Deducing molecular structure of plastics from flow data using Deep Learning

Kevin Timothy Muller
201779539

Supervised by Prof. Daniel J Read

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

September 2024

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

UNIVERSITY OF LEEDS

School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name : Kevin Timothy Muller

Student ID : 201779539

Abstract

With plastic usage rates increasing by the day, it is of utmost importance that we, as humans, not only take responsibility and reduce our use of plastics, but also innovate our technology to aid us in this battle of keeping mother nature safe. One such way is to increase the efficiency at which plastics are recycled and reused. After a plastic mass is recycled, it is difficult to instinctively gauge the material properties of the plastic, without which the plastic cannot be reused immediately or effectively.

This dissertation project aims to solve this issue with a Deep Learning approach of analyzing the rheological flow curves of the plastic mass and predict it's Molecular Weight Distribution (MWD), with which one would be able to determine it's material properties. The dataset used for analysis is from a tube model developed by Das & Read (2023) which mimics the trends of plastics and generates the flow curves for various values of MWD for us to analyze.

The types of plastics that are analyzed are only Linear Polymers following a simple unimodal log-normal or bimodal log-normal distribution. This dissertation gets satisfactory (highly accurate) results on the generated datasets with the help of LSTMs and Fully connected layers.

The unimodal dataset is modeled and predicted with high accuracies. Then, we reduce the frequencies of flow curve information and also induce artificial noise into the dataset and still receive accuracies above the industry standard. The deep learning model is then tested for robustness, which it does prove to be when cleaned on noisy data. Lastly, we conduct spectral analysis and analyze which frequencies of the flow curves predict the MWD characteristics better.

The Bimodal dataset, however, fails to be predicted with higher accuracies than the industry standard in some scenarios, and the restriction in frequency ranges and addition of artificial errors do not provide the required accuracies in most scenarios. However, with the analysis of the error trends of each target variable, the report ends with a few suggestions to rectify the low accuracies given by this deep learning approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Plastics are a resource used in abundance all over the world. In the year 2023, an impressive 400 million metric tons of this material was produced for various purposes such as packaging, construction, and automotive design among many others in the world. There are two types of plastics produced, namely Single-use plastics and Multi-use plastics. Single-use plastics are plastics intended for short-term use which typically comprise of packaging items such as bags, bottles, wrappers, straws, and disposable utensils. Multi-use plastics, however, are plastics with longer lifespans such as construction materials, automotive parts, and household goods. 40% (D. et al. 2023) of the world's total plastic production, are made up of single-use plastics. Many issues would arise if single-use plastics are not managed or recycled effectively, such as :

- **Environmental pollution :** Single-use plastics are typically made from petrochemical-based polymers that are not biodegradable.
- **Wildlife Harm :** These plastics often end up in oceans and other natural environments, where they can harm wildlife (Garcês & Pires 2024). Marine animals, birds, and terrestrial species can ingest or become entangled in plastic waste, leading to injury or death.
- **Microplastic Formation :** Over time, single-use plastics break down into smaller particles known as microplastics, which can be ingested by a wide range of organisms, entering the food chain and potentially affecting human health.

There is hence an imperative need to recycle and reuse plastics at an efficient pace in order to prevent the overfilling of landfills, causing ecological damage.

Plastics are typically collected and sorted manually (Ruj et al. 2015) by type (e.g., PET, HDPE, PVC), color, and sometimes by resin identification codes. This is done manually. Then, they are cleaned and shredded into smaller pieces; after which, they are sorted to ensure purity by separating different polymers that may have been missed in earlier steps. This step may involve processes like density separation (using water) or air classification.

However, in order to gain a more precise understanding of the composition of the plastic material, it is important to understand it's Molecular Weight Distribution (MWD) (Tan et al. 2019), which, in essence, describes the range and distribution of molecular weights of the chains

within a polymer sample. Understanding a polymer materials' MWD helps in the understanding of it's material properties. Based on the MWD of the polymer plastic, it is classified into a polymer type.

There are several methods employed in the industry to find the MWD of a polymer plastic. Some of them are discussed below :

Gel Permeation Chromatography (GPC) (vec 2024) or Size Exclusion Chromatography (SEC) involves dissolving the polymer plastic sample in a solvent and pouring the same into a column filled with porous beads. The smaller molecules will penetrate deeper into the pores and take longer to remove, while larger molecules pass through more quickly. The result is a chromatogram that reflects the MWD of the polymer plastic. However, this method is time-consuming and costly.

Another method that is used to achieve the same goal is Differential Scanning Calorimetry (DSC) (Singh & Singh 2022), which involves measuring the amount of heat absorbed or released by a polymer plastic sample as it is heated or cooled on top of a DSC pan. This data can reveal important information about the polymer plastic's thermal transitions, such as glass transition temperature (Tg), melting temperature (Tm), and crystallization temperature (Tc). These thermal properties are often related to the MWD of the polymer plastic through analysis of the DSC curve. The downsides to this method are that - it makes use of indirect measurement, the information provided is limited, and it is only suitable for the analysis of small samples, that is, it cannot be used to monitor large volumes in industrial production. Aside from this, this method is fast and efficient.

The method discussed in this dissertation report is called Rheology (Read 2015). It is similar in nature to DSC, wherein it involves the data analysis of a curve related to a certain polymeric attribute. In our case, this corresponds to the flow properties of the polymer plastic. Despite its similarities with DSC, Rheology is better suited for our use case as the flow properties of the plastic mass can be quickly assessed and it can be applied to a sample size making it practical to monitor large volumes in industrial production. Another advantage of Rheology is that it analyzes the test material without altering its properties, making the process non-destructive in nature.

But how reliable and accurate is data analytics in predicting the MWD of polymer plastics from flow data? How susceptible is it to risk? Can such results be easily reproduced? These are the questions the report aims to answer.

However, before doing so, it is important to first fully understand what the MWD is, and why this distribution can help us classify polymer plastics into their polymer types effectively.

1.1 Molecular Weight Distribution (MWD)

1.1.1 MWD in Polymers

Polymers are large molecules composed of repeating structural units called monomers, which are chemically bonded to form long chains. These materials can be natural, like DNA, proteins, and cellulose, or synthetic, like plastics (polyethylene, polystyrene) and synthetic fibers (nylon, polyester). The arrangement and composition of these monomers give polymers their unique properties.

Now, there exist in polymers different types of monomers as well. Monomers can either be linear or branched. Linear monomers have a straightforward, chain-like structure without any side branches. They consist of repeating units arranged in a straight line. Some examples of polymers made up with linear monomers are Ethylene and Vinyl Chloride. Branched monomers however, have side chains or branches attached to the main backbone. Isobutylene and Styrene in conjunction with a branching agent are some example of polymers with branched monomers.

In this dissertation report, the study focuses on the types of polymers which make up most plastics (constantly referred to in the report as "polymer plastics"), which are polymers made up of linear monomers (such as Polyethylene (PE), Polypropylene (PP), and Polyvinyl Chloride (PVC)). These types of polymers are called Linear Polymers. The MWD of Linear Polymers can be of any arbitrary form, but in this dissertation, we will analyze only those that follow a unimodal-normal distribution (i.e. have a single molecular weight) and a bimodal-normal distribution (i.e. having two distinct molecular weights).

1.1.2 MWD : Derivation

Consider a polymer plastic that consists of molecules of different weights m.

(Mead et al. 2018) The number-biased probability distribution is notated by $P(m)$ and is defined as the probability that a molecule chosen randomly from the polymer plastic has a weight equal to m. Similarly, $P(m)\,dm$ is the probability that a molecule randomly selected from the polymer plastic weighs between m and $m + dm$.

The weight-biased probability distribution is notated by $W(m)$ and is defined as the probability of a monomer picked at random is on a chain with a molecule of molecular weight m. This probability is biased more towards larger molecules in the polymer plastic sample. The two probabilities are related as follows :

$$W(m) = \frac{mP(m)}{m_n} \quad \text{(where } m_n \text{ is the number average molecular weight)} \tag{1.1}$$

m_n can be expressed as :

$$m_n = \int mP(m)\,dm \tag{1.2}$$

In the way that the number average molecular weight relates to the number-biased probability, the weight average molecular weight notated by m_w is expressed as :

$$m_w = \int mW(m)\,dm \tag{1.3}$$

$$\implies \quad m_w = \frac{\int m^2 P(m)\,dm}{\int mP(m)\,dm} \quad \text{(from equations 1.1 and 1.2)} \tag{1.4}$$

However, when plotting the weight average molecular weight, it is found to generally normalize on a logarithmic scale. So, let

$$x = \log(m) = \frac{\ln(m)}{\ln(10)} \tag{1.5}$$

$$\frac{dx}{dm} = \frac{1}{m\ln(10)} \tag{1.6}$$

$$\implies \quad dm = dx(m\ln(10)) \tag{1.7}$$

Now,

$$P(m)\,dm = P(x)\,dx \tag{1.8}$$

$$\implies \quad P(x) = P(m)\,\frac{dm}{dx} \tag{1.9}$$

So the number-biased probability on the logarithmic scale can be given by,

$$P(\log(m)) = mP(m)\ln(10) \quad \text{(from equations 1.5 and 1.6)} \tag{1.10}$$

And the weight-biased probability on the logarithmic scale can hence be given by,

$$W(\log(m)) = mW(m)\ln(10) \tag{1.11}$$

$$\implies \quad W(\log(m)) = \frac{m^2 P(m)\ln(10)}{m_n} \quad \text{(from equation 1.1)} \tag{1.12}$$

1.1.3 MWD : Target Parameters

The molecular weight distribution (MWD) which we aim to predict using the rheology curves is the distribution of $W(\log(m))$ with respect to $\log(m)$. It is only in the log parameter space that the MWD is ideally unimodal-normal in nature (in the case of linear polymers with a single dominating molecular weight). In this case, there are only two parameters that need to be predicted, the mean and the variance, which are denoted by $\log(Z)$ and $\log(PDI)$.

Z is a dimensionless quantity that indicates the degree of entanglement within a polymer plastic sample. If Z is much greater than 1, it means that the polymer plastic chains are long enough to be well entangled with each other. This means that the polymer plastic typically has a high amount of strength, toughness and elasticity. When the Z value is close to 1 or much less than 1, it implies that the polymer plastic chains are short and the material has lower strength,

making it behave more like a viscous liquid with little or no entanglement. Formulaically, $\log(Z)$ is found to be :

$$\log(Z) = \log\left(\frac{m_w}{m_e}\right) \tag{1.13}$$

where m_e is known as the entanglement molecular weight and is defined as the molecular weight at which the polymer plastic's chains in a melt or concentrated solution become entangled with one another (Fetters et al. 1999). This value is unique to each plastic material type (that is, polyethylene (PE), polypropylene (PP), and polyvinyl chloride (PVC) have distinct values of m_e).

PDI is an abbreviation for Polydispersity Index and provides insight into the breadth of the MWD of the polymer plastic sample. It indicates how uniform (or non-uniform) the polymer plastic chains are in terms of their length. A PDI of 1 indicates that the polymer plastic has chains of equal molecular weight. A PDI greater than 1 indicates a distribution of molecular weights. The higher the PDI, the broader the distribution. $\log(PDI)$ can be formulaically shown by :

$$\log(PDI) = \log\left(\frac{m_w}{m_n}\right) \tag{1.14}$$

From this point on, this dissertation report will refer to $\log(Z)$ and $\log(PDI)$ as $\overline{Z}$ and $\overline{PDI}$. When these variables are used, it is understood that the parameters measured are in the log parameter space and any properties mentioned are only viable in the log parameter space.

1.1.4 MWD : Effects on Polymer Plastics

- **Mechanical Properties :** Polymer plastics with low $\overline{PDI}$ tend to have more uniform mechanical properties, like strength and elasticity. Polymer plastics with high $\overline{PDI}$ can offer better toughness and impact resistance because the variety of chain lengths can distribute stress more effectively.

- **Processing Behavior :** The ease of processing (e.g., molding, extrusion) is affected by MWD. Polymer plastics with high $\overline{PDI}$ often have better processability due to a mix of low and high molecular weight chains, which help in flow and stability during processing (Ahn & Kim 2002).

- **Thermal Properties :** MWD affects the melting and crystallization behavior of polymer plastics, which is critical for applications that involve heat exposure.

- **Chemical Resistance and Durability :** Variations in chain length affect how the polymer interacts with chemicals, influencing its resistance to degradation, solvents, and environmental factors.

- **End-Use Performance :** Understanding MWD helps in tailoring polymer plastics for specific applications, ensuring the right balance of properties like strength, flexibility, and durability (Clarke-Pringle & Macgregor 1998).
- **Quality Control and Reproducibility :** Manufacturers can use an MWD to maintain consistency in polymer plastic production, ensuring that each batch meets the desired specifications.

1.2 Rheology

1.2.1 Obtaining flow data

Plastics in their molten state adopt the characteristics of viscoelastic liquids. It is in this form that the plastic is subjected to the rheology experiment. This experiment involves studying the flow properties (Sun & Sahinidis 2021), that is, the trend of the storage modulus (G') and loss modulus (G'') for the change in angular frequency (ω) of the test plastic.

The storage modulus (G') represents the solid-like nature of the polymer plastic. It measures the material's ability to store energy during deformation and reflects its stiffness. The loss modulus (G'') represents the liquid-like nature of the polymer plastic. It measures the material's ability to flow and lose energy, indicating its internal friction. The Angular Frequency (ω) is the rate of oscillation applied during rheological testing.

The first step in conducting the rheological experiment is to first load the material onto a rotational rheometer. For an oscillatory linear rheology experiment, the rheometer applies an oscillating strain (i.e. deformation) to the material and measures the stress (i.e. force per unit area) required to deform the material.

The Strain (γ) applied can be given by the formula :

$$\gamma = d/h \tag{1.15}$$

where d is the maximum horizontal displacement that is plastic is going through and h is the height of the measuring plate on which the plastic material is loaded. However, for an oscillatory rheology experiment, the applied strain is of form :

$$\gamma = \gamma_0 \sin(\omega t) \tag{1.16}$$

where γ_0 is a constant and ω is the angular frequency, which is varied.

In a rheological experiment, the Stress (σ) to deform the material depends on what type of material it is. When the material is a solid,

$$\sigma = G'\gamma \tag{1.17}$$

Where G' is the storage modulus and γ is the strain that the material is put through.

From equation 1.16, σ can be written as :

$$\sigma = G'\gamma_0 \sin(\omega t) \tag{1.18}$$

In the case of liquid materials,

$$\sigma = \eta \left(\frac{d(\gamma)}{dt} \right) \tag{1.19}$$

where η is the viscosity of the liquid.
Once again, from equation 1.16, we get,

$$\sigma = \eta\omega \cos(\omega t) \tag{1.20}$$

Now, Viscoelastic liquids have properties intermediate between solid and this behaviour varies with the frequency (Domone & Illston 2017). Typically materials are more solid like at high frequency and more liquid like at low frequency. The stress of these materials upon accounting for both a solid and liquid response can be written in a single equation as follows :

$$\sigma = \gamma_0 \left[G'(\sin(\omega t)) + G''(\cos(\omega t)) \right] \tag{1.21}$$

where G' is the storage modulus and G'' is the loss modulus.

After the values of G' and G'' are recorded at various frequencies ω, they are plotted in the log parameter space. They produce in general a graph as shown in (Figure 1.1) which is what is commonly known as flow data.

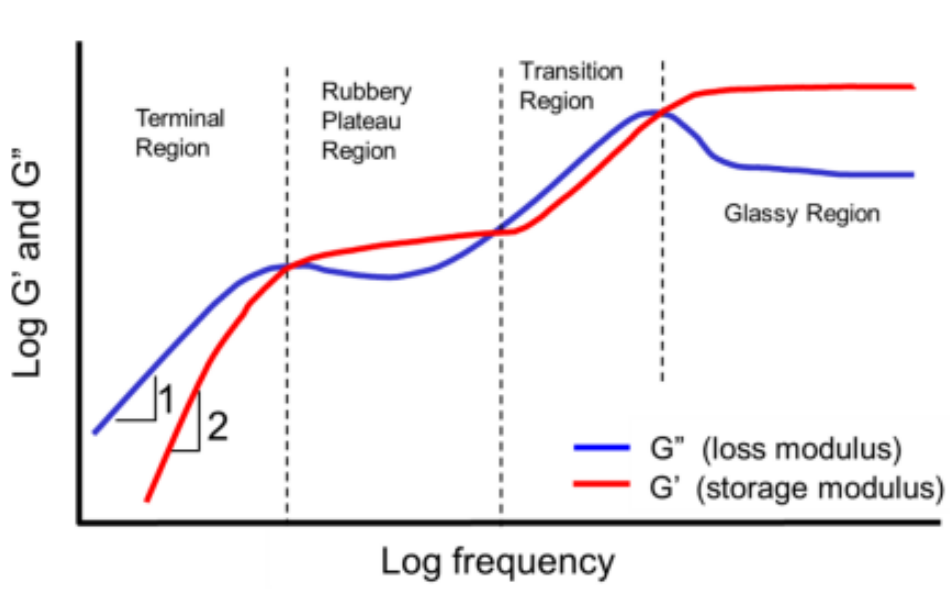

Figure 1.1: Storage modulus and Loss modulus as a function of angular frequency (log space) along the full viscoelastic spectrum (Source : A Basic Introduction to Rheology *(2016))*

From this point on, this report will refer to $\log(G')$ and $\log(G'')$ as $\overline{G'}$ and $\overline{G''}$. $\log(\omega)$ will

also be referred to as $\overline{\omega}$.

1.2.2 Understanding flow data

There are 4 distinct parts (Rao et al. 1982) to graph from Figure 1.1, which are differentiated by the intersection of $\overline{G'}$ and $\overline{G''}$ plots.

- **Low Frequency Region (Terminal Region) :**

  At low frequencies, the log storage modulus $\overline{G'}$ is much lower than the log loss modulus $\overline{G''}$, indicating that the material behaves mostly like a liquid. This is because, at low frequencies, the material has enough time to relax and flow.

- **Plateau Region (Rubbery Plateau Region) :**

  In this region, the storage modulus is consistently greater than the loss modulus, so the material behaves more like an elastic solid. In polymers, this region is known as the rubbery plateau as in this region, the polymer chains are entangled, giving the material its elastic properties.

- **Transition Region (Viscoelastic Region) :**

  In this region, the storage modulus and loss modulus become comparable in magnitude. Hence, the material exhibits both viscous and elastic behaviour. The material does not have enough time to relax at these frequencies fully, so the elastic component starts to become significant.

- **High Frequency Region (Glass Transition Region) :**

  At very high frequencies, there is a steep rise in the storage modulus and an eventual slow decline in the loss modulus due to which the material enters into a glassy state and behaves like a solid.

