid-alt-[bis(2-oxazoline)]}	10
Polypeptoid gamma	10	Poly(3-hydroxyheptanoate)	10	Poly {cyclohexanediacetic acid-alt-[bis(2-oxazoline)]}	10
Polymethylpentene	10, 30, 45	Poly(3-hydroxyoctanoate)	10	Poly {phenylsuccinic acid-alt-[bis(2-oxazoline)]}	10
Polyetherketoneketone	10	Poly(3-hydroxynonanoate)	10	Polycyclohexylenedimethylene terephthalate	10
Polyketone	10, 30, 50	Poly(3-hydroxydecanoate)	10	P(Phe-alt-G)	78, 90
Polypropylene carbonate	10	Poly(3-hydroxyundecanoate)	10	P(Val-alt-G)	55, 93
Polybutylene terephthalate	10	Poly(3-hydroxydodecanoate)	10	P(Ile-alt-G)	55
Polytrimethylene terephthalate	10, 20, 30	Poly(3-hydroxytetradecanoate)	10	Polyvinylpyrrolidone	50, 55, 65
Epichlorohydrin rubber	50				

## 2 Results and discussion

### 2.1 Descriptors selection

#### 2.1.1 Pearson correlation coefficient

Pearson correlation coefficient (PCC) was applied for analyzing the correlation between descriptors and polymer SPs in the dataset. When the magnitude of correlation coefficient  $r_{xy}$  is in the range of 0.5 to 0.75, it indicates a moderate correlation between descriptors, and greater than 0.75 could be considered strongly correlated.<sup>[82]</sup> For datasets in current work no descriptors with  $r_{xy}$  greater than 0.75 were observed and all descriptors with  $r_{xy}$  greater than 0.5 were left in the datasets. After PCC processing, 17 and 22 Padel descriptors were retained for dataset 1 and dataset 2, respectively. These were then combined with quantum chemical and chain-related descriptors to create final datasets containing 23 and 28 descriptors, respectively.

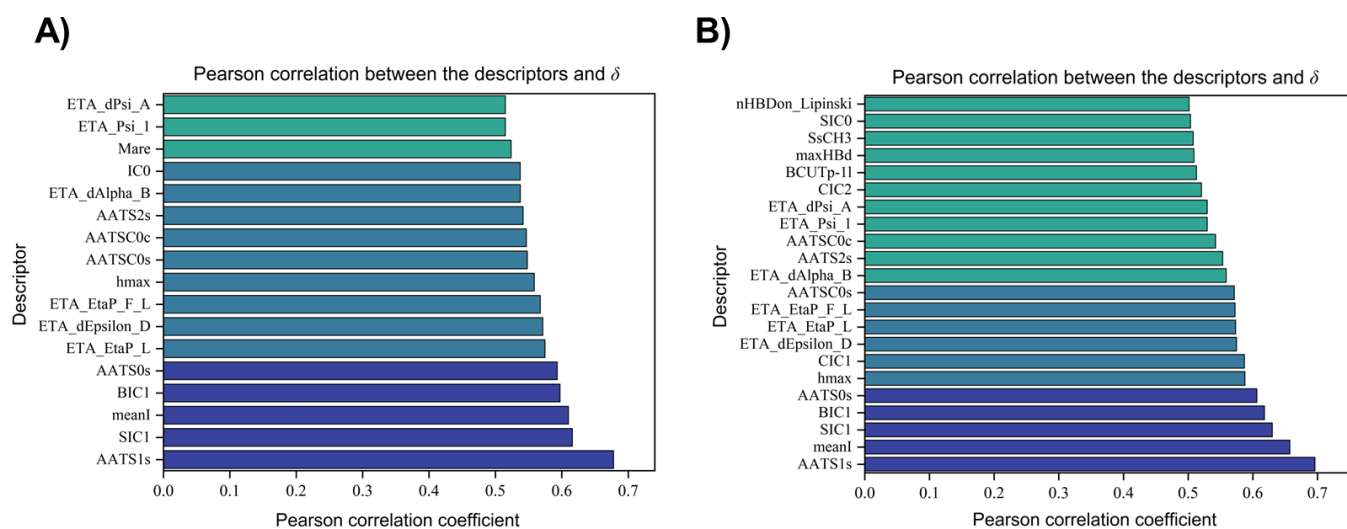


Figure 12. PCC analysis for A) dataset 1 and B) dataset 2.

#### 2.1.2 Principal component analysis

In principal component analysis (PCA), the amount of dimensionality reduction has a

significant impact on the quality of the processed dataset. Without dimensionality reduction, the performance of the ML model may be limited, but if the target dimension is too low, a huge loss of information can occur. The number of the principal component (PC) can be determined by examining the cumulative explained variance rate  $v$  associated with the number of components, in figure 13 the red bars show the percentage explained variance for each PC and the blue line shows  $v$ .

For the datasets in current work, it can be observed that the first PC (PC1) explains 39.49 % of the variance of the dataset. Then the first 10 PCs explained 74.91% and approximately 50 PCs were needed to describe close to 100% of the variance. The performance of the final model is influenced not only by the number of PCs, but also by other hyperparameters of the particular ML model. Therefore, a grid search procedure (**Gridsearch** function in **Scikit-learn**) combined with the hyperparameters of the corresponding ML model was applied. The optimal number of components was determined as 23 for both datasets, which would retain 90% of the variance.

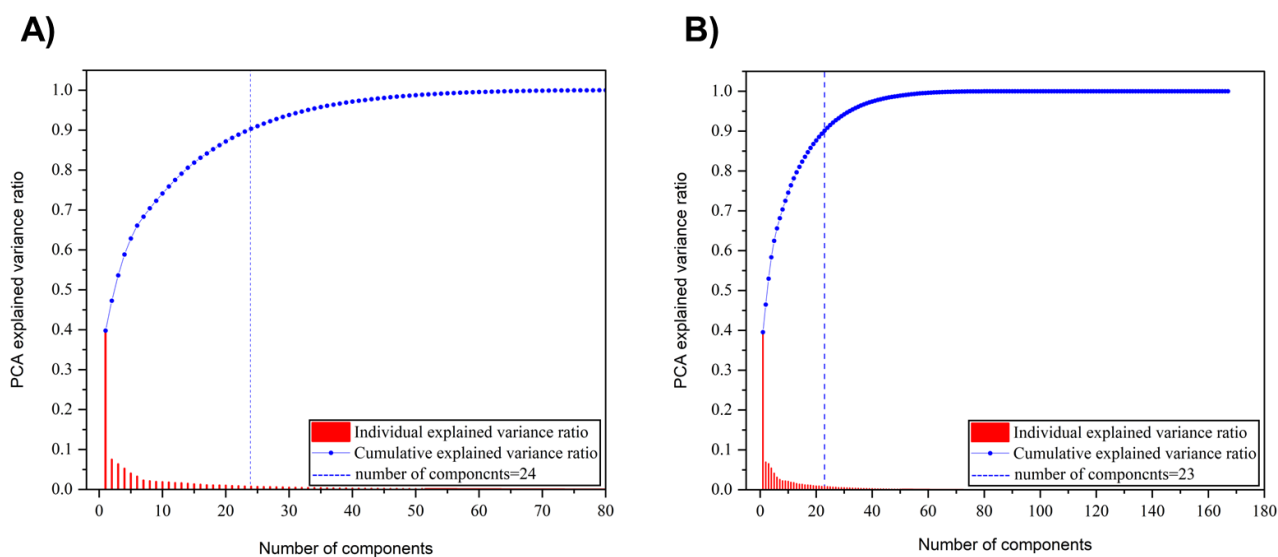
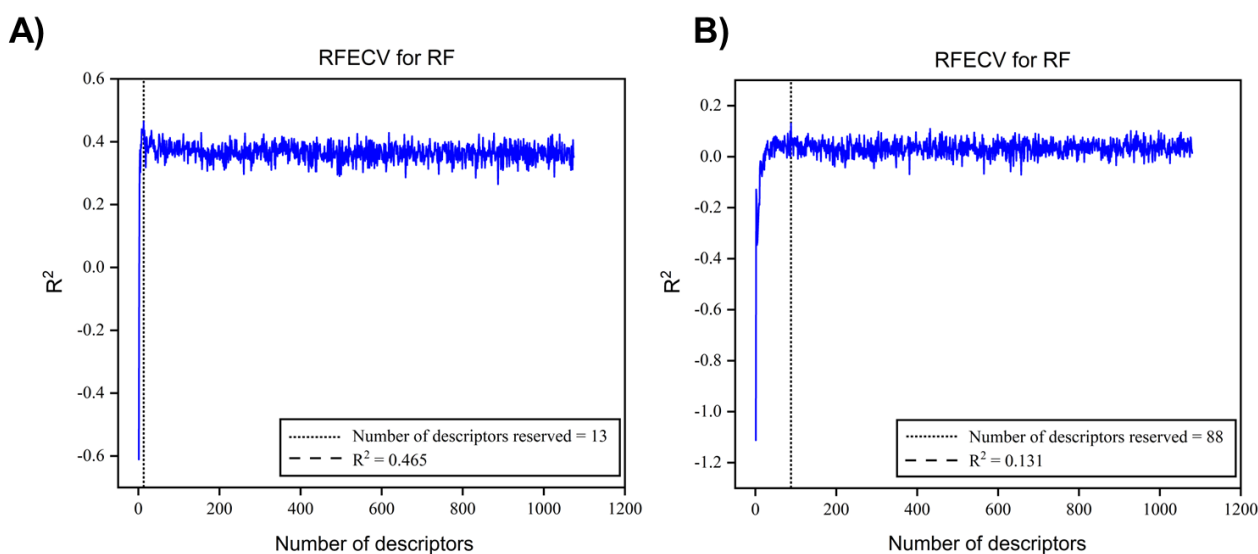


Figure 13. PCA results for A) dataset 1 and B) dataset 2.

### 2.1.3 Recursive feature elimination

As explained in section 1.5.3, recursive feature elimination (RFE) extracts the subsets with different combination of descriptors from the original datasets and select the most

suitable subset as the sub-dataset for a given ML models. Therefore, unlike the two feature selection methods mentioned above, the sub-datasets after RFE processing were not identical for different ML models. Figure 14 shows an example that the cross-validation test scores of a Random forest model (RF) against the number of descriptors. On the basis of the mechanism of RFE, it is able to rank the filtered-out descriptors with the order of elimination, but not for the descriptors still in the set. The selected descriptors for all ML models are listed in Appendix Ch.I.4.



**Figure 14.** RFE results based on RF model for **A)** dataset 1 and **B)** dataset 2.

## 2.2 ML algorithm and Hyperparameters optimization

### 2.2.1 Support vector machine regression (SVR)

SVR is a regression model of the Support Vector Machine (SVM) algorithm, which is a relatively simple supervised ML algorithm. The main idea behind SVM is to find a hyperplane with largest margin that separates the different classes in the training data. Once the hyperplane is determined, new (unknown) data can be classified by determining on which side of the hyperplane it falls.<sup>[57]</sup> Therefore, based on this mechanism, SVM was initially used for classification tasks. More than that, its regression model can also effectively handle continuous multidimensional data. By



adding kernel functions, kernelized SVM can also effectively handle nonlinearly separable data, and the choice of kernel functions is one of the hyperparameters of the SVM algorithm.<sup>[49]</sup>

In **Scikit-learn** package, three hyperparameters of SVR can be adjusted: **kernel**, **Gamma** and **C**. **Kernel** means the choice of kernel functions, there are several kernel functions available: linear, polynomial, rbf and sigmoid. **Gamma** decides the influence range of a single training example reaches during transformation, which in turn affects how tightly the decision boundaries end up surrounding points in the input space. **C** controls the amount of regularization applied to the data.

### 2.2.2 Random forest (RF)

RF is a popular machine learning algorithm which combines the output of multiple decision trees to reach a single result. In short, the RF algorithm randomly generates a certain number of uncorrelated decision tree models from the same input dataset through some mechanism.<sup>[58]</sup> Once the forest is created, all decision tree models will provide their respective predictions based on the training data, and for the regression task, the average of the individual decision trees will be used as the final prediction of the RF. Since many randomly generated decision trees can produce a fairly large number of predictions, the hyperparameters of the RF model are reasonably easy to tune, and in some cases good performance can be obtained even without tuning. In current work, only two hyperparameters were in focus: **max\_depth** and **n\_estimators**. **Max\_depth** controls the split depth of single decision tree and **n\_estimators** is the number of trees in the forest.

### 2.2.3 Lasso

Least Absolute Shrinkage and Selection Operator, or Lasso in short is an improvement of the multi-linear regression. The aim of multi-linear regression is to minimize the cost function or residual sum of squares (RSS)

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (29)$$

where  $y_i$  is the true value of the input data and  $\hat{y}_i$  is the predicted value. The performance of linear regression can be easily affected by large number of colinear features in the dataset and over-fitting problem is difficult to avoid. To overcome such problems, an extra L1-regularization term  $\alpha \sum_{j=1}^p |\beta_j|$  is added

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |\beta_j|, \quad (30)$$

where  $\alpha$  controls the strength of regularization,  $p$  is the number of variables that are available in the dataset and  $\beta_j$  is the factor of the  $j$ -th variable. When  $\alpha = 0$ , the model will again become an ordinary linear regression, and when  $0 < \alpha < \infty$ , the regression model is Lasso. Thus,  $\alpha$  is the only hyperparameter for Lasso, notated as **alpha** in **Scikit-learn**.

#### 2.2.4 Elastic net

Elastic net regression (ENR) is another popular improvement on multi-linear regression that adds L1 and L2-regularization terms  $\frac{1-\alpha}{2} \beta_j^2$  to the cost function.

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \beta_j^2, \quad (31)$$

when  $\alpha = 0$ , the model becomes regression model with L2-regularization (ridge regression), and if  $\alpha = 1$ , the model then changes to a Lasso model. Therefore, ENR is able to combine the Lasso and ridge regression, in order to find a better balance between variance and bias. In Scikit learn, the  $RSS$  is re-write into

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \text{alpha} * \text{l1\_ratio} \sum_{j=1}^p |\beta_j| + 0.5 * \text{alpha} * (1 - \text{l1\_ratio}) \beta_j^2, \quad (32)$$

where **alpha** and **l1\_ratio** are the hyperparameters for ENR.

The optimized hyperparameters of each ML model for datasets 1 and 2 were listed in Table 2 and table 3, respectively.

**Table 2.** Hyperparameters for machine learning models based on dataset 1

Selection method	Model	Hyperparameter
PCC	SVR	kernel = linear, C = 2.2
	RF	max_depth = 50, n_estimators = 200
	Lasso	alpha = 0.1
	ENR	alpha = 0.9, l1_ratio = 0.29
PCA	SVR	kernel = rbf, C = 10.2, $\gamma = 0.0005$
	RF	max_depth = 20, n_estimators=50
	Lasso	alpha = 0.3
	ENR	alpha = 0.1, l1_ratio = 1.0
RFE	SVR	kernel = linear, C = 4.5
	RF	max_depth = 20, n_estimators=50
	Lasso	alpha = 0.1
	ENR	alpha = 0.5, l1_ratio = 0

**Table 3.** Hyperparameters for machine learning models based on dataset 2

Selection method	Model	Hyperparameter
PCC	SVR	kernel = linear, C = 0.2
	RF	max_depth = 50, n_estimators = 50
	Lasso	alpha = 0.1
	ENR	alpha = 0.1, l1_ratio = 0.05
PCA	SVR	kernel = rbf, C = 14.9, $\gamma = 0.002$
	RF	max_depth = 30, n_estimators=50
	Lasso	alpha = 0.3
	ENR	alpha = 1.0, l1_ratio = 0.22
RFE	SVR	kernel = linear, C = 8.5
	RF	max_depth = 20, n_estimators=100
	Lasso	alpha = 1
	ENR	alpha = 1.0, l1_ratio = 0

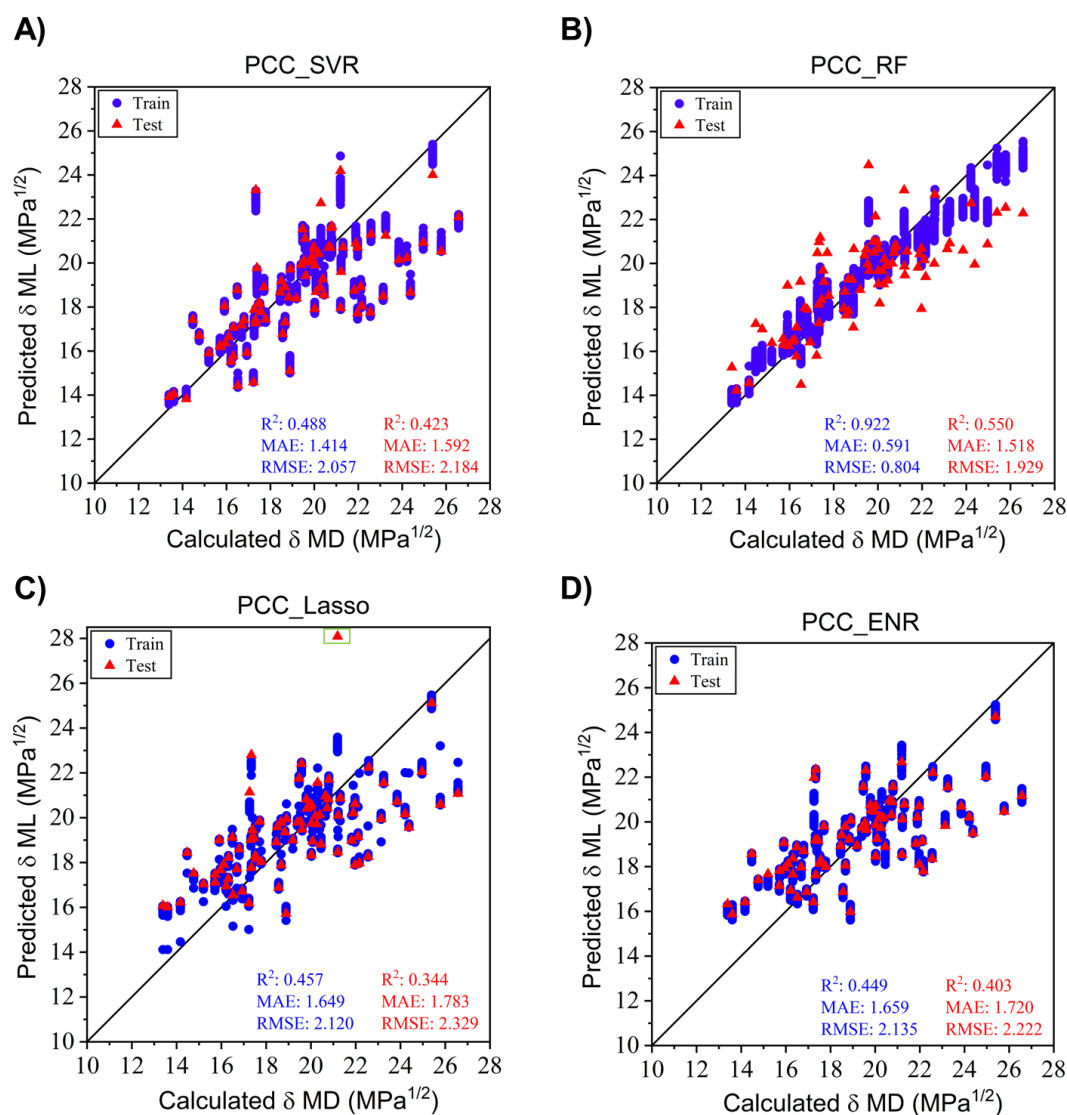
## 2.3 ML performance for dataset 1

In our earlier work, ML approach was applied to predict the heat of vaporization or cohesive energy  $E_{\text{coh}}$  of small molecules with literature data.<sup>[53]</sup> Then the Hildebrand

solubility parameter  $\delta$  of polymer REs was calculated with predicted  $E_{cov}$  using equation 6 and correlated to the  $\delta$  of polymers. Acceptable correlation could be observed but there were three major problems. First, literature data were collected from different sources and under different measurement conditions. Consequently, the data were less reliable and had unknown noise when used for machine learning. Second, the approach is a two-step procedure: ML model predicts one quantity and then correlates with another by further calculations. The whole procedure was complicated, and the final performance for a real task was suspected. Third, the dataset contained only structural information of small molecules or RE without any information about polymer chain. Hence, whether these data alone can adequately reflect the polymer structure was an issue.

