UNIVERSITÀ
DI SIENA
1240

Department of
Information Engineering and Mathematics

M. Sc. Program
ELECTRONICS AND COMMUNICATIONS ENGINEERING

Fantastic and Extraordinary Title of Your Incredibly Long and Beautiful Thesis

Supervisor:
Prof. Name Surname

Adjunt Supervisors:
Prof. Name Surname
Dr. Name Surname

Candidate:
Name Surname
Mat. 002234

Academic Year 2021-2022

Contents

Introduction

The contamination of aquatic ecosystems by pharmaceutical compounds has become an urgent environmental challenge. Among the most promising remediation strategies is the use of specialized polymer sponges capable of adsorbing these pollutants from wastewater. However, discovering the optimal polymer for a specific drug molecule is a complex and time–consuming process, traditionally driven by trial–and–error laboratory experiments. In recent years, computational methods and machine learning (ML) have emerged as powerful tools to accelerate the discovery of new materials [1]. This thesis explores a dual–track machine learning approach to address this challenge, focusing on both the generation of novel molecular structures and the prediction of polymer–molecule adsorption efficiency.

In line with the methodological priorities of this research, the thesis begins by introducing the fundamental machine learning architectures employed. First, we examine the Multi–Layer Perceptron (MLP), a robust feedforward neural network used for predictive tasks. Next, we present minGPT, a lightweight implementation of the Generative Pre–trained Transformer architecture, which serves as the backbone for generative modeling. The primary objective of deploying these models is to explore the complex chemical space of polymer–molecule interactions, using algorithmic pattern recognition to bridge the gap between chemical structures and physical properties.

To anchor these machine learning models in chemical reality, substantial effort

was devoted to data collection and feature engineering. A comprehensive dataset was assembled by systematically extracting information from scientific literature on the use of polymers for the adsorption of molecules in wastewater. The key variables collected include the solution pH, the initial concentration of the pharmaceutical (mg/L), and the adsorption capacity (mg/g), defined as the amount of pollutant adsorbed per unit mass of polymer. Given the well–known scarcity of data describing these specific interaction triplets, the dataset was further augmented through interpolation techniques. In addition, to convert chemical information into a machine–readable representation, computational chemistry libraries — specifically RDKit, Polymetrix, and tblite — were employed to derive a rich set of features from the raw data.

The experimental phase of this work is structured around two main contributions. The first consists of a generative pipeline based on a custom wrapper of minGPT, designed to produce synthetic yet chemically valid polymer SMILES (Simplified Molecular Input Line Entry System) strings, thereby enabling a theoretical expansion of the candidate chemical space. The second contribution addresses the predictive task: employing the extracted features and a Multi–Layer Perceptron (MLP) to estimate the adsorption capacity of a given polymer–molecule pair. To assess the model's ability to generalize under conditions of limited real–world data, extensive experiments were conducted using k–fold cross–validation in combination with systematic hyperparameter tuning.

Ultimately, this thesis evaluates the current potential of machine learning approaches for predicting environmental remediation metrics, highlighting both the promise of molecular generation techniques and the persistent challenges associated with data scarcity in predictive chemical modeling.

The remainder of this thesis is organized as follows: Chapter 1 introduces the chemical context and selected features. Chapter 2 details the mathematical foundations of the MLP and minGPT architectures. Chapter 3 outlines the data scraping,

interpolation, and feature extraction pipeline. Chapter 4 presents the generative modeling approach for P–SMILES. Chapter 5 discusses the predictive experiments, cross-validation results, and current limitations of the proposed approach. Finally, Chapter 6 provides concluding remarks and directions for future research.

Chapter 1

Problem Description

This chapter introduces the fundamental chemical concepts underlying the study, as well as the issue of drug disposal in wastewater, which constitutes the central focus of this thesis. Indeed, before introducing the machine learning architectures used for generative and predictive tasks, it is crucial to understand the physical problem being solved. We first discuss how chemical structures are represented for computational analysis, followed by an overview of the wastewater remediation process using polymer adsorbents. Finally, we define the key physicochemical features that form our dataset and review the state of the art in predicting polymer–molecule adsorption.

1.1 Representation of Molecules and Polymers — SMILES and P–SMILES

To apply machine learning algorithms to chemical problems, physical molecules and macroscopic polymer structures must first be translated into a machine–readable format. While humans traditionally rely on 2D structural diagrams to understand

chemical connectivity, computational models — particularly those based on Natural Language Processing (NLP) approaches — require data to be formatted as one–dimensional sequences or numerical arrays. The Simplified Molecular–Input Line–Entry System (SMILES) has become the de facto standard for this translation [2].