1.2.3 Generating flow data using Time-Temperature Superposition Theorem

In real life, there are limitations to the range of angular frequency we can reliably measure the flow data for. Typically, the range of $\overline{\omega}$ values are between -1 and 2. This range of frequencies only covers either the low frequency or the plateau region of the flow curves and there is a lot more information to be extracted from the rest of the frequency spectrum.

To acquire values of $\overline{\omega}$ lesser than -1, a scientist would have to wait a minimum of $t = (2\pi \cdot 10^{\overline{\omega}})^{-1}$ seconds which in the case of -2 would be :

$$t = (2\pi \cdot 0.01)^{-1} \approx 15.92 \textit{ seconds}$$

just to record one value of $\overline{G'}$ and $\overline{G''}$. Indeed, the limiting factor in obtaining values lower than $\overline{\omega} = -1$, is patience. It would simply take too much time to reliably record the low-

frequency range of every plastic mass as for most plastics, the low-frequency range lies below a $\overline{\omega}$ value of -2.

In order to acquire values of $\overline{\omega}$ greater than 2, the time period ends up being very small and the force needed to accelerate the metal parts of the rheometer starts to become larger than the forces required to deform the material. Hence, it becomes harder to measure stress reliably.

$$t = (2\pi \cdot 1000)^{-1} \approx 1.592 \times 10^{-4} \textit{ seconds}$$

In these situations, some materials can make use of the time-temperature superposition theorem (Ljubic et al. 2014) and can use of changes to temperature as a proxy for changing frequency. Low-frequency information can be acquired by the increase in temperature to allow for more mobility between the polymer chains, hence increasing the viscosity. On the other hand, high-frequency information can be obtained by experimenting with lower temperatures, giving the polymer chains more rigidity.

In this way, by making changes to the temperature, it is possible to effectively explore a wider frequency range and hence record values of $\overline{G'}$ and $\overline{G''}$ without experiencing the aforementioned limitations.

1.3 Tube Model

However, in order to create a deep-learning approach to predict the characteristics of the MWD, a data source is required to train and test the data. Unfortunately, there exists no free database with reliable credibility.

Das & Read (2023) have developed a tube model that, for a given polymer plastic and MWD characteristics, can be used to reliably predict rheology curves. The tube model is built on firm assumptions and theories and hence can be considered to mimic the values and trends of experimentally recorded rheology curves.

The dataset used to train our deep learning model will be from the rheology curves generated as outputs for $\overline{Z}$ and $\overline{PDI}$ values from this tube model. Since the data is generated and not sourced from actual real-life experimentation, it is possible to acquire as large of a dataset as we want, to make the deep learning model as accurate as possible.

1.4 Overall Aim of the Dissertation

This dissertation aims to primarily verify whether a deep learning methodology is sufficient to accurately predict the MWD of a polymer plastic given its flow data. By rapidly determining the distribution of molecules in each batch, a manufacturer can decide whether, and if so, how, the recycled plastic can be used. Aside from this, the dissertation aims to answer the following questions :

- How accurate are the predictions?

- Is it possible to restrict the number of frequencies measured and still characterize the material?
- What is the effect of experimental noise or uncertainty?
- How will the model deal with real-life data?
- How robust is the model?
- Does spectral analysis reveal trends stated by rheological concepts?
- Is it possible to uncover unique insights that will provide a solid foundation for future work?

In this dissertation, the focus is on linear polymer plastics. As discussed earlier, the parameters to be predicted for the MWD of linear polymer plastics with a predominant single molecular weight are the mean and variance of a unimodal log-normal distribution given by $\overline{Z}$ and $\overline{PDI}$ respectively.

In the case of linear polymer plastics consisting of two distinct molecular weights predominantly make up the polymer plastic, the MWD follows a bimodal log-normal distribution. Considering a bimodal log-normal distribution to be the combination of two unimodal log-normal distributions,let the mode with the smaller mean be represented as S and the mode with the larger mean be represented as L. Assuming that the ratio of coverage of mode L is given by $\overline{\phi_l}$, there are totally 5 parameters to predict. The mean and variance of mode S ($\overline{Z_s}$ and $\overline{PDI_s}$), the mean and variance of mode L ($\overline{Z_l}$ and $\overline{PDI_l}$), and the ratio of coverage of mode L ($\overline{\phi_l}$).

Chapter 2

Preliminary Analysis

2.1 Unimodal Data

The layout of the Unimodal data given in the dataset
`uni_PE_wint_2_low_uniform_200000_training.csv` is as shown in Table 2.1 :

$\overline{Z}$	$\overline{PDI-1}$	$\overline{G'(0)}$	$\overline{G'(1)}\ldots\ldots\overline{G'(94)}$	$\overline{G''(0)}$	$\overline{G''(1)}\ldots\ldots\overline{G''(94)}$
$\overline{z}_1$	$\overline{pdi-1}_1$	$\overline{g'(0)}_1$	$\overline{g'(1)}_1 \ldots \overline{g'(94)}_1$	$\overline{g''(0)}_1$	$\overline{g''(1)}_1 \ldots \overline{g''(94)}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{z}_1$	$\overline{pdi-1}_{450}$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{z}_2$	$\overline{pdi-1}_1$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{z}_2$	$\overline{pdi-1}_{450}$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{z}_{450}$	$\overline{pdi-1}_1$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overline{z}_{450}$	$\overline{pdi-1}_{450}$	$\overline{g'(0)}_{202500}$	$\overline{g'(1)}_{202500} \ldots \overline{g'(94)}_{202500}$	$\overline{g''(0)}_{202500}$	$\overline{g''(1)}_{202500} \ldots \overline{g''(94)}_{202500}$

Table 2.1: Layout of the given Unimodal Dataset

The unimodal dataset consists of 202500 generated Rheology curves for 450 unique $\overline{Z}$ and $\overline{PDI-1}$ values. For each unique value of $\overline{Z}$, there are 450 $\overline{PDI-1}$ values for which the Rheology curves are generated. Hence, there are 450×450 curves generated in the dataset which totals 202500.

The column containing the $\overline{Z}$ values is labeled "Z" and the column containing the $\overline{PDI-1}$ values is labeled "PDI", which is misleading but done so for the sake of simplicity. The $\overline{Z}$ values are evenly spread out between 0.48 and 3. The $\overline{PDI-1}$ values are exponentially spread

out across -1.63 and 0.95. However, once converted into this $\overline{PDI}$, this column is also evenly distributed, the range being 0 to 1. These are the target variables for our deep learning model.

For our explanatory variables, there are 95 columns for $\overline{G'}$ and 95 columns for $\overline{G''}$ corresponding to the 95 values of $\overline{\omega}$ at which the storage modulus and loss modulus have been recorded. The columns in the dataset are named "G'(0)", "G'(1)", "G'(2)" and so on until "G'(94)" for the 95 $\overline{G'}$ columns and "G"(0)", "G"(1)", "G"(2)" and so on until "G"(94)" for the 95 $\overline{G''}$ columns. The range of $\overline{G'}$ values are -17.89 to 7.42 that of $\overline{G''}$ are -5.96 to 7.61.

The numbers within the brackets, that is, 0,1,2, etc. are not the values of $\overline{\omega}$, but correspond to the serial number of the ω value given in a separate text file named `run0.dat`. The ω values range from 10^{-4} up to 1.41×10^{10}.

The plastic used in this batch of generated data is Polyethylene with an m_e value of 820.

In order to observe the trends of the flow curves at different values of $\overline{Z}$ and $\overline{PDI}$, plots at $\overline{Z}$ and $\overline{PDI}$ percentile values of $0, 25, 50, 75$ and 100 are made. The plotted graphs are showcased below (Figure 2.1) :

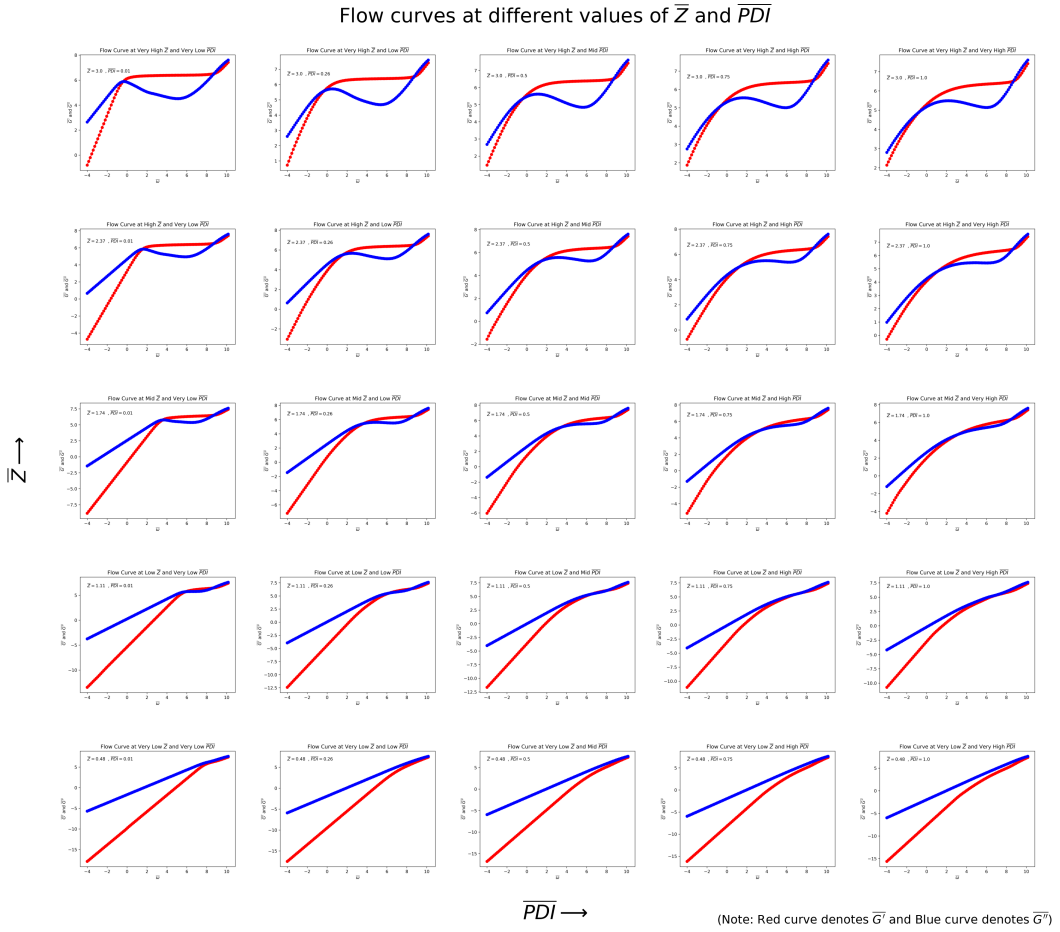

Figure 2.1: Flow curves at 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile values of $\overline{Z}$ and $\overline{PDI}$

Upon inspection of the graphs in Figure 2.1, we see that $\overline{Z}$ has a huge impact on the outcome

of the flow curve. As $\overline{Z}$ increases, the intersection between the storage modulus ($\overline{G'}$) and loss modulus ($\overline{G''}$) occurs earlier at all values of $\overline{PDI}$. With the increase $\overline{Z}$, the area taken up by the rubbery plateau region decreases as well.

Although not noticeable at first glance, $\overline{PDI}$ also has a huge impact on the outcome of the flow curve. The lowest ($\overline{G'}$) and ($\overline{G''}$) values recorded for the curves in the low frequency region increases with increase in $\overline{PDI}$. This effect is more noticeable for the storage modulus curve ($\overline{G'}$) than the loss modulus curve ($\overline{G''}$).

2.2 Bimodal Data

The layout of the Bimodal data given in the dataset
`bi_PE_5param_wint_2_400000_training` is as shown in Table 2.2 :

$\overline{Z_s}$	$\overline{Z_l}$	$\overline{PDI_s}$	$\overline{PDI_l}$	$\overline{\phi_l}$	$\overline{G'(0)}$ ... $\overline{G'(94)}$	$\overline{G''(0)}$ ... $\overline{G''(94)}$
$\overline{z}_{s1}$	$\overline{z}_{l1}$	$\overline{pdi}_{s1}$	$\overline{pdi}_{l1}$	$\overline{\phi}_{l1}$	$\overline{g'(0)}_1$ ... $\overline{g'(94)}_1$	$\overline{g''(0)}_1$ ... $\overline{g''(94)}_1$
⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	
$\overline{z}_{s408050}$	$\overline{z}_{l408050}$	$\overline{pdi}_{s408050}$	$\overline{pdi}_{l408050}$	$\overline{\phi}_{l408050}$	$\overline{g'(0)}_{408050}$ ... $\overline{g'(94)}_{408050}$	$\overline{g''(0)}_{408050}$ ... $\overline{g''(94)}_{408050}$

Table 2.2: Layout of the given Bimodal Dataset

The bimodal dataset consists of 408050 generated Rheology curves for several different combinations of $\overline{Z_s}$, $\overline{Z_l}$, $\overline{PDI_s}$, $\overline{PDI_l}$, and $\overline{\phi_l}$, which are the target variables. The columns containing these variables are named "Zs", "Zl", "PDI_s", "PDI_l" and "Phi_l" respectively. The combinations are determined by a random number generator but at the same time, are uniformly distributed in a 5D space with a few restrictions.

A bimodal curve can be thought of as two unimodal curves joined together. Let the mode with the lower mean be S and the mode with the larger mean L. The mean of the mode L ($\overline{Z_l}$) must at least be twice as large as the mean of mode S ($\overline{Z_s}$). There are no restrictions on variance ($\overline{PDI_s}$ and $\overline{PDI_l}$). $\overline{\phi_l}$ is the ratio of area covered by the large curve L on a scale of 0 to 1. Logically, $\overline{\phi_l}$ can take up values within $(0, 1)$ but in order to be able to differentiate between the two unimodal curves clearly, $\overline{\phi_l}$ is restricted to take up values between $[0.05, 0.95]$.

The explanatory variables, that is, the points on the storage modulus and loss modulus curves are 95 each in both cases and are represented the exact same way as done in the unimodal case. The frequencies (ω) at which these points are recorded are also the exact same as that of the unimodal case, ranging from 10^{-4} up to 1.41×10^{10}. The range of $\overline{G'}$ values are -17.47 to 7.42 and for $\overline{G''}$ are -5.92 to 7.61.

The plastic used in this batch of generated data is also Polyethylene with an m_e value of 820.

Now, in order to observe the changes to the shapes of the flow curves with respect to the 5 target parameters, some restrictions are developed in order to clearly differentiate between the effect of each parameter. Since there are conditions placed on $(\overline{Z_l})$ to be atleast twice as large as $(\overline{Z_s})$, there is no need to observe the impact the simultaneous change of these two parameters on the flow curves. The same is the case with $\overline{PDI_s}$ and $\overline{PDI_l}$, with a fixed value of $\overline{PDI_l}$ implying a certain range of values that can be taken up by $\overline{PDI_s}$. In order to first observe the impact of the means and variances of the two modes on the flow curves, $\overline{\phi_l}$ is fixed at a constant value close to 0.5. Once again, the comparison will be made at 5 significant percentile values (that is, at the 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile) of each variable.

- $\overline{Z_s}$ vs $\overline{PDI_s}$:

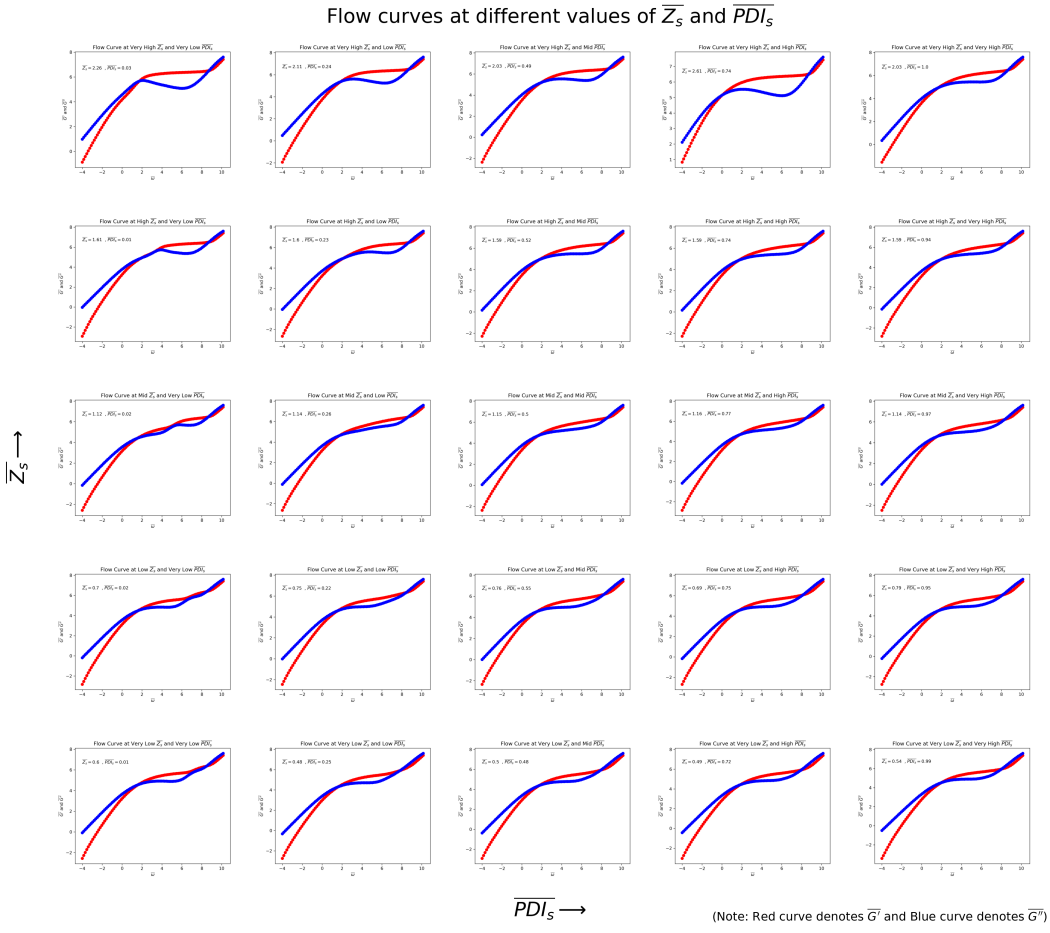

Figure 2.2: Flow curves at 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile values of $\overline{Z_s}$ and $\overline{PDI_s}$

Mode S, being the mode with the smaller mean of the two, has less of an impact on the flow curves (Figure 2.2). At very low $\overline{PDI_s}$, while there are some noticeable changes

to the curves with change in $\overline{Z_s}$, this is not the case at other values of $\overline{PDI_s}$. Change in $\overline{PDI_s}$ with respect to a constant value of $\overline{Z_s}$ does not change the curve significantly.