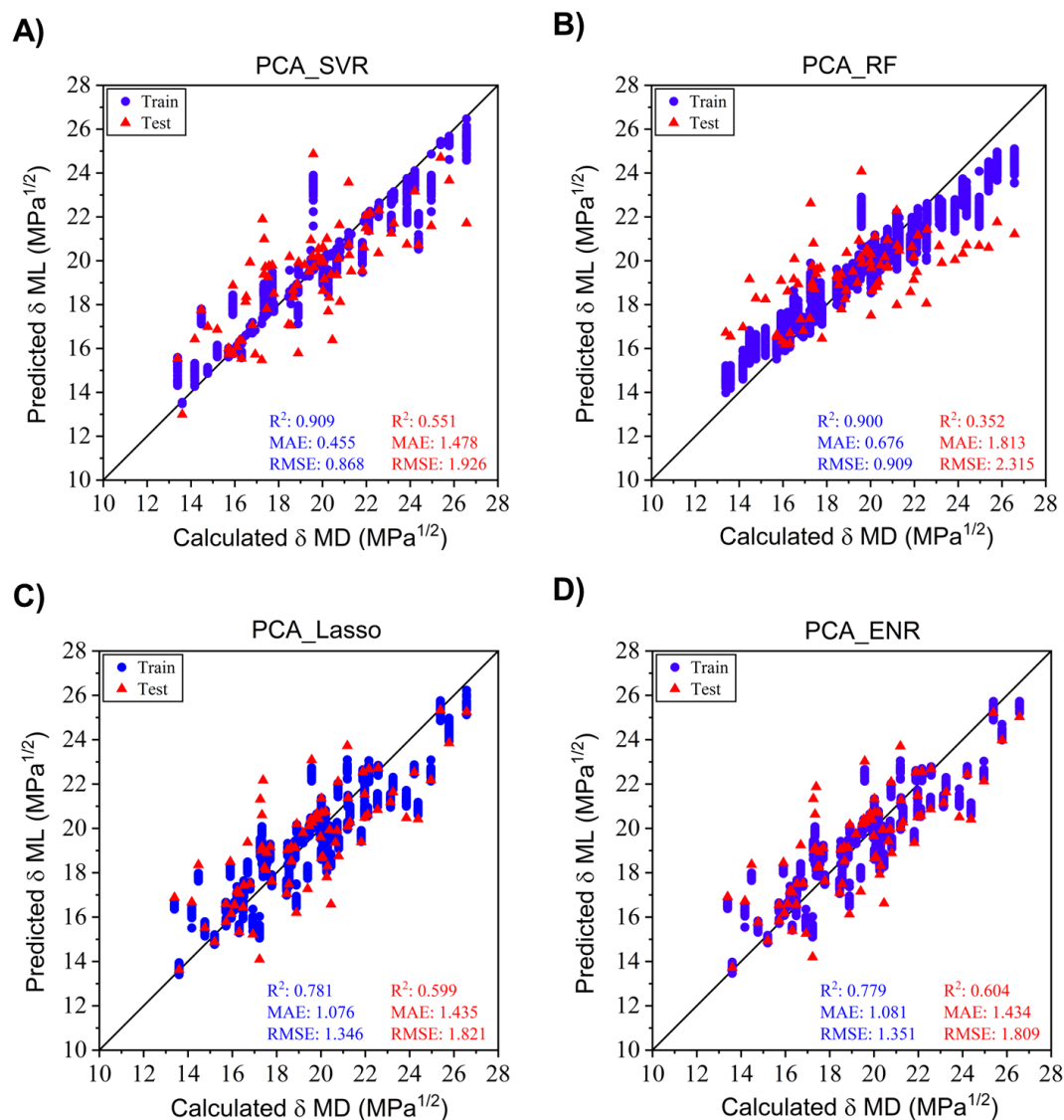
Therefore, an improved one-step ML approach was developed, and the data were all collected from simulations with same process. Dataset 1 was built in the same manner as the previous work, but with the addition of more samples. Dataset 2 contained more information extracted from the polymer chain structure, and the size of the dataset was further expanded as data on polymers with various chain lengths became available.

Figure 15 shows the comparison of  $\delta$  predicted by the ML model using PCC for descriptor selection with  $\delta$  simulated by MD. When using Leave-Out Cross-Validation (LOOCV), one sample from the dataset is used to validate the model in each round of training, and the rest of the samples are used for training until all the samples are used for validation at once. The blue traces show the change in predictions during the training with LOOCV. Overall, none of the models showed good performance after LOOCV and some of them with serious problems. RF model (figure 15B) achieved relatively good results in training, but not in testing, suggesting an over-fitting problem.<sup>[83]</sup> Two linear models, Lasso and ENR, were at the bottom of the ranking and Lasso even had an outlier (figure 15C).



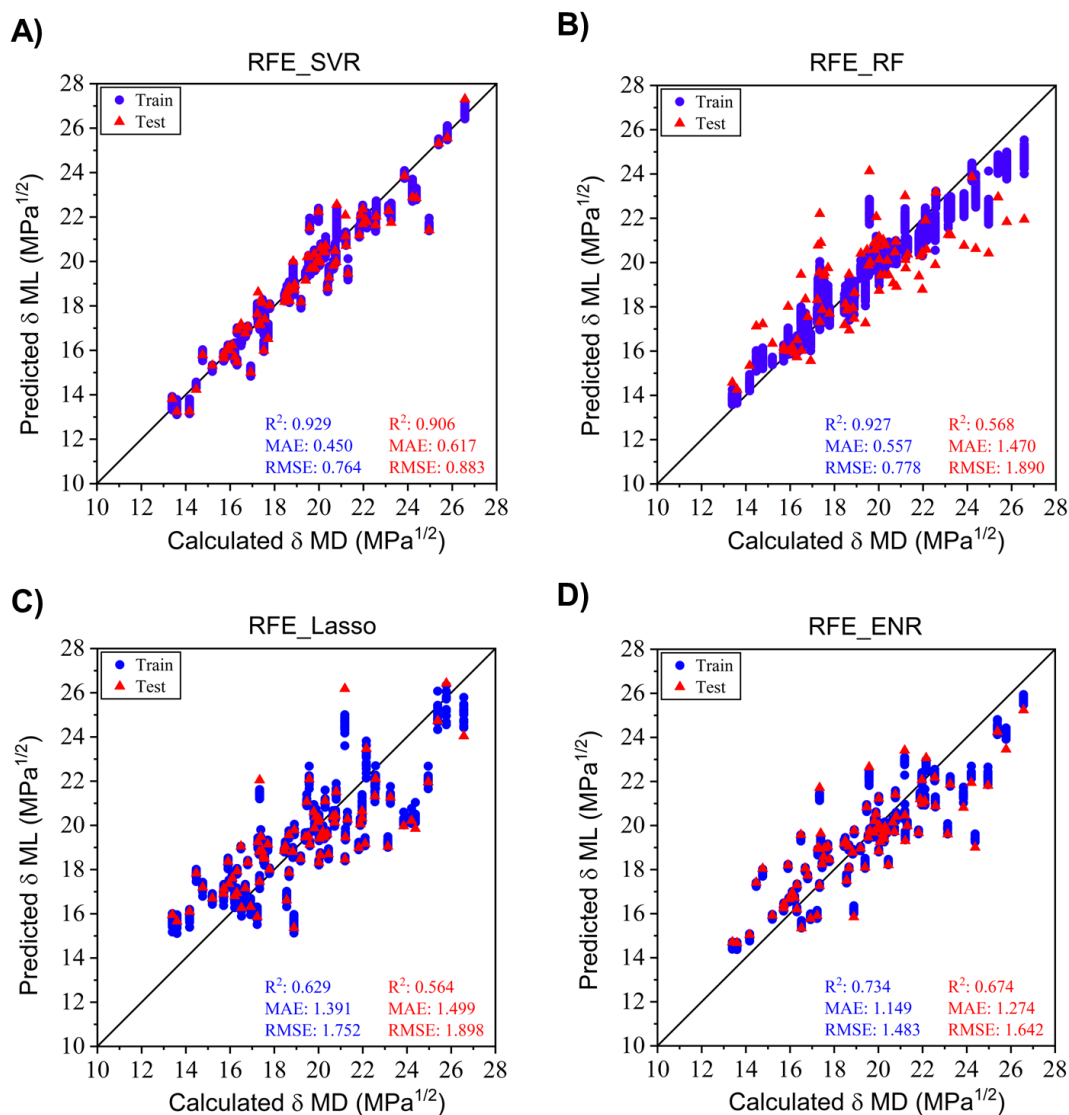
**Figure 15.** Performance comparison of ML models with PCC and MD results: **A)** SVR, **B)** RF, **C)** Lasso, **D)** ENR. MAE and RMSE in  $\text{MPa}^{1/2}$ . The blue traces show the change in predictions during training with LOOCV, where each sample is selected for validation until all samples have been selected once.

ML models using PCA gave much better results than those using PCC, but they were still not satisfactory (figure 16). The SVR and RF models scored high in training but too low in testing, which indicated a high risk of over-fitting (figures 16A and B). The Lasso and ENR models were not over-fitted, just under-performing.



**Figure 16.** Performance comparison of ML models with PCA and MD results: **A)** SVR, **B)** RF, **C)** Lasso, **D)** ENR. MAE and RMSE in MPa<sup>1/2</sup>.

The best result for dataset 1 was found in the SVR model with RFE (figure 17A). RFE\_SVR achieved good scores without overfitting not only in training but also in testing. The very short blue traces for training represented very consistent predictions in LOOCV. RFE\_RF was comparable to SVR in training, but failed in testing, with a higher potential for over-fitting. The ENR model outperformed the Lasso model and the consistency during LOOCV was very satisfactory.



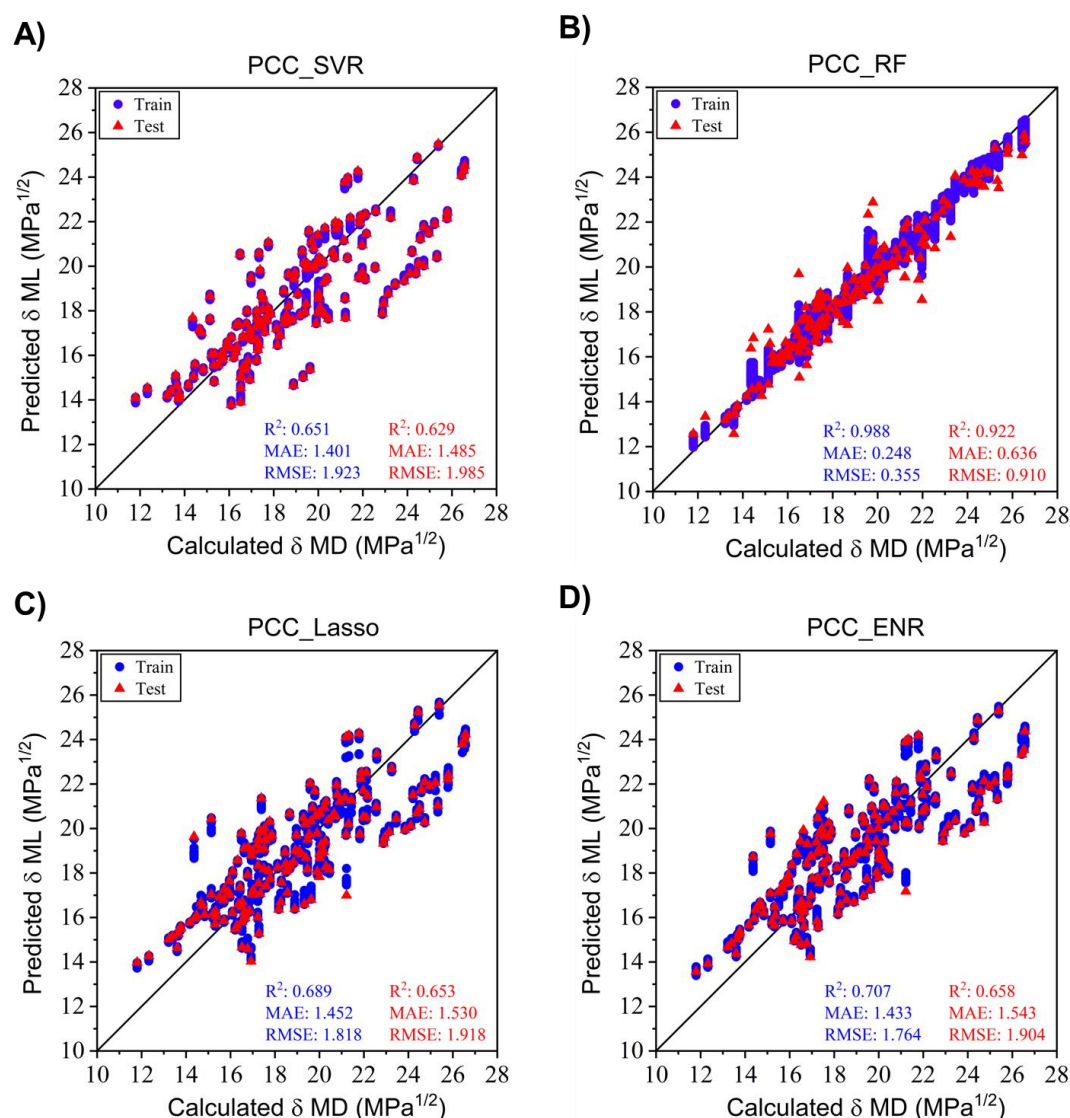
**Figure 17.** Performance comparison of ML models with RFE and MD results: **A)** SVR, **B)** RF, **C)** Lasso, **D)** ENR. MAE and RMSE in  $\text{MPa}^{1/2}$ .

To summarize, only ML models using RFE for descriptor selection showed acceptable performance, which was possibly due to the mechanism of RFE. As a wrapper-type algorithm, the selected ML model will become the core of RFE, and the descriptor selection procedure is more tailored. Nevertheless, direct prediction of solubility parameter  $\delta$  via one-step ML has been demonstrated to be feasible.

## 2.4 ML performance for dataset 2

Based on the data structure of dataset 1, dataset 2 was further extended to include more

polymer chain descriptors, which already introduced in previous sections. The performance of ML models will be presented in the same manner as those for dataset 1. In general, based on experience with practical cases, the performance of ML models improves as the amount of data increases. As shown in figure 18, all the models using PCC have improved considerably, especially the RF model (figure 18B). The good  $R^2$  scores in the test indicated that the RF no longer had over-fitting problems and that the model was essentially ready for practical use.

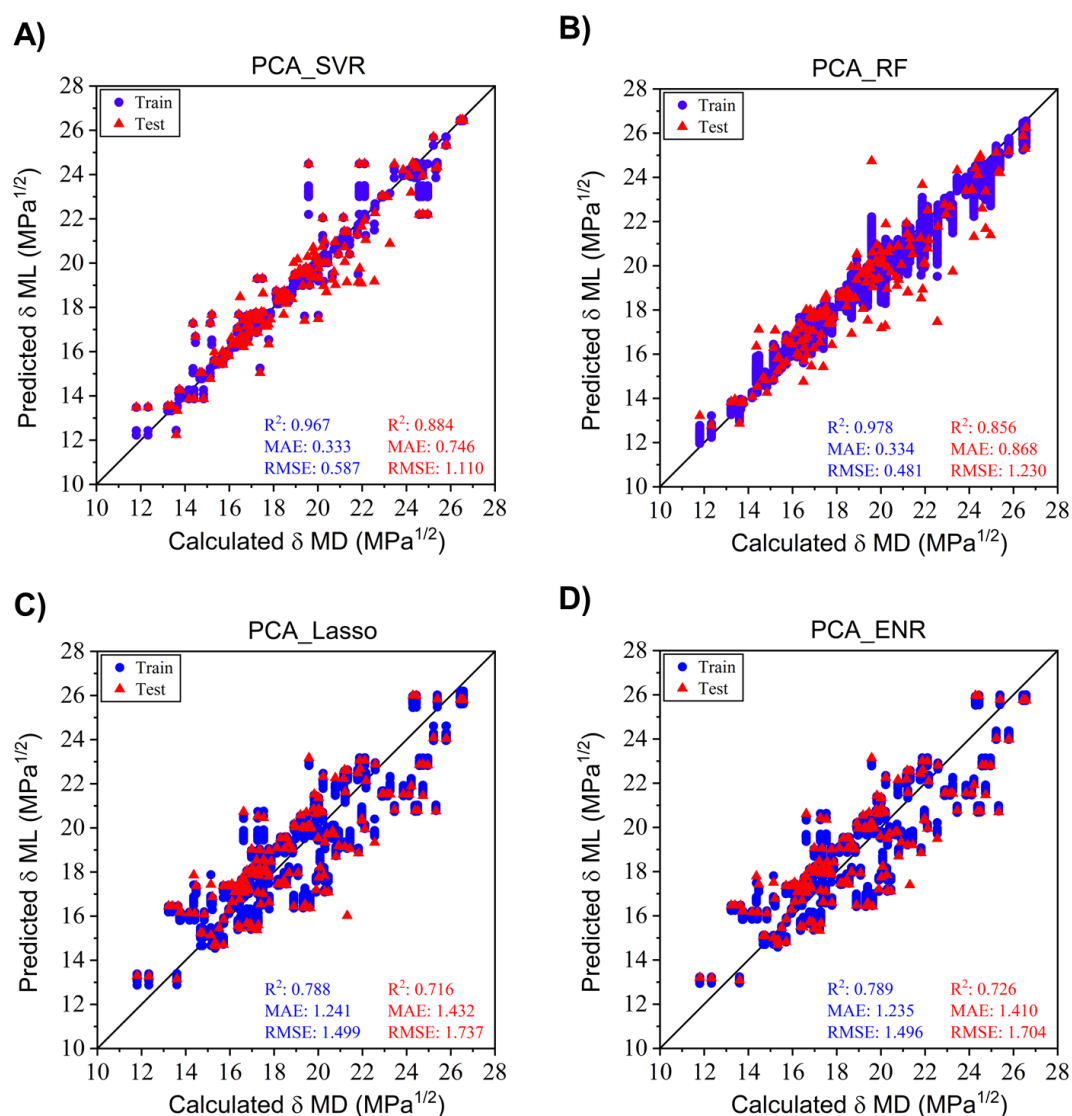


**Figure 18.** Performance comparison of ML models with PCC and MD results: A) SVR, B) RF, C) Lasso, D) ENR. MAE and RMSE in MPa<sup>1/2</sup>.

Similar to the PCC models, all PCA ML models showed significant improvements over the models for dataset 1. The SVR and RF models no longer have over-fitting



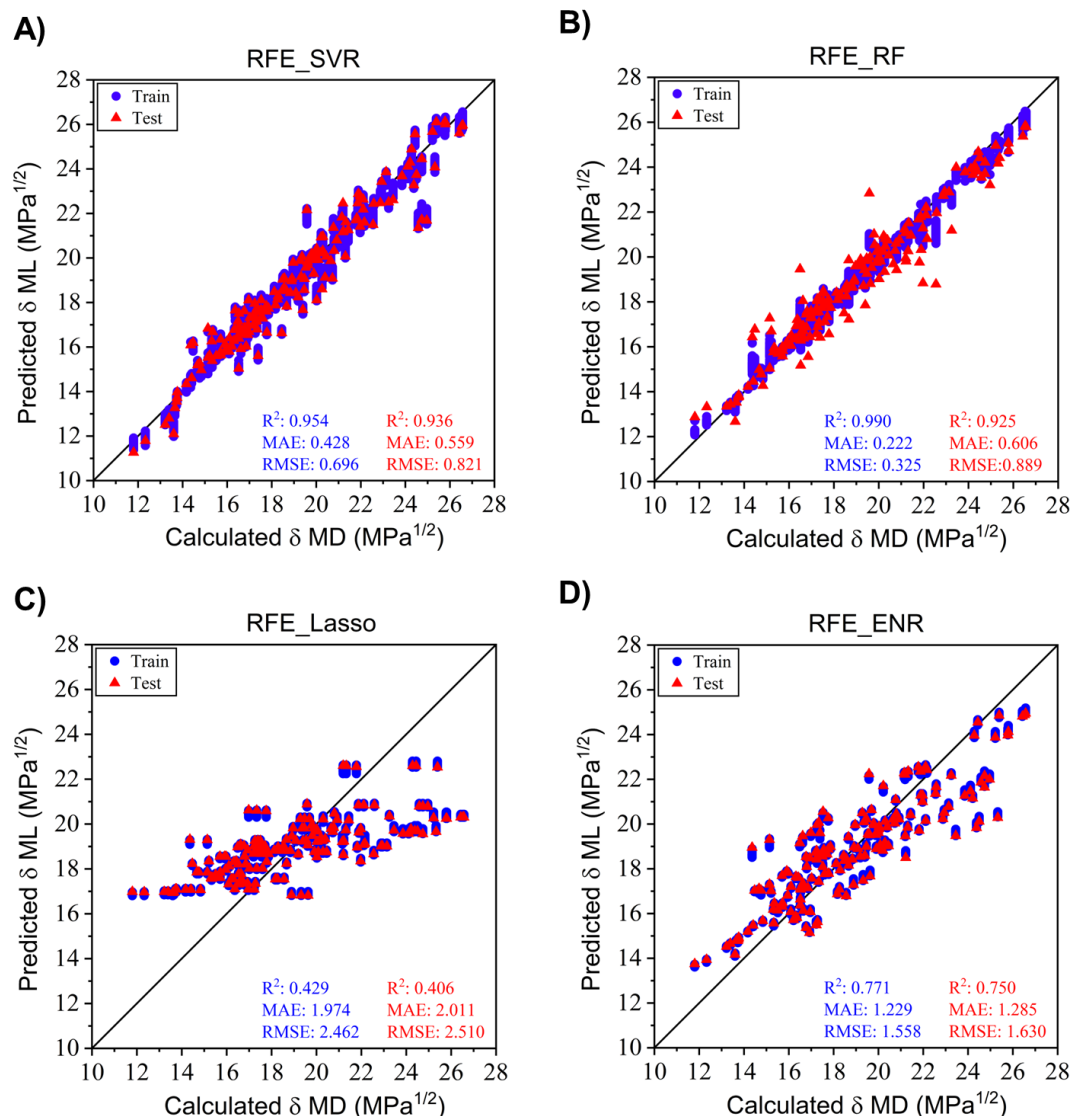
issues (figure 19A and B), and the R2 scores for the test were fairly good but did not reach 0.9. For SVR and RF, it is clear to see that the majority of the samples in the dataset have  $\delta$  concentrated in the range of 16 to 22, which makes this portion of the data more effective in training. If there are more data with  $\delta$  greater than 22, ML models may further refine the issue of too much scatter. The Lasso and ENR models did not show much improvement in training, but the test scores were much better than those for dataset 1.