SMILES: Simplified Molecular–Input Line–Entry System

SMILES is a line notation method that represents the structure of a chemical species using short ASCII strings. Developed in the late 1980s, it allows for a highly compact and easily searchable representation of chemical graphs. The system is governed by a set of formal grammar rules:

- **Atoms** are represented by their standard atomic symbols (e.g., `C` for carbon, `O` for oxygen, `N` for nitrogen). Aromatic atoms are typically written in lowercase (e.g., `c`, `n`, `o`);
- **Bonds** are denoted by specific symbols, such as `=` for double bonds and `#` for triple bonds. Single bonds are generally implicitly assumed between adjacent atoms to save space;
- **Branches** in the molecular graph are enclosed in parentheses `()`;
- **Rings** are represented by breaking one bond in the ring and assigning a matching numerical digit to the two atoms where the ring was broken (e.g., Benzene is represented as `c1ccccc1`).

P–SMILES vs. BigSMILES: Adapting Line Notation for Polymers

While standard SMILES is highly effective for discrete, small molecules (such as the pharmaceutical pollutants targeted in this study), representing polymers presents a

fundamentally different challenge. Unlike small molecules, which have a fixed atomic composition and exact molecular weight, polymers are macromolecules composed of repeating structural units (monomers) and exhibit a stochastic distribution of chain lengths.

To represent polymers within a chemoinformatics framework, extensions of the SMILES syntax have been developed, most notably Polymer SMILES (P–SMILES) and BigSMILES [3]. P–SMILES captures the simple topological repeating unit using explicit attachment points — usually represented by asterisk symbols (`*`), dummy atoms (like `[At]`), or specific bracket enclosures. For example, the repeating unit of polyethylene, which consists of a simple $-CH_2-CH_2-$ backbone, can be concisely represented in P–SMILES as `*CC*`.

In contrast, BigSMILES provides a much more comprehensive, object–oriented grammar capable of describing stochastic polymers, diverse end–groups, and complex macromolecular ensembles. Theoretically, BigSMILES contains richer structural information and would be better suited for the advanced Machine Learning approaches utilized in this work. However, its adoption is currently hindered by severe data scarcity. Furthermore, many software tools designed to parse BigSMILES are either deprecated or no longer actively maintained. Consequently, in this thesis, we opted to use the simpler but much more robustly supported P–SMILES representation for polymer encoding.

Limitations of Line Notations

Despite their widespread utility, SMILES and P–SMILES representations possess inherent limitations. Primarily, they reduce three–dimensional molecular topologies into one–dimensional strings, implicitly discarding complex spatial information such as precise atomic coordinates, conformational folding, and stereochemical nuances.

Furthermore, because a SMILES string traces a path through a molecular graph, a single molecule can often be represented by multiple valid strings depending on the starting atom of the traversal. This necessitates the use of strict canonicalization algorithms to ensure consistency across datasets. For polymers specifically, the simplified P–SMILES representation struggles to capture macroscopic properties that drastically affect adsorption, such as cross–linking density, branching frequency, or specific block copolymer arrangements.

Line Notations for Generative and Predictive Modeling

The use of SMILES and P–SMILES is critical for the computational pipeline developed in this thesis. Nonetheless, because both representations reduce chemical topologies to 1D strings of text, they directly enable the use of Transformer–based architectures, such as minGPT. By treating the SMILES/P–SMILES grammar as a synthetic language, the minGPT model can learn the underlying sequence probabilities, which allows it to autoregressively generate new SMILES and P–SMILES strings. It is crucial to note, however, that the generative model simply outputs character sequences based on learned distributions; it does not inherently guarantee chemical validity. In practice, a significant portion of the raw generated strings are syntactically or chemically invalid. Therefore, the generated sequences must be subsequently parsed and validated using RDKit [4], making this post–generation filtering step essential. Furthermore, valid representations serve as the foundational input for feature extraction. Using tools like Polymetrix and tblite, the P–SMILES of the polymer and the SMILES of the target pollutant can be parsed to compute geometric, thermodynamic, and electronic descriptors. These descriptors subsequently form the feature vectors used by the Multi–Layer Perceptron (MLP) to predict the adsorption capacity of the polymer–molecule pair in wastewater applications.

1.2 Polymer–Molecule Interactions in Wastewater Treatment

The presence of pharmaceutical compounds in aquatic ecosystems poses a severe and increasing environmental threat. These molecules, which enter waterways through agricultural runoff, industrial discharge, and inadequate municipal wastewater treatment, are often highly resistant to natural degradation. Consequently, they accumulate in the environment, leading to adverse effects on aquatic life and potentially entering the human food chain.

To combat this problem, adsorption utilizing porous polymer networks — often referred to as polymer sponges or hydrogels — has emerged as a highly effective remediation strategy. Polymers are particularly well–suited for this task because their chemical backbones can be functionalized to attract specific classes of pollutants.

When a polymer is introduced into contaminated water, the pharmaceutical molecules interact with the polymer matrix through various intermolecular forces, including hydrogen bonding, electrostatic interactions, van der Waals forces, and $\pi - \pi$ stacking. The efficacy of this capture process is highly dependent on the complementary structural features of both the polymer repeating unit and the target molecule. Discovering the optimal polymer for a specific pollutant involves navigating a massive combinatorial space of chemical interactions, a task traditionally limited by the slow pace of trial–and–error laboratory synthesis.