- $\overline{Z_l}$ vs $\overline{PDI_s}$:

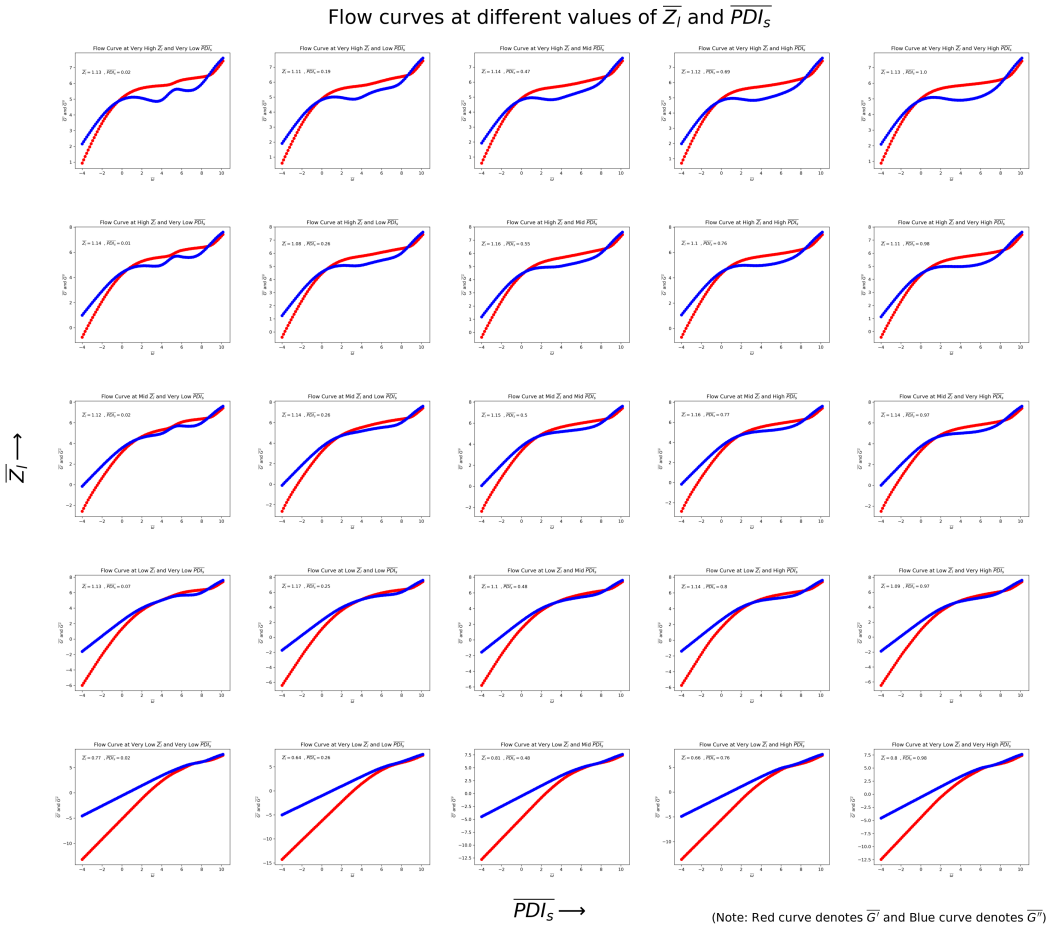

Figure 2.3: Flow curves at 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile values of $\overline{Z_l}$ and $\overline{PDI_s}$

The shape of the flow curves (Figure 2.3) at each value of $\overline{Z_l}$ at any certain value of $\overline{PDI_s}$ seem very distinct from each other, which should allow for accurate predictions of $\overline{Z_l}$. A higher value of $\overline{Z_l}$ generally means a lower value of $\overline{\omega}$ at which the flow curves first intersect. A higher value of $\overline{Z_l}$ also generally shows a higher slope of $\overline{G'}$ in the terminal region. However, when having a constant value of $\overline{Z_l}$, a change in $\overline{PDI_s}$ values does not seem to impact the flow curves significantly.

- $\overline{Z_s}$ vs $\overline{PDI_l}$:

Here, the flow curves show the exact opposite response to that in the previous case (Figure 2.3). In Figure 2.4, since the PDI value is from mode L and the Z value from node S, the change in $\overline{Z_s}$, seems to have little to no impact on the outcome of the flow curve, unlike the previous case where $\overline{Z_l}$ had a lot of impact on the outcome of the flow curve.

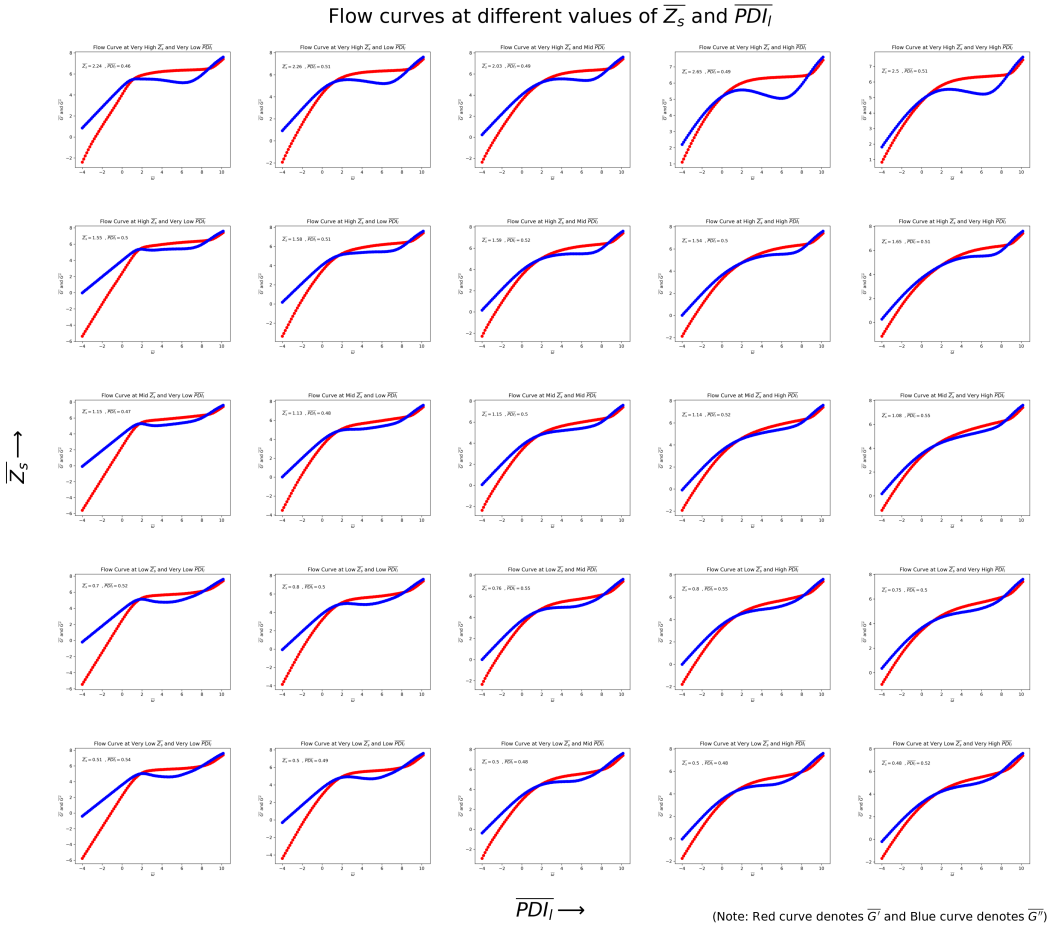

Figure 2.4: Flow curves at 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile values of $\overline{Z_s}$ and $\overline{PDI_l}$

In this case, for any fixed $\overline{Z_s}$, the increase in $\overline{PDI_l}$ shows a decrease in slope of $\overline{G'}$ and an increase in slope of $\overline{G''}$ in the terminal region. This also implies an increase in the y-intercept value of $\overline{G'}$ and a decrease in y-intercept value of $\overline{G''}$.

- $\overline{Z_l}$ vs $\overline{PDI_l}$:

  In figure 2.5, both Z and PDI are taken from the mode with the larger mean L. Hence, it is observed that change in both $\overline{Z_l}$ and $\overline{PDI_l}$ have a significant impact on the flow curves. For a fixed value of $\overline{PDI_l}$, the increase in the value of $\overline{Z_l}$ implies the first intersection of the flow curves marking the end of the terminal region to occur at a lower frequency $\overline{\omega}$. At the same time, an increase in the area enclosed by the flow curves in the rubbery plateau region is also observed. When keeping $\overline{Z_l}$ fixed and varying $\overline{PDI_l}$, the increase in $\overline{PDI_l}$ shows a decrease in the slope of $\overline{G'}$ and an increase in the slope of $\overline{G''}$ in the terminal region, an increase in the an increase in the y-intercept values of $\overline{G'}$ and a decrease in y-intercept values of $\overline{G''}$.

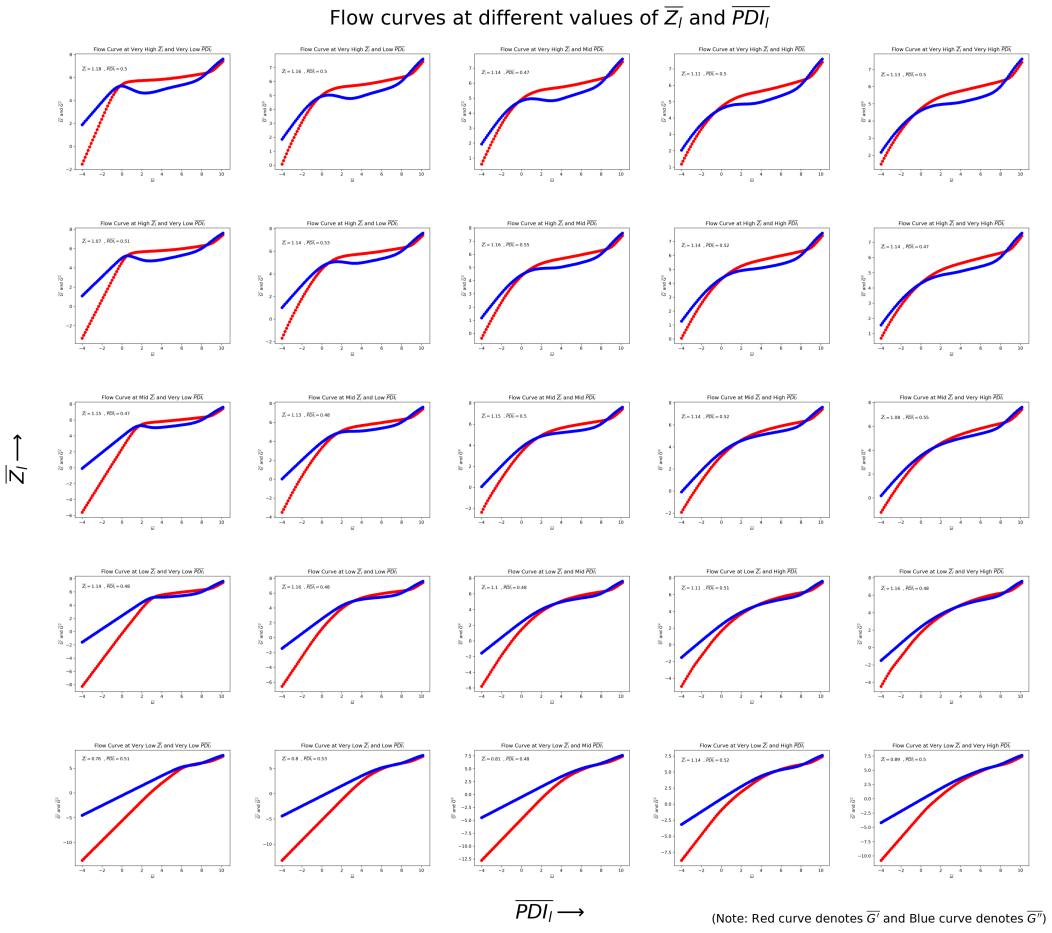

Figure 2.5: Flow curves at 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentile values of $\overline{Z_l}$ and $\overline{PDI_l}$

Chapter 3

Methodology

The deep learning approach for this problem must be one that can learn graphical data which is sequential in nature, and provide numerical outputs. An Recurrent Neural Network (RNN) architecture is perfect to model this dataset as the data has temporal properties which can be captured with such an architecture, allowing for better test accuracies.

3.1 LSTM

The explanatory variables are in the form of 2 curves with 95 points each. When running a model with so many coordinates, it is as important to retain all the initially scanned features as it is to scan new features, to allow for every scanned feature to affect the weights and biases in the neural network. For this reason, we make use of a Long Short Term Memory (LSTM) layer to overcome these shortcomings (Józefowicz et al. 2015). Unlike traditional RNNs, LSTMs do not suffer from the vanishing gradient problem due to the existence of a forget gate which decides on whether a certain piece of information should be kept or thrown away. The architecture of an LSTM is shown in Figure 3.1

Upon accepting an input $\mathbf{x}_t$, it and the previous hidden state $\mathbf{h}_{t-1}$ determine which parts of the previous cell state need to be forgotten in the forget gate $\mathbf{f}_t$ (Ohno & Kumagai 2021).

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \tag{3.1}$$

Then, the input gate $\mathbf{i}_t$ decides on which parts of the new information to update from a set of candidate values $\tilde{\mathbf{c}}_t$.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \tag{3.2}$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \tag{3.3}$$

Now, the new cell state $\mathbf{c}_t$ is updated after partially forget the unimportant old values and adding in the new candidate values.

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \tag{3.4}$$

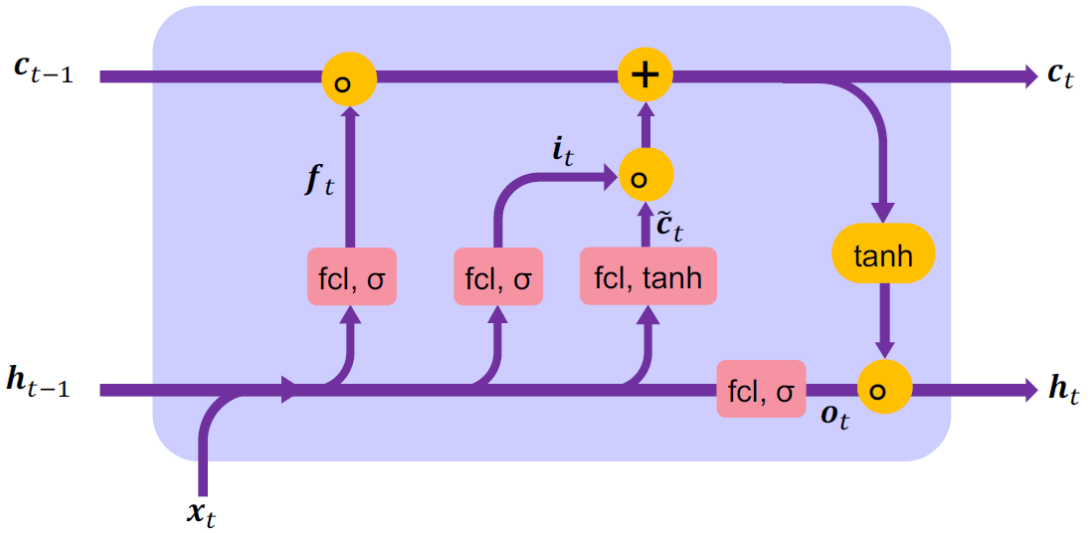

Figure 3.1: LSTM Architecture (Source : Ali et al. (2024))

Finally, the output gate decides what the next input should be by defining the next set of hidden values $\mathbf{h}_t$.

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \tag{3.5}$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) \tag{3.6}$$

This process continues in a cycle until the model is satisfactorily trained in which case $\mathbf{h}_t$ is given as the output. In this way, an LSTM in conjunction with fully connected (Dense) layers will suitably yield the necessary results.

3.2 Packages and Dependencies

- **`TensorFlow/Keras`** : The deep learning framework used to build, train, and evaluate the model. `TensorFlow` provides the core functionalities, while `Keras` (part of `TensorFlow`) (Joseph et al. 2021) provides a high-level API for neural network construction.
- **`NumPy`** : An open source library used to perform numerical operations on arrays. It is crucial for data preparation, reshaping, and manipulation.
- **`Matplotlib/Seaborn`** : Libraries helpful in visualizing training results, loss curves, and data distributions.
- **`scikit-learn`** : An open source library used for data preprocessing, such as scaling, splitting data, and evaluating model performance.

Chapter 4

Prediction on Unimodal Data

4.1 Model Development

Upon reading the csv file : `uni_PE_wint_2_low_uniform_200000_training.csv`, the csv is first checked to see if the datatype along all rows in each column is consistent, that is, of numerical datatype. The second check is to see if there are any missing values in the dataset. Both these checks return in favor of a clean dataset that is ready to be analyzed.

The distribution of the target variables are checked. The column "Z" is uniformly distributed but column "PDI" is not.

This is obviously not ideal for modeling as since the $\overline{PDI}$ values are skewed to the left, a model trained on such data would also predict a similar pattern. An ideal training set of data would look equally distributed as shown in $\overline{Z}$. The column "PDI" contains values $\overline{PDI - 1}$. Since we know that the column "PDI" is uniformly distributed in $\overline{PDI}$, each value in this column is converted accordingly and the column "PDI" is now uniformly distributed.

Now, the dataset is ready for modeling. Before the modeling process begins, it is important to first recognize what are the acceptable accuracies and the metrics we will use to predict the target variables : $\overline{Z}$ and $\overline{PDI}$.

The main metrics used to measure model performance is Mean Absolute Error (MAE) and Percentage Mean Absolute Error ($PMAE$) which are expressed by the formulae :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.1}$$

where n is the number of test values, y_i is the original value and $\hat{y}_i$ is the predicted value.

$$PMAE = \frac{MAE_y}{Range_y} \times 100\% \tag{4.2}$$

where MAE_y is the Mean Absolute Error of the target variable y and $Range_y$ is the difference between the highest and lowest value of y.

Industry standards dictate that a maximum error of 15% in the prediction of PDI (linear

space) and Z (linear space) is acceptable for the adequate separation of recycled polymers into polymer types. Since the variables in our dataset are in the logarithmic space, we will first derive the relation between absolute errors in the logarithmic space and percentage errors in the linear space.