**Figure 19.** Performance comparison of ML models with PCA and MD results: **A)** SVR, **B)** RF, **C)** Lasso, **D)** ENR. MAE and RMSE in  $\text{MPa}^{1/2}$ .

Of all the ML models for dataset 2, the SVR and RF models with RFE performed the best (figure 20A and B). The RFE\_RF model showed a tremendous level of

improvement over the model for dataset 1, and no over-fitting was detected. However, the  $R^2$  score of RFE\_SVR was even higher than the RF in the test, although the training score was slightly worse.



**Figure 20.** Performance comparison of ML models with RFE and MD results: **A)** SVR, **B)** RF, **C)** Lasso, **D)** ENR. MAE and RMSE in  $\text{MPa}^{1/2}$ .

It is also interesting to compare the Lasso model with the ENR model because the hyperparameters of ENR indicating that the model was an L1-regularized linear model identical to the Lasso model, but the two models eventually showed different performance. This actually remains because of the mechanism by which RFE operates. As described in section of the dataset 1, the RFE uses selected ML model as the core for filtering the descriptors, with different filtering results for different models. Two

models were trained with different filtered datasets and then produced different final results.

The RFE\_SVR model was found to be the best model for dataset 2, not only because of the highest test  $R^2$  scores, but also considering the computational demands. The critical point is that training SVR requires much less computational resources than RF, which is the very reason that SVR becomes the best model for dataset 2 in the whole work. The computing time of all models for dataset 1 and 2 is provided in Appendix Ch.I.5.

Through this work, several ML models with satisfactory performance were found for dataset 2. These models were essentially ready for practical use and their performance may be further optimized as the dataset expands in future work. The inclusion of descriptors obtained from the polymer chains improved the ML models significantly, and this is supposed to solve the third major problem from our previous work. At this point, a ML procedure has been completely constructed based on simulation data, while the dataset contains structural information about repeating units (REs) as well as polymer chains. This procedure comprehensively addresses the key problems summarized in previous work and has the capability of continuous refinement.

### 3 Conclusion

In this study, a polymer database was constructed containing structural information on polymer REs and polymer chains. The Hildebrand solubility parameters  $\delta$  of the polymers were obtained through molecular dynamics simulations as target values for machine learning. Two datasets were extracted from the database with different descriptors selected. Three descriptor selection methods were combined with four ML algorithms to generate 12 ML models for each of the two datasets. The relatively good performance of the ML model for dataset 2 demonstrated that it is feasible to predict polymer  $\delta$  using a one-step ML procedure.

With the addition of chain-related descriptors, most of the models have been significantly improved and PCC\_RF, RFE\_RF and RFE\_SVR were made available for practical applications. During the work, it was found that the choice of descriptor selection methods has a significant impact on the model performance, especially the RFE. RFE\_SVR was considered to be the best model for the current work, not only because of its performance, but also because the computation time was much shorter than the RF model during LOOCV. However, as the dataset expands, LOOCV will become more computationally intensive, so other cross-validation methods will have to be used and the computational cost of RF would be different at that time. As a linear model, elastic net (ENR) showed much better results than Lasso and it has the potential to be further improved. One of the most important things about linear models is that they are simple and highly explainable<sup>[84]</sup>. Powerful but complex models are sometimes not a good choice for scientific research because scientists are more interested in understanding the path or reason for the model's decisions or predictions<sup>[84]</sup>. Moreover, complex models such as neural networks often require large amounts of data for training, which is also very difficult for many scientific tasks<sup>[49]</sup>. If the quality of the dataset is high enough, good results can also be obtained using simple and explainable models, which is one of the goals that this work wants to pursue.

The search for new descriptors is one of the most important tasks for the future, and there are several possibilities. Much information still can be extracted and encoded from polymer chains, for example radius of gyration, viscosity and other thermodynamic properties. In the current work, only two quantum chemical descriptors were chosen, but other chemical quantities describing reactivity or interactions, such as chemical hardness and polarizability,<sup>[85]</sup> can be obtained from DFT calculations, and these could potentially be added to future datasets. It is also important to collect stable and reliable data from experimental characterizations, as simulations sometimes suffer from a lack of chemical accuracy, which can be corrected by data from real experiments.<sup>[49]</sup> In addition, many experimental characterization results are shown by various spectrum such as Differential Scanning Calorimeter (DSC),<sup>[86]</sup> Fourier

Transform Infrared Spectroscopy (FTIR),<sup>[87]</sup> and Nuclear Magnetic Resonance Spectroscopy (NMR).<sup>[88]</sup> A growing number of publications introduce ML methods for predicting material properties using a various spectrum, some of which are adequate for the scientific task.<sup>[89, 90]</sup> If used in the right way, the spectrum can be a treasure chest of information and become high-quality training material for ML. The dataset will thus not only contain structure-related information but will also include more statistical information, which is helpful for the improvement of ML performance and generalization. Of course, such an endeavor requires assembling more manpower and resources to get started and can never be accomplished by a few scientists or a mini team.

## Chapter II

# Prediction of polymer-drug miscibility based on the Flory-Huggins theory

---

A favorable interaction of the drug with the carrier material is significant for an efficient encapsulation. In this chapter, the miscibility of selected polymer-drug mixtures was analyzed by MD simulation results. Polyester amides (PEAs) are a class of polymers that are considered potential competitors to poly(lactic-*co*-glycolic acid) (PLGA) for drug delivery applications. In this work, two methods based on the Flory-Huggins theory were applied to predict miscibility and to compare the miscibility of certain PEAs with PLGA. L-valine (Val) and L-isoleucine (Ile) based PEAs showed comparable performance to PLGA and different miscibility trends were found by MD simulations. The theoretical predictions were in good agreement with the experimental observations, which confirms the reliability of the MD simulations.

## 1 Materials and methods

### 1.1 Polymer candidates and reference drugs

At the request of the collaborative experimental group, several polymers were selected from table 4 and their miscibility with specific drug molecules was predicted on the basis of Flory-Huggins mean field theory.<sup>[27]</sup> The information of polymers and corresponding drug molecules were listed in table 4 and due to the complexity of chemical names, abbreviated names are used for all polymers in this section.<sup>[30, 91]</sup>

**Table 4.** The summary of selected polymers and drugs

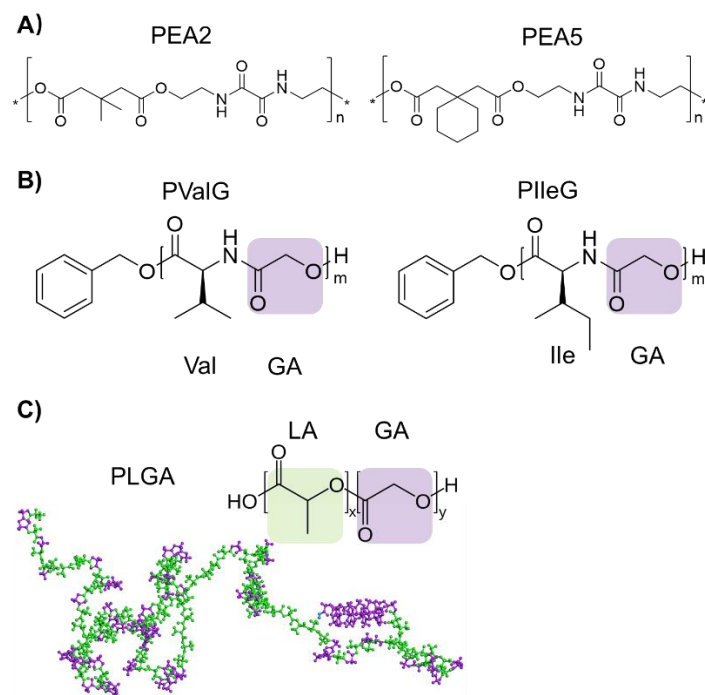
<b>Polymer (Degree of polymerization)</b>	<b>Drug (number of molecules)</b>
PEA2 (10) <sup>[28]</sup>	Indomethacin (IMC) (1) <sup>[28]</sup>
PEA5 (10) <sup>[28]</sup>	Indomethacin (IMC) (1) <sup>[28]</sup>
PValG (55) <sup>[91]</sup>	BRP-187 (1) <sup>[91]</sup>
PIleG (55) <sup>[91]</sup>	BRP-187 (1)
PValG (55) <sup>[91]</sup>	BRP-201 (1) <sup>[91]</sup>
PIleG (55) <sup>[91]</sup>	BRP-201 (1)
PLGA (184)	BRP-187 (1)
PLGA (184)	BRP-201 (1)

Figure 21 shows the structural formular of all polymers. Random copolymer poly(lactic-*co*-glycolic acid) (PLGA) is considered as one of the most effective biodegradable polymer for drug delivery.<sup>[92]</sup> Due to the low toxicity and biocompatibility with tissue and cells, PLGA has been approved by US FDA to use in drug delivery systems.<sup>[93]</sup> But PLGA still has drawbacks including the limited tuning ability of the nanoparticles, e.g., regarding limited drug encapsulation capacities and the occurrence of free drug precipitates in addition to the nanoparticles.<sup>[94]</sup>

Polyester amides (PEAs) are a class of polymers that growing number of studies have supported as an alternative to PLGA.<sup>[28, 30, 95]</sup> Additional amide moieties in the PEA structures (figures 21A and B) can contribute to hydrogen bonding between the polymer and the drug or between the polymer and the polymer, thereby influencing the properties of the polymeric drug carrier, such as degradation behavior and encapsulation ability. L-valine (Val) and L-isoleucine (Ile) based PEAs (PValG and PIleG in figure 21B) were selected as the candidates because of the similarity of the glycolic acid block (GA) in their structures to PLGA, which theoretically has comparable performance as a drug carrier.<sup>[96]</sup>

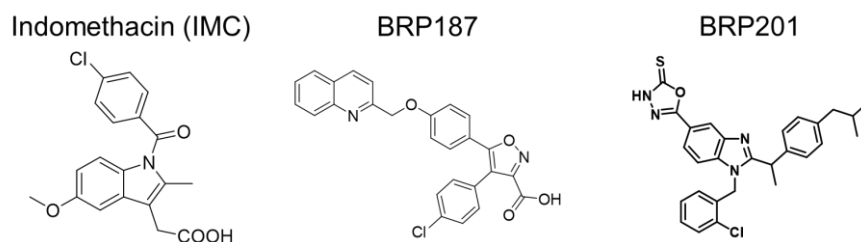
Based on the data from the PLGA used by the experimental partner, a PLGA chain

with 48.3/51.7 block ratio of lactic- (LA) and glycolic acid (GA) was built for MD simulations. The two repeating units of the PLGA molecule were randomly combined according to the reactivity ratios (LA: 4.5, GA: 0.32) from the literature.<sup>[97]</sup> The constructed polymer chain contains a total of 184 repeating units, terminates in a hydroxyl group and weighs about 12000 g/mol (figure 21C).



**Figure 21.** Structural formular of **A)** PEA2 and PEA5, and **B)** PValG and PIlleG, and **C)** atomistic model of PLGA random block copolymer with 48.3/51.7 molar ratio of LA/GA. Green: lactic acid (LA) with reactivity ratio 4.5 and purple: glycolic acid (GA) with reactivity ratio 0.32. Abbreviations: **Val**: L-valine and **Ile**: L-isoleucine.

Figure 22 shows the structural formular of all drug molecules. Indomethacin is a nonsteroidal anti-inflammatory drug (NSAID) used to treat mild to moderate acute pain<sup>[28]</sup> and relieve symptoms of arthritis (osteoarthritis and rheumatoid arthritis) or gout, such as inflammation, swelling, stiffness, and joint pain.



**Figure 22.** Structural formular of drug molecules



Two water-insoluble compounds BRP 187 and BRP 201 are lipophilic anti-inflammatory drugs, which are representatives of the novel drug class of dual inhibitors that target the 5-lipoxygenase-activating protein (FLAP) and the microsomal prostaglandin E2 synthase-1 (mPGES 1).<sup>[98]</sup> These two drugs were selected as the reference drugs, in order to evaluate the potential and stability of above-mentioned polymers as drug carrier materials.

## 1.2 Simulations and calculations

Firstly, the cohesive energy density (*CED*) or Hildebrand solubility parameter  $\delta$  for pure compounds and mixture were obtained through MD simulations. The simulation process was same as the process applied in chapter I. After collecting *CED*, segment volume  $V_m$  and volume fraction  $\varphi_s$  of the compounds ( $\varphi_p$  for polymer and  $\varphi_s$  for drug), Flory-Huggins parameter  $\chi_{p-s}$  would be calculated by two methods (equation 33 and equation 34)<sup>[28, 33]</sup>

$$\text{Method I} \quad \chi_{p-s}^{\text{I}} = \frac{V_m}{RT} (\delta_p - \delta_s)^2, \quad (33)$$

$$\text{Method II} \quad \chi_{p-s}^{\text{II}} = \frac{V_m}{RT} (\varphi_p CED_p + \varphi_s CED_s - CED_{p-s}). \quad (34)$$

It is clear that method I only needs the  $\delta$  from the simulations of pure compounds but method II needs the simulation results from exact mixture systems. Thus, Method I is considered a rapid approximation, and Method II can be used for accurate predictions, but tends to be much more computationally expensive than Method I. Then, the Gibbs free energy change of mixing  $\Delta G_{\text{mix}}$  could be further calculated using both  $\chi$  (equation 2)

$$\Delta G_{\text{mix}} = RT(n_1 \ln(\varphi_1) + n_2 \ln(\varphi_2) + n_1 \varphi_2 \chi_{1-2}),$$

for convenience, the equation can be further derived as follows

$$\Delta G_{\text{mix}} = RT \left( \frac{\varphi_p}{N_p} \ln(\varphi_p) + \frac{\varphi_s}{N_s} \ln(\varphi_s) + \varphi_p \varphi_s \chi_{p-s} \right), \quad (35)$$

where  $N_i$  is the number of segments of the compounds.

## 2 Results and discussion

### 2.1 Simulations for PEA-Indomethacin mixture

Table 5 shows the results of simulations for pure compounds and mixtures, including Hildebrand solubility parameters and Gibbs free energy change of mixing.

**Table 5.** Results of MD simulations: cohesive energy density,  $CED$ , Hildebrand solubility parameters,  $\delta$ , energy of mixing,  $\Delta E_{\text{mix}}$ , Flory-Huggins parameters,  $\chi_{\text{sim}}$ , and Gibbs free energy change of mixing,  $\Delta G_{\text{mix}}$ .  $\chi_{\text{sim}}$  and  $\Delta G_{\text{mix}}$  were calculated with methods I and II as described in the text.

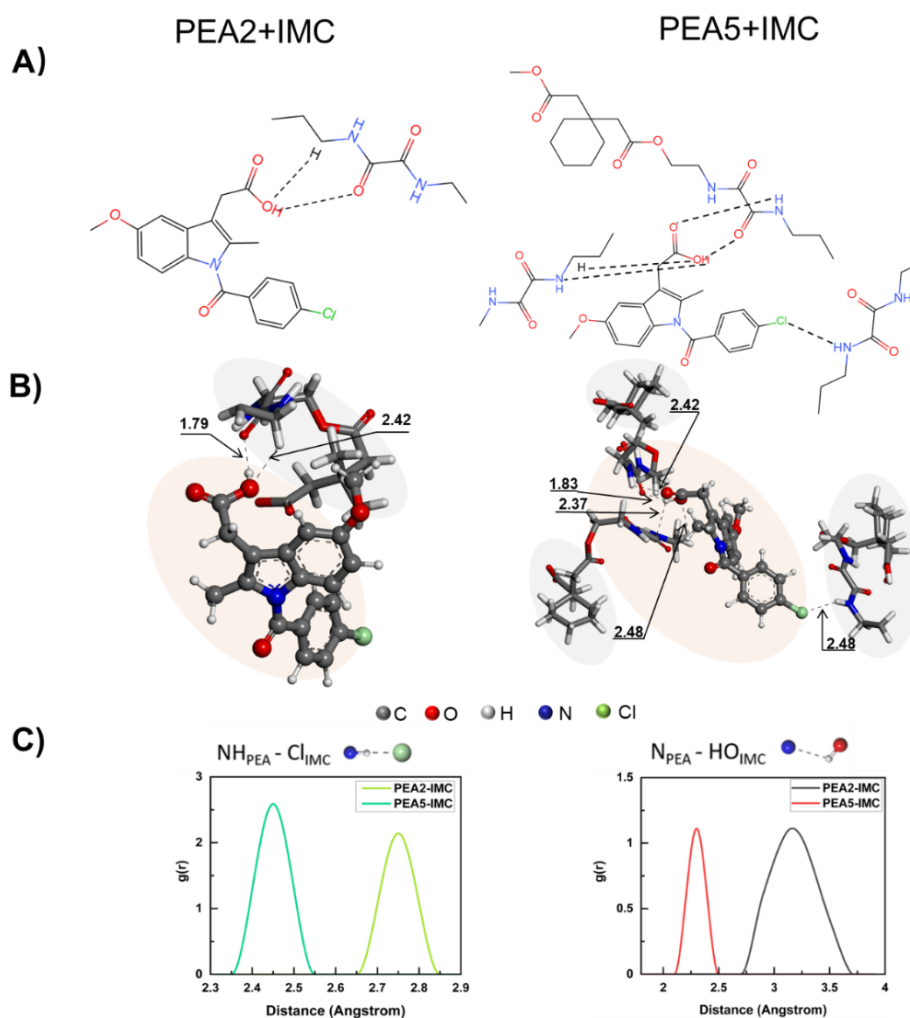
	$CED$ [J cm <sup>-3</sup> ]	$\delta$ [ $\sqrt{\text{MPa}}$ ]	$\chi_{\text{sim}}$		$\Delta G_{\text{mix}}$ [J mol <sup>-1</sup> ]	
			I	II	I	II
PEA2+IMC	420.58	20.51	1.13	0.15	-65.95	-82.15
PEA5+IMC	389.24	19.73	2.10	-0.46	-50.06	-92.08
PEA2	421.05	20.52	-	-	-	-
PEA5	383.94	19.59	-	-	-	-
Indomethacin	570.95	23.89	-	-	-	-

It is clear to see that  $\Delta G_{\text{mix}}$  calculated by Method I and II for PEA2+IMC and PEA5+IMC were both negative, which indicated miscibility of IMC with PEA2 as well as PEA5. However, the values from the two methods showed opposite trends. The values of  $\Delta G_{\text{mix}}$  obtained with Method I indicated better solubility of IMC in PEA2 than in PEA5, which is in disagreement with experimental observations.<sup>[28]</sup> In contrast, method II correctly predicted higher solubility of IMC in PEA5 as compared to PEA2. Method I provided a simple empirical estimation of  $\chi_{\text{sim}}$  based on pure compounds, which may account for the poor performance as it did not take into account specific molecular interactions such as hydrogen bonding.<sup>[33]</sup>

Figure 23A and B show that multiple hydrogen bonds<sup>[99]</sup> were formed between the

IMC molecule and the surrounding polymer chains with bond distances from 1.72 to 2.48 Å. More pronounced hydrogen bonding in PEA5+IMC compared to PEA2+IMC was evident from the RDF plots (figure 23C). In the case of the intermolecular atomic  $\text{NH}_{\text{PEA}} - \text{Cl}_{\text{IMC}}$  pairs the maximum was 2.45 and 2.75 Å for PEA5+IMC and PEA2+IMC, respectively. Therefore, the chlorine atom of IMC was closer to hydrogen atoms of PEA5 compared to PEA2 (figure 23C, left) contributing stronger to the attractive heteromolecular forces. The same trend was found for  $\text{N}_{\text{PEA}} - \text{OH}_{\text{IMC}}$  (figure 23C, right). Intermolecular RDF plots for other types of hydrogen bonds that were additionally found in the simulated cells appeared similar in strength (Appendix Ch.II.2).

According to MD simulations, stronger hydrogen bonding in PEA5+IMC compared to PEA2+IMC possibly accounts for the better solubility of IMC in PEA5, which provides theoretical support for the experimental observations of the cooperative experimental group.<sup>[28]</sup>



**Figure 23.** Snapshots of MD simulation trajectories showing examples of hydrogen bonds between the indomethacin molecule and PEA chains PEA2-IMC and PEA5-IMC: **A)** chemical structure representation **B)** ball-and-stick representation (grey background for PEA, orange background IMC). **C)** Intermolecular radial distribution function (RDF) plots of the two hydrogen bonding types possibly responsible the different solubility behavior. The interaction distance is in Angstrom ( $\text{\AA}$ ).

## 2.2 Simulations for New PEAs-BRP drugs

Table 6 shows the results of the simulations based on the two methods for the Flory-Huggins parameters, Gibbs free energy change of mixing and the predicted solubility limits for PValG, PileG and the corresponding mixtures with BRP-187 and BRP-201. Simulation results of PLGA random copolymer were also included as the comparison. For the four simulated mixtures of PValG and PileG,  $\Delta G_{\text{mix}}$  was negative incorporating roughly 1 wt% of drug, using both the simpler Method I ( $\chi_1$ ) and the more

accurate Method II ( $\chi_{II}$ ). Thus, the PValG and PIlleG are considered to be miscible with the drugs at a mass fraction that is commonly applied for the formulation of drugs into nanoparticles.<sup>[91]</sup> Although the simulation results for PLGA also indicate the miscibility of two drugs, but different trend in thermodynamic properties can be seen, comparing with two polyester amides.