1.3 Selected Features and Targets

To bridge physical chemistry and predictive machine learning, the adsorption process must be quantitatively characterized. In this thesis, the dataset used to train the MLP collects three fundamental features extracted from the scraped literature.

pH of the Solution

The pH of the wastewater is a critical environmental variable that governs the ionization state of both the polymer surface and the pharmaceutical molecule. Changes in pH can protonate or deprotonate functional groups, drastically altering the electrostatic interactions between the adsorbent and the pollutant. Capturing the pH is therefore essential for the model to understand the environmental context of the adsorption.

Initial Concentration (mg/L)

The initial concentration represents the starting amount of the pharmaceutical pollutant in the water before the polymer is added, measured in milligrams per liter (mg/L). This feature serves as the thermodynamic driving force for the adsorption process. Higher initial concentrations typically increase the probability of pollutant molecules colliding with and adhering to the available binding sites on the polymer matrix.

Adsorption Capacity (mg/g)

The adsorption capacity is the primary target variable for our predictive modeling. It measures the mass of the pollutant adsorbed per unit mass of the polymer material, expressed in milligrams per gram (mg/g). This feature definitively answers the question of efficiency: how much of the pollutant can a specific amount of the polymer actually remove? Accurately predicting this capacity for novel, unseen polymer–molecule pairs is the central predictive goal of this work.

1.4 Related Work

Historically, the discovery of novel adsorbents for wastewater treatment has been empirically driven. Researchers synthesize a candidate polymer, expose it to a specific pollutant under controlled laboratory conditions, and measure the resulting adsorption capacity. While this approach provides high–fidelity data, it is inherently unscalable given the vastness of the chemical space.

More recently, computational chemistry approaches such as Molecular Dynamics (MD) simulations [5] and Density Functional Theory (DFT) [6] have been employed to model polymer–molecule interactions at the atomic level. While these methods provide deep mechanistic insights, they are computationally expensive and too slow to screen thousands of candidates rapidly.

The application of Machine Learning to predict adsorption capacity represents the current frontier in this field. However, the primary bottleneck in training robust ML models for this specific domain remains data scarcity. Most published literature reports only a handful of experimental data points for highly specific conditions, lacking the standardization required for a large–scale data analysis.

Recent works have highlighted the need for unified databases and automated literature scraping to aggregate these fragmented experimental results. By systematically extracting data from the scientific literature, deriving features using computational chemistry libraries such as RDKit and Polymetrix [7], and augmenting sparse datasets through interpolation techniques, this thesis builds upon the emerging paradigm of data–driven materials discovery to address these traditional limitations.

Chapter 2

Machine Learning Architectures

This chapter introduces the theoretical foundations of the machine learning algorithms employed in this thesis. Two fundamentally different architectures were utilized to tackle the dual objectives of this research: Generative Pre–trained Transformers (specifically minGPT) were used for the generative task of discovering novel polymer structures, while Multi–Layer Perceptrons (MLP) were deployed for the regression task of predicting the adsorption capacity. Additionally, we discuss K–Fold Cross–Validation, a critical statistical method used to ensure the robustness of our predictive models given the constraints of a limited dataset.

2.1 Artificial Neural Networks and Multi–Layer Perceptrons

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the biological neural networks that constitute animal brains. The most fundamental and widely used feedforward ANN is the Multi–Layer Perceptron (MLP).

An MLP consists of at least three layers of interconnected nodes, or "neurons":

an input layer, one or more hidden layers, and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. The data flows in a single direction — from the input to the output — hence the designation "feedforward."

Mathematically, the calculation carried out at a given hidden layer l can be described as:

$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{2.1}$$

where $\mathbf{h}^{(l-1)}$ is the input from the previous layer, $\mathbf{W}^{(l)}$ is the weight matrix, $\mathbf{b}^{(l)}$ is the bias vector, and σ represents a non–linear activation function, such as the Rectified Linear Unit (ReLU) or the Sigmoid function.

During the training phase, the network learns to map inputs to outputs by adjusting its weights and biases to minimize a predefined loss function (e.g., Mean Squared Error for regression tasks). This optimization is typically achieved using the Backpropagation algorithm combined with gradient descent methods like Adam [8].

In the context of this thesis, the MLP is tasked with predicting a continuous variable: the adsorption capacity (mg/g) of a polymer–molecule pair. The input layer receives a high dimensional feature vector containing topological, thermodynamic, and electronic descriptors extracted from the SMILES and P–SMILES strings. Through its hidden layers, the MLP learns the complex, non–linear mappings between these chemical features and the target efficiency metric.

2.2 Generative Pre–trained Transformers (GPT) and minGPT

While MLPs excel at finding patterns in fixed–size numerical arrays, they are not naturally suited for processing sequential data or generating new sequences. For the

task of generating novel chemical structures, we turn to the Transformer architecture, introduced by Vaswani et al. in 2017 [9].