Let y be a value in linear space. $\overline{y}$ is the corresponding value of y in the log space.

$$\overline{y} = \log(y) \tag{4.3}$$

Let the percentage error in the linear space be p. So for a value y, the predicted value in the linear space cannot be greater than y', which can be written as :

$$y' = y\left(1 + \frac{p}{100}\right) \tag{4.4}$$

Now, let's consider the corresponding error in log space. The predicted value in log space given by $\overline{y}'$ is :

$$\overline{y}' = \log(y') = \log\left(y\left(1 + \frac{p}{100}\right)\right) \quad \text{(from equations 4.3 and 4.4)} \tag{4.5}$$

$$\implies \overline{y}' = \log(y) + \log\left(1 + \frac{p}{100}\right) \tag{4.6}$$

Since $\overline{y} = \log(y)$, the absolute error in log space ($\Delta\overline{y}$) is :

$$\Delta\overline{y} = \overline{y}' - \overline{y} = \log\left(1 + \frac{p}{100}\right) \quad \text{(from equations 4.3 and 4.6)} \tag{4.7}$$

Substituting the value $p = 15$ into equation 4.7, we get,

$$\Delta\overline{y}_{max} = \log\left(1 + \frac{15}{100}\right) \tag{4.8}$$

$$\implies Acceptable\ Error\ (AE) = 0.06 \tag{4.9}$$

Hence, a percentage error of 15% in the linear space translates to an absolute error of 0.06 in the log space. We shall hence build models with these accuracies in mind; while having a reasonable model runtime.

The RNN architecture consists of one LSTM layer followed by 2 fully connected layers. This seems to be the perfect middle-ground between striving for high accuracy and committing an adequate amount of time to the model training. The per-layer architecture is shown below :

```
model = Sequential()
model.add(LSTM(64, input_shape=(1,190)))
model.add(Dense(16, activation='relu'))
model.add(Dense(2, activation='linear'))
```

Other hyper-parameters that have been tuned for this specific model training are :

- **Loss function :** The loss function chosen is the `mean_squared_error` which adequately monitors how well-trained a model is on numerical data.

- **Learning rate :** This value is optimally set at 10^{-4} as at lower values the model does not pick up on significantly more information. At higher values, the loss function reaches a local minima and is unable to find the global minima.

- **Optimizer :** `Adam`

- **Batch size :** Set to an optimal value of 32. A batch size lower than 32 tends to overfit the training data and takes too long to finish training.

- **Number of epochs :** Set to an optimal value of 100, as the training is not expected to require more than set amount of epochs.

- **Patience :** The package `EarlyStopping` from `tensorflow.keras.callbacks` is used to monitor the validation loss, in our case, the validation mean squared error. When the lowest amount of validation loss is achieved, it waits for 10 epochs to see if the RNN model can learn anything further, that is, achieve a lower validation loss. In this case, the lowest recorded validation loss is set to this new value and Early Stopping waits for another 10 epochs of training. If, however, if the model is trained for 10 more epochs and a lower validation loss is not obtained, this function resets the model weights to that which gave the lowest validation score and terminates the model training.

  The number of epochs `EarlyStopping` must wait for to receive a lower validation loss before terminating the model training is called the patience parameter. This can be any value determined by the user. In this case, it is set to 10.

- **Train:test ratio :** Set to a value of $80:20$ as it is the recommended standard for modeling. All results shown with regard to a certain dataset are trained and tested on the exact same values with a `random_state = 42` under `train_test_split()`.

4.2 Prediction on Synthetic Unimodal Dataset

On a set of uniformly distributed test sets, the RNN model is trained and gives a loss of 8.11×10^-6 on the test set, which is very low while having an accuracy score of 99.93%. In order to gain a better understanding of the accuracy of the predictions, the MAE (from 4.1) and $PMAE$ (from 4.2) of each target variable is calculated.

$$MAE_{\overline{Z}} = 1.62 \times 10^{-3} \tag{4.10}$$

$$MAE_{\overline{PDI}} = 2.06 \times 10^{-3} \tag{4.11}$$

So,

$$PMAE_{\overline{Z}} = \frac{1.62 \times 10^{-3}}{(3 - 0.48)} \times 100\% \tag{4.12}$$

$$\implies PMAE_{\overline{Z}} \approx 0.06\% \tag{4.13}$$

$$PMAE_{\overline{PDI}} = \frac{2.06 \times 10^{-3}}{(1.00 - 0.01)} \times 100\% \tag{4.14}$$

$$\implies PMAE_{\overline{PDI}} \approx 0.21\% \tag{4.15}$$

These MAE and $PMAE$ values are extremely low and show that the model has learned the data very well. While the mean errors are low, there could still be a high variance in the errors produced by the model, leading to some cases where the $\overline{Z}$ and $\overline{PDI}$ values might have errors above the industry standard leading to polymer type misclassification.

In order to make sure that this is not the case, the highest error out of all the predictions (e_{max}) made is checked to see if it is in fact lesser than or equal to the acceptable error. Then, the model can be undeniably accepted.

$$e_{\overline{Z}max} = 0.02 < 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.16}$$

$$e_{\overline{PDI}max} = 0.04 < 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.17}$$

Hence, the model is undeniably acceptable and can be used in the industry to predict on similar data.

However, as previously discussed, the model will predict different parts of a target variable with higher accuracies than others. Observing the trends in errors against target variables will yield valuable insights.

For a target variable y_i, the $MAE_{\overline{y_i}}$ vs y_j is plotted for all combinations of y (in this case : Z and PDI) at intervals of 0.125.

Analysis of MAE trends :

$\overline{Z}\ errors$ **:**

The error trends (Figure 4.1) in $\overline{Z}$ showcase that the lowest errors are seen in $\overline{Z}$ values between 1.5 and 2.5, and in $\overline{PDI}$ values between 0.4 and 0.8 and at 0.2 as well. From the graph in Figure 2.1, we see why this is the case. The rows showcasing the 25^{th}, 50^{th} and 75^{th} percentiles of $\overline{Z}$ have a greater difference in the shape of the flow curves, allowing for a better distinguishability. This means that the flow data has features which are significantly affected by those specific $\overline{Z}$ values allowing for a more accurate prediction. It also means that the flow data has features which are significantly affected by those specific $\overline{PDI}$ values which result in a more accurate prediction of $\overline{Z}$.

However, the reduced MAEs at certain areas of $\overline{PDI}$ should not be taken too seriously as when looked at from the perspective of a percentage drop, the $\overline{Z}$ values between 1.5 and 2.5 give a drop in error of 50% while the $\overline{PDI}$ values between 0.4 and 0.8 and at 0.2 give a drop in

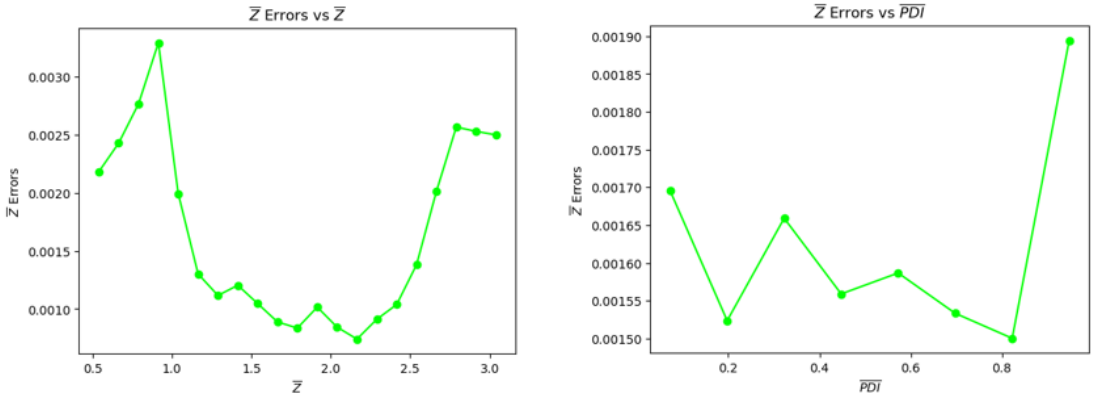

Figure 4.1: Plots of $MAE_{\overline{Z}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Synthetic Unimodal Data

error of 6.67%.

$$Percentage\ Error\ Drop\ (PED_{Z,Z}) = \frac{(0.2\text{–}0.1) \times 10^{-2}}{0.2 \times 10^{-2}} \times 100 = 50\% \tag{4.18}$$

$$Percentage\ Error\ Drop\ (PED_{Z,PDI}) = \frac{(0.165\text{–}0.154) \times 10^{-2}}{0.165 \times 10^{-2}} \times 100 = 6.67\% \tag{4.19}$$

On the flip side, it can also be observed that the $\overline{Z}$ errors in predicting $\overline{Z}$ values of 0.5 to 1 and 2.5 to 3 have much higher errors. High $\overline{PDI}$ values also seem to struggle with predicting $\overline{Z}$ in general. It is possible that we find trends and alternate strategies by which we can reduce the error of $\overline{Z}$ in these areas in upcoming sections of the report.

$\overline{PDI}$ *errors* **:**

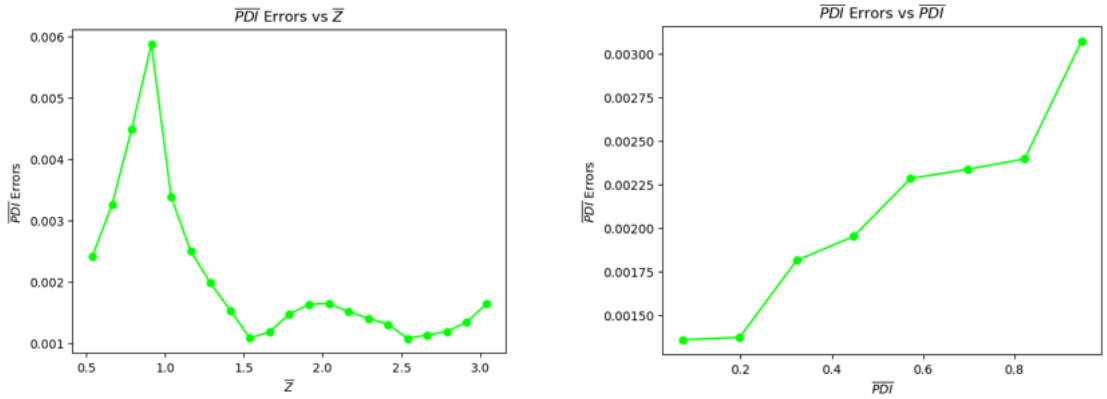

Figure 4.2: Plots of $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Synthetic Unimodal Data

The error trend (Figure 4.2) of $\overline{PDI}$ seems to be mostly positively correlated to the $\overline{PDI}$

value itself, that is, it is easier for the model to predict lower $\overline{PDI}$ values than high $\overline{PDI}$ values. By observing the graph in Figure 2.1, we can notice that the columns with the 75^{th} and 100^{th} percentile values of $\overline{PDI}$ have flow curves similar to each other at a certain value of $\overline{Z}$, however the y-intercept of the 0^{th} and 25^{th} percentile values of $\overline{PDI}$ are very much different from each other, hence lending to a significant increase in accuracy between the bottom half (0 to 0.5) and the top half (0.5 to 1) of $\overline{PDI}$ values.

$$PED_{PDI,PDI} = \frac{(0.265\text{–}0.165) \times 10^{-2}}{0.265 \times 10^{-2}} \times 100 = 37.736\% \tag{4.20}$$

However, there is a more significant drop in error in the trend observed in the $\overline{PDI}$ errors vs $\overline{Z}$. $\overline{Z}$ values greater than 1.5 predict $\overline{PDI}$ a lot more accurately than $\overline{Z}$ values less than 1.5.

$$PED_{PDI,Z} = \frac{(0.35\text{–}0.15) \times 10^{-2}}{0.35 \times 10^{-2}} \times 100 = 57.142\% \tag{4.21}$$

This percentage drop is even more significant than the drop observed in the $\overline{Z}$ errors. This is because the curves in the 25^{th} percentile values of $\overline{Z}$ are all similar even with change in the value of $\overline{PDI}$. Hence, the flow curves pertaining to that region of $\overline{Z}$ do not contain the features to predict $\overline{PDI}$ values accurately.

Analysis of Density plots :

For the last step of analysis, a density plot of the errors against the target parameters will provide the information on which regions have the most outliers and the general spread in error of the predictions.

Every black point on the density plots represents an error value of a prediction made. The darker patches of black represent a higher concentration of points while lighter patches of black represent a lower concentration of points. The purple contour encircling the majority of points signifies the level in which 90% of the points lie. This means that the points that lie outside the purple contour are in the highest 10% of error values and can be considered outliers. The green graph line is the MAE graph line in which each green point signifies the mean error of the points lying $0.125/2$ to the left and right of it. The 3 red lines parallel to the X – axis are the 25^{th}, 50^{th} and 75^{th} percentile from bottom to top respectively.

$\overline{Z}\ errors$ **:**

The density plot (Figure 4.3) of $\overline{Z}$ errors vs $\overline{Z}$ shows outliers following the trend of the MAE graph. However, the high end of $\overline{Z}$ (2.75 to 3) does not have as many outliers as does the low end (0.75 to 1) although both areas of $\overline{Z}$ have similar MAE. This means that although $\overline{Z}$'s errors are spread out in the low end of the spectrum, it has the chance to be predicted more accurately than a curve at a high $\overline{Z}$ value (this can be verified visually on the graph as well). The $\overline{Z}$ errors with respect to $\overline{PDI}$ do not follow much of a trend.

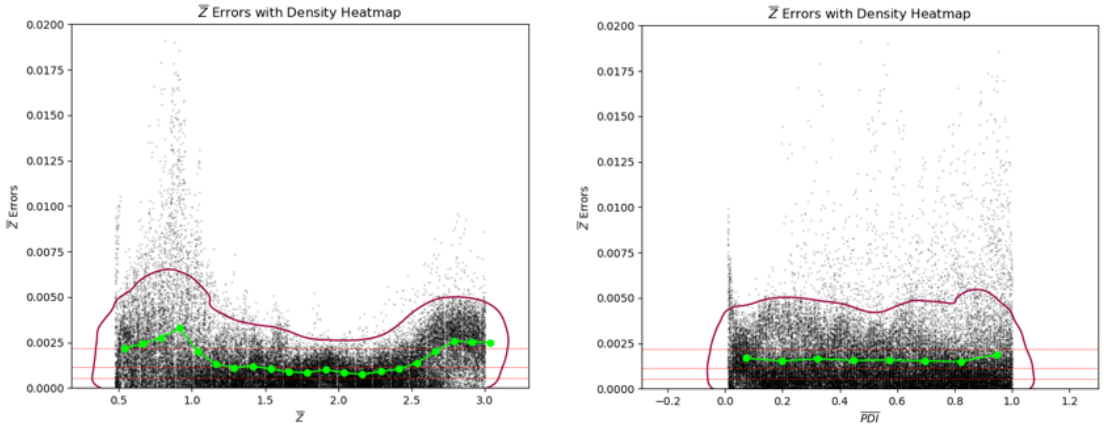

Figure 4.3: Density Plots of $\overline{Z}$ errors vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Synthetic Unimodal Data

$\overline{PDI}$ *errors* **:**

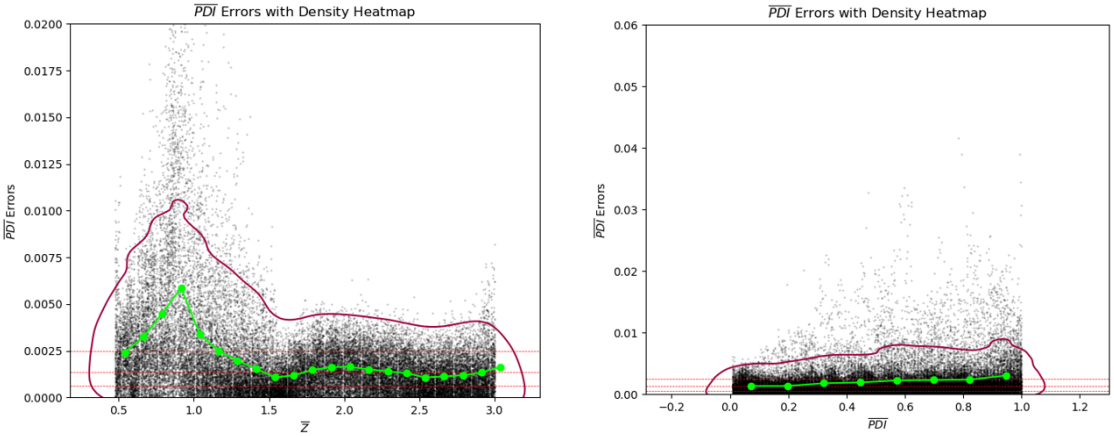

Figure 4.4: Density Plots of $\overline{PDI}$ errors vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Synthetic Unimodal Data

Similar to the $\overline{Z}$ errors, even the $\overline{PDI}$ errors (Figure 4.4) follow the trend of the MAE graph except in the extremely high end where there are a little more outliers than expected. The same is true for the $\overline{PDI}$ errors with respect to $\overline{PDI}$.

Conclusion : The flow data of curves with low values of $\overline{Z}$ (0 to 1.5) do not have the significant features necessary to predict both $\overline{Z}$ and $\overline{PDI}$ accurately. On the other hand, curves with $\overline{Z}$ values of 1.5 to 2.5 are able to predict their respective values of $\overline{Z}$ and $\overline{PDI}$ accurately. Low $\overline{PDI}$ values (0 to 0.2) are able to give accurate predictions of $\overline{Z}$ and $\overline{PDI}$.

$\overline{Z}$ values between 2.75 and 3 do not have as many outliers as expected considering the

MAE at that region.

4.3 Effects of Restricted Frequency Range

As earlier discussed in the section related to Time Temperature Superposition Theorem, it is not feasible in real life due to practical limitations, to measure the flow curve values ($\overline{G'}$ and $\overline{G''}$) beyond the range of -1 and 2. This section focuses on the possibility of modeling the MWD using just the flow data from the frequency values ($\overline{\omega}$) between -1 and 2.

Doing so brings the number of flow curve data points from 190 down to 54 (27 $\overline{G'}$ and 27 $\overline{G''}$).

Testing on uniformly distributed test sets gives a test loss of 4.2×10^{-5} and a test accuracy of 99.72%. The MAE (from 4.1) and $PMAE$ (from 4.2) values are as follows :

$$MAE_{\overline{Z}} = 3.33 \times 10^{-3} \tag{4.22}$$

$$MAE_{\overline{PDI}} = 5.43 \times 10^{-3} \tag{4.23}$$

So,

$$PMAE_{\overline{Z}} = \frac{3.33 \times 10^{-3}}{(3 - 0.48)} \times 100\% \tag{4.24}$$

$$\implies PMAE_{\overline{Z}} \approx 0.12\% \tag{4.25}$$

$$PMAE_{\overline{PDI}} = \frac{5.43 \times 10^{-3}}{(1.00 - 0.01)} \times 100\% \tag{4.26}$$

$$\implies PMAE_{\overline{PDI}} \approx 0.55\% \tag{4.27}$$

These values of MAE and $PMAE$ are once again, very low and well within the AE value of 0.06. Once again however, in order to check whether the model predicts ideally within the AE, the highest error (e_{max}) is compared against the AE.

$$e_{\overline{Z}max} = 0.032 < 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.28}$$

$$e_{\overline{PDI}max} = 0.058 < 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.29}$$

Hence, the model is very likely to predict all values of $\overline{Z}$ and $\overline{PDI}$ accurately.