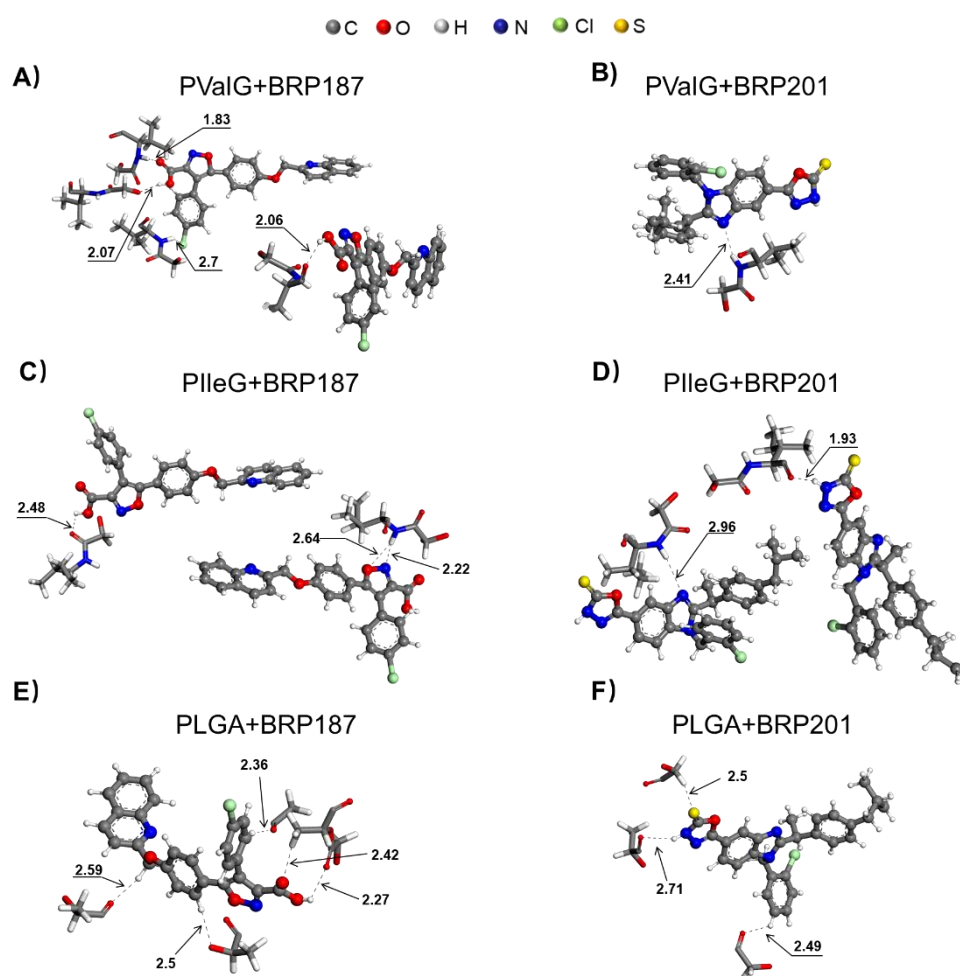
**Table 6.** Flory-Huggins (FH) parameters  $\chi_I$  and  $\chi_{II}$  calculated with the method I and II, respectively, the corresponding free energy of mixing (J/mol) and solubility limits (wt%) for the simulated mixtures of PValG, PIlleG and PLGA with BRP-187 and BRP-201.

Polymer	Drug	Drug weight ratio [wt%]	FH parameters		$\Delta G_{\text{mix}} (\chi_I)$ [J mol <sup>-1</sup> ]	Solubility limit ( $\chi_I$ ) [wt%]	$\Delta G_{\text{mix}} (\chi_{II})$ [J mol <sup>-1</sup> ]	Solubility limit ( $\chi_{II}$ ) [wt%]
			$\chi_I$	$\chi_{II}$				
PValG	BRP-187	1	0.73	2.63	-39.1	5.0	-26.7	1.5
PValG	BRP-201	1.1	0.13	3.85	-49.6	20.2	-16.1	1.5
PIlleG	BRP-187	1	1.31	2.68	-39.1	3.5	-26.7	1.4
PIlleG	BRP-201	1	0.39	3.23	-47.3	8.7	-21.8	1.6
PLGA	BRP-187	0.8	0.099	0.035	-18.22	21.5	-18.39	38.1
PLGA	BRP-201	0.8	1.6e <sup>-4</sup>	-4.6e <sup>-3</sup>	-18.49	$\infty$	-18.5	$\infty$

Theoretically, the values of the FH parameters ( $\chi$ ) above 0.5 indicated generally limited miscibility between the drug and the polymers, as shown in the results of PValG and PIlleG.<sup>[27]</sup> However, calculations based on Method II for all mixtures at 1 wt% drug concentration showed that although  $\chi$  values for PLGA mixtures were much smaller than those for polyester amide mixtures, the difference in  $\Delta G_{\text{mix}}$  was very small. In addition, the  $\chi$  and  $\Delta G_{\text{mix}}$  values of the PValG and PIlleG mixtures obtained by the two methods were also significantly different, but the results for the PLGA mixtures were very close. The differences between the two methods can be explained by the conclusion of simulations for the PEA-Indomethacin mixtures that specific interactions between the polymer and the drug were not taken into account in Method I.<sup>[28]</sup> Moreover, the absence of greatly differing results suggests that such interactions may not play a role in PLGA mixtures, and it will be discussed later.

The interaction between drug and polymer is shown for each system in figure 24. Indeed, hydrogen bonding between drug and polyester amides contributed to the

miscibility (figure 24A-D). As both drugs feature hydrogen bond donors (the carboxylic acid functionality in BRP-187 and one nitrogen atom in the oxadiazole-based heterocycle in BRP-201), interactions with ester and amide oxygen atoms of the PEA were seen. In addition, the amide moieties of the PEA act as hydrogen bond donors as they are in close proximity to various hydrogen bond acceptors of the drug (*e.g.*, the isoxazole ring and chlorine atom of BRP-187, or the imidazole ring of BRP-201).



**Figure 24.** Ball-and-stick representation of polymer segment – drug interactions. **A)** PValG with BRP-187, **B)** PileG with BRP 201, **C)** PileG with BRP-187, **D)** PileG with BRP-201, **E)** PLGA with BRP-187 and **F)** PLGA with BRP-201. The interaction distance is in Angstrom ( $\text{\AA}$ ).

Hydrogen bonding was also observed in mixtures of PLGA and two BRP drugs, and both PLGA repeating units can form this interaction with the BRP drug molecule. According to the literature,<sup>[99]</sup> interactions cannot generally be considered as "strong"

hydrogen bonding if the interaction distance is greater than 2.5 Å. The same situation of hydrogen bonding can be found in the mixture of PValG with BRP201 and the values of  $\Delta G_{\text{mix}}$  based on Method II were also very close. Overall, the relatively weak hydrogen bonding between PLGA and BRP drugs could be the reason for the minor differences in the  $\Delta G_{\text{mix}}$  from two calculation methods, because such interaction didn't play a role in the miscibility of PLGA with BRP drugs. This also indicates that the driving force for miscibility in PLGA could be different from that of Polyester amides when mixing with BRP drugs.

As shown in table 6, the less accurate Method I predicted the solubility limit of BRP drugs in PEAs to be between 3.5 and 20.2 wt%. However, the more accurate calculations using Method II consistently predict a miscibility limit of about 1.5 wt%, which is in good agreement with experimental measurements.<sup>[91]</sup> For PLGA mixtures, the solubility limits of BRP drugs showed very different trends, especially for PLGA-BRP201. According to theory, BRP201 can mix infinitely in PLGA at 300 K because of a negative  $\chi$ , but there is no experimental evidence to support this prediction. Additionally, it should be noted that  $\chi$  is a concentration-dependent quantity, and in the current work,  $\chi$  was obtained by performing MD simulations on mixtures with only one drug molecule. As the concentration of the drug in the mixture increases,  $\chi$  could be affected and the predictions might be different. In this study, it was preferred to analyze the miscibility by the comparison of  $\Delta G_{\text{mix}}$  at a certain drug concentration, and the solubility limit could be used as a theoretical estimate only. For better and more reliable predictions, it is recommended that the mixtures be simulated on the basis of known drug concentrations close to the experiments, as in the case of the PEA-BRP mixtures.

The simulation results showed comparable performance of polyester amides to PLGA in miscibility with both BRP drugs with about 1 wt% drug concentration and this was also in good agreement with experimental observations by encapsulation test of BRP drugs with polymer nanoparticles.<sup>[91]</sup> However, it should be noted that the miscibility mechanism in reality is more complex and involves more compounds in the

preparation of polymer nanoparticles and drug encapsulation. Simulations can only provide a theoretical perspective on the potential influencing factors of polymer-drug interactions and point out a possible way to improve the design of polymeric drug carriers.

### **3 Conclusion**

The *in silico* miscibility prediction based on Flory-Huggins theory has been demonstrated to be a reliable tool for describing the miscibility behavior of polyester amides with specific drugs, in good agreement with experimental observations.<sup>[28, 30]</sup> Based on the comparison of the simulation results of PEA2+IMC and PEA5+IMC mixtures, hydrogen bonding between PEA and drug molecules was considered to be the key factor for miscibility. The same phenomenon occurred in the simulations of PValG and PIlleG mixtures, but the simulations of PLGA mixtures indicated that hydrogen bonding played a minor role in miscibility with BRP187 and BRP201. Therefore, to further improve the accuracy of the prediction, not only hydrogen bonding interactions but also other types of host-guest interactions (e.g., van der Waals forces and hydrophobic interactions) must be taken into account in accurate simulations<sup>[100]</sup> and the absence of consideration of host-guest interactions is the reason for the inaccuracy of the fast calculation method.

Combined with experimental measurements, it can be confirmed that the performance of PValG and PIlleG mixed with BRP187 and BRP201 is comparable to that of PLGA, but it is important to note that in practice, hydrophobic PEA and PLGA are typically used as the hydrophobic block of larger block copolymers, e.g. poly(2-ethyl-2-oxazoline)-block-poly(ester amide) and poly(ethylene glycol) -block-PLGA.<sup>[91]</sup> Thus, miscibility can be affected not only by the interaction between the polymer and the drug, but also by many other factors. Still, simulations can provide theoretical support for improving the performance of polymeric drug carriers. Based on the differences in the



results obtained from the two calculation methods, it is clear that simulations of the exact mixing system are still necessary in order to be more consistent with the experimental observations. However, the computational cost of such simulations may become an issue when more complex polymer candidates come into sight. Since machine learning approaches have been applied to predict the solubility parameters of polymers and have shown good accuracy and agreement with MD simulations (Chapter I), and the MD simulation workflow for mixture systems in current work is almost identical to that for pure polymers and drugs. Therefore, the simulation results of mixtures are available for machine learning approaches and miscibility prediction based on a mixture database can be realized in the near future.

## Chapter III

# DPD simulations of modified block copolymer micelles with specific therapeutic agents

---

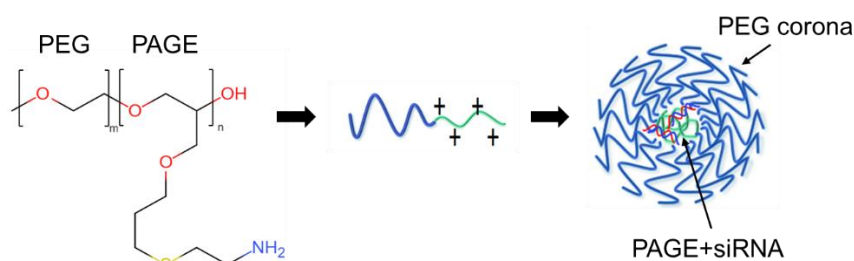
In this chapter, the encapsulation ability of modified block copolymers with specific therapeutic agents was investigated using dissipative particle dynamics (DPD) simulations to identify the key factors for improving the performance of polymer polyplex micelles. In difference to the polymers used in previous chapters, a series of block copolymers were chosen to build nano micelles to encapsulate small interfering RNAs (siRNAs) as pharmaceutical agents.<sup>[101]</sup> The poly(ethylene glycol)-*block*-poly(allyl glycidyl ether) (PEG-*b*-PAGE) diblock copolymer carries amino moieties in the side chain, which theoretically interact with the phosphate group of the siRNA molecule to enhance encapsulation. The molecular weights or block ratios of PEG blocks are thought to affect the properties of polyplex micelles made from PEG-*b*-PAGE, therefore a series of block copolymers with different PEG block ratios were constructed. The aim of this study was to evaluate the relationship between the PEG block ratio and encapsulation, both experimentally and by simulation, and find the most suitable polymer formulation for siRNA delivery.

## 1 Materials and methods

### 1.1 Polymers and small interfering RNA

In the current study, the PAGE block was functionalized by amino moieties, which imparted cationic charges to the molecule (PEG-*b*-PAGE<sub>NH2</sub>).<sup>[101]</sup> PAGE segment with

cationic charges allows polyion complexation with anionically charged siRNA, forming the core of the polyplex micelles with PEG as a corona. The ratio between PEG corona and siRNA complexed PAGE core of polyplex micelles was chemically varied by altering the degree of polymerization of PAGE<sub>NH2</sub>.



**Figure 25.** Structural formular of PEG-*b*-PAGE<sub>NH2</sub> and the theoretical structure of the polyplex micelle encapsulating siRNA.

Table 7 lists micelles prepared with four different PEG-*b*-PAGE<sub>NH2</sub> polymer chains. All polymer chains contain PEG blocks of the same length, but the PAGE<sub>NH2</sub> blocks are different. The Sequence of siRNA molecule for polyplex micelles is shown in table 8.<sup>[101]</sup>

**Table 7.** Physicochemical properties of polyplex micelles and Mn ratio of PEG:PAGE

Polyplex micelles	DP of PEG	Dp of PAGE	Mn of PAGE	Mn ratio of PEG:PAGE
EN15	42	15	2808	0.66
EN29	42	29	5556	0.34
EN60	42	60	11485	0.16
EN76	42	76	14579	0.13

DP: degree of polymerization

**Table 8.** The sequence of siRNA used for the encapsulation test

siRNA	Sequence (5' to 3')
Sense (22 base pairs) $T_m = 54^\circ\text{C}$	[2OMeU]CAUAUCAAAGAUACACCCCC
Antisense (22 base pairs)	[2OMeG]GGGGUGUAUCUUUGAAUAUGA

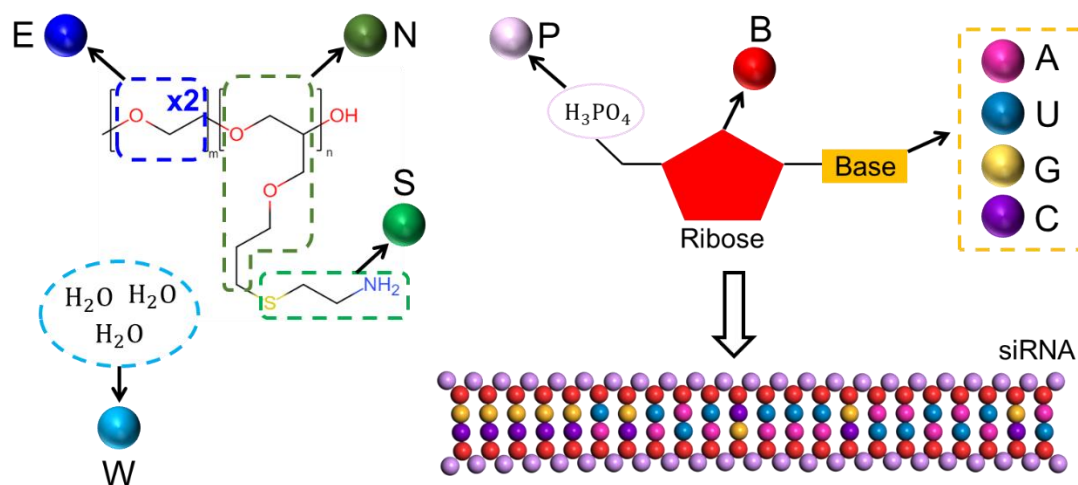
$$T_m = 54^\circ\text{C}$$

---

## 1.2 Dissipative particle dynamics simulation

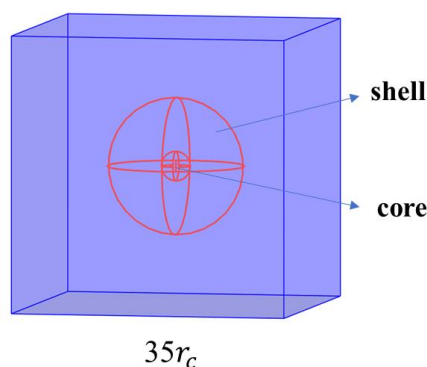
All DPD simulations were performed with the program **Materials Studio** (MS) and the **Mesocite** module.<sup>[54]</sup> The polyplex micelle systems simulated in this study were composed of siRNA, PEG<sub>42</sub>-*b*-PAGE<sub>NH2</sub> block copolymer and water. Coarse-grained models of all species involved in these systems are shown in figure 26. Every two PEG repeat units were coarse-grained into one bead (denoted as E), one PAGE<sub>NH2</sub> repeat unit was comprised of two beads N and S (S bead is positively charged because of amino moieties). The siRNA molecule was coarse-grained into six different types of beads (denoted as A, B, C, G, P, and U), in which A, U, G, and C represented the four different bases of siRNA, B and P represented the ribose and phosphoric acid (charged). The bead of water W contained three water molecules. In the DPD simulation all beads share the same mass and size, and the average volume, mass and radius of each bead was set to 177 Å<sup>3</sup>, 113 amu and 3.5 Å, respectively.

The coarse-grained models of polymer chains were built based on the experimental DP values (table 7) and the siRNA model was built based on the TLR4 sequence (table 8). Since simulations of actual micelles would be too computationally demanding even for DPD, we used smaller models that contained only one siRNA molecule and enough polymer material for its complete encapsulation. Nevertheless, such models are still expected to provide a qualitative picture of siRNA-polymer arrangement in the micelles.<sup>[101]</sup>



**Figure 26.** Coarse grained structures of PEG<sub>42</sub>-*b*-PAGE<sub>NH<sub>2</sub></sub> block copolymer, siRNA and water.

The length and time scale of DPD simulations in current work,  $r_c$  and  $t_c$ , were set to 8.1 Å and 0.0047 ns as 1 time step. Cubic simulation cells in the size of  $25 \times 25 \times 25 r_c^3$  were built. Using reduced units, spring constant was determined as 4.0 and dissipative force parameters as 4.5 for all models. The siRNA molecule was placed in the cell as the core and PEG<sub>42</sub>-*b*-PAGE<sub>NH<sub>2</sub></sub> molecules formed a shell with radius from 15 to 70 Å (figure 27). The rest of the space in the cell was filled with water. All cells were simulated for 20000 time steps. The repulsive parameter  $a_{ij}$  for all beads were obtained based on the MD simulations in Materials studio and the details are provided in Appendix Ch.III.2.



**Figure 27.** Schematic diagram of simulation cell. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

To investigate the effect of charges on the performance of polyplex micelles, DPD simulations of charged micelles were carried out for selected PEG<sub>42</sub>-*b*-PAGE<sub>NH2</sub> molecules. In charged bead micelle systems, the S beads of PAGE block carried positive charges and the P beads of the siRNA molecule carried negative charges. The nitrogen to phosphate (N/P) ratio was set to 5 according to the experiment.<sup>[101]</sup> An appropriate number of W beads were replaced by charged W beads with negative charge to maintain neutrality of the whole system. Particle-Particle-Particle-Mesh (PPPM) summation method was applied for electrostatic interactions of beads.

The DPD simulation for a charged system is usually more complicated than a neutral system and needs more sophisticated adjustments, for example the balance between charge values and spring constant between the beads.<sup>[41, 102]</sup> The influence will be discussed in section 2: Results and discussion.

According to the experience of DPD simulations in the soft matter systems,<sup>[103]</sup> the charges and the form of charges for DPD simulations have relatively little impact on qualitative analysis. While sophisticated tuning of specific interactions can lead to more rational microstructures, it is not cost-effective for mesoscale qualitative studies.<sup>[103]</sup>

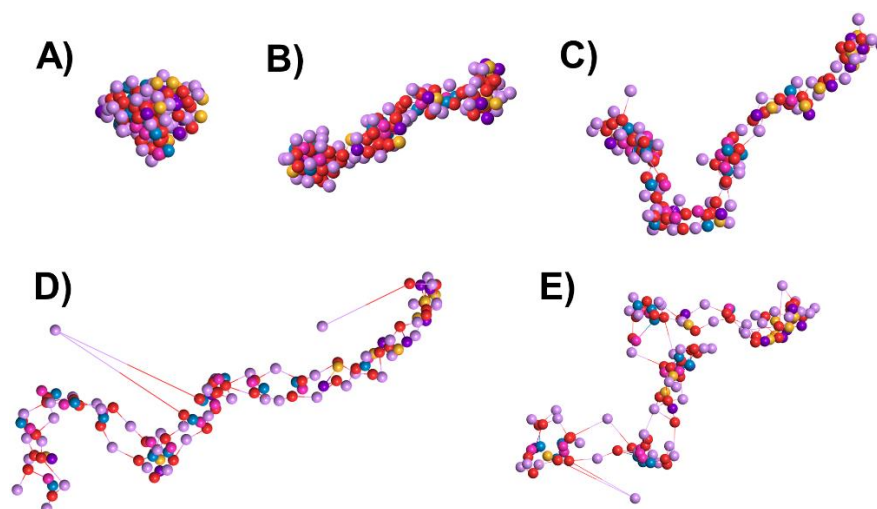
Therefore, it is preferable to implicitly include the electrostatic interactions between the PAGE block and siRNA in the DPD force field,<sup>[101]</sup> while all beads remain electrically neutral, resulting in a less consuming simulation. The comparison between charged and neutral micelles will be shown in next section.