Transformers rely entirely on an "attention mechanism" to draw global dependencies between input and output sequences, dispensing with the recurrence and convolutions used in prior sequence models. Specifically, the Generative Pre–trained Transformer (GPT) family [10] utilizes a decoder–only architecture designed for autoregressive tasks. In autoregressive generation, the model predicts the next token in a sequence given all previous tokens.

The core computational unit of the Transformer is the self–attention mechanism, computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

where Q, K, and V represent the Query, Key, and Value matrices respectively, and d_k is the dimension of the keys. Conceptually, the Query represents "what am I looking for?", the Key represents "what do I contain?", and the Value represents "what information should I pass along?". The scaling factor $\sqrt{d_k}$ prevents vanishing gradients for large key dimensions. This mechanism allows every position in the sequence to attend to all other positions simultaneously, enabling the model to learn long–range dependencies — a crucial feature when dealing with the rigid syntax of chemical line notations.

2.2.1 Transformer Architecture Components

Beyond the attention mechanism, the Transformer architecture comprises several interconnected components that together enable effective sequence modeling.

- **Positional Encoding:** Unlike recurrent networks that process sequences step-by-step, the Transformer has no inherent notion of token order. Positional encodings inject sequence position information by adding a deterministic pattern

to the token embeddings. The original Transformer uses sinusoidal functions of different frequencies to allow the model to attend to relative positions effectively:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.3}$$

where pos is the position and i is the dimension index.

- **Multi-Head Attention:** Rather than performing a single attention function, the Transformer employs multiple attention heads in parallel. Each head learns different aspects of the relationships between tokens, allowing the model to jointly attend to information from different representation subspaces at different positions. The overall operation is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.4}$$

where each individual head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.5}$$

Here, Q, K, and V are the input Query, Key, and Value matrices. The terms W_i^Q, W_i^K, and W_i^V represent learned parameter matrices that linearly project the original inputs into a smaller, head-specific subspace. Finally, W^O is another learned weight matrix used to project the concatenated output of all the heads back to the original model dimension.

- **Feed-Forward Networks (FFN):** After each attention layer, a position-wise feed-forward network is applied to perform non-linear transformations on each

position independently:

$$\text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{2.6}$$

where σ is typically the GELU activation function, and W_1, W_2, b_1, and b_2 are learned weights and biases.

- **Residual Connections and Layer Normalization:** Each sub-layer (attention and feed-forward) is wrapped with a residual connection followed by layer normalization:

$$x \leftarrow \text{LayerNorm}(x + \text{Sublayer}(x)) \tag{2.7}$$

These components act to stabilize training and enable gradient flow through very deep networks.

- **Encoder versus Decoder (and Causal Masking):** The original Transformer comprised both an encoder (processing the full input sequence) and a decoder (generating the output autoregressively). GPT, however, employs a decoder-only architecture. It uses *causal masking* to ensure that predictions for position i can only depend on positions less than i. This is implemented as a triangular mask applied to the attention scores before the softmax operation, preventing future tokens from influencing current predictions.

2.2.2 minGPT Implementation

In this work, we utilize **minGPT**, a minimal, highly readable, and customizable implementation of the GPT architecture created by Andrej Karpathy. The library implements the complete GPT model in approximately 300 lines of clear PyTorch code, making it ideal for understanding the internals of transformer-based language models.

The minGPT architecture consists of the following components:

- **Token Embedding Layer:** Converts integer token indices into dense vectors of dimension d_{model}.
- **Position Embedding Layer:** Adds positional information to the token embeddings.
- **Transformer Blocks:** Stacks of attention and feed-forward layers with residual connections. Each block applies LayerNorm, performs multi-head causal self-attention, applies LayerNorm again, and then processes through the feed-forward network.
- **Language Modeling Head:** A linear layer that projects from the final embedding dimension back to the vocabulary size, producing logits (raw, unnormalized scores) over the next-token prediction.

The model is available in several sizes: `gpt-nano` (3 layers, 48 dimensions), `gpt-micro` (4 layers, 128 dimensions), `gpt-mini` (6 layers, 192 dimensions), and larger variants up to `gpt2-xl`. In this thesis, we employ the `gpt-nano` configuration to balance model capacity with computational efficiency for the polymer generation task.

By treating the generation of synthetic P–SMILES strings as a natural language modeling problem, we train the minGPT model on a dataset of valid polymer strings. The model learns the complex grammatical rules of P–SMILES (e.g., balancing parentheses for branches, matching ring closure digits) and can consequently generate entirely new, syntactically valid polymer structures that could serve as potential wastewater adsorbents.

2.2.3 Autoregressive Generation

During inference, the model generates sequences autoregressively. Starting from an initial prompt (or a special start token), the model computes a probability distribution over the entire vocabulary for the next token:

$$p(\text{token}_{t+1} \mid \text{token}_1, \ldots, \text{token}_t) = \text{softmax}\left(\frac{\text{logits}_t}{T}\right) \tag{2.8}$$

where T is the *temperature* parameter controlling the randomness of sampling. Low temperatures (e.g., $T < 1$) make the distribution sharper, favoring high-probability tokens, while higher temperatures increase diversity at the risk of less coherent output.