Analysis of MAE trends :

$\overline{Z}\ errors$ **:**

Figure 4.5 shows a similar response to figure 4.1 in terms of the shape of the graph. The only difference is that of the magnitude and range of errors produced. When comparing Figure 4.5 with Figure 2.1, we see why this is the case. $\overline{Z}$ values between 1.5 and 2.5 generate flow curves with high variability in the slope of $\overline{G'}$ in the frequency range of -1 to 2. Hence, the model predicts with lower errors in this region.

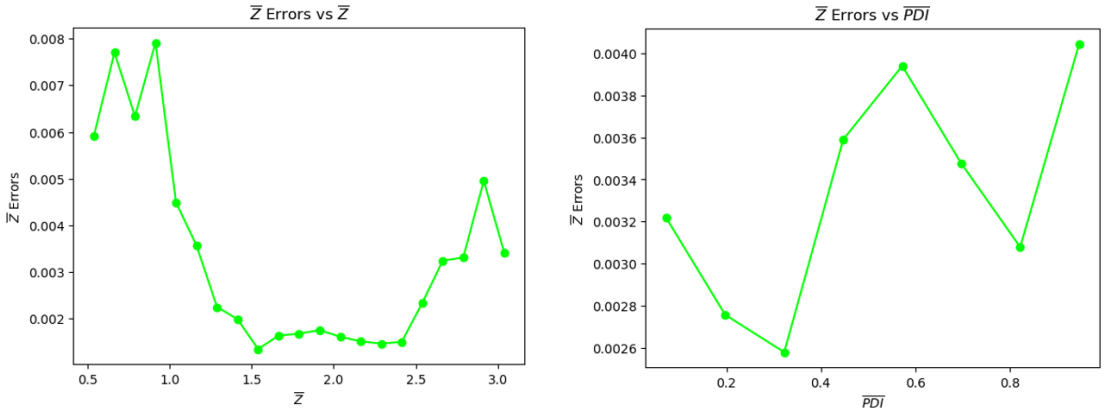

Figure 4.5: Plots of $MAE_{\overline{Z}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Frequency Restricted Unimodal Data

The $\overline{Z}$ errors vs $\overline{PDI}$ has such a much lower variance than $\overline{Z}$ errors vs $\overline{Z}$ that the graph is not worth reading too much into, although it does show an interesting trend.

$\overline{PDI}$ *errors* **:**

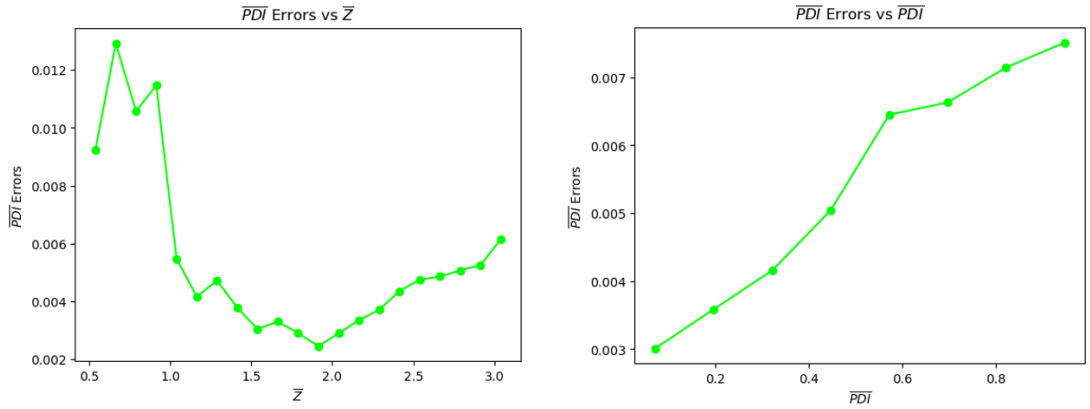

Figure 4.6: Plots of $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Frequency Restricted Unimodal Data

Once again, Figure 4.6 shows a similar response to Figure 4.2 in terms of the shape of the graph. Slight differences are observed in the $\overline{Z} = 2.5$ range where the $\overline{PDI}$ errors continues on an upward trajectory. This is because the main quantity which defined the value of $\overline{PDI}$ is the area occupied between the flow curves in the rubbery plateau region, which is lost in the $\overline{\omega}$ range of -1 to 2. In the graph $\overline{PDI}$ errors vs $\overline{PDI}$, a similar bump in errors is observed in the $\overline{PDI} = 0.8$ range. In this range, it seems as if the model is very reliant on the y-intercept of the curve $\overline{G'}$. This value lying outside the frequency range gives it it's increased error.

4.4 Effects of Induced Artificial Errors

In conducting the rheology experiment, there are a couple of reasons due to which one can expect an error in measurement of the flow curves. One is the random error which is caused due to the human operating the rheometer and another is caused by the rheometer itself, which scales the entire measurement of the flow curves by a constant k. In order to emulate the effects of such errors in the clean dataset provided to us, we will manually impute these errors into the dataset.

A random error can be expected to be of around 10% in the linear scale and the error induced by the rheometer is also around 10% in the linear scale. Converting an error of 10% in the linear scale to the log scale is :

$$\Delta\overline{y} = \log\left(1 + \frac{p}{100}\right) \tag{4.30}$$

Substituting $p = 10$, we get,

$$\Delta\overline{y} = \log(1.1) = 0.04 \tag{4.31}$$

So, a normally distributed random error in the range of ± 0.04 (from 4.31) is imputed into each cell of the explanatory variables (i.e. $\overline{G'}$ and $\overline{G''}$). This error is due to the human error caused in measurement. For the error caused by the rheometer in use, a random error in the range of ± 0.04 (from 4.31) is added to each row of the explanatory variables.

Training a model on this dataset gives a test loss of 5.95×10^{-5} and a test accuracy of 99.69%. The MAE (from 4.1) and $PMAE$ (from 4.2) values are as follows :

$$MAE_{\overline{Z}} = 3.9 \times 10^{-3} \tag{4.32}$$

$$MAE_{\overline{PDI}} = 6.9 \times 10^{-3} \tag{4.33}$$

So,

$$PMAE_{\overline{Z}} = \frac{3.33 \times 10^{-3}}{(3 - 0.48)} \times 100\% \tag{4.34}$$

$$\implies PMAE_{\overline{Z}} \approx 0.14\% \tag{4.35}$$

$$PMAE_{\overline{PDI}} = \frac{6.9 \times 10^{-3}}{(1.00 - 0.01)} \times 100\% \tag{4.36}$$

$$\implies PMAE_{\overline{PDI}} \approx 1.84\% \tag{4.37}$$

These values of MAE and $PMAE$ are once again, very low and well within the AE value of 0.06. Once again however, in order to check whether the model predicts ideally within the AE, the highest error (e_{max}) is compared against the AE.

$$e_{\overline{Z}max} = 0.053 < 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.38}$$

$$e_{\overline{PDI}max} = 0.069 > 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.39}$$

Hence, the model is very likely to predict all values of $\overline{Z}$ but not that of $\overline{PDI}$ accurately. However, it is only 7 out of the 40500 test values which have an $\overline{PDI}$ error value greater than 0.06. It is very negligible and can be ignored.

Analysis of MAE trends :

$\overline{Z}$ *errors* **:**

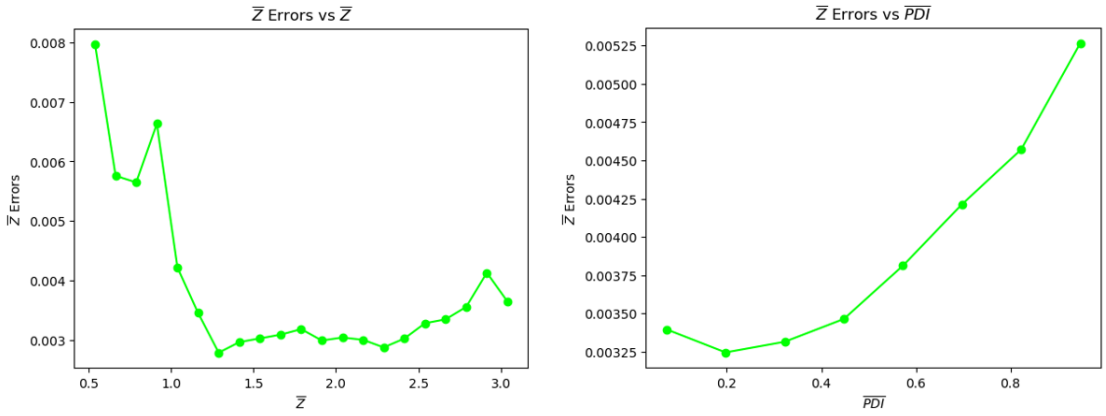

Figure 4.7: Plots of $MAE_{\overline{Z}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Induced Artificial Error Unimodal Data

The $\overline{Z}$ errors vs $\overline{Z}$ curve shown in Figure 4.7 shows that the very low $\overline{Z}$ values are struggling to get predicted. Curves with low $\overline{Z}$ values generate flow curves with only a terminal region. In adding errors to this region of the curve, the model finds it harder to predict $\overline{Z}$ values due to the lack of explanatory parameters.

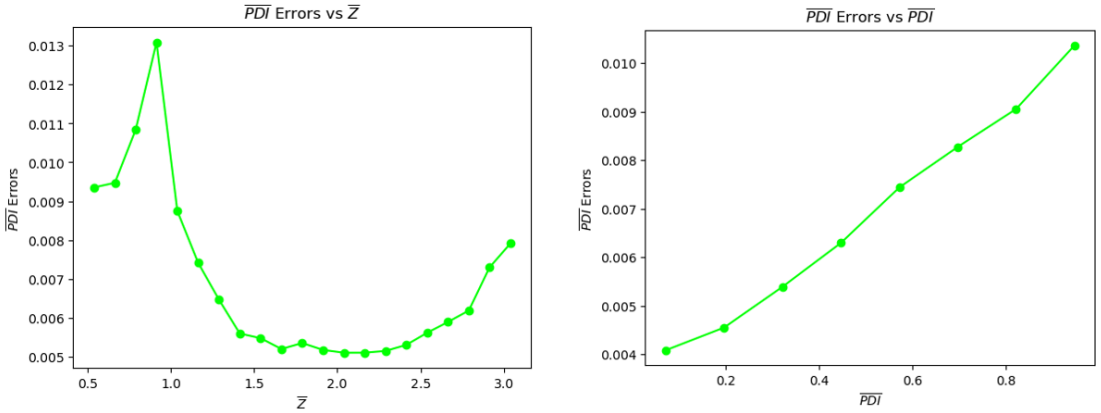

Figure 4.8: Plots of $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Induced Artificial Error Unimodal Data

In general, the addition of errors smoothens the graphs (Figure 4.7 and Figure 4.8) in com-

parison to the graphs (Figure 4.1 and Figure 4.2) trained on models without errors.

4.5 Prediction on Pseudo-Realistic Unimodal Dataset

Pseudo-Realistic data is data that is artificially changed to emulate real life data. Real life data in our case comprises of a restriction in the frequency range and also a 20% error in measurement (random and rheometer based). The results shown below indicate the performance of the model on such data.

The model when predicting on uniformly distributed test data gives a test loss of 9.46×10^{-4} and a test accuracy of 98.58%. The MAE (from 4.1) and $PMAE$ (from 4.2) values are as follows :

$$MAE_{\overline{Z}} = 1.6 \times 10^{-2} \tag{4.40}$$

$$MAE_{\overline{PDI}} = 2.7 \times 10^{-2} \tag{4.41}$$

So,

$$PMAE_{\overline{Z}} = \frac{1.6 \times 10^{-2}}{(3 - 0.48)} \times 100\% \tag{4.42}$$

$$\implies PMAE_{\overline{Z}} \approx 0.67\% \tag{4.43}$$

$$PMAE_{\overline{PDI}} = \frac{2.7 \times 10^{-2}}{(1.00 - 0.01)} \times 100\% \tag{4.44}$$

$$\implies PMAE_{\overline{PDI}} \approx 7.2\% \tag{4.45}$$

These values of MAE and $PMAE$ are low and well within the AE value of 0.06. Once again however, in order to check whether the model predicts ideally within the AE, the highest error (e_{max}) is compared against the AE.

$$e_{\overline{Z}max} = 0.12 > 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.46}$$

$$e_{\overline{PDI}max} = 0.23 > 0.06 = AE. \quad \text{(from equation 4.9)} \tag{4.47}$$

1.59% of $\overline{Z}$ values are predicted with errors greater than 0.06 (i.e. the acceptable error (AE)), and 11.17% of $\overline{PDI}$ values are predicted with errors greater than 0.06. Although this trend is seen in the cases with restricted frequency range and inducing artificial errors, it is still surprising to see the jump in magnitude of error values in $\overline{PDI}$ in comparison to $\overline{Z}$. This clearly verifies the initial observation that $\overline{Z}$ is overall easier to predict due to the changes in flow data to various $\overline{Z}$ values.

Analysis of MAE trends :

$\overline{Z}$ $errors$ **:**

The one anomaly in the trend of $\overline{Z}$ errors vs $\overline{Z}$ (Figure 4.9) is the rise in errors in the $\overline{Z}$

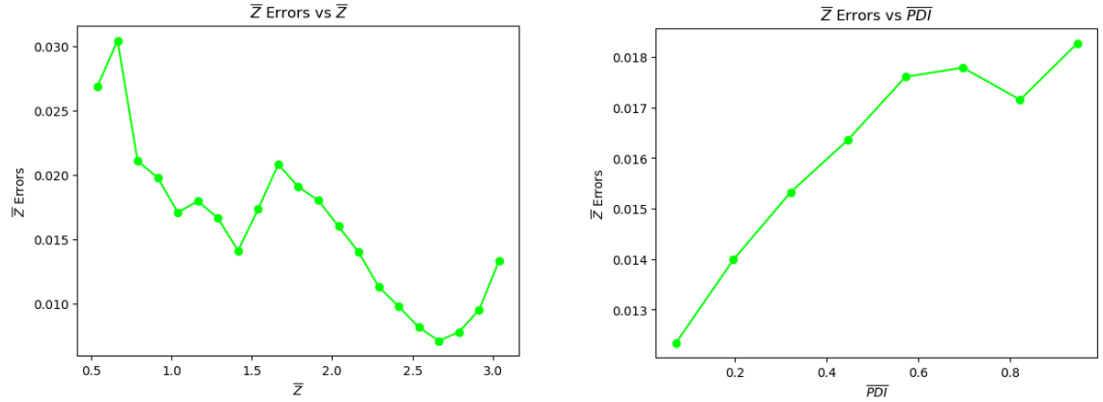

Figure 4.9: Plots of $MAE_{\overline{Z}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Pseudo Realistic Unimodal Data

range of 1.5 to 2.0. This region of $\overline{Z}$ normally has the least amount of error. This was the case when training on the clean original dataset, the restricted frequency dataset and the dataset with artificial errors induced. However, that is not the case here. It is also interesting that the region which normally had a spike in errors (i.e. $\overline{Z}$ = 2.5 to 3.0) now has the least errors and predicts with the highest accuracy. This shows that the very high values of $\overline{Z}$ are the least prone to artificial errors and that the frequency range ($\overline{\omega}$) of -2 to 1 contains most of the information required to accurately predict these values.

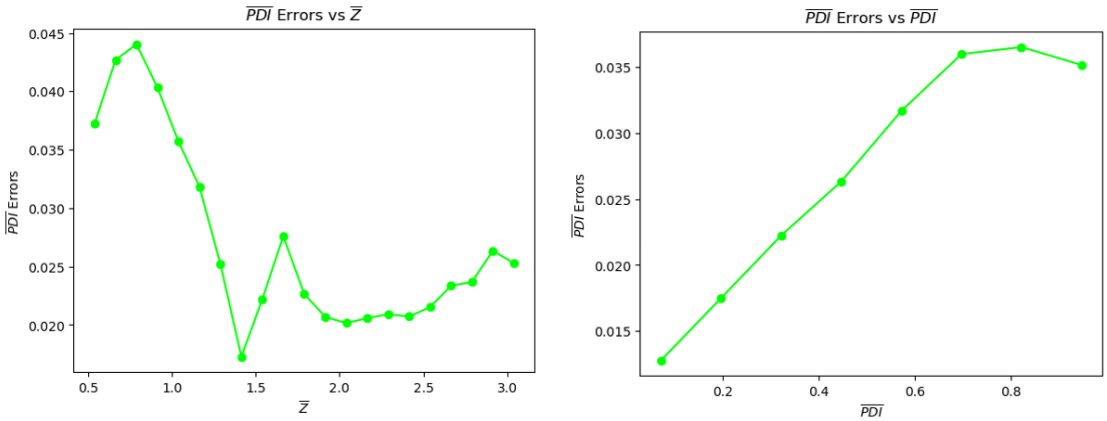

Figure 4.10: Plots of $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) using the LSTM model trained on Pseudo Realistic Unimodal Data

Similar to the previous case, even $\overline{PDI}$ errors (Figure 4.10) has a spike in errors in the range $\overline{Z}$ = 1.5 to 2. Other trends however, are consistent with previous predictions.

4.6 Evaluation of Model Robustness

Model Robustness is the ability of a model to maintain it's performance even in unforseen circumstances. A model predicting accurately on unclean data when trained on clean data and a model predicting accurately on clean data when trained on unclean data are both examples of robust models. Situations may arise when a model trained on a certain type of plastic, used a certain distribution of errors is given the task to classify plastics of a different type, with higher error values. In such scenarios, a robust model would predict with as little error as possible.

In order to evaluate the robustness of the RNN model, the original dataset and the dataset with artificially induced errors are made use of. The maximum error value induced artificially into the dataset is 0.64. The mean error value of artificially induced errors is 0.11. With these values in mind, we can measure the robustness of our RNN model.

Analysis of MAE trends :

The first scenario involves the comparison of MAE values on predicting unclean data, when the model is trained on clean data.