## **2 Results and discussion**

### **2.1 DPD simulations for charged siRNA**

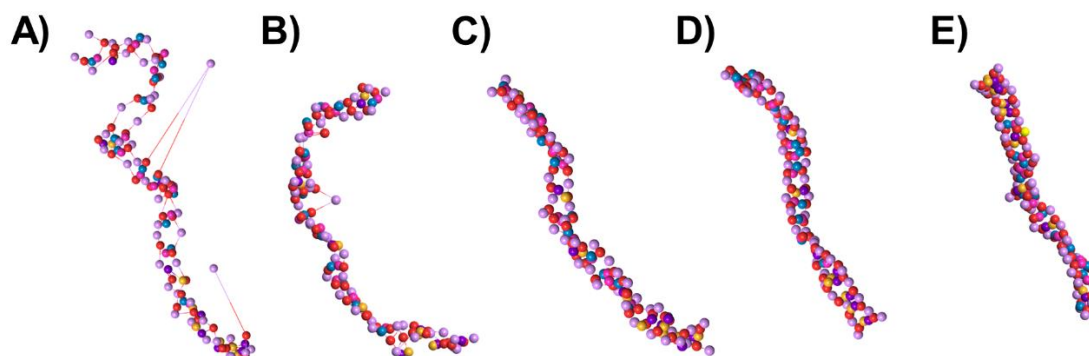
As explained in previous section, the DPD simulations for charged systems need more tunings on the beads. The tuning began with setting the appropriate charge values for the beads and constructing a simulation cell containing only siRNA and water for

testing. Figure 28 shows the structure of siRNA with different charge values and the spring constant  $C$  was set to default value ( $C = 4.0$ ) for all simulations.



**Figure 28.** Structure of the charged siRNA molecule with spring constant  $C = 4.0$  and bead charges: **A)**  $0e$ , **B)**  $0.1e$ , **C)**  $0.2e$ , **D)**  $0.3e$  and **E)**  $0.4e$ . Water beads are hidden. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

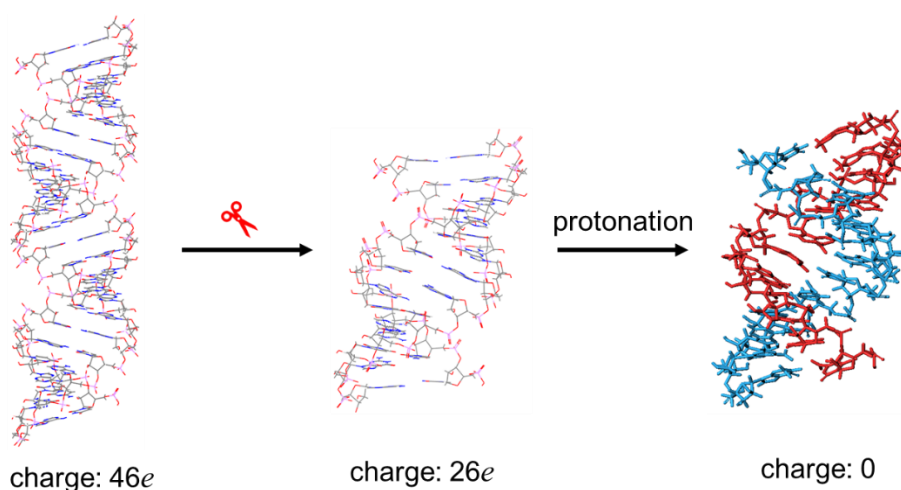
It is clearly seen that the repulsive force between the P beads increased with increasing charge values, resulting in the stretching of the siRNA molecule until a non-physical morphology was formed. A default charge value of  $0.3e$  was used for the rest of the simulation, and the spring constant  $C$  was then adjusted to overcome the strong repulsive forces. Figure 29 shows the structure of charged siRNA ( $0.3e$ ) with different spring constant  $C$  and the siRNA molecule became more physical with the increasing  $C$ , but when the spring force was too strong, siRNA molecule was going to be squeezed together.



**Figure 29.** Structure of siRNA molecule with  $0.3e$  charge and different values of the spring constant  $C$ :

A)  $C = 4$ , B)  $C = 12$ , C)  $C = 20$ , D)  $C = 28$ , and E)  $C = 36$ . Water beads are hidden. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

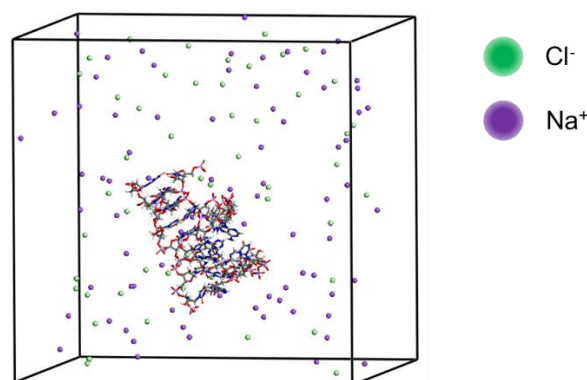
To have a look at the connection between MD and DPD simulations, we also performed two molecular dynamics simulations of protonated and charged siRNA molecules. The simulation procedure was based on the work from Ziebarth et al. but replaced with the COMPASS force field to maintain consistency throughout the project.<sup>[104]</sup> To reduce the computational cost, only half of the siRNA sequence was used to construct the atomistic model of siRNA (figure 30) and a simulation cell with size about  $75 \times 75 \times 75 \text{ \AA}^3$  were built. For protonated siRNA, the simulation cell contained only one siRNA molecule and the rest of space was filled by water.



**Figure 30.** Atomistic model of siRNA with half of the full sequence. Sense: CCCCCACAUAGAA, Antisense: GGGGGUGUAUCUU.

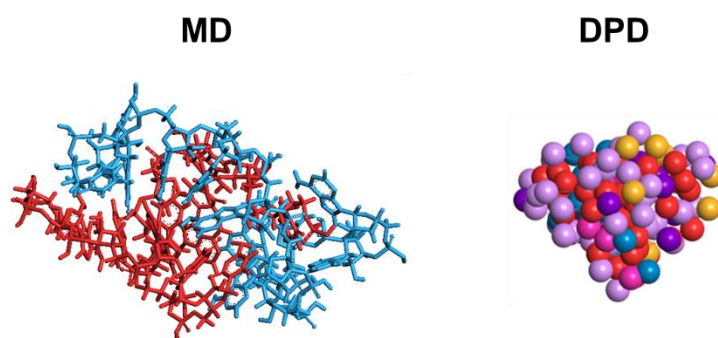
For charged siRNA the simulation cell as mor complicated. In order to maintain the electroneutrality of the simulation cell, ionized sodium chloride was additionally added according to literature operations. In the cell, in total 78 sodium cations and 52 chloride anions were added and the concentration of sodium chloride reached 194 mM (figure 31), which is similar to the literature settings.<sup>[104]</sup> Further MD simulation procedure see Appendix Ch. III.3.





**Figure 31.** Simulation cell filled by charged siRNA molecule, Na<sup>+</sup>, Cl<sup>-</sup> and water. Water molecules were hidden.

Different from single-stranded RNAs such as messenger RNA (mRNA), siRNAs have a double-stranded structure, with base pairs built up between the two strands by hydrogen bonding, which is rather closer to the DNA molecules. Figure 32 shows the equilibrated structure of protonated siRNA, and it is clear that the molecule was unable to maintain an ordered double-stranded structure. Some base pairs were gradually broken, but the two chains are not completely decoupled, which is very close to the morphology in the DPD simulations, albeit at half the length.

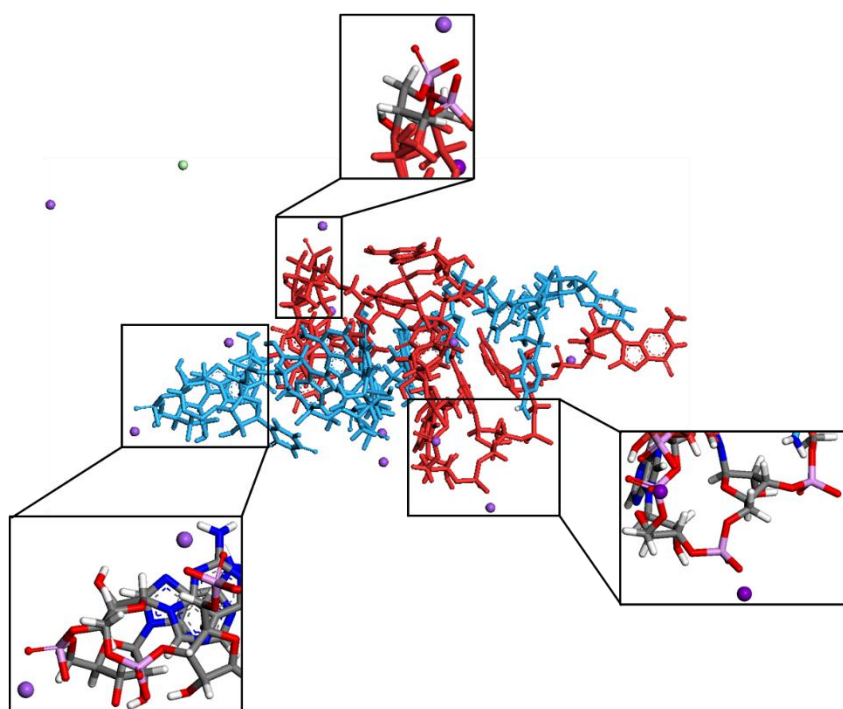


**Figure 32.** The equilibrated structure of protonated siRNA from MD and DPD simulations

The simulation for charged siRNA was not satisfying (figure 33) because the two strands almost broke up. The disintegration of base pairs could result from too weak a hydrogen bond, together with excessive electrostatic forces, might be responsible for the structural instability. Hydrogen bonding in base pairs cannot be described by ordinary hydrogen bonding, and the COMPASS force field may underestimate the

strength of these interactions. The situation is very similar to that of the DPD simulations, but since the COMPASS force field series is a commercial product, it is not possible to adjust its parameters.

In the work of Ziebarth et al, refined AMBER force field ff12SB was applied to simulate DNA and siRNA molecules,<sup>[104]</sup> which is considered to be the best force field for protein and nucleic acids at the time. The COMPASS force field series has proven to be an accurate and powerful force field system for the simulations of organic compounds, inorganic small molecules, and polymers,<sup>[105]</sup> but COMPASS may not be available for the simulations of nucleic acid molecules without additional modifications.<sup>[106]</sup> This may pose difficulties in the future in integrating the entire project and using the same force field and simulation procedures in a uniform manner.

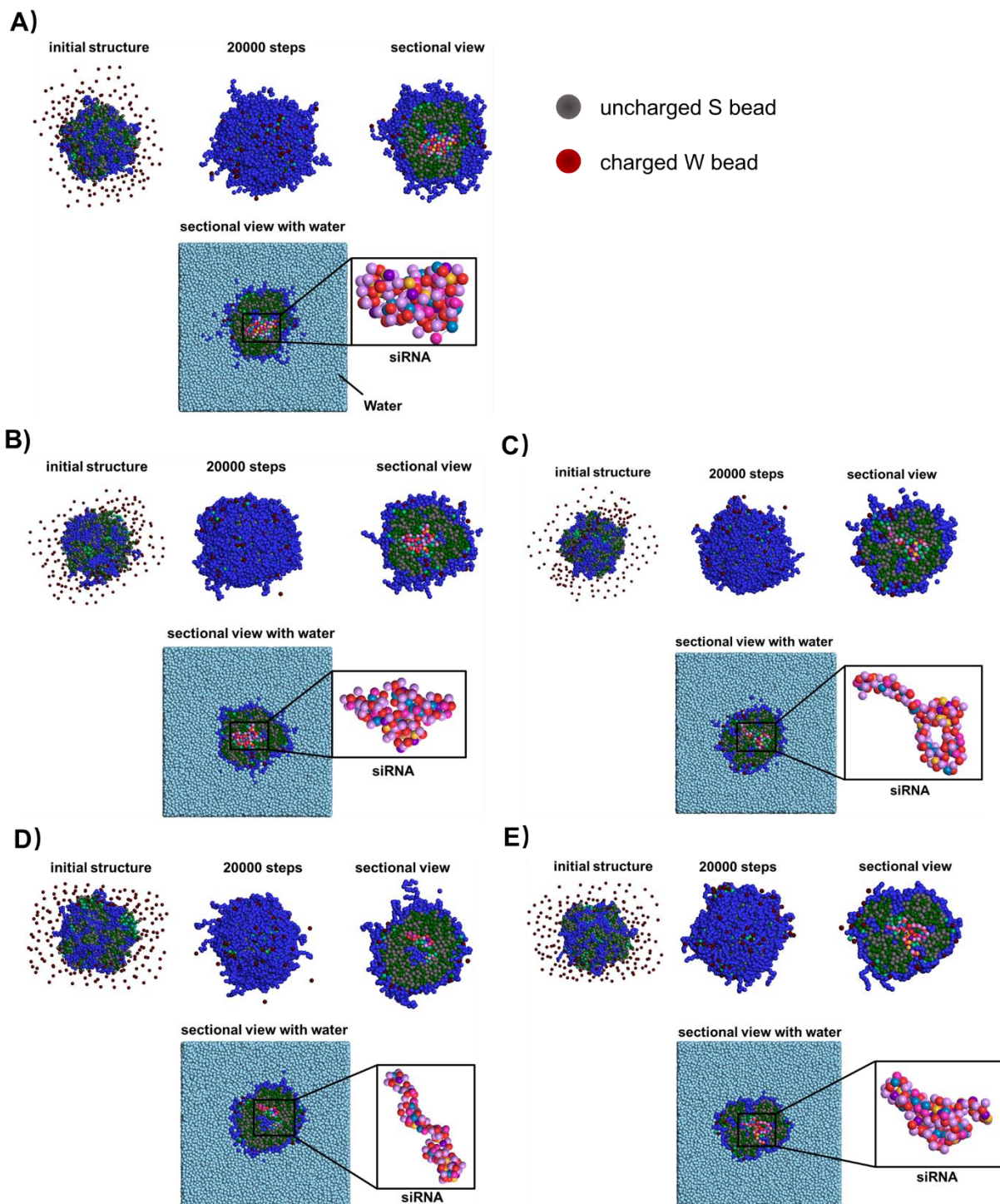


**Figure 33.** The equilibrated structure of siRNA with Na<sup>+</sup> and Cl<sup>-</sup>. Water molecules were hidden.

## 2.2 DPD simulations for charged polyplex micelle EN15

Following the tests on siRNA, the EN15 charged polyplex micelle with the largest PEG

weight ratio was selected for the next simulation. The morphologies of EN15 micelle with different  $C$  are presented in figure 34.



**Figure 34.** DPD simulations of EN15 micelle system with charge of  $0.3e$  after 20000 time steps  
A)  $C = 4$ , B)  $C = 12$ , C)  $C = 20$ , D)  $C = 28$  and E)  $C = 36$ .

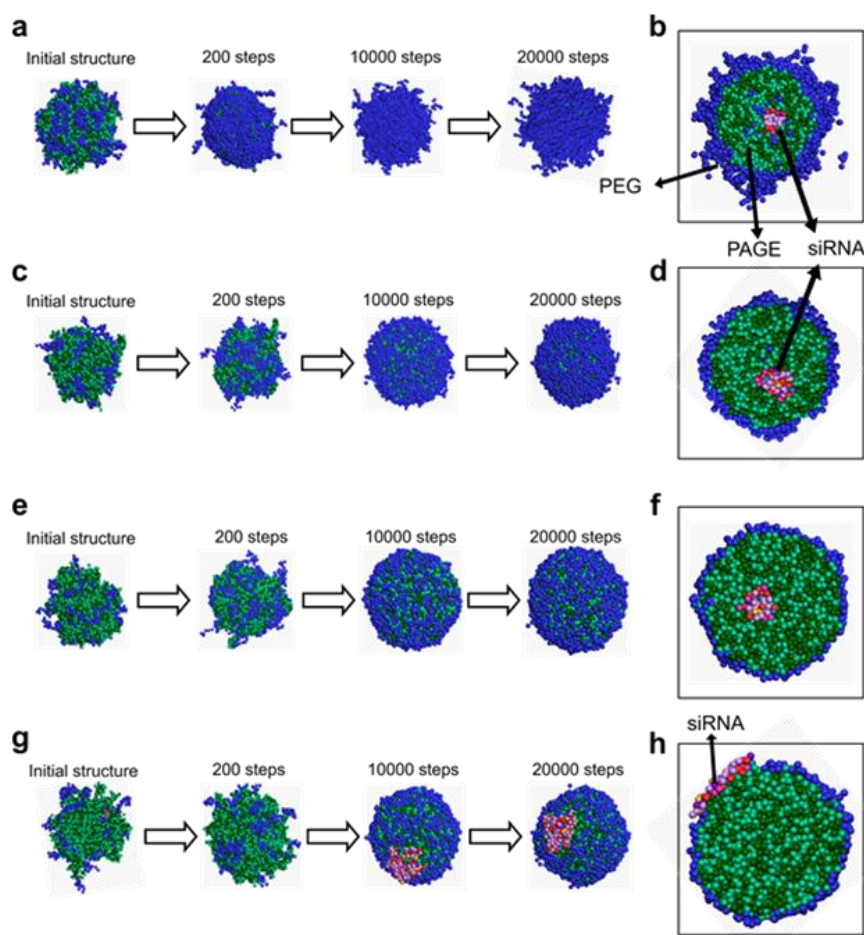
With the increase of  $C$  value, the structure of siRNA changed as in the previous

simulations, but no significant change was observed in the trend of encapsulation capacity and the morphology of micelles. Before reached  $C = 28$ , the siRNA molecule could stretch within the micelles, but as  $C$  continues to increase, the siRNA core was again compressed.

It is concluded that  $C = 28$  was best suited for modeling charged systems with  $0.3e$ , since the morphology of siRNA was the most physical and could even show a double helix structure. At the same time, no obvious changes were found in the morphology of polyplex micelle, indicating that the qualitative behavior for the micelle system was not affected in an apparent manner, which in agreement with the findings and suggestions from other published works on organic soft matters.<sup>[103]</sup> Therefore, charge neutral beads were used with default spring constant  $C = 4$  for the rest of the DPD simulations.

### **2.3 DPD simulations for neutral polyplex micelles**

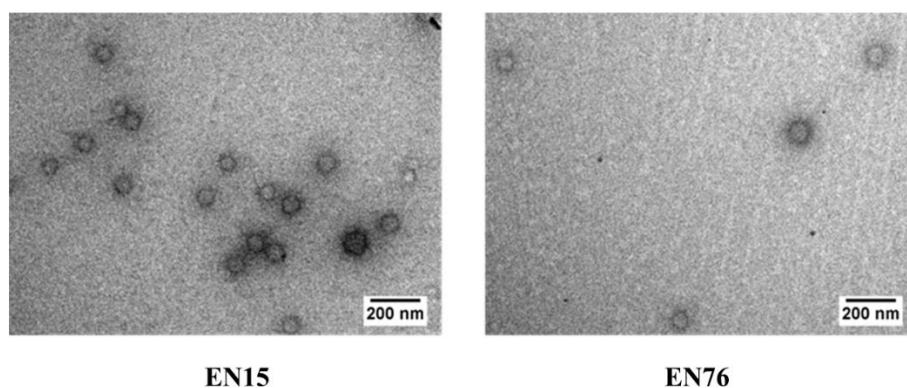
Figure 35 shows the DPD results of charge neutral polyplex micelles from EN15 to EN76, respectively and different trends were observed. The first thing to note is that for EN15 there was no significant difference between the charged and neutral models, comparing with figure 34. This indicates that the short-range electrostatic interactions that are included implicitly in the DPD force field parameters have the strongest influence on the structure of the micelles.