Additional generation strategies include *top-k sampling*, which restricts sampling to the k most probable tokens at each step, and *top-p (nucleus) sampling*, which dynamically selects the smallest set of tokens whose cumulative probability exceeds a threshold p. These techniques balance the trade-off between generating plausible, grammatically correct structures and exploring novel chemical space.

2.3 K–Fold Cross–Validation

A pervasive challenge in applying machine learning to specific physicochemical domains — such as polymer–based wastewater remediation — is data scarcity. Training complex models like MLPs on limited data frequently leads to overfitting, where the model memorizes the training examples but fails to generalize to unseen chemical combinations.

To rigorously evaluate model performance and tune hyperparameters under these constraints, we employed K–Fold Cross–Validation. In this resampling procedure, the entire dataset is randomly partitioned into K equal–sized, mutually exclusive

subsets or "folds."

The cross–validation process then proceeds in K iterations. In each iteration i:

1. The i–th fold is retained as the validation set;
2. The remaining $K - 1$ folds are combined to form the training set;
3. The model is trained from scratch on the training set and evaluated on the validation set.

Once all K iterations are complete, the evaluation metrics (such as the Mean Absolute Error or R^2 score) are averaged across all folds to produce a single, reliable performance estimate. This technique ensures that every data point in our carefully curated dataset is used for both training and validation exactly once. By systematically rotating the test set, K–Fold Cross–Validation provides a robust and unbiased assessment of the MLP's predictive feasibility, giving us confidence in the model's true generalization capabilities despite the underlying data scarcity.

Chapter 3

Data Acquisition and Feature Engineering

A robust machine learning pipeline for polymer-molecule interaction prediction requires both a reliable dataset and informative numerical representations of chemical structures. This chapter details the methodology employed to collect, augment, and transform experimental data into machine-readable feature vectors suitable for downstream modeling tasks.

3.1 Literature Scraping for Polymer-Molecule Data

The primary bottleneck in developing predictive models for polymer-based wastewater remediation is the scarcity of standardized experimental datasets. Unlike general molecular property databases, there exists no comprehensive public repository documenting the adsorption capacities of specific polymer-molecule pairs under controlled environmental conditions. To address this limitation, a systematic literature review was conducted to extract experimental data from peer-reviewed publications.

The data collection effort focused on identifying studies that reported quantita-

tive measurements of pharmaceutical adsorption onto polymer materials in aqueous solutions. From each relevant publication, three key variables were extracted for each experimental condition: the pH of the solution, the initial concentration of the pharmaceutical pollutant measured in milligrams per liter (mg/L), and the resulting adsorption capacity expressed in milligrams per gram (mg/g). These values respectively encode the environmental conditions of the experiment, the thermodynamic driving force for adsorption, and the efficiency metric that characterizes the polymer's effectiveness.

The scraped data were organized into a structured CSV (Comma-Separated Values) format, where each row represents a unique experimental observation identified by the polymer used, the target pharmaceutical molecule, and the specific experimental conditions. Importantly, the original source of each data point was preserved, enabling traceability back to the primary literature. This dataset, referred to as the Polymer Drug Concentration Capacity (PDCC) dataset, serves as the foundational resource for all subsequent machine learning experiments in this thesis.

The challenges encountered during data collection are noteworthy. Experimental results are often reported in non-standardized formats, with adsorption capacities expressed using varying units or normalization schemes. Furthermore, many studies report only a limited number of data points for highly specific polymer-molecule pairs under fixed conditions, making cross-study comparisons difficult. These factors underscore the need for careful data curation and standardization before the data can be utilized for machine learning applications.

3.2 Overcoming Data Scarcity: Interpolation and Augmentation

The PDCC dataset, while carefully curated, remains comparatively small relative to typical machine learning benchmarks. Data scarcity poses a fundamental challenge: models trained on limited data risk overfitting to the specific experimental conditions represented in the training set, failing to generalize to novel polymer-molecule combinations. To mitigate this issue, a data augmentation strategy based on interpolation was developed and systematically evaluated.

The rationale for interpolation is grounded in the physicochemical nature of the adsorption process. When a polymer is introduced into contaminated water, the mass of pollutant adsorbed per unit mass of polymer (capacity) typically exhibits a monotonic relationship with the initial pollutant concentration, approaching a saturation plateau at sufficiently high concentrations. This behavior is consistent with adsorption isotherm models such as the Langmuir or Freundlich models. Consequently, the capacity values at intermediate concentrations can be reasonably approximated by interpolation between observed data points.

The augmentation pipeline implemented in this work follows a two-stage approach. First, linear interpolation is performed between pairs of adjacent concentration-capacity data points belonging to the same polymer-molecule pair. This process generates intermediate synthetic data points that represent plausible adsorption behaviors at concentrations not explicitly measured in the original experiments. The number of interpolated points between each pair of observed values is configurable, allowing control over the degree of dataset expansion.