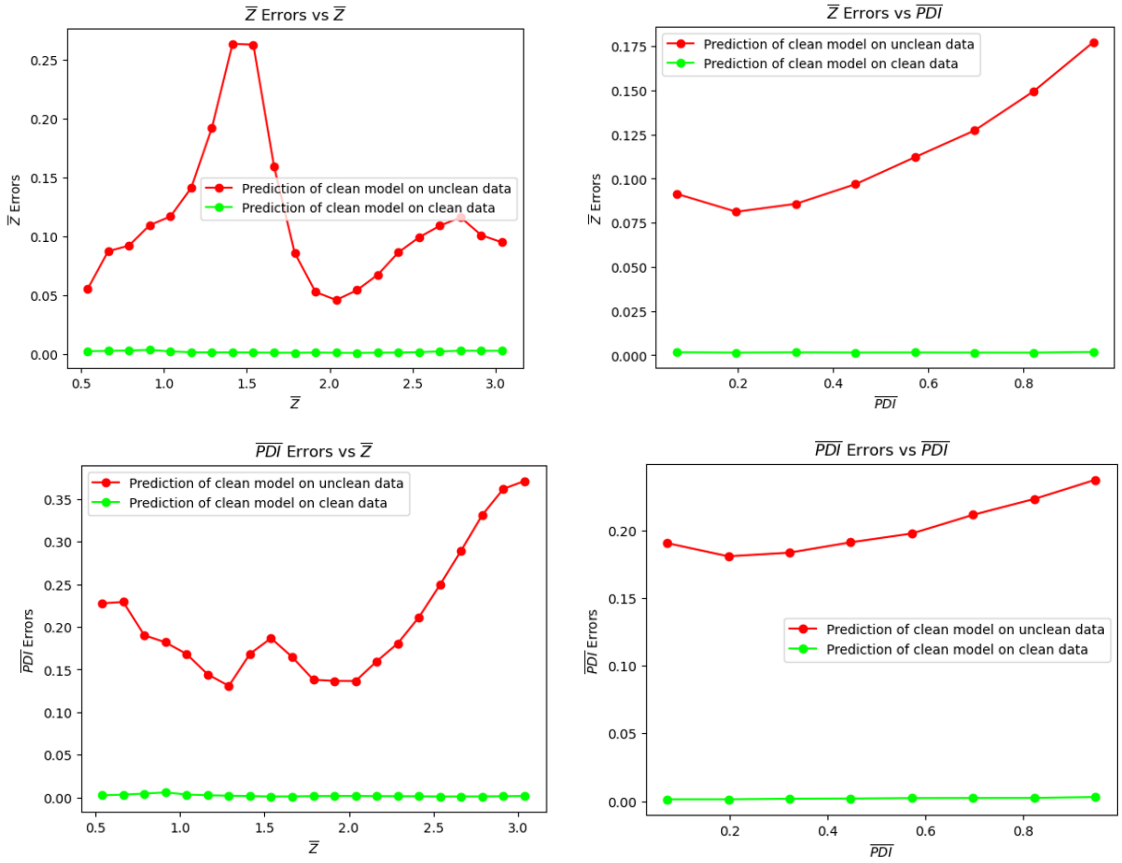

Figure 4.11: Plots of $MAE_{\overline{Z}}$ and $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) of the LSTM model trained on the original dataset and predicting on artificial error dataset (red curve) and trained on the original dataset and predicting on the original dataset (green curve)

From Figure 4.11, we observe that there is a clear difference in the error curves. The clean model is definitely not suitable to predict $\overline{Z}$ and $\overline{PDI}$ values from unclean data. This is to be expected however, as since the original dataset follows clear trends which the RNN model has picked up on. Deviation from these trends leads to clear errors in predictions.

$$MAE_{\overline{Z}} = 0.12 \tag{4.48}$$

$$MAE_{\overline{PDI}} = 0.21 \tag{4.49}$$

These values are clearly above the AE of 0.06 and hence we can say that the model is not suitably robust.

The second scenario involves the comparison of MAE values on predicting clean data, when the model is trained on unclean data.

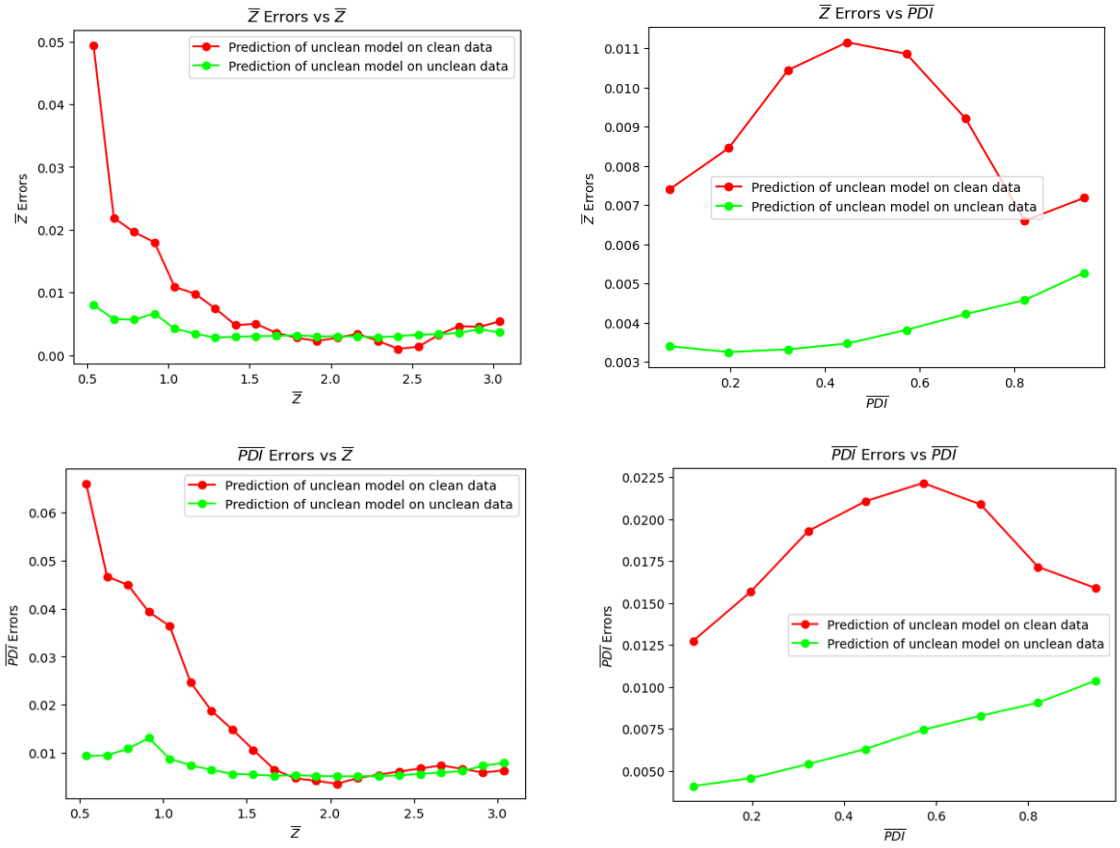

Figure 4.12: Plots of $MAE_{\overline{Z}}$ and $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) of the LSTM model trained on the artificial error dataset and predicting on the original dataset (red curve) and trained on the artificial error dataset and predicting on the artificial error dataset (green curve)

The graphs shown in Figure 4.12 are indeed very interesting. The $\overline{Z}$ errors of the unclean model on the clean data follow similar trends to that of $\overline{Z}$ with artificial errors induced. However, when $\overline{Z}$ is greater than 1.75, the model trained on unclean data predicts clean data better than

the model trained on clean data in certain scenarios. Once again, it is clear that the higher values of $\overline{Z}$ and $\overline{PDI}$ are least affected by artificially induced errors.

$$MAE_{\overline{Z}} = 10^{-2} \tag{4.50}$$

$$MAE_{\overline{PDI}} = 1.75 \times 10^{-2} \tag{4.51}$$

These values are clearly below the AE of 0.06 and hence we can say that the model is suitably robust.

4.7 Evaluation of Spectral Components

As discussed earlier in the introductory section pertaining to rheology, there are 4 distinct sections to flow curves, which are differentiated by the intersection of $\overline{G'}$ and $\overline{G''}$ plots.

The 4 sections are namely the terminal region, the rubbery plateau region, the transition region, and the glassy region. In the flow curves given in the dataset, there exist curves with only the first 3 regions (i.e. the terminal region, the rubbery plateau region, and the transition region). Now, there are characteristics of each of the regions which contribute to the prediction of $\overline{Z}$ and $\overline{PDI}$. In this section, we aim to definitively deduce which sections of the flow curve, on average, contribute to the prediction of which variable. This is done by firstly splitting the curves into their 3 parts. Upon doing so the sections of each curve are seperately modelled and the MAE graphs are drawn to see which section of which curve predicts a certain region of a target variable the best.

When splitting the curves, it is obvious that some curves will not have a transition region, or sometimes even a plateau region. To mitigate this restriction as much as possible, the points of intersection in every curve are averaged out to decide the points at which the curves are split. So, every curve in the dataset is evenly split whether or not their next region starts at that point, for ease of modeling.

The low frequency range lies in the range $\overline{\omega} = [-4, 2.92]$ (94 explanatory variables ($\overline{G'}$ and $\overline{G''}$)).

The mid frequency range lies in the range $\overline{\omega} = [3.07, 8.49]$ (74 explanatory variables ($\overline{G'}$ and $\overline{G''}$)).

The high frequency range lies in the range $\overline{\omega} = [8.49, 10.14]$ (22 explanatory variables ($\overline{G'}$ and $\overline{G''}$)).

Analysis of MAE trends :

The graphs shown in Figure 4.13 are very clear in expressing that the high frequency information of the flow curves do not contain as valuable information as do the low and mid frequency region. In $\overline{Z}$ values between 0.5 and 1 however, even the high frequency information gives a comparable accuracy to the remaining curves. However, this is only because at very low

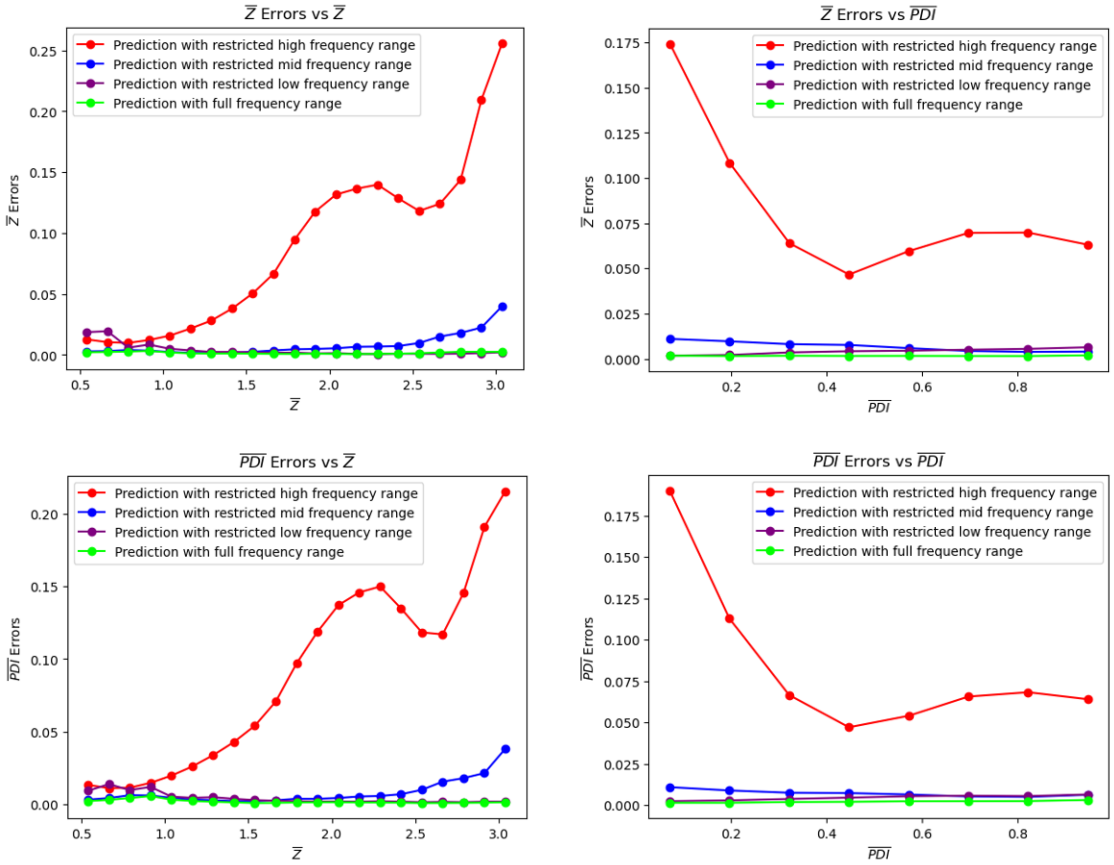

Figure 4.13: Plots of $MAE_{\overline{Z}}$ and $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) of the LSTM model trained on a dataset restricted to different sets of frequencies and tested on the same dataset

values of $\overline{Z}$, the high frequency information from the flow curve is normally found at low or mid frequency ranges at other values of $\overline{Z}$.

More importantly however is the fact that certain sections of the purple or blue lines coincide and even have lower errors than that of the green line which are predictions made by the model trained on the full frequency range of data. In general, the low frequency range seems to have enough information to predict $\overline{Z}$, but, at high values of $\overline{Z}$, both the terminal and plateau region of the flow curves exist in this frequency range resulting in low MAE similar to that of the full frequency spectrum. At very low values of $\overline{Z}$, the mid frequency range consists of only the point of intersection between $\overline{G'}$ and $\overline{G''}$ and the slope of $\overline{G'}$ and $\overline{G''}$. It seems as if this information is enough for the model to predict low values of $\overline{Z}$ while the rest of the spectrum ends up being noise. A similar explanation is the case for the better prediction of $\overline{PDI}$ values as well.

Figure 4.14 shows the change in error trends if artificial errors were added to the dataset. When comparing these graphs to the graphs in Figure 4.13, it seems that data trained on higher frequencies (i.e. the red and blue curves) have a greater increase in percentage error than that of data trained on low frequencies (i.e. violet curve).

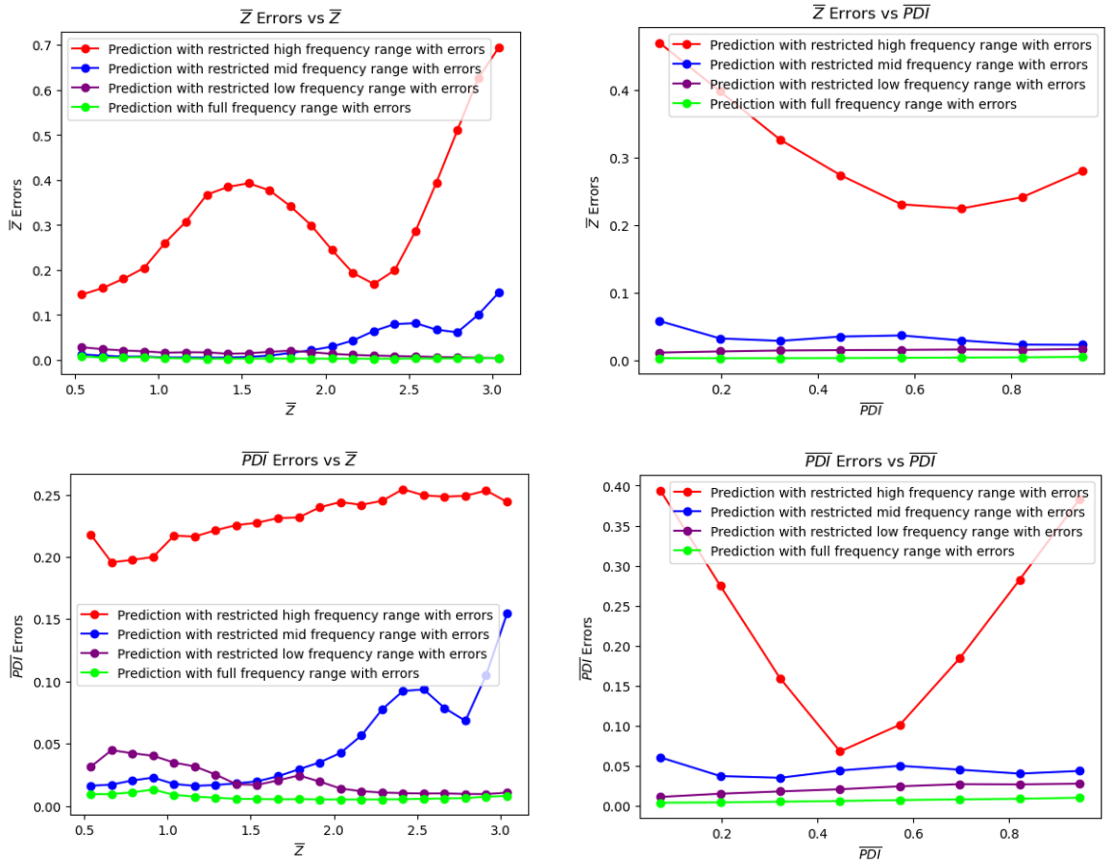

Figure 4.14: Plots of $MAE_{\overline{Z}}$ and $MAE_{\overline{PDI}}$ vs $\overline{Z}$ (left) and $\overline{PDI}$ (right) of the LSTM model trained on a dataset restricted to different sets of frequencies with artificial errors and tested on the same dataset

Chapter 5

Prediction on Bimodal Data

5.1 Model Development

Upon reading the csv file : `bi_PE_5param_wint_2_400000_training.csv`, the csv is first checked to see if the datatype along all rows in each column is consistent, that is, of numerical datatype. The second check is to see if there are any missing values in the dataset. Both these checks return in favor of a clean dataset that is ready to be analyzed.

The distribution of the 5 target variables are checked. They are all uniformly distributed in the 5D space.

Similar to the modeling of the unimodal dataset, we shall aim to predict the target parameters at a MAE no more than 0.06 in the log space which translates to a percentage error of 15% in the linear space (from 4.9).

The RNN architecture consists of 4 LSTM layers followed by 4 fully connected layers. This seems to be the perfect middle-ground between striving for high accuracy and committing an adequate amount of time to the model training. The per-layer architecture is shown below :

```
model = Sequential()
model.add(LSTM(128, input_shape=(1,190),return_sequences=True))
model.add(LSTM(64,return_sequences=True))
model.add(LSTM(64,return_sequences=True))
model.add(LSTM(32))
model.add(Dense(32, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(5, activation='linear'))
```

Other hyper-parameters that have been tuned for this specific model training are :

- **Loss function :** The loss function chosen is the `mean_squared_error` which adequately monitors how well-trained a model is on numerical data.
- **Learning rate :** This value is optimally set at 10^{-4} as at lower values the model does not

pick up on significantly more information. At higher values, the loss function reaches a local minima and is unable to find the global minima.