**Figure 35.** Snapshots of DPD simulations showing the formation and structure of polyplex micelles complexed with siRNA (water beads are hidden for clarity): **(a)** EN15 and **(b)** cross-sectional view of the final micelle; **(c)** EN29 and **(d)** cross-sectional view of the final micelle; **(e)** EN60 and **(f)** cross-sectional view of the final micelle; **(g)** EN76 and **(h)** cross-sectional view of the final micelle. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

In the initial stage of DPD simulations, all polyplex micelles with siRNA formed disordered distribution of beads. As the simulations progressed, EN15, EN29, and EN60 gradually formed a layered micelle, with the hydrophilic PEG beads building an outer layer and the PAGE beads accumulating around the siRNA core. After 10000 time steps, the micelle structure became stable and maintained its stability for the remaining 10000-time steps (20000-time steps for the entire DPD simulation). The cross-sectional views in figures 35b, 35d and 35f demonstrated that EN15, EN29, and EN60 fully encapsulated the siRNA molecule in the core. The cross-sectional view also revealed a gradual decrease in the PEG layer thickness from EN15, EN29 to EN60. EN76 (figure

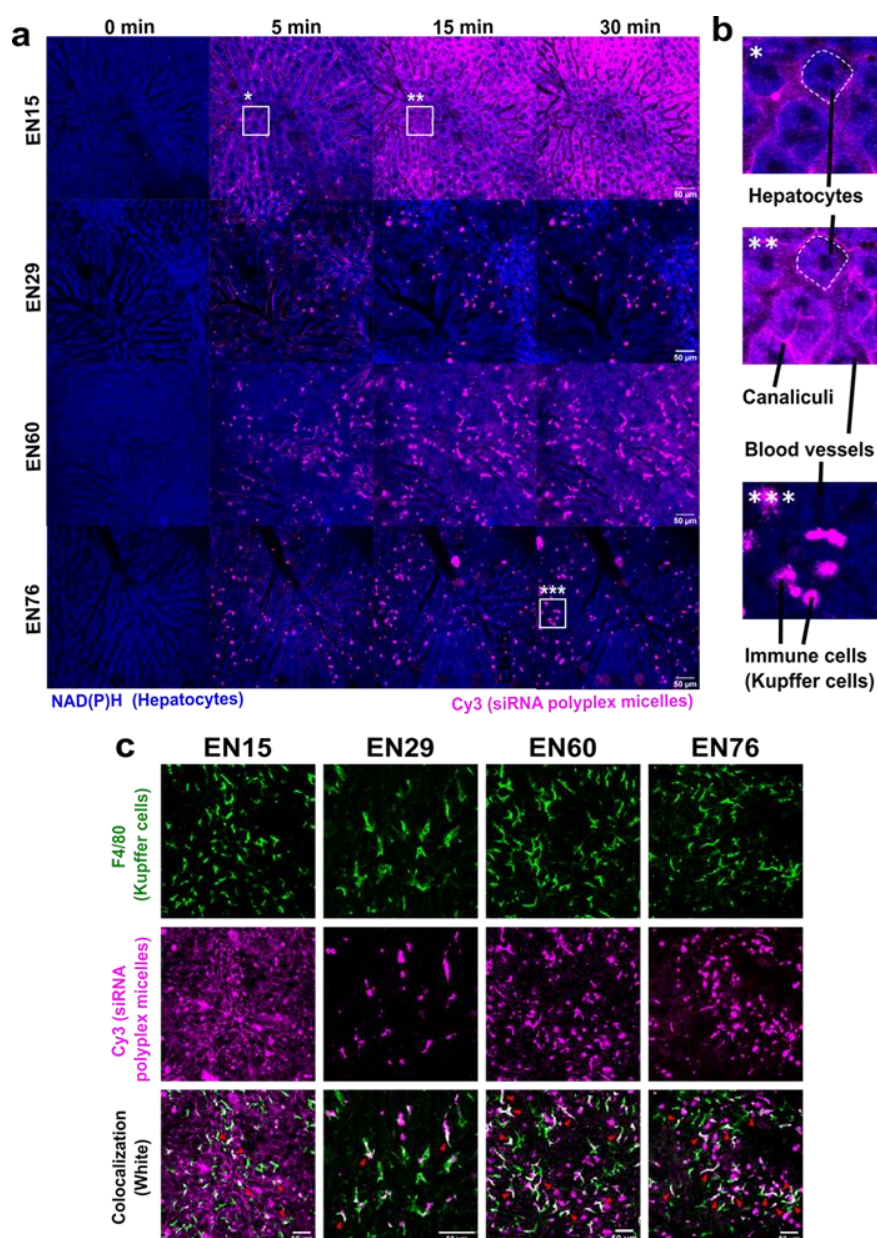
35g) showed similarly disordered structures in the initial simulation stages as other micelles. However, a disrupted PEG corona formed in the final micelle structure, with the siRNA partially exposed on the surface of the micelle (figures 35g and h). All micelles maintained spherical morphologies, which is in good agreement with experimental observations (figure 36).



**Figure 36.** TEM images suggest a spherical shape of all polyplex micelles. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

The failure of EN76 micelles to encapsulate siRNA was inconsistent with theoretical assumptions since, as the polymer chain with longest PAGE<sub>NH2</sub> block, the interaction between the PAGE<sub>NH2</sub> and siRNA should be the strongest. Moreover, further experimental measurements confirmed that EN76 could not encapsulate siRNA (figure 37), which cross-verified that the electrostatic interactions between PAGE<sub>NH2</sub> block and siRNA were not a key factor affecting the stability of polyplex micelles. The intravital microscopy images demonstrated the injection process of four micelles containing fluorescently labeled siRNA in the mouse liver. Take EN 15 as an example, 5 minutes after injection, the magenta fluorescence could be observed in liver canaliculi (tiny channels among liver cells for substance exchange), indicating the ongoing transportation of the siRNA. 30 minutes after injection, fluorescence was all over the observation area, indicating that the siRNA survived for 30 minutes in the liver and were not fully captured by immune cells. In contrast, the fluorescence of EN76 encapsulated siRNA did not change after 15 minutes and immune cells could be found in the blood vessel area. This difference in images was straightforward evidence of the

stealth ability of EN15-siRNA and confirmed the protective effect of EN15 micelle. Micelles with higher PEG weight ratio, i.e., thicker hydrophilic outer shell showed higher stability both in simulations and experimental observations. In combination with other experimental tests and measurements, it is confirmed that the weight ratio of PEG is more influential on the performance of the polyplex micelles than the specific interaction between  $PAGE_{NH_2}$  and siRNA.



**Figure 37.** Different PEG ratio of siRNA polyplex micelles significantly influences the biodistribution profiles. **(a)** Intravital microscopy of the liver. Hepatocytes are identified by their strong NAD(P)H autofluorescence (blue). Each polyplex micelles were complexed with Cy3-siRNA (magenta) and

injected through a tail vein catheter. A representative set of images from a time series (0 min before injection to 30 min after injection) is presented. **(b)** Magnification of some representative aeriols in the images, depicting hepatocytes, blood vessels (sinusoids), the post-hepatocellular canaliculi (indicating elimination of Cy3), and immune cells, in particular, Kupffer cells (local macrophages), **(c)** F4/80-FITC (green) antibody was injected after IVM evaluation to counterstain Kupffer cells. Colocalization (white) between Cy3-siRNA polyplex micelles (magenta) and Kupffer cells (green) was observed (red ▲). N/P ratio of 5 was used for micelles in the experiments. (Reprinted from [101] open access: CC BY 4.0 DEED, Copyright 2023 the Authors.)

### 3 Conclusion

In this study, DPD simulations successfully provided the qualitative analysis on the performance of series of polyplex micelles with specific therapeutic agents, which in good agreement with experimental measurements. For PEG-*b*-PAGE<sub>NH2</sub> micelles encapsulating siRNA, it is confirmed that the weight ratio of PEG is one of the key factors for the performance and stability of polyplex micelle. This may provide some guidance for improving such nanostructures in a cheaper route, as polymerization of PEG is much easier than enhancing electrostatic interactions between the polymers and the therapeutic agents.

For the DPD simulation itself, the charge-neutral model with implicitly included short-range electrostatic interactions may be sufficient for a qualitative task. Adding charge and further tuning of parameters may bring the physical morphology closer to reality, but this is not cost-effective.<sup>[16, 103]</sup> Furthermore, it is noted that the repulsive parameter  $a_{ij}$  were calculated with Flory-Huggins parameter  $\chi$  from MD simulations, indicating that using the Flory-Huggins parameter as a bridge could enable a multiscale simulation approach from the atomic scale to the mesoscale. However, the differences in simulating siRNA molecules with DPD and MD suggest that there are still many barriers to be crossed, as larger scale simulations will always mean that more simplified and empirical parameters will be involved,<sup>[102, 103, 107]</sup> and maintaining the expression of the molecular structure across the range of scales can be a great challenge.



## Conclusion and outlook

---

In this study, computer-aided methods around thermodynamic properties proved to be a powerful tool for investigating the properties of nanoparticulate polymeric drug carriers. The good agreement with experimental characterizations and observations demonstrates that the *in silico* approach is able to provide theoretical support from the atomic scale to the mesoscale.

A machine learning procedure was constructed based on molecular dynamics simulations to predict the Hildebrand solubility parameters  $\delta$  of polymers. As a first step, a database containing information on the structure of polymer repeating units and polymer chains was created, in which the repeating units of the polymers were replaced by saturated organic small molecules (repeating elements, REs) with structural similarities. Four ML algorithms were then combined with three descriptor selection methods to produce a series of ML models. After training, some models showed satisfactory performance, and the computation time was much smaller than that of MD simulations. The addition of chain-related descriptors significantly improved the ML model compared to our previous work.<sup>[53]</sup> Along this direction, a focus of future work will be to further extend the database to collect a wider range of descriptors, such as structural information on polymer chains, reliable experimental data, and various spectra from experimental measurements, etc.<sup>[87, 90]</sup> It is also important to focus on the explainability of the model to ease the involvement of scientists from different fields.<sup>[84]</sup> Moreover, this ML procedure could theoretically be migrated to the thermodynamic investigation of small organic molecules (solvents, drugs, etc.) as well, provided that high-quality database is available.

Subsequently, miscibility predictions based on the Flory-Huggins parameter  $\chi$

and the Gibbs free energy change of mixing  $\Delta G_{\text{mix}}$  were indicated to be accurate in selecting suitable polymers for specific drug molecules. Theoretical calculations were in good agreement with experimental observations, but the problem is that high-precision simulations of mixtures tended to be too resource-consuming and inefficient. Since the ML procedure for polymers has already been built, if it can also be applied to small molecules as described above, then a larger scale or multi-level ML model can be built with inputs for the target polymer as well as the drug to directly predict compatibility. If such a procedure is indeed established in the future as a means of primary screening, it could save a great deal of time and effort, even if the accuracy rate is not extremely high.

Following this, DPD simulation successfully revealed the key factors for the encapsulation of siRNA by PEG-*b*-PAGE<sub>NH2</sub> polyplex micelles. The contribution of the electrostatic interaction of the PAGE block with siRNA to the encapsulation ability was not as important as expected. Instead, the weight ratio of PEG (i.e., the thickness of the hydrophilic shell) had a greater effect on encapsulation performance, a finding that was in agreement with experimental observations. During the study, it was found that, as far as qualitative conclusions are concerned, the simulation of the charged system did not help much, even though the structure in the simulation was closer to the image of scientific intuition but did not affect the conclusions. One of the most important parameters in the DPD simulation: the repulsive parameter  $a_{ij}$  was calculated from  $\chi$ , while in this work the  $\chi$  could be calculated from  $\delta$  obtained from the MD simulations, which made the  $\delta$  an important bridge linking the three main components in this work. Still, it should be noted that it is not easy to maintain consistency between scales. The settings of force fields and other parameters would produce different effects at different scales. ML also can play a role in bridging scales and reduce the computational cost.<sup>[18, 108]</sup>

In conclusion, simulation and machine learning around Hildebrand solubility parameter can play a role in all stages of polymeric drug carrier design. From screening polymers for specific drugs to analyzing the encapsulation properties of polymer

nanostructures up to investigating the mechanism of drug release from carriers, there is a space for theoretical computation combined with machine learning. It is even possible to construct multi-scale, full-process machine learning models from molecule screening to subsequent drug release. Some countries and research institutions have already done some preparatory work in this area,<sup>[18, 108]</sup> but it is important to note that the data and manpower required to build models on such a large scale are enormous and not affordable for any individual or small team. However, it is feasible to build a full-process ML model by narrowing down the focus of the search (e.g., by first identifying the drugs and potential polymers to be studied), but it still requires a sustained long-term commitment.

# Appendix

---

## Chapter I

### 1 MD simulation procedure

The detailed procedure of MD simulation is presented as figure Ch.I.1 and the specific simulation parameters for the work of chapter I are listed as follows:

1. The initial density of the system is set to 1 g/cm<sup>3</sup>.
2. Each unit cell will generate 10 to 20 configurations. The appropriate number of configurations will be adjusted according to size of the polymer.
3. 5 to 10 cells with lowest energy will be selected.
4. During the simulated annealing, the system is equilibrated at  $T = 300$  K. Subsequently, the temperature will be step-wisely increased to 1000 K and then decreased to 300 K. At each step, the temperature is increased/decreased by 100 K, followed by a 5 ps equilibration.
5. Duration of sampling process is set as 200 ps for both *NPT* and *NVT* simulation.
6. During *NPT* simulation, cell has 200 ps to equilibrate, while during *NVT* the duration is set as 50 ps.

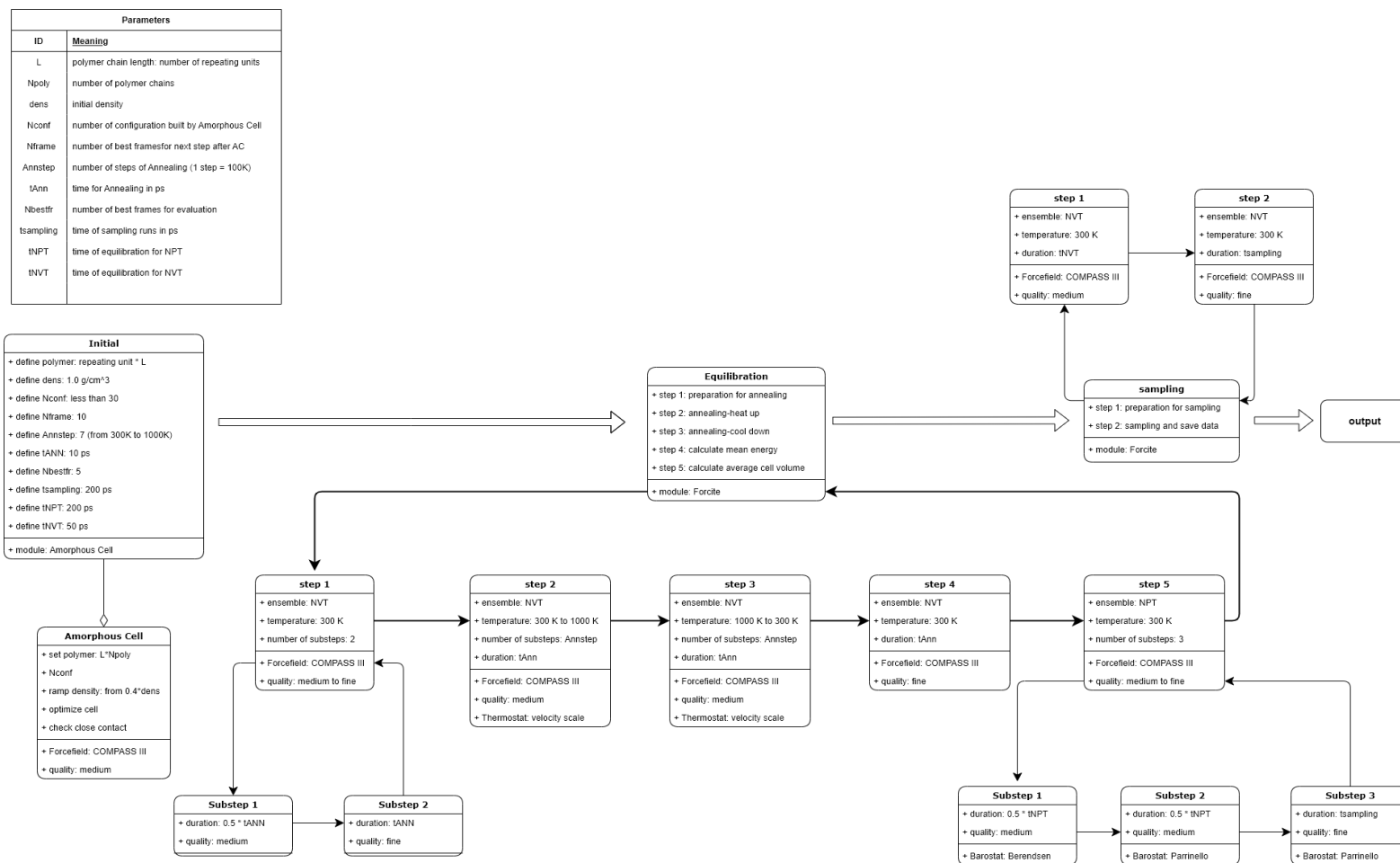


Figure Ch.I.1. Workflow of MD simulation for calculating Hildebrand solubility parameter

## 2 Polymer REs and SMILES

Due to too much data, the summary of polymers is not convenient to present. All 82 polymers and corresponding REs can be found in .xlsx document through the link:

<https://nas.cms.uni-jena.de:5501/sharing/z70kHFQw3>

Note that the names of some REs were automatically generated by the software based on the IUPAC nomenclature because as fictional molecules there are no real names for them.

## 3 Datasets and descriptors

Two datasets can be found in .csv file through following link:

Dataset 1: <https://nas.cms.uni-jena.de:5501/sharing/7JsKr97No>

Dataset 2: <https://nas.cms.uni-jena.de:5501/sharing/MPtvCIVE7>

The description of padel descriptors in .xlsx:

<https://nas.cms.uni-jena.de:5501/sharing/qtXXYYMSij>

## 4 RFE selection of descriptors

### For dataset 1

SVR: in total 30 descriptors

AATS5s	VE1_Dzm	SdsCH	mindsCH	maxHCsats
AATSC8c	VE3_Dze	minwHBd	maxHBd	hmax
AATSC3p	VE1_Dzp	minHBint2	maxwHBd	ETA_EtaP_L
MATS4i	SpMin7_Bhm	minHBint5	maxHsNH2	ETA_EtaP_F_L
GATS6v	SpMin5_Bhs	minHsNH2	maxHssNH	BIC0
GATS1e	SHsNH2	minHCsatu	maxHCHnX	JG11

RF: in total 13 descriptors

AATS1s	maxsCH3	IC0	MLFER_BO
AATSC0c	ETA_dEpsilon_D	SIC0	
SpMin3_Bhe	ETA_EtaP	CIC1	
CrippenLogP	ETA_EtaP_F_L	MDEC-12	

Lasso: in total 3 descriptors

AATS1s	SHsNH2	ETA_dEpsilon_D
--------	--------	----------------

ENR: in total 19 descriptors

number of repeating units	nHBint2	meanI	nHBDon_Lipinski
AATS1s	SHsNH2	ETA_dEpsilon_D	SIC1
AATSC4s	SsCH3	ETA_EtaP	CIC1
BCUTp-11	minHsNH2	ETA_EtaP_L	XLogP
SpMin8_Bhm	maxHsNH2	ETA_EtaP_F_L	

**For dataset 2**

SVR: in total 30 descriptors

number of repeating units	SpMin6_Bhm	AVP-5
AATSC3m	SpMin7_Bhm	ndCH2
MATS5i	SpMin7_Bhv	SdsCH
GATS1c	SpMax7_Bhp	StsC
GATS1e	SpMax2_Bhs	minwHBd
VE1_Dzp	C1SP3	minHBa
minHCHnX	maxsNH2	VR2_D
mintsC	gmax	JGI3

## Appendix

maxHBint2	ETA_dEpsilon_D	maxHCHnX
maxHBint5	ETA_EtaP_F_L	MDEO-22

RF: in total 88 descriptors

number of repeating units	AATSC4p	VR1_Dzs	ETA_dEpsilon_D
Molecular weight	AATSC1i	BCUTc-11	ETA_EtaP
Connectivity index 0X	AATSC0s	BCUTp-11	ETA_EtaP_F
Connectivity index 1X	MATS3c	SpMax8_Bhe	ETA_EtaP_F_L
Connectivity index 1Xv	MATS1m	SpMin3_Bhe	nHBDOn_Lipinski
ALogP	MATS1e	SpMax8_Bhs	IC0
AATS0v	MATS3p	SpMin3_Bhs	SIC0
AATS1i	GATS1c	SpMin6_Bhs	SIC1
AATS4i	GATS4c	SpMin7_Bhs	SIC2
AATS0s	GATS2m	Mp	CIC1
AATS1s	GATS1e	CrippenLogP	CIC3
ATSC1c	GATS2e	SHCsats	CIC4
ATSC3c	GATS1p	SsCH3	BIC1
ATSC1m	GATS3i	minHBa	BIC4
ATSC1v	GATS1s	minssCH2	MIC0
ATSC3p	GATS3s	maxsCH3	MIC5
ATSC4p	VE2_Dzm	meanI	MDEC-12
ATSC3s	VE1_Dzp	hmin	MLFER_BH
AATSC0c	VE2_Dzp	LipoaffinityIndex	MLFER_BO
AATSC3c	SM1_Dzs	ETA_Epsilon_5	MLFER_S
AATSC2v	VE2_Dzs	ETA_dEpsilon_A	RotBFrac
JGI2	VE3_D	TopoPSA	XLogP



Lasso: in total 2 descriptors

AATS1s	ETA_dEpsilon_D
--------	----------------

ENR: in total 13 descriptors

nN	maxHsNH2	SIC1
AATS1s	meanI	BIC1
AATSC4s	ETA_dEpsilon_D	MDEC-12
nBase	ETA_EtaP_L	
SHsNH2	ETA_EtaP_F_L	

## 5 Computing time of ML models

The SVR, RF and Lasso models were computed on the platform with following specs:

Processor: AMD Ryzen 9 5950X 16-Core (3.40 GHz)

RAM: 16.0 GB (DDR 4)