Second, origin points are added to each polymer-molecule group. Specifically, for each unique combination of polymer and target molecule (and accounting for pH when applicable), a synthetic observation at concentration equal to zero with

capacity equal to zero is appended. This addition reflects the physically intuitive boundary condition that no pollutant can be adsorbed when none is present in the solution, and it anchors the interpolated curves at the origin.

Four distinct augmentation strategies were implemented and tested: pure interpolation, interpolation followed by origin point addition, origin point addition followed by interpolation, and origin point addition only. Each method produces a different augmented dataset, and the optimal strategy was determined empirically through the cross-validation experiments described in Chapter 6.

3.3 Feature Extraction using RDKit, Polymetrix, and tblite

Raw SMILES and P-SMILES strings, while sufficient for generative modeling, must be converted into numerical feature vectors before they can be used as inputs to the Multi-Layer Perceptron. This section describes the feature extraction pipeline that transforms chemical line notations into chemically meaningful descriptors.

The feature extraction workflow leverages three established computational chemistry libraries: RDKit for molecular property calculation and canonicalization, Polymetrix for advanced polymer-specific descriptors, and tblite for quantum chemical calculations. Each library contributes complementary information about the molecular structure.

For both polymers and small molecules, the following descriptors are computed. The partition coefficient (logP) estimates the lipophilicity of the compound, which is critical for understanding how the molecule partitions between the aqueous phase and the polymer matrix. The distribution coefficient (logD) extends logP by accounting for ionization effects at a given pH, making it particularly relevant for wastewater

applications where pH varies. HOMO-LUMO energies in electron volts (eV) characterize the electronic structure of the molecule, specifically the energy of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). These values are central to predicting intermolecular interactions through Frontier Molecular Orbital theory. The net charge at a specified pH represents the ionization state of the molecule under environmental conditions, influencing electrostatic attraction or repulsion with the polymer surface.

Additionally, Morgan fingerprints are computed to encode the topological structure of the molecule as a fixed-length binary vector. These fingerprints capture local chemical environments and enable the model to recognize structural similarities between molecules.

The Polymetrix library provides a suite of polymer-specific features that extend beyond standard molecular descriptors. These include the topological polar surface area (TPSA), which quantifies the total surface area of polar atoms in the molecule and correlates with hydrogen bonding capacity. The synthetic accessibility (SA) score estimates how easily a given polymer structure could be synthesized in practice, which is essential for real-world applicability. Additional Polymetrix features include counts of hydrogen bond donors and acceptors, the number of rotatable bonds, ring counts (aromatic and non-aromatic), molecular weight, and various VSA (van der Waals Surface Area) descriptors.

A notable technical challenge arises when computing features for polymers. Standard cheminformatics libraries, including RDKit, are primarily designed for discrete small molecules and may not correctly handle the non-standard atom types and connectivity patterns present in P-SMILES notation. To overcome this limitation, a capping strategy was employed: for each polymer, the repeating unit is first extracted and treated as a discrete molecular fragment. This fragment is then capped with hydrogen atoms at the attachment points, converting it into a valid

molecule that can be parsed by RDKit. The features computed from this capped monomer are used as proxies for the properties of the full polymer chain. This approximation is chemically reasonable because the repeating unit determines much of the polymer's bulk chemical behavior.

The full feature extraction pipeline produces a high-dimensional numerical vector for each polymer-molecule pair, encoding structural, thermodynamic, electronic, and environmental information. These feature vectors serve as the input to the MLP described in Chapter 5, where they are scaled using a StandardScaler to ensure that all features contribute equally to the learning process.

Chapter 4

Generative Modeling for Molecular Discovery

Beyond predictive modeling, this thesis explores the complementary task of generating novel polymer structures that could serve as effective wastewater adsorbents. Rather than relying solely on known polymers, a generative approach enables the systematic exploration of a theoretically infinite chemical space. This chapter describes the implementation and evaluation of a generative model for producing valid P-SMILES strings.

4.1 The minGPT Wrapper Implementation

The Generative Pre-trained Transformer (GPT) architecture, originally developed for natural language processing, has demonstrated remarkable capabilities in learning complex sequential patterns. By treating chemical notation as a specialized language, the same principles can be applied to generate novel molecular structures. In this work, the minGPT framework—a lightweight, modular implementation of the GPT architecture created by Andrej Karpathy—was adapted to serve as the backbone for

polymer generation.

The core conceptual shift in applying language models to molecular generation is the reframing of chemical validity as grammatical correctness. Just as a sentence in English must follow syntactic rules to be grammatically valid, a SMILES or P-SMILES string must conform to strict chemical grammar rules. The autoregressive nature of the GPT architecture makes it particularly well-suited for this task: the model learns the conditional probability distribution of the next token given all previous tokens in the sequence, effectively learning the grammar of molecular notation.