- **Optimizer :** `Adam`
- **Batch size :** Set to an optimal value of 32. A batch size lower than 32 tends to overfit the training data and takes too long to finish training.
- **Number of epochs :** Set to an optimal value of 100, as the training is not expected to require more than set amount of epochs.
- **Patience :** The package `EarlyStopping` from `tensorflow.keras.callbacks` is used to monitor the validation loss, in our case, the validation mean squared error. When the lowest amount of validation loss is achieved, it waits for 10 epochs to see if the RNN model can learn anything further, that is, achieve a lower validation loss. In this case, the lowest recorded validation loss is set to this new value and Early Stopping waits for another 10 epochs of training. If, however, if the model is trained for 10 more epochs and a lower validation loss is not obtained, this function resets the model weights to that which gave the lowest validation score and terminates the model training.

  The number of epochs `EarlyStopping` must wait for to receive a lower validation loss before terminating the model training is called the patience parameter. This can be any value determined by the user. In this case, it is set to 10.
- **Train:test ratio :** Set to a value of $80:20$ as it is the recommended standard for modeling. All results shown with regard to a certain dataset are trained and tested on the exact same values with a `random_state = 42` under `train_test_split()`.

5.2 Prediction on Synthetic Bimodal Dataset

On a set of uniformly distributed test sets, the RNN model is trained and gives a loss of 3×10^{-3} on the test set, while having an accuracy score of 99.83%. In order to gain a better understanding of the accuracy of the predictions, the MAE (from 4.1) and $PMAE$ (from 4.2) of each target variable is calculated. Moreover, the highest and lowest errors are also calculated for each variable to gain an understanding of the spread of errors.

Ideally, all the predicted values in the test set do not have an error greater than the AE of 0.06. However, if this happens to be the case, the percentage of values predicted with an error greater than AE (given by the variable $PUAE$ (Percentage of UnAcceptable Errors)) are also calculated. All the above values are tabulated and represented in table 5.1 :

From the table we observe that all target variables have an e_{max} value greater than AE. Further analysis is required in the error trends of individual target variables in order to confirm whether this model is ready for use in the industry.

	$\overline{Z_s}$	$\overline{PDI_s}$	$\overline{Z_l}$	$\overline{PDI_l}$	$\overline{\phi_l}$
MAE	0.036	0.027	0.026	0.021	0.039
$PMAE$	1.63	2.79	1.19	2.18	4.33
e_{min}	3.67×10^{-7}	1.31×10^{-7}	3.25×10^{-7}	4.69×10^{-8}	2.21×10^{-7}
e_{max}	0.72	0.66	0.46	0.45	0.87
$PUAE$	16.56%	11.18%	11.65%	7.48%	17.19%

Table 5.1: Model performance statistics on Synthetic Bimodal Dataset

For a target variable y_i, the $MAE_{\overline{y_i}}$ vs y_j is plotted for all combinations of y at intervals of 0.125.

Analysis of MAE trends :

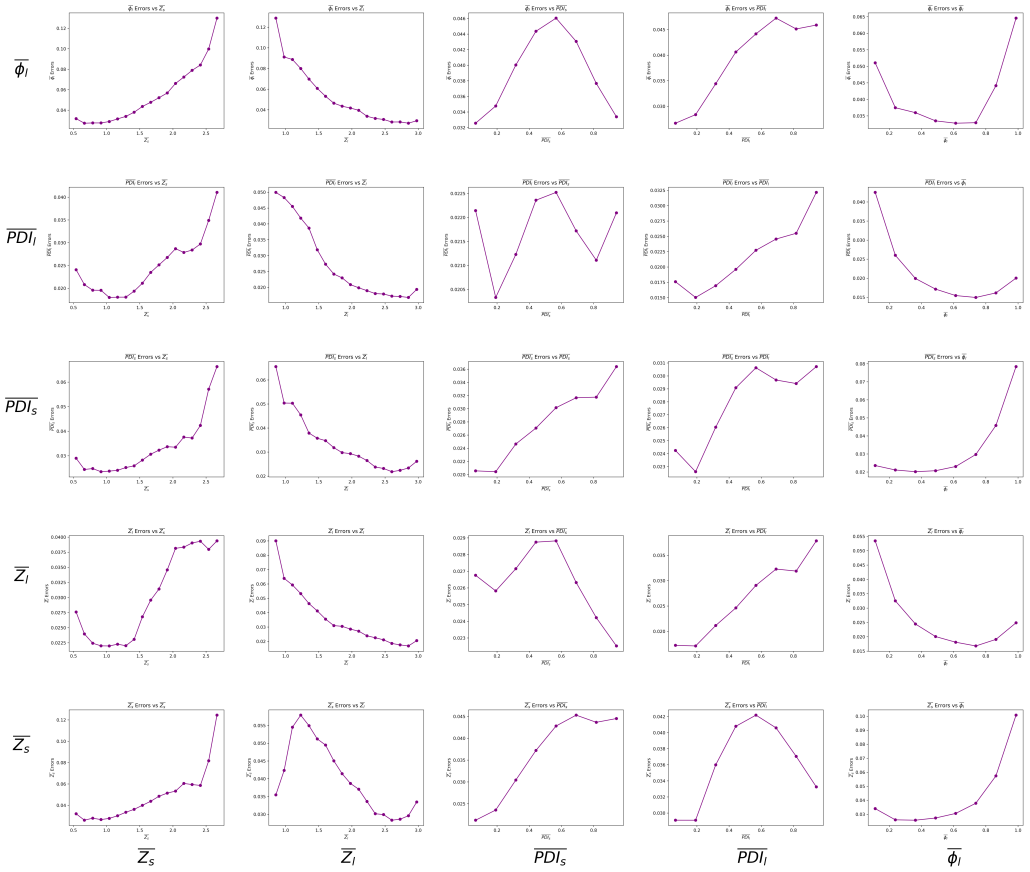

Figure 5.1: MAE trends of all target variables against each other using the LSTM model trained on the Synthetic Bimodal Dataset

From Figure 5.1, we observe that the MAE trends of other target variables against $\overline{\phi_l}$ (5^{th} column) are consistent as when $\overline{\phi_l}$ is lesser than 0.5, the mode S covers more area than mode L,

hence allowing for more accurate predictions of $\overline{Z_s}$ and $\overline{PDI_s}$. When $\overline{\phi_l}$ is greater than 0.5, the inverse is true.

In general, as discussed in the preliminary analysis of Bimodal Data, mode S does not impact the flow curves as much as mode L. Hence, the MAEs of $\overline{Z_l}$ and $\overline{PDI_l}$ are smaller than $\overline{Z_s}$ and $\overline{PDI_s}$ respectively.

Some obvious trends from observing the graphs are that the error trend of every target increases with increase in $\overline{Z_s}$ and decreases with increase in $\overline{Z_l}$. This is because of the restriction that when $\overline{Z_s}$ takes up higher values, $\overline{Z_l}$ will also have to be high as $\overline{Z_l}$ has to be a minimum of $2 \times \overline{Z_s}$. From Figure 2.5, we see why low values of $\overline{PDI_l}$ always give low errors. The trends of $\overline{PDI_s}$ and $\overline{PDI_l}$ errors against themselves are similar to the trends observed in the unimodal case. When observing trends between $\overline{Z_l}$ errors vs $\overline{PDI_s}$ and $\overline{Z_s}$ errors vs $\overline{PDI_l}$, we observe that the trend is like that of an inverted 'U' due to the flow curves being distinctly prominent at very low and very high PDI values as shown in the 1^{st} and 5^{th} columns of Figures 2.3 and 2.4 respectively.

Conclusion :

In its raw form, this RNN model cannot be used to model real life plastics following the trends as given in this dataset. However, considering some of the error trends and factoring some known quantities such as low $\overline{PDI_l}$, low $\overline{PDI_s}$, low $\overline{Z_s}$, high $\overline{Z_l}$ or $\overline{\phi_l}$ values close to 0.5, might make this model viable for real life usage.

5.3 Effects of Restricted Frequency Range

We now observe the effects of a limited $\overline{\omega}$ value of -1 to 2 and whether it is feasible to predict the target variables with an error lower than AE. On the test set, the model gives a loss of 0.026 and an accuracy of 99.81%.

	$\overline{Z_s}$	$\overline{PDI_s}$	$\overline{Z_l}$	$\overline{PDI_l}$	$\overline{\phi_l}$
MAE	0.13	0.20	0.06	0.05	0.09
$PMAE$	5.99	20.7	2.88	5.84	10.70
e_{min}	2.21×10^{-7}	8.55×10^{-7}	6.18×10^{-7}	4.49×10^{-8}	8.99×10^{-9}
e_{max}	1.35	0.80	0.64	0.88	0.84
$PUAE$	66.75%	81.54%	36.03%	30.85%	47.55%

Table 5.2: Model performance statistics on Frequency Restricted Bimodal Dataset

From Table 5.2, we observe even the MAE values of $\overline{Z_s}$, $\overline{PDI_s}$ and $\overline{\phi_l}$ being greater than AE. Furthermore, the $PUAE$ values of all variables have gained a significant increase. It is very unlikely that such data can be reliably used for real life applications.

Analysis of MAE trends :

Even when the frequency range is restricted, the MAE trends (Figure 5.2) of $\overline{\phi_l}$ still remain

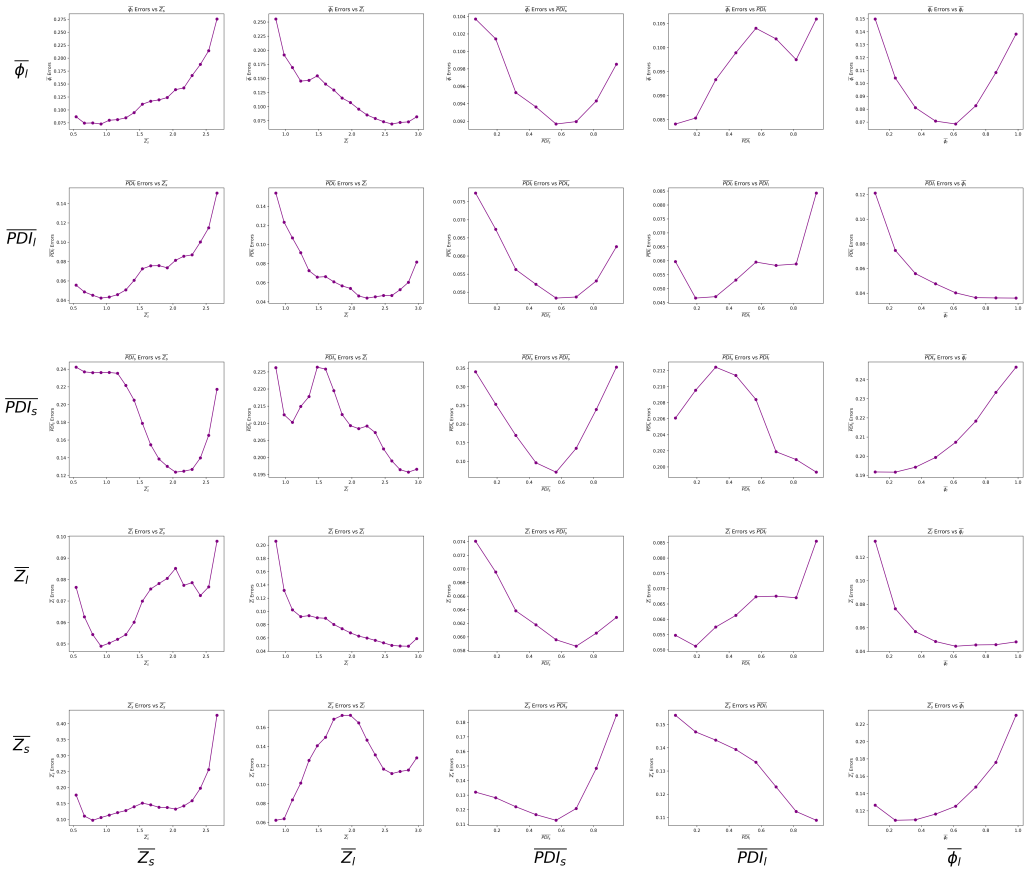

Figure 5.2: MAE trends of all target variables against each other using the LSTM model trained on the Frequency Restricted Bimodal Dataset

the same. However, this is not the case for other target variables which rely on the very low and high frequency data for accurate predictions.

One such case is that of $\overline{Z_l}$ errors vs $\overline{PDI_s}$ and $\overline{Z_s}$ errors vs $\overline{PDI_l}$. When the model had the information over the whole spectrum of data, the MAE trend was that of an inverted 'U'. Now it is almost a straight line of negative slope. Both $\overline{Z_l}$ and $\overline{Z_s}$ rely greatly on the flow curve intersections at the high frequency range. Another noticeable change of the MAE trends is that in $\overline{PDI_s}$ errors vs $\overline{Z_s}$. At low values of $\overline{Z_s}$, the low range of $\overline{PDI_s}$ values rely on the flow curves' rubbery plateau region for their accurate prediction.

5.4 Effects of Induced Artificial Errors

Training a model on a dataset of induced artificial errors of ± 0.04 element wise (human error) and ± 0.04 row wise (rheometer error) gives a test loss of 0.0125 and a test accuracy of 99.81%. The table below reveals other information with regard to model performance :

	$\overline{Z_s}$	$\overline{PDI_s}$	$\overline{Z_l}$	$\overline{PDI_l}$	$\overline{\phi_l}$
MAE	0.09	0.09	0.05	0.04	0.07
$PMAE$	4.29	9.63	2.13	3.68	8.30
e_{min}	2.09×10^{-6}	1.58×10^{-6}	6.10×10^{-7}	2.10×10^{-7}	6.45×10^{-7}
e_{max}	1.36	0.80	0.64	0.84	0.88
$PUAE$	47.35%	53.93%	25.03%	18.28%	36.49%

Table 5.3: Model performance statistics on Artificial Error Bimodal Dataset

Overall, the artificial errors do not seem to affect the model performance as much restricting the frequency range did (Table 5.3). This highlights the importance of information given by higher frequencies than that in the case of unimodal data where artificial errors affected the model performance much more than the restriction of frequency. However, this does not mean that this model can be used for industrial classification of plastics as the MAE values of $\overline{Z_s}$, $\overline{PDI_s}$ and $\overline{\phi_l}$ are greater than AE; albeit to a lesser degree than the previous case of restricted frequencies.

Analysis of MAE trends :

From the graphs in Figure 5.3, we observe that most MAE trends are the same as that of the original dataset. This is to be expected as a model predicting 5 parameters over 2 would definitely be more resistant to noise. The trends which do change drastically due to noise are trends which depend on small nuances in the flow curves for their prediction. $\overline{PDI_s}$ errors vs $\overline{Z_l}$ is one such case. At the 25^{th} and 50^{th} percentile values of $\overline{Z_l}$ in Figure 2.3, artificial errors would make it hard to differentiate between the rubbery plateau region of those flow curves.

5.5 Prediction on Pseudo-Realistic Bimodal Dataset

Pseudo-Realistic data is data that is artificially changed to emulate real life data. Real life data in our case comprises of a restriction in the frequency range ($-1 \leq \overline{\omega} \leq 2$) and also a 20% error in measurement (random and rheometer based). The results shown below indicate the performance of the model on such data.

The model when predicting on uniformly distributed test data gives a test loss of 0.04 and a test accuracy of 99.81%. The MAE, $PMAE$, e_{max} and $PUAE$ for each of the target variables are as follows :

It is unsurprising that out of all 4 datasets, the MAE values are the highest when modeling this pseudo-realistic dataset as it contains both errors and frequency restrictions making it difficult for the RNN model to accurate predict the 5 target parameters (Table 5.4). The high $PUAE$ values also echo the same consensus that the model is not ready for industrial deployment.

Analysis of MAE trends :

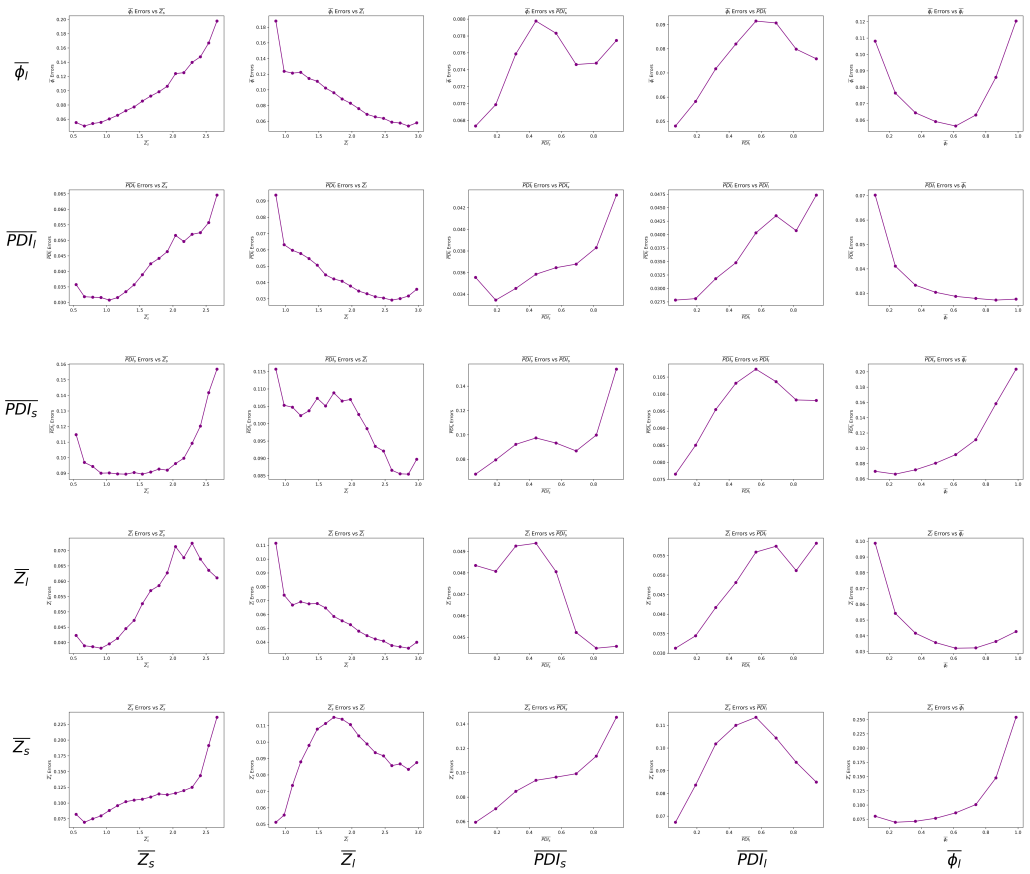

Figure 5.3: MAE trends of all target variables against each other using the LSTM model trained on the Artificial Error Bimodal Dataset

	$\overline{Z_s}$	$\overline{PDI_s}$	$\overline{Z_l}$	$\overline{PDI_l}$	$\overline{\phi_l}$
MAE	0.22	0.22	0.07	0.07	0.12
$PMAE$	9.93	21.9	3.11	6.83	12.99
e_{min}	3.27×10^{-6}	1.41×10^{-7}	1.57×10^{-6}	1.65×10^{-7}	9.80×10^{-8}
e_{max}	1.78	0.82	0.67	0.88	0.83
$PUAE$	77.26%	83.68%	39.10%	40.13%	59.45%

Table 5.4: Model performance statistics on Pseudo Realistic Bimodal Dataset

From the graphs in Figure 5.4, we notice a similarity between these graphs and those in Figures 5.2 and 5.3. The graphs in Figure 5.4 follow most of the MAE trends of the frequency restricted dataset (Figure 5.2) while also following some of the dissimilarities between the error trends of the original dataset (Figure 5.1) and that of the artificially induced errors (Figure 5.2). All this takes place with an obvious up scale in MAE values.