OS: Windows 10 professional x64

**Table Ch.I.1.** Summary of computing time for SVR, RF and Lasso

ML model	Dataset 1 (min)	Dataset 2 (min)	ML model	Dataset 1 (min)	Dataset 2 (min)	ML model	Dataset 1 (min)	Dataset 2 (min)
PCC-SVR	0.58	2.36	PCA-SVR	6	21.36	RFE-SVR	6.77	14.36
PCC-RF	60.06	172.91	PCA-RF	60.23	163.49	RFE-RF	89.2	443.14
PCC-Lasso	0.16	0.44	PCA-Lasso	0.13	0.31	RFE-Lasso	0.15	0.37

ENR models were computed on the platform with following specs:

Processor: Intel Core i5 12500H 12-Core (2.50 GHz)

RAM: 32.0 GB (DDR 4)

OS: Windows 11 professional x64

**Table Ch.I.2.** Summary of computing time for ENR

<b>ML model</b>	<b>Time for Dataset 1 (min)</b>	<b>Time for Dataset 2 (min)</b>
PCC-ENR	4.3	9.43
PCA-ENR	9.6	13.8
RFE-ENR	6.6	15.2

## Chapter II

### 1 Details of polymers

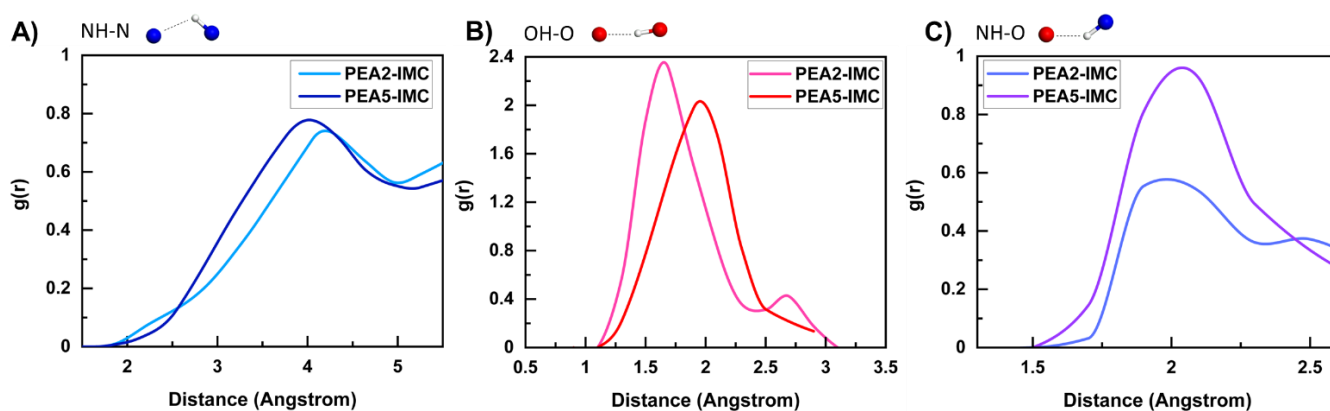
**Table Ch.II.1.** Summary of polymers for MD simulations in chapter II

Polymer	DP	Molar mass of single chain g/mol	SP $\delta$ MPa <sup>1/2</sup>
PEA2	10	3005.13	20.52
PEA5	10	3405.78	19.59
PValG	55	8752.43	18.91
PIleG	55	9523.92	19.89
PLGA	184	12029.2	21.35

**Table Ch.II.2.** Summary of drugs for MD simulations in chapter II

Drug	Molar mass g/mol	SP $\delta$ MPa <sup>1/2</sup>
Indomethacin	357.793	23.89
BRP187	456.885	23.54
BRP201	503.64	21.44

### 2 Intermolecular RDF plots of H bonds (PEA+IMC)

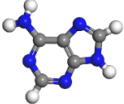
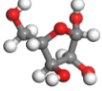
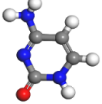
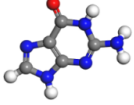

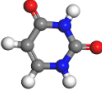
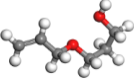


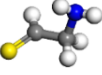


**Figure Ch.II.1.** PEA+IMC intermolecular RDF plots of A) NH-N, B) OH-O and C) NH-O

## Chapter III

### 1 Atomistic structures of all beads

**Table Ch.III.1.** Atomistic structures and Hildebrand solubility parameters  $\delta$  of DPD beads

Beads	Atomistic structure	$\delta$ MPa <sup>1/2</sup>
A		30.1
B		32.8
C		41.0
G		40.3
P		43.6
U		34.6
N		22.4
E		20.7 <sup>a</sup>
W		47.8 <sup>a</sup>
S		19.4


 ○ H   ● C   ● N   ● O   ● S   ● P

a) SP from [16]

## 2 Repulsive parameters $\alpha_{ij}$ for all beads

**Table Ch.III.2.** Interaction parameters  $\alpha_{ij}$  between different DPD beads

	A	B	C	G	P	U	N	E	S	W
A	78.00									
B	79.05 <sup>a</sup>	78.00								
C	94.67 <sup>a</sup>	87.36 <sup>a</sup>	78.00							
G	92.75 <sup>a</sup>	85.94 <sup>a</sup>	78.06 <sup>a</sup>	78.00						
P	103.57 <sup>a</sup>	94.28 <sup>a</sup>	78.95 <sup>a</sup>	79.48 <sup>a</sup>	78.00					
U	80.90 <sup>a</sup>	78.46 <sup>a</sup>	83.67 <sup>a</sup>	82.57 <sup>a</sup>	89.26 <sup>a</sup>	78.00				
N	88.48	96.15	131.59	128.10	146.81	102.40	78.00			
E	90.40	98.65	135.82	132.20	151.59	105.28	78.08	78.00		
S	94.15	103.41	143.62	139.76	160.36	110.72	78.61	78.25	78.00	
W	122.16	109.61	84.57	85.87	80.52	102.44	175.68 <sup>a</sup>	78.98	191.71	78.00

a) Parameters taken from from [16]

## 3 MD simulation procedure for siRNA

### Basic settings:

Force field: COMPASS III

Simulation cell: 75 x 75 x 75 Å<sup>3</sup>

Time step: 2 fs

Constraint algorithm: RATTLE on H covalent bonds

Algorithm for electrostatic interaction: PPPM (accuracy 0.0001)

Algorithm for Van der Waals interaction: atom based

Thermostat: NHL

Barostat: Berendsen

### Procedure:

1. NVT simulation: 2000 steps with harmonic restraints  
(restrain constant: 10 kcal/mol/Å<sup>2</sup>)
2. NVT simulation: 1000 steps without restraints
3. NPT simulation: 20 ps at 1K -> 100 K -> 200 K -> 300 K

4. NPT simulation: 20 ns with restraints (restrain constant: 10 kcal/mol/ Å<sup>2</sup>)
5. NPT simulation: 400 ps without restraints

## List of Abbreviations

---

<b>0D</b>	Zero-dimensional
<b>1D</b>	One-dimensional
<b>2D</b>	Two-dimensional
<b>AE</b>	Atomization energy
<b>AI</b>	Artificial intelligence
<b>B3-LYP</b>	The Becke 3-parameter Lee–Yang–Parr exchange–correlation functional
<b>CED</b>	Cohesive energy density
<b>def2-TZVP</b>	Triple zeta valence plus polarization basis sets
<b>DFT</b>	Density functional theory
<b>PES</b>	Potential energy surface
<b>DP</b>	Degree of polymerization
<b>DPD</b>	Dissipative particle dynamics
<b>ENR</b>	Elastic net regression
<b>FH</b>	Flory-Huggins interaction parameter
<b>FLAP</b>	5-lipoxygenase-activating protein
<b>GA</b>	Glycolic acid
<b>HF</b>	Hartree-Fock
<b>Ile</b>	L-isoleucine
<b>IMC</b>	Indomethacin
<b>LA</b>	Lactic acid
<b>Lasso</b>	Least absolute shrinkage and selection operator
<b>LOOCV</b>	leave-one-out cross validation

<b>MAE</b>	The mean absolute error
<b>MD</b>	Molecular dynamics
<b>ML</b>	Machine learning
<b>mPGES 1</b>	Microsomal prostaglandin E2 synthase-1
<b>N/P ratio</b>	The nitrogen to phosphate ratio
<b>NSAID</b>	Nonsteroidal anti-inflammatory drug
<b>PC</b>	Principal component
<b>PCA</b>	Principal component analysis
<b>PCC</b>	Pearson correlation coefficient
<b>PEA</b>	Polyester amides
<b>PEG</b>	Polyethylene glycol
<b>PEG-<i>b</i>-PAGE</b>	Poly(ethylene glycol)- <i>block</i> -poly(allyl glycidyl ether)
<b>PEI</b>	Polyethylenimine
<b>PGA</b>	Polyglycolic acid
<b>pHEMA</b>	Poly(2-hydroxyethyl methacrylate)
<b>PIIeG</b>	L-isoleucine based Polyester amide
<b>PLGA</b>	Poly(lactic-co-glycolic acid)
<b>PNIPA</b>	Poly(N-isopropylacrylamide)
<b>PValG</b>	L-valine based Polyester amide
<b>QM</b>	Quadrupole moment
<b>QSAR</b>	Quantitative structure-activity relationship
<b>QSPR</b>	Quantitative structure-property relationship
<b>R<sup>2</sup></b>	The coefficient of determination
<b>RE</b>	Repeating element
<b>RF</b>	Random Forest
<b>RFE</b>	Recursive feature elimination
<b>RFECV</b>	The cross-validation method used in RFE



## List of Abbreviations

---

<b>RMSE</b>	The root mean square error
<b>RSS</b>	Residual sum of squares
<b>siRNA</b>	Small interfering RNA
<b>SMILES</b>	Simplified Molecular Input Line Entry Specification
<b>SP</b>	Hildebrand solubility parameter
<b>SVM</b>	Support Vector Machine
<b>SVR</b>	Support vector machine
<b>TEM</b>	Transmission electron microscopy
<b>Val</b>	L-valine
<b>WHO</b>	World Health Organization

## Reference list

---

- (1) World Health Organization (WHO), *Cardiovascular diseases*. 2023.  
[https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\\_1](https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1) (accessed 2023.12.14).
- (2) The United States Food and Drug Administration (US FDA), *Finding and Learning about Side Effects (adverse reactions)*. 2023.  
<https://www.fda.gov/drugs/information-consumers-and-patients-drugs/finding-and-learning-about-side-effects-adverse-reactions> (accessed 2023.12.14).
- (3) Allen, T. M.; Cullis, P. R. Drug delivery systems: Entering the mainstream. *Science* 2004, 303 (5665), 1818-1822. DOI: 10.1126/science.1095833.
- (4) Qin, X.; Li, Y. S. Strategies To Design and Synthesize Polymer-Based Stimuli-Responsive Drug-Delivery Nanosystems. *Chembiochem* 2020, 21 (9), 1236-1253. DOI: 10.1002/cbic.201900550, Qiu, L. Y.; Bae, Y. H. Polymer architecture and drug delivery. *Pharm. Res.* 2006, 23 (1), 1-30. DOI: 10.1007/s11095-005-9046-2.
- (5) Dolgin, E. The Tangled History of Mrna Vaccines. *Nature* 2021, 597 (7876), 318-324. DOI: 10.1038/d41586-021-02483-w.
- (6) Schoenmaker, L.; Witzigmann, D.; Kulkarni, J. A.; Verbeke, R.; Kersten, G.; Jiskoot, W.; Crommelin, D. J. A. mRNA-lipid nanoparticle COVID-19 vaccines: Structure and stability. *Int. J. Pharm.* 2021, 601. DOI: 10.1016/j.ijpharm.2021.120586.
- (7) Liechty, W. B.; Kryscio, D. R.; Slaughter, B. V.; Peppas, N. A. Polymers for Drug Delivery Systems. *Annu. Rev. Chem. Biomol. Eng.* 2010, 1, 149-173. DOI: 10.1146/annurev-chembioeng-073009-100847.
- (8) Ulbrich, K.; Holá, K.; Subr, V.; Bakandritsos, A.; Tucek, J.; Zboril, R. Targeted Drug Delivery with Polymers and Magnetic Nanoparticles: Covalent and Noncovalent Approaches, Release Control, and Clinical Studies. *Chem. Rev.* 2016, 116 (9), 5338-5431. DOI: 10.1021/acs.chemrev.5b00589.
- (9) Pillai, O.; Panchagnula, R. Polymers in drug delivery. *Curr. Opin. Chem. Biol.*

- 2001, 5 (4), 447-451. DOI: 10.1016/S1367-5931(00)00227-1.
- (10) Wang, Y. C.; Zheng, Y.; Zhang, L.; Wang, Q. W.; Zhang, D. R. Stability of nanosuspensions in drug delivery. *J. Control. Release* 2013, 172 (3), 1126-1141. DOI: 10.1016/j.jconrel.2013.08.006.
- (11) Arifin, D. Y.; Lee, L. Y.; Wang, C. H. Mathematical modeling and simulation of drug release from microspheres: Implications to drug delivery systems. *Adv. Drug Deliv. Rev.* 2006, 58 (12-13), 1274-1325. DOI: 10.1016/j.addr.2006.09.007, Li, Y. Y.; Hou, T. J. Computational Simulation of Drug Delivery at Molecular Level. *Curr. Med. Chem.* 2010, 17 (36), 4482-4491. DOI: 10.2174/092986710794182935.
- (12) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* 1953, 21 (6), 1087-1092. DOI: 10.1063/1.1699114.
- (13) Adekoya, O. C.; Adekoya, G. J.; Sadiku, E. R.; Hamam, Y.; Ray, S. S. Application of DFT Calculations in Designing Polymer-Based Drug Delivery Systems: An Overview. *Pharmaceutics* 2022, 14 (9). DOI: 10.3390/pharmaceutics14091972.
- (14) Mollazadeh, S.; Sahebkar, A.; Shahlaei, M.; Moradi, S. Nano drug delivery systems: Molecular dynamic simulation. *J. Mol. Liq.* 2021, 332. DOI: 10.1016/j.molliq.2021.115823, Rezaeisadat, M.; Bordbar, A. K.; Omidyan, R. Molecular dynamics simulation study of curcumin interaction with nano-micelle of PNIPAAm-b-PEG co-polymer as a smart efficient drug delivery system. *J. Mol. Liq.* 2021, 332. DOI: 10.1016/j.molliq.2021.115862.
- (15) Ahmad, S.; Johnston, B. F.; Mackay, S. P.; Schatzlein, A. G.; Gellert, P.; Sengupta, D.; Uchegbu, I. F. In silico modelling of drug-polymer interactions for pharmaceutical formulations. *J. R. Soc. Interface.* 2010, 7, S423-S433. DOI: 10.1098/rsif.2010.0190.focus.
- (16) Xie, X. N.; Xu, S. P.; Pi, P. H.; Cheng, J.; Wen, X. F.; Liu, X.; Wang, S. N. Dissipative Particle Dynamic Simulation on the Assembly and Release of siRNA/Polymer/Gold Nanoparticles Based Polyplex. *AIChE J.* 2018, 64 (3), 810-821. DOI: 10.1002/aic.15961.
- (17) Groot, R. D.; Warren, P. B. Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* 1997, 107 (11), 4423-4435. DOI: 10.1063/1.474784.

- (18) Alber, M.; Tepole, A. B.; Cannon, W. R.; De, S.; Dura-Bernal, S.; Garikipati, K.; Karniadakis, G.; Lytton, W. W.; Perdikaris, P.; Petzold, L.; Kuhl, E. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *Npj Digit. Med.* 2019, 2. DOI: 10.1038/s41746-019-0193-y.
- (19) Bannigan, P.; Bao, Z. Q.; Hickman, R. J.; Aldeghi, M.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine learning models to accelerate the design of polymeric long-acting injectables. *Nat. Commun.* 2023, 14 (1). DOI: 10.1038/s41467-022-35343-w.
- (20) Yang, J.; Knape, M. J.; Burkert, O.; Mazzini, V.; Jung, A.; Craig, V. S. J.; Miranda-Quintana, R. A.; Bluhmki, E.; Smiatek, J. Artificial neural networks for the prediction of solvation energies based on experimental and computational data. *Phys. Chem. Chem. Phys.* 2020, 22 (42), 24359-24364. DOI: 10.1039/D0CP03701J.
- (21) Yousefinejad, S.; Hemmateenejad, B. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemom. Intell. Lab. Syst.* 2015, 149, 177-204. DOI: 10.1016/j.chemolab.2015.06.016.
- (22) Liu, P. X.; Long, W. Current Mathematical Methods Used in QSAR/QSPR Studies. *Int. J. Mol. Sci.* 2009, 10 (5), 1978-1998. DOI: 10.3390/ijms10051978.
- (23) Wei, J.; Chu, X.; Sun, X. Y.; Xu, K.; Deng, H. X.; Chen, J. G.; Wei, Z. M.; Lei, M. Machine learning in materials science. *Infomat* 2019, 1 (3), 338-358. DOI: 10.1002/inf2.12028.
- (24) Belmares, M.; Blanco, M.; Goddard, W. A.; Ross, R. B.; Caldwell, G.; Chou, S. H.; Pham, J.; Olofson, P. M.; Thomas, C. Hildebrand and Hansen solubility parameters from molecular dynamics with applications to electronic nose polymer sensors. *J. Comput. Chem.* 2004, 25 (15), 1814-1826. DOI: 10.1002/jcc.20098.
- (25) Stephenson, R. M.; Malanowski, S. Handbook of the thermodynamics of organic compounds; Springer Science & Business Media, 1987. DOI: 10.1007/978-94-009-3173-2.
- (26) Erlebach, A.; Muljajew, I.; Chi, M. Z.; Buckmann, C.; Weber, C.; Schubert, U. S.; Sierka, M. Predicting Solubility of Small Molecules in Macromolecular Compounds for Nanomedicine Application from Atomistic Simulations. *Adv. Theory Simul.* 2020, 3 (5), 2000001. DOI: 10.1002/adts.202000001.

- (27) Teraoka, I. *Polymer Solutions: An Introduction to Physical Properties*; John Wiley & Sons, Inc., 2002. DOI: 10.1002/0471224510.
- (28) Muljajew, I.; Chi, M. Z.; Vollrath, A.; Weber, C.; Beringer-Siemers, B.; Stumpf, S.; Hoepfener, S.; Sierka, M.; Schubert, U. S. A combined experimental and in silico approach to determine the compatibility of poly(ester amide)s and indomethacin in polymer nanoparticles. *Eur. Polym. J.* 2021, 156. DOI: 10.1016/j.eurpolymj.2021.110606.
- (29) Hansen, C. M. 50 Years with solubility parameters - past and future. *Prog. Org. Coat.* 2004, 51 (1), 77-84. DOI: 10.1016/j.porgcoat.2004.05.004.
- (30) Muljajew, I.; Erlebach, A.; Weber, C.; Buchheim, J. R.; Sierka, M.; Schubert, U. S. A polyesteramide library from dicarboxylic acids and 2,2'-bis(2-oxazoline): synthesis, characterization, nanoparticle formulation and molecular dynamics simulations. *Polym. Chem.* 2020, 11 (1), 112-124. DOI: 10.1039/c9py01293a.
- (31) Venkatram, S.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Critical Assessment of the Hildebrand and Hansen Solubility Parameters for Polymers. *J. Chem. Inf. Model.* 2019, 59 (10), 4188-4194. DOI: 10.1021/acs.jcim.9b00656.
- (32) Bapat, S.; Kilian, S. O.; Wiggers, H.; Segets, D. Towards a framework for evaluating and reporting Hansen solubility parameters: applications to particle dispersions. *Nanoscale Adv.* 2021, 3 (15), 4400-4410. DOI: 10.1039/d1na00405k.
- (33) Erlebach, A.; Muljajew, I.; Chi, M.; Bückmann, C.; Weber, C.; Schubert, U. S.; Sierka, M. Predicting Solubility of Small Molecules in Macromolecular Compounds for Nanomedicine Application from Atomistic Simulations. *Adv. Theory Simul.* 2020, 3 (5). DOI: 10.1002/adts.202000001.
- (34) Carvalho, S. P.; Lucas, E. F.; Gonzalez, G.; Spinelli, L. S. Determining Hildebrand Solubility Parameter by Ultraviolet Spectroscopy and Microcalorimetry. *J. Braz. Chem. Soc.* 2013, 24 (12), 1998-2007. DOI: 10.5935/0103-5053.20130250.
- (35) Erlebach, A.; Ott, T.; Otzen, C.; Schubert, S.; Czaplowska, J.; Schubert, U. S.; Sierka, M. Thermodynamic compatibility of actives encapsulated into PEG-PLA nanoparticles: In Silico predictions and experimental verification. *J. Comput. Chem.* 2016, 37 (24), 2220-2227. DOI: 10.1002/jcc.24449.
- (36) Bernholc, J. Computational materials science: The era of applied quantum mechanics. *Phys. Today* 1999, 52 (9), 30-35. DOI: 10.1063/1.882840.