A critical component of the implementation is the custom tokenizer developed for P-SMILES notation. Unlike natural language tokenizers that operate on subword units, the chemical tokenizer must handle atomic symbols, bond types, branching parentheses, and ring closure digits as atomic units. The tokenizer builds a vocabulary from the training dataset, assigning a unique integer identifier to each distinct token observed in the training data. Special tokens are introduced to mark the beginning and end of each sequence.

The training procedure follows the standard language modeling objective: given a sequence of tokens representing a valid P-SMILES string, the model learns to predict the next token in the sequence. During training, the model is exposed to thousands of valid polymer structures, gradually learning the statistical regularities that characterize chemically plausible polymers. The loss is computed as the cross-entropy between the predicted token distribution and the actual next token, and gradients are computed via backpropagation through the transformer layers.

The hyperparameters of the model include the number of transformer layers (determining the model's capacity to learn complex dependencies), the number of attention heads per layer, the embedding dimension, and the context length (the maximum sequence length the model can handle). These were tuned through preliminary

experiments to balance model expressiveness with computational efficiency.

Training was performed on two datasets. The primary training set for polymer generation was the PI1M dataset, a comprehensive collection of synthetic polymer structures represented in P-SMILES notation. For comparison, a secondary model was trained on the "ZINC_base" dataset to evaluate the generation of standard molecular SMILES strings.

4.2 Generating Valid Synthetic Polymer SMILES (P-SMILES)

Once trained, the model can generate new polymer structures through autoregressive sampling. Starting from a special start-of-sequence token, the model predicts the probability distribution over the next token in the vocabulary. A token is then sampled from this distribution and appended to the sequence. This process repeats iteratively until a special end-of-sequence token is generated or the maximum sequence length is reached.

The sampling procedure supports several generation strategies. Temperature scaling controls the randomness of the sampling process. At low temperatures (e.g., 0.8), the model tends to produce highly probable tokens, yielding conservative generations that closely resemble the training data. At higher temperatures (e.g., 1.2), the model is more likely to explore lower-probability tokens, producing more diverse and potentially novel structures. Top-k sampling, which restricts the candidate tokens to the k most probable options at each step, can be combined with temperature scaling to further control the trade-off between quality and diversity.

Following generation, each output string must be validated and canonicalized. This post-processing step is essential because the generative model produces strings

based on learned statistical patterns and does not inherently guarantee chemical validity. Invalid strings may result from sampling errors or may represent syntactically correct but chemically impossible structures. Validation is performed using RDKit, which parses the generated string and checks whether it corresponds to a chemically valid molecular graph. For polymers, additional validation specific to P-SMILES grammar is applied to ensure proper attachment point notation.

Novelty checking is performed by comparing generated structures against the training set. A generated polymer is considered novel if its canonicalized representation does not appear in the training data. This step ensures that the generative model is not merely reproducing memorized structures but is genuinely creating new candidates.

4.3 Evaluation of Generated P-SMILES

The quality of the generated polymers was assessed through multiple metrics addressing validity, novelty, and structural diversity.

Validity rates provide a measure of how often the model produces syntactically correct output. When generating standard molecular SMILES from the ZINC_base-trained model, approximately 43.2% of the 12,800 generated strings were successfully parsed and validated by RDKit. For P-SMILES generation using the PI1M-trained model, the validity rate was notably higher at approximately 57.6%. The difference reflects the relative complexity of standard molecular grammar compared to the P-SMILES notation used for polymers.

Novelty analysis further confirmed that a substantial fraction of valid generations represent structures not present in the training data. Among the valid SMILES generated, all 5,525 were confirmed to be novel with respect to the training set. Similarly, all 7,373 valid P-SMILES were novel compared to the PI1M training data.

These results demonstrate that the model has learned to generalize beyond simple memorization and can generate genuinely new chemical structures.

The generation results are summarized in Table 4.1. These metrics indicate that the generative model successfully produces a mixture of valid and novel structures, providing a diverse pool of candidate polymers for subsequent filtering and evaluation.

Table 4.1: Summary of Generative Model Evaluation Results

Output Type	**Generated**	**Valid**	**Novel**
SMILES (ZINC_base)	12,800	5,525 (43.2%)	5,525 (100%)
P-SMILES (PI1M)	12,800	7,373 (57.6%)	7,373 (100%)

The high novelty rate is particularly significant for the drug discovery pipeline envisioned in this thesis. A generative model that merely reproduces known structures would offer limited value for discovering novel adsorbents. The ability to generate truly novel polymers, combined with the filtering and prediction stages described in subsequent chapters, enables a closed-loop workflow for computational polymer discovery.