If $Graph_O$ are the MAE trend graphs from predictions of the original dataset, $Graph_R$ are the MAE trend graphs from predictions of the frequency restricted dataset, $Graph_A$ are

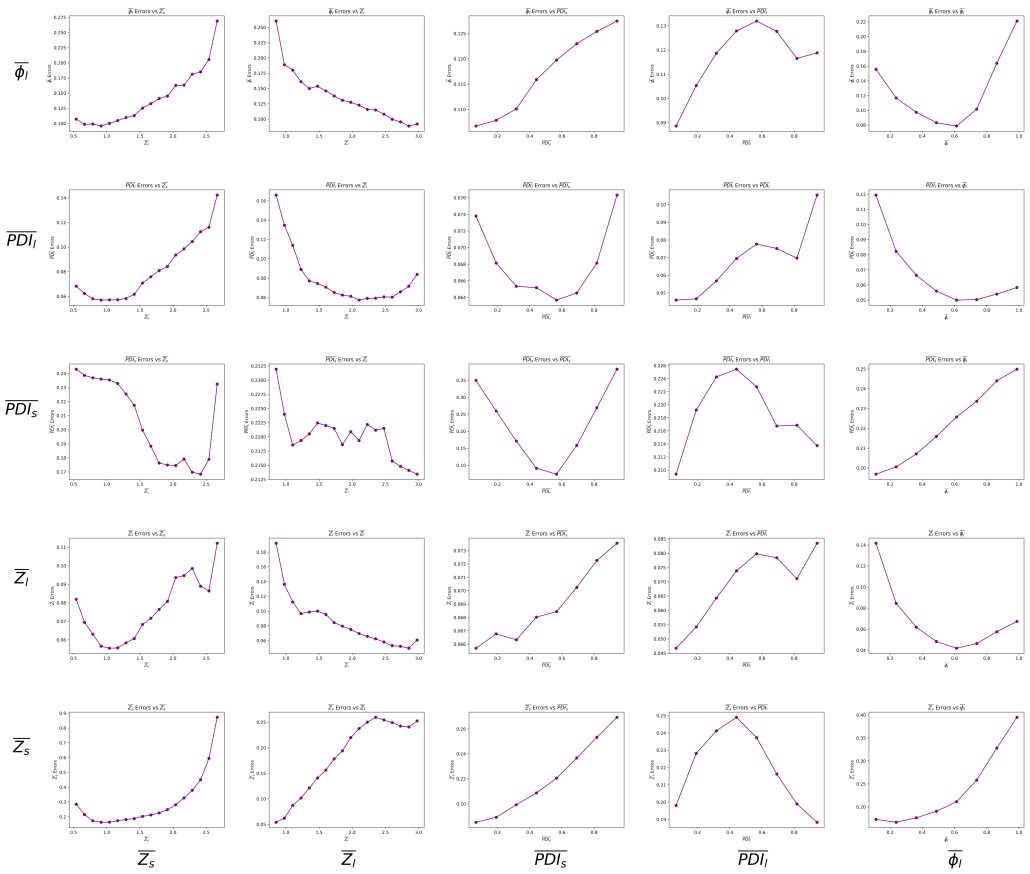

Figure 5.4: MAE trends of all target variables against each other using the LSTM model trained on the Pseudo Realistic Bimodal Dataset

the MAE trend graphs from predictions of the artificial error induced dataset and $Graph_{PR}$ are the MAE trend graphs from predictions of the pseudo-realistic dataset, $Graph_{PR}$ can be represented simply as :

$$Graph_{PR} = f \times (Graph_R + Graph_A - Graph_O) \tag{5.1}$$

where f is a scaling factor greater than 1, which scales the magnitude of the graphs.

The only exception to this rule is the graph $\overline{Z_l}$ errors vs $\overline{PDI_s}$, in which both $Graph_R$ and $Graph_A$ are downward facing curves but $Graph_{PR}$ has an upward trajectory. In this case, the combination of the restriction of frequency and artificial errors seem to have a negative impact on the factors of the flow curve using which $\overline{Z_l}$ is predicted.

Chapter 6

Conclusions and Future Work

In conclusion, yes, a deep learning approach is viable for the prediction of MWDs of linear polymers. In this way, polymer plastics can be rapidly characterized into polymer types and be put to use in the appropriate applications swiftly and effectively.

6.1 Unimodal Data

The trends of the resultant accuracies from all the predictions are plotted in the graph below (Figure 6.1):

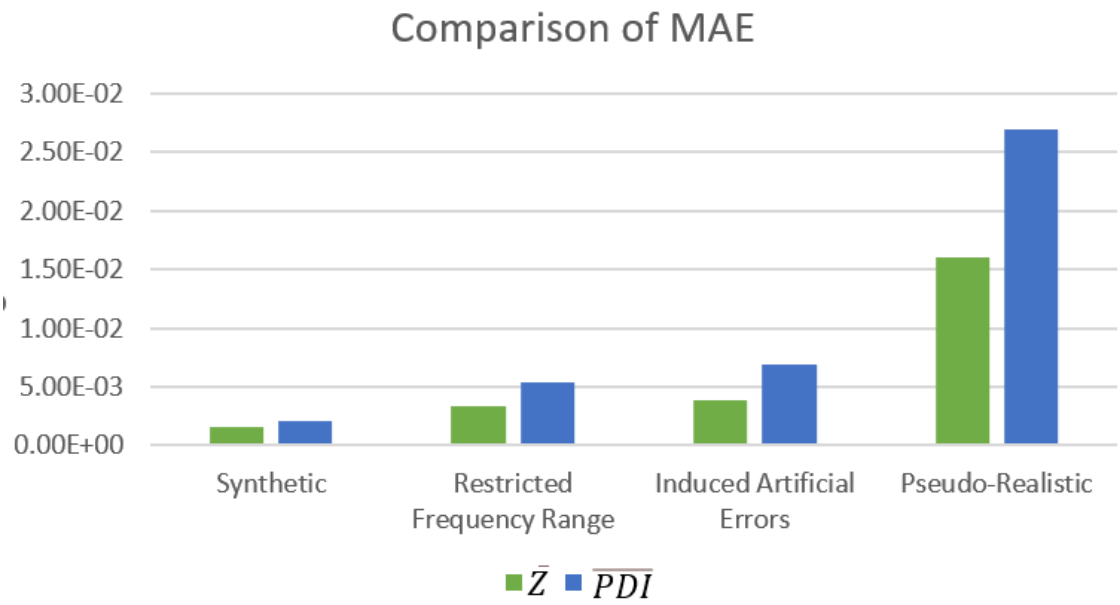

Figure 6.1: Comparison of MAE (Unimodal data)

In the case of unimodal polymer plastics, we can infer from the graph that the reduction in the number of frequencies measured and artificial noise do not hinder the model to the point where the MAE exceeds the AE of 0.06. Although there were some errors which exceeded

the AE in predicting pseudo real life data, it is only a small percentage and is still viable to be used in industrial categorization of plastics. As for the effect of restriction in the number of frequencies measured and artificial noise, Table 6.1 describes the range (Low, Mid or High) at which the prediction of each target variable had it's error trends least offset to to the original error trends. Essentially, target variables at these ranges are least susceptible to increased error even in the midst of manipulated data.

	Restricted Frequency Range	Induced Artificial Errors	Pseudo-Realistic
$\overline{Z}$	Mid	High	High
$\overline{PDI}$	Low	Low	Low

Table 6.1: Ranges at which Unimodal target variables are least susceptible to error changes

We also concluded that the RNN model trained on clean data is not robust enough to predict accurately on unclean data, however, in values of $\overline{Z}$ ranging from 1.5 to 3, the RNN model trained on unclean data is robust and can predict as accurately on clean data as an RNN model trained on clean data can. Considering that the MAE value of this RNN model trained on unclean data is still lower than AE (from 4.9), it might be better to use this model in industrial categorization, where there is no prior information on the incoming plastic samples.

Spectral analysis revealed that models trained on some frequency ranges alone can predict target variables in certain ranges better than models trained on the full frequency range of data.

6.2 Bimodal Data

The trends of the resultant accuracies from all the predictions are plotted in the graph below (Figure 6.2):

In the case of bimodal polymer plastics, we can infer from the graph that even the synthetic data has MAE values exceeding the AE of 0.06, making it not ideal to use in industrial polymer characterization (from 4.9). However, there are clear cut trends which have been uncovered through the analysis of error trends which would help in the future work of modeling this data. The trends are that if any target variable predicts with low error at a certain range, every other target variable will also predict with low error if it is also in that range. From the error graphs, we see that low $\overline{PDI_l}$, low $\overline{PDI_s}$, low $\overline{Z_s}$, high $\overline{Z_l}$ and $\overline{\phi_l}$ values close to 0.5 predict with low errors. So, for example, if a flow curve is known to have a $\overline{\phi_l}$ value close to 0.5 (middle range of $\overline{\phi_l}$ values) , any other target variable will also be predicted with low error if it is in it's middle range.

Implementing these findings into modeling using feature engineering (Duboue 2020) is very likely to yield fruitful results and a great baseline for future work. Another line of thinking which could be explored would be to try and use CNNs as well in the model architecture, making it a CNN-RNN hybrid architecture (Hsu et al. 2017) (Wang et al. 2016) (Kazmi et al. 2023). This has not been explored in this dissertation due to time constraints in training the data for a suitable

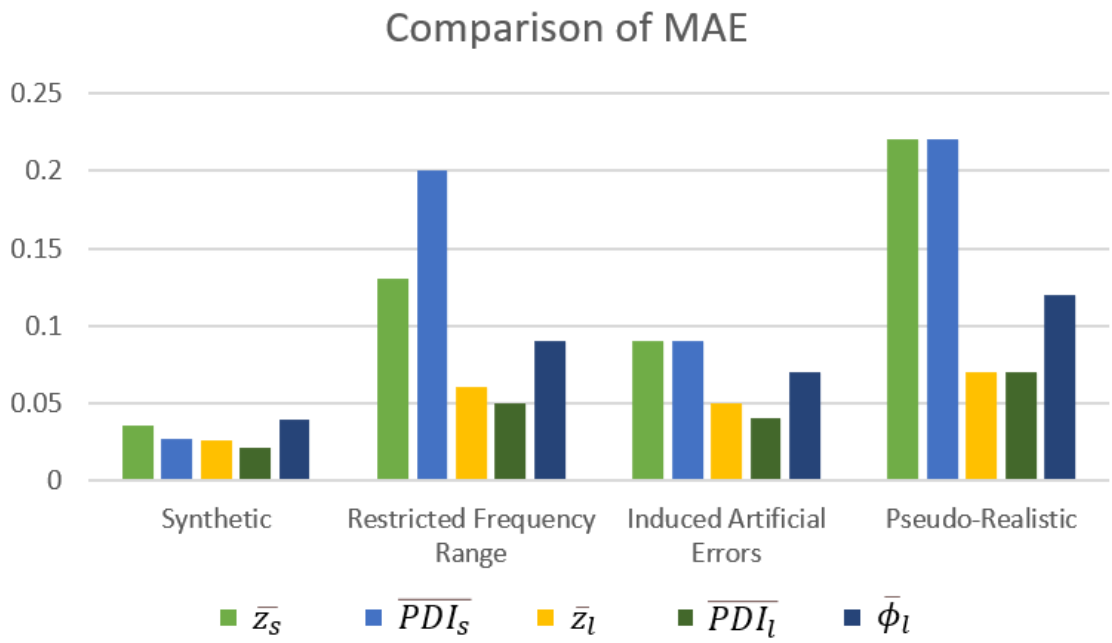

Figure 6.2: Comparison of MAE (Bimodal data)

number of epochs. Now, although it might be possible to use the model predicting data similar to the given synthetic data, it is definitely unsafe to use this model to characterize flow data with restricted frequencies and artificial errors. Table 6.2 describes the range (Low, Mid or High) at which the prediction of each target variable had it's error trends least offset to to the original error trends. Target variables at these ranges are least susceptible to increased error even in the midst of manipulated data.

	Restricted Frequency Range	Induced Artificial Errors	Pseudo-Realistic
$\overline{Z_s}$	Low	Low	Low
$\overline{PDI_s}$	High	High	High
$\overline{Z_l}$	High	High	High
$\overline{PDI_l}$	Low	Low	Low
$\overline{\phi_l}$	Mid	Mid	Mid

Table 6.2: Ranges at which Bimodal target variables are least susceptible to error changes

Other future work could involve attempting to find a deep learning solution to predict the parameters of a MWD of any distribution (Vandal et al. 2018) (Botella et al. 2018), not limited to just unimodal normal or bimodal normal. Some of the most common plastics are linear polymers, however, the classification of any plastic of any MWD would be an ideal to strive toward.

Bibliography

A Basic Introduction to Rheology (2016), *Malvern Instruments Limited* .
URL: *https://cdn.technologynetworks.com/TN/Resources/PDF/WP160620BasicIntroRheology.pdf*

Ahn, Y.-C. & Kim, H.-J. (2002), 'A study on the rheological properties and processability of polycarbonate', *Journal of Applied Polymer Science* **86**, 2921–2929.

Ali, S., Ordyniak, S. & Martinez, J. S. (2024), 'Comp5625m deep learning', *University of Leeds Module Resources* .
URL: *https://minerva.leeds.ac.uk/ultra/courses/*$_{5}54264_{1}/outline/file/_{1}1878286_{1}$

Botella, C., Joly, A., Bonnet, P., Monestiez, P. & Munoz, F. (2018), A deep learning approach to species distribution modelling, *in* 'Multimedia Tools and Applications for Environmental & Biodiversity Informatics'.

Clarke-Pringle, T. & Macgregor, J. F. (1998), 'Optimization of molecular-weight distribution using batch-to-batch adjustments', *Industrial & Engineering Chemistry Research* **37**, 3660–3669.

D., M. Z., Aranda, F. L. & Rivas, B. L. (2023), 'Polymers recycling: Upcycling techniques. an overview', *Journal of the Chilean Chemical Society* .

Das, C. & Read, D. J. (2023), 'A tube model for predicting the stress and dielectric relaxations of polydisperse linear polymers', *Journal of Rheology* **67**(3), 693–721.

Domone, P. & Illston, J. M. (2017), 'Liquids, viscoelasticity and gels', *Construction Materials* .

Duboue, P. (2020), The art of feature engineering.

Fetters, L. J., Lohse, A., Milner, S. T. & Graessley, W. W. (1999), 'Packing length influence in linear polymer melts on the entanglement, critical, and reptation molecular weights', *Macromolecules* **32**, 6847–6851.

Garcês, A. & Pires, I. (2024), 'The detrimental impacts of plastic pollution on wildlife', *Research in Ecology* .

Hsu, S. T., Moon, C., Jones, P. & Samatova, N. F. (2017), A hybrid cnn-rnn alignment model for phrase-aware sentence classification, *in* 'Conference of the European Chapter of the Association for Computational Linguistics'.

Joseph, F. J. J., Nonsiri, S. & Monsakul, A. (2021), 'Keras and tensorflow: A hands-on experience', *Advanced Deep Learning for Engineers and Scientists* .

Józefowicz, R., Zaremba, W. & Sutskever, I. (2015), An empirical exploration of recurrent network architectures, *in* 'International Conference on Machine Learning'.

Kazmi, S., Görgülü, B., Cevik, M. & Baydogan, M. G. (2023), 'A concurrent cnn-rnn approach for multi-step wind power forecasting', *ArXiv* **abs/2301.00819**.

Ljubic, D., Stamenović, M., Smithson, C. S., Nujkić, M., Medjo, B. & Putic, S. (2014), Time - temperature superposition principle - application of wlf equation in polymer analysis and composites.

Mead, D. W., Monjezi, S. & Park, J. (2018), 'A constitutive model for entangled polydisperse linear flexible polymers with entanglement dynamics and a configuration dependent friction coefficient. part i: Model derivation', *Journal of Rheology* **62**, 121–134.

Ohno, K. & Kumagai, A. (2021), 'Recurrent neural networks for learning long-term temporal dependencies with reanalysis of time scale representation', *2021 IEEE International Conference on Big Knowledge (ICBK)* pp. 182–189.

Rao, K. P., Doraivelu, S. M. & Gopinathan, V. (1982), 'Flow curves and deformation of materials at different temperatures and strain rates', *Journal of Mechanical Working Technology* **6**, 63–88.

Read, D. J. (2015), 'From reactor to rheology in industrial polymers', *Journal of Polymer Science Part B* **53**, 123–141.

Ruj, B., Pandey, V., Jash, P. & Srivastava, V. (2015), 'Sorting of plastic waste for effective recycling', *International Journal of Applied Science and Engineering Research* **4**, 564–571.

Singh, M. K. & Singh, A. (2022), 'Thermal characterization of materials using differential scanning calorimeter', *Characterization of Polymers and Fibres* .

Sun, Y. & Sahinidis, N. (2021), Design of polymer configuration and flow.

Tan, R., Zhou, D., Liu, B., Sun, Y., Liu, X., Ma, Z., Kong, D., He, J., Zhang, Z. & hui Dong, X. (2019), 'Precise modulation of molecular weight distribution for structural engineering', *Chemical Science* **10**, 10698 – 10705.

Vandal, T. J., Kodra, E., Dy, J. G., Ganguly, S., Nemani, R. R. & Ganguly, A. R. (2018), 'Quantifying uncertainty in discrete-continuous and skewed data with bayesian deep learning', *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* .

vec, F. (2024), 'Size exclusion chromatography has been around for sixty years', *Chemické listy* .

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. & Xu, W. (2016), 'Cnn-rnn: A unified framework for multi-label image classification', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 2285–2294.