- (37) Nordlund, K. Historical review of computer simulation of radiation effects in materials. *J. Nucl. Mater.* 2019, 520, 273-295. DOI: 10.1016/j.jnucmat.2019.04.028.
- (38) Ciccotti, G.; Dellago, C.; Ferrario, M.; Hernández, E. R.; Tuckerman, M. E. Molecular simulations: past, present, and future (a Topical Issue in EPJB) (vol 95, 3, 2022). *Eur. Phys. J. B* 2022, 95 (1). DOI: 10.1140/epjb/s10051-022-00278-0.
- (39) Mao, Q.; Feng, M. Y.; Jiang, X. Z.; Ren, Y. H.; Luo, K. H.; van Duin, A. C. T. Classical and reactive molecular dynamics: Principles and applications in combustion and energy systems. *Prog. Energy Combust. Sci.* 2023, 97. DOI: 10.1016/j.pecs.2023.101084.
- (40) Truszkowski, A.; van den Broek, K.; Kuhn, H.; Zielesny, A.; Epple, M. Mesoscopic Simulation of Phospholipid Membranes, Peptides, and Proteins with Molecular Fragment Dynamics. *J. Chem. Inf. Model.* 2015, 55 (5), 983-997. DOI: 10.1021/ci5006096.
- (41) Nie, S. Y.; Zhang, X. F.; Gref, R.; Couvreur, P.; Qian, Y.; Zhang, L. J. Multilamellar Nanoparticles Self-Assembled from Opposite Charged Blends: Insights from Mesoscopic Simulation. *J. Phys. Chem. C* 2015, 119 (35), 20649-20661. DOI: 10.1021/acs.jpcc.5b03833.
- (42) Xie, X.; Xu, S.; Pi, P.; Cheng, J.; Wen, X.; Liu, X.; Wang, S. Dissipative particle dynamic simulation on the assembly and release of siRNA/polymer/gold nanoparticles based polyplex. *AIChE J.* 2017, 64 (3), 810-821. DOI: 10.1002/aic.15961.
- (43) Glotzer, S. C.; Paul, W. Molecular and mesoscale simulation methods for polymer materials. *Annu. Rev. Mater. Res.* 2002, 32, 401-436. DOI: 10.1146/annurev.matsci.32.010802.112213.
- (44) Praprotnik, M.; Delle Site, L.; Kremer, K. Multiscale simulation of soft matter: From scale bridging to adaptive resolution. *Annu. Rev. Phys. Chem.* 2008, 59, 545-571. DOI: 10.1146/annurev.physchem.59.032607.093707, Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.* 2016, 37, 29-45. DOI: 10.1016/j.sbi.2015.11.011.
- (45) Chernatynskiy, A.; Phillpot, S. R.; LeSar, R. Uncertainty Quantification in Multiscale Simulation of Materials: A Prospective. *Annu. Rev. Mater. Res.* 2013, 43, 157-182. DOI: 10.1146/annurev-matsci-071312-121708.

- (46) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press, 2002. DOI: 10.1016/B978-0-12-267351-1.X5000-7.
- (47) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825-2830.
- (48) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. . In *Advances in Neural Information Processing Systems 2019*.
- (49) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* 2018, 559 (7715), 547-555. DOI: 10.1038/s41586-018-0337-2.
- (50) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* 1996, 96 (3), 1027-1043. DOI: 10.1021/cr950202r.
- (51) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH, 2009. DOI: 10.1002/9783527628766.
- (52) Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Perez-Sanchez, H.; Benediktsson, J. A.; Thapa, A.; Barr, A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Syst. Appl.* 2017, 72, 151-159. DOI: 10.1016/j.eswa.2016.12.008.
- (53) Chi, M. Z.; Gargouri, R.; Schrader, T.; Damak, K.; Maalej, R.; Sierka, M. Atomistic Descriptors for Machine Learning Models of Solubility Parameters for Small Molecules and Polymers. *Polymers* 2022, 14 (1). DOI: 10.3390/polym14010026.
- (54) BIOVIA, Dassault Systèmes, Materials Studio, 23.1.0.308, San Diego: Dassault Systèmes, [2020-2023]. (accessed).
- (55) Furche, F.; R., A.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. TURBOMOLE. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2013, 4 (2), 91-100. DOI: 10.1002/wcms.1162.
- (56) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* 2011, 32 (7), 1466-1474. DOI: 10.1002/jcc.21707.

- (57) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* 1995, 20 (3), 273-297. DOI: 10.1023/A:1022627411411.
- (58) Ho, T. K. Random decision forests. In 3rd international conference on document analysis and recognition. , Montreal QC Canada, 1995; IEEE: pp 278-282. DOI: 10.1109/ICDAR.1995.598994.
- (59) Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., B: Stat. Methodol.* 1996, 58 (1), 267-288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
- (60) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc., B: Stat. Methodol.* 2005, 67, 301-320. DOI: 10.1111/j.1467-9868.2005.00503.x.
- (61) Akkermans, R. L. C.; Spenley, N. A.; Robertson, S. H. Monte Carlo methods in Materials Studio. *Mol. Simul.* 2013, 39 (14-15), 1153-1164. DOI: 10.1080/08927022.2013.843775.
- (62) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys* 1984, 81 (8), 3684-3690. DOI: 10.1063/1.448118.
- (63) Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* 1980, 45 (14), 1196-1199. DOI: 10.1103/PhysRevLett.45.1196.
- (64) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* 2019, 5. DOI: 10.1038/s41524-019-0221-0.
- (65) Van Nhu, N.; Singh, M.; Leonhard, K. Quantum mechanically based estimation of perturbed-chain polar statistical associating fluid theory parameters for analyzing their physical significance and predicting properties. *J. Phys. Chem. B* 2008, 112 (18), 5693-5701. DOI: 10.1021/jp7105742.
- (66) Zheng, J. J.; Xu, X. F.; Truhlar, D. G. Minimally augmented Karlsruhe basis sets. *Theor. Chem. Acc.* 2011, 128 (3), 295-305. DOI: 10.1007/s00214-010-0846-z.
- (67) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* 2011, 32 (7), 1456-1465. DOI: 10.1002/jcc.21759 From NLM Medline.
- (68) Buckingham, A. D.; Disch, R. L.; Dunmur, D. A. Quadrupole Moments of Some



- Simple Molecules. *J. Am. Chem. Soc.* 1968, 90 (12), 3104-+. DOI: DOI 10.1021/ja01014a023.
- (69) Applequist, J. Traceless Cartesian Tensor Forms for Spherical Harmonic-Functions - New Theorems and Applications to Electrostatics of Dielectric Media. *J. Phys. A Math. Gen.* 1989, 22 (20), 4303-4330. DOI: 10.1088/0305-4470/22/20/011.
- (70) Lin, C.; Liu, L. W.; Liu, Y. J.; Leng, J. S. The compatibility of polylactic acid and polybutylene succinate blends by molecular and mesoscopic dynamics. *Int. J. Smart Nano Mater.* 2020, 11 (1), 24-37. DOI: 10.1080/19475411.2020.1729274.
- (71) Bicerano, J. Prediction of polymer properties; CRC Press, 2002. DOI: 10.1201/9780203910115.
- (72) Dworakowska, S.; Bogdał, D. Application of Connectivity Indices in Polymer Chemistry. In *The 16th International Electronic Conference on Synthetic Organic Chemistry 2012*.
- (73) Sedgwick, P. Pearson's correlation coefficient. *Br. Med. J.* 2012, 344. DOI: 10.1136/bmj.e4483.
- (74) Feng, W. L.; Zhu, Q. Y.; Zhuang, J.; Yu, S. M. An expert recommendation algorithm based on Pearson correlation coefficient and FP-growth. *Cluster Comput.* 2019, 22, S7401-S7412. DOI: 10.1007/s10586-017-1576-y.
- (75) Greenacre, M.; Groenen, P. J. F.; Hastie, T.; D'Enza, A. L.; Markos, A.; Tuzhilina, E. Principal component analysis. *Nat. Rev. Methods Primers* 2022, 2 (1). DOI: 10.1038/s43586-022-00184-w.
- (76) Maćkiewicz, A.; Ratajczak, W. Principal Components-Analysis (PCA). *Comput. Geosci.* 1993, 19 (3), 303-342. DOI: 10.1016/0098-3004(93)90090-R.
- (77) Li, F.; Yang, Y. Analysis of recursive feature elimination methods. In *The 28th ACM/SIGIR International Symposium on Information Retrieval Salvador Brazil, 2005*; Association for Computing Machinery: pp 633-634. DOI: 10.1145/1076034.1076164.
- (78) Chen, X.-w.; Jeong, J. C. Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications, Cincinnati OH USA, 2007*; IEEE: pp 429-435. DOI: 10.1109/ICMLA.2007.35.
- (79) Browne, M. W. Cross-validation methods. *J. Math. Psychol.* 2000, 44 (1), 108-132. DOI: 10.1006/jmps.1999.1279.
- (80) Akhtar, F.; Li, J.; Pei, Y.; Xu, Y.; Rajput, A.; Wang, Q. Optimal Features Subset

- Selection for Large for Gestational Age Classification Using GridSearch Based Recursive Feature Elimination with Cross-Validation Scheme. Singapore, 2020; Springer Singapore: pp 63-71.
- (81) Bengfort, B.; Bilbro, R. Yellowbrick: Visualizing the scikit-learn model selection process. *J. Open Source Softw.* 2019, 4 (35), 1075. DOI: 10.21105/joss.01075.
- (82) Mukaka, M. M. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med. J.* 2012, 24 (3), 69-71, Koo, T. K.; Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* 2016, 15 (2), 155-163. DOI: 10.1016/j.jcm.2016.02.012 From NLM.
- (83) Dietterich, T. Overfitting and undercomputing in machine learning. *ACM Comput. Surv.* 1995, 27 (3), 326-327. DOI: 10.1145/212094.212114.
- (84) Zhong, X. T.; Gallagher, B.; Liu, S. S.; Kailkhura, B.; Hiszpanski, A.; Han, T. Y. J. Explainable machine learning in materials science. *Npj Comput. Mater.* 2022, 8 (1). DOI: 10.1038/s41524-022-00884-7.
- (85) Pearson, R. G. The Electronic Chemical-Potential and Chemical Hardness. *J. Mol. Struct. Theochem* 1992, 87, 261-270.
- (86) Lv, X.; Wang, S. Y.; Shan, P.; Zhao, Y. L.; Zuo, L. A machine learning based method for automatic differential scanning calorimetry signal analysis. *Measurement* 2022, 187. DOI: 10.1016/j.measurement.2021.110218.
- (87) Lansford, J. L.; Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat. Commun.* 2020, 11 (1). DOI: 10.1038/s41467-020-15340-7.
- (88) Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* 2020, 58 (6), 512-519. DOI: 10.1002/mrc.4989.
- (89) Stein, H. S.; Guevarra, D.; Newhouse, P. F.; Soedarmadji, E.; Gregoire, J. M. Machine learning of optical properties of materials - predicting spectra from images and images from spectra. *Chem. Sci.* 2019, 10 (1), 47-55. DOI: 10.1039/c8sc03077d.
- (90) Chen, P. Y.; Shibata, K.; Hagita, K.; Miyata, T.; Mizoguchi, T. Prediction of the Ground-State Electronic Structure from Core-Loss Spectra of Organic Molecules by Machine Learning. *J. Phys. Chem. Lett.* 2023, 14 (20), 4858-4865. DOI:

- 10.1021/acs.jpcelett.3c00142, Jia, H. L.; Wang, C. H.; Wang, C.; Clancy, P. Machine Learning Approach to Enable Spectral Imaging Analysis for Particularly Complex Nanomaterial Systems. *Acs Nano* 2023, 17 (1), 453-460. DOI: 10.1021/acs.nano.2c08884.
- (91) Behnke, M.; Vollrath, A.; Dahlke, P.; Larios, F. P.; Chi, M.; Tsarenko, E.; Jordan, P. M.; Weber, C.; Dirauf, M.; Czaplewska, J. A.; et al. PEtOxylated polyesteramide nanoparticles for the delivery of anti-inflammatory drugs. *Mater. Today Chem.* 2024, 35. DOI: 10.1016/j.mtchem.2023.101848.
- (92) Szlek, J.; Paclawski, A.; Lau, R.; Jachowicz, R.; Kazemi, P.; Mendyk, A. Empirical search for factors affecting mean particle size of PLGA microspheres containing macromolecular drugs. *Comput. Methods Programs Biomed.* 2016, 134, 137-147. DOI: 10.1016/j.cmpb.2016.07.006.
- (93) Wang, Y.; Qin, B.; Xia, G.; Choi, S. H. FDA's Poly (Lactic-Co-Glycolic Acid) Research Program and Regulatory Outcomes. *AAPS J.* 2021, 23 (4). DOI: 10.1208/s12248-021-00611-y.
- (94) Mir, M.; Ahmed, N.; Rehman, A. U. Recent applications of PLGA based nanostructures in drug delivery. *Colloids Surf. B* 2017, 159, 217-231. DOI: 10.1016/j.colsurfb.2017.07.038.
- (95) Winnacker, M.; Rieger, B. Poly(ester amide)s: recent insights into synthesis, stability and biomedical applications. *Polym. Chem.* 2016, 7 (46), 7039-7046. DOI: 10.1039/c6py01783e.
- (96) Brunacci, N.; Neffe, A. T.; Wischke, C.; Naolou, T.; Nöchel, U.; Lendlein, A. Oligodepsipeptide (nano)carriers: Computational design and analysis of enhanced drug loading. *J. Control. Release* 2019, 301, 146-156. DOI: 10.1016/j.jconrel.2019.03.004, Burton, T. F.; Pinaud, J.; Giani, O. Rapid and Controlled Organocatalyzed Ring-Opening Polymerization of 3S-(Isobutyl)morpholine-2,5-dione and Copolymerization with Lactide. *Macromolecules* 2020, 53 (15), 6598-6607. DOI: 10.1021/acs.macromol.0c00940.
- (97) Kuehster, L.; Jhon, Y. K.; Wang, Y.; Smith, W. C.; Xu, X. M.; Qin, B.; Zhang, F.; Lynd, N. A. Stochastic and Deterministic Analysis of Reactivity Ratios in the Partially Reversible Copolymerization of Lactide and Glycolide. *Macromolecules* 2022, 55 (16), 7171-7180. DOI: 10.1021/acs.macromol.2c00757.

- (98) Banoglu, E.; Çelikoglu, E.; Völker, S.; Olgaç, A.; Gerstmeier, J.; Garscha, U.; Çaliskan, B.; Schubert, U. S.; Carotti, A.; Macchiarulo, A.; Werz, O. 4,5-Diarylisoxazol-3-carboxylic acids: A new class of leukotriene biosynthesis inhibitors potentially targeting 5-lipoxygenase-activating protein (FLAP). *Eur. J. Med. Chem.* 2016, 113, 1-10. DOI: 10.1016/j.ejmech.2016.02.027.
- (99) Steiner, T. The hydrogen bond in the solid state. *Angew. Chem. Int. Ed.* 2002, 41 (1), 49-76. DOI: 10.1002/1521-3773(20020104)41:1<48::aid-anie48>3.0.co;2-u.
- (100) Khan, S. B.; Lee, S. L. Supramolecular Chemistry: Host-Guest Molecular Complexes. *Molecules* 2021, 26 (13). DOI: ARTN 3995 10.3390/molecules26133995.
- (101) Foo, W.; Cseresnyés, Z.; Rössel, C.; Teng, Y. F.; Ramoji, A.; Chi, M. Z.; Hauswald, W.; Huschke, S.; Hoepfner, S.; Popp, J.; et al. Tuning the corona-core ratio of polyplex micelles for selective oligonucleotide delivery to hepatocytes or hepatic immune cells. *Biomaterials* 2023, 294. DOI: 10.1016/j.biomaterials.2023.122016.
- (102) Gavrilov, A. A.; Chertovich, A. V.; Kramarenko, E. Y. Dissipative particle dynamics for systems with high density of charges: Implementation of electrostatic interactions. *J. Chem. Phys.* 2016, 145 (17). DOI: 10.1063/1.4966149, Terron-Mejia, K. A.; Lopez-Rendon, R.; Goicochea, A. G. Electrostatics in dissipative particle dynamics using Ewald sums with point charges. *J. Phys. Condens. Matter.* 2016, 28 (42). DOI: 10.1088/0953-8984/28/42/425101.
- (103) Bänsch, F.; Steinbeck, C.; Zielesny, A. Notes on the Treatment of Charged Particles for Studying Cyclotide/Membrane Interactions with Dissipative Particle Dynamics. *Membranes* 2022, 12 (6). DOI: 10.3390/membranes12060619.
- (104) Ziebarth, J. D.; Kennetz, D. R.; Walker, N. J.; Wang, Y. M. Structural Comparisons of PEI/DNA and PEI/siRNA Complexes Revealed with Molecular Dynamics Simulations. *J. Phys. Chem. B* 2017, 121 (8), 1941-1952. DOI: 10.1021/acs.jpcc.6b10775.
- (105) Sun, H.; Jin, Z.; Yang, C.; Akkermans, R. L.; Robertson, S. H.; Spenley, N. A.; Miller, S.; Todd, S. M. COMPASS II: extended coverage for polymer and drug-like molecule databases. *J. Mol. Model.* 2016, 22 (2), 47. DOI: 10.1007/s00894-016-2909-0 From NLM Medline, Akkermans, R. L. C.; Spenley, N. A.; Robertson, S. H. COMPASS III: automated fitting workflows and extension to

ionic liquids. *Mol. Simul.* 2021, 47 (7), 540-551. DOI: 10.1080/08927022.2020.1808215.

- (106) Martin, M. G. Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor-liquid coexistence curves and liquid densities. *Fluid Ph. Equilibria.* 2006, 248 (1), 50-55. DOI: 10.1016/j.fluid.2006.07.014.
- (107) Li, Y. M.; Guo, Y. Y.; Bao, M. T.; Gao, X. L. Investigation of interfacial and structural properties of CTAB at the oil/water interface using dissipative particle dynamics simulations. *J. Colloid Interface Sci.* 2011, 361 (2), 573-580. DOI: 10.1016/j.jcis.2011.05.078.
- (108) Ingólfsson, H. I.; Bhatia, H.; Aydin, F.; Ooppelstrup, T.; López, C. A.; Stanton, L. G.; Carpenter, T. S.; Wong, S. R.; Di Natale, F.; Zhang, X. H.; et al. Machine Learning-Driven Multiscale Modeling: Bridging the Scales with a Next-Generation Simulation Infrastructure. *J. Chem. Theory Comput.* 2023, 19 (9), 2658-2675. DOI: 10.1021/acs.jctc.2c01018.

## List of Publications

---

Below is the list of publications that I have either already published or am currently in the process of publishing during my PhD. All of these works are directly related to the research presented in this thesis.

6. *Machine learning approach for fast prediction of polymer properties (working title)*  
M. Chi, C. Zhu, M. Sierka  
*Manuscript in preparation*
  
5. *PEtOxylated polyesteramide nanoparticles for the delivery of anti-inflammatory drugs*  
M. Behnke, A. Vollrath, P. Dahlke, F. P. Larios, M. Chi, E. Tsarenko, P. M. Jordan, C. Weber, M. Dirauf, J. A. Czaplewska, B. Beringer-Siemers, S. Stumpf, C. Kellner, C. Kretzer, S. Hoepfner, I. Nischang, M. Sierka, C. Eggeling, O. Werz, U. S. Schubert  
*Mater. Today Chem.* **2024**, 35
  
4. *Tuning the Corona-core Ratio of Polyplex Micelles for Selective Oligonucleotide Delivery to Hepatocytes or Hepatic Immune Cells*  
W. Foo, Z. Cseresnyés, C. Rössel, Y. Teng, A. Ramoji, M. Chi, W. Hauswald, S. Huschke, S. Hoepfner, J. Popp, F. H. Schacher, M. Sierka, M. T. Figge, A. T. Press, M. Bauer  
*Biomaterials* **2023**, 294, 122016
  
3. *Atomistic Descriptors for Machine Learning Models of Solubility Parameters for Small Molecules and Polymers*  
M. Chi, R. Gargouri, T. Schrader, K. Damak, R. Mañej, M. Sierka  
*Polymers* **2022**, 14, 26
  
2. *A combined experimental and in silico approach to determine the compatibility of poly(ester amide)s and indomethacin in polymer nanoparticles*  
I. Muljajew, M. Chi, A. Vollrath, C. Weber, B. Beringer-Siemers, S. Stumpf, S. Hoepfner, M. Sierka, U. S. Schubert  
*Eur. Polym. J.* **2021**, 156, 110606

1. *Predicting Solubility of Small Molecules in Macromolecular Compounds for Nanomedicine Application from Atomistic Simulations*  
A. Erlebach, I. Muljajew, M. Chi, C. Bückmann, C. Weber, U. S. Schubert, M. Sierka  
*Adv. Theory Simul.* **2020**, 3, 2000001

## **Declaration of Authorship**

I solemnly declare that I have independently prepared and completed the presented work without undue assistance from third parties or the use of references not cited in literature. Data and ideas which were included directly or indirectly from other sources are cited by referencing the original source. All contents prepared with the assistance of or in collaboration with other people, respectively, are clearly stated in the respective text passages and summarized in the acknowledgments.

No further individuals were involved in the preparation of the presented thesis with respect to content and materials. In particular, I did not receive paid assistance or consulting services from PhD consultant or others.

Nobody has received direct or indirect monetary benefits for work associated with the content of this submitted thesis. This work has to date not been submitted domestically or abroad in the current or in a similar version to any other examination board. I am aware of the applicable regulations for Doctoral Studies of the Faculty of Physics and Astronomy. I declare on my honor that I have told the truth to the best of my knowledge and have not concealed anything.

Place, Date

Signature



## **Ehrenwörtliche Erklärung**

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Alle Inhalte, die unter Mitwirkung bzw. in Zusammenarbeit mit anderen Personen erstellt wurden, sind in den jeweiligen Textpassagen deutlich gekennzeichnet und in den Danksagungen zusammengefasst.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlung bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen.

Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Die geltende Promotionsordnung der Physikalisch- Astronomischen Fakultät ist mir bekannt. Ich versichere ehrenwörtlich, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Ort, Datum

Unterschrift