4.4 Integration with the Discovery Pipeline

The generative model serves as the foundation of the broader computational pipeline developed in this thesis. Generated polymers enter a multi-stage filtering and evaluation process. First, validity checking ensures that only chemically valid structures proceed. Second, heuristic filters based on chemical properties (discussed in Chapter 5) eliminate structures that are unlikely to be effective adsorbents. Third, for the most promising candidates, the predictive model estimates the expected adsorption capacity against the target pharmaceutical. This integrated approach

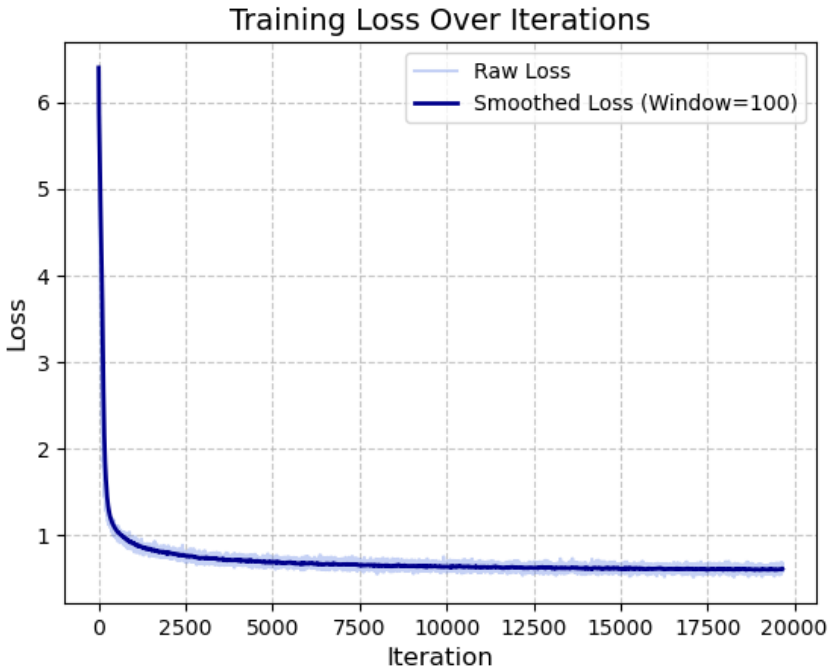

Figure 4.1: Training loss of the minGPT model over training iterations.

represents a shift from traditional trial-and-error laboratory synthesis toward a data-driven, computationally accelerated discovery workflow.

Chapter 5

Predictive Modeling of Adsorption Capacity

Having established the ability to generate novel polymer structures, this chapter addresses the complementary challenge of predicting their effectiveness. The central predictive task is to estimate the adsorption capacity of a given polymer-molecule pair under specified environmental conditions.

5.1 Experimental Setup for the MLP

The Multi-Layer Perceptron employed for capacity prediction is a feedforward neural network with a carefully chosen architecture designed to capture the nonlinear relationships between chemical features and adsorption efficiency. The network consists of an input layer sized to match the dimensionality of the feature vector, multiple hidden layers with progressively decreasing widths (specifically: 16, 8, 4, 4, and 4 neurons in each respective hidden layer), and a single output neuron that produces the predicted capacity value.

The feature vector input to the MLP encompasses all relevant information about

the polymer, the target pharmaceutical molecule, and the experimental conditions. This includes the thermodynamic and electronic descriptors extracted as described in Chapter 4: logP and logD values for both polymer and molecule, HOMO-LUMO energies in electron volts, net charges at the relevant pH, and topological descriptors such as TPSA and molecular weight. The Morgan fingerprint representation provides additional structural information, while the initial concentration and water pH encode the experimental context.

Preprocessing of the input features is critical for stable and effective training. The feature vector is scaled using a StandardScaler, which transforms each feature to have zero mean and unit variance. This ensures that features measured in different units (such as energies in eV versus dimensionless logP values) contribute proportionally to the learning process. A separate MinMaxScaler is applied to the target capacity values, mapping them to the range [0, 1] to facilitate gradient-based optimization.

The output activation function requires special consideration. Because adsorption capacity is inherently non-negative, a SoftPlus activation function is applied to the network output. SoftPlus is a smooth, differentiable approximation of the ReLU function, defined as $\text{SoftPlus}(x) = \log(1 + e^x)$. Unlike a linear output, which could theoretically produce negative predictions, SoftPlus guarantees non-negative outputs that are consistent with the physical nature of the target variable.

Training is performed using the Adam optimizer with mean squared error (MSE) as the loss function. The MSE loss is appropriate for regression tasks where larger errors should be penalized proportionally more than smaller errors. The model is trained for a fixed number of epochs with early stopping based on validation loss when a validation set is available.

5.2 Addressing the Small Dataset Challenge

The limited size of the PDCC dataset presents the most significant challenge for predictive modeling. With only a few hundred experimental observations available, standard machine learning practices that assume abundant data must be adapted. Several strategies were employed to maximize the effective information content of the available data and to obtain reliable performance estimates.

Leave-One-Out Cross-Validation (LOOCV), the extreme case of K-fold cross-validation where the number of folds equals the number of data points, serves as the primary evaluation methodology. Each fold contains a single observation, and the model is trained on all remaining data. While computationally expensive, LOOCV is particularly valuable for small datasets because it maximizes the amount of training data available in each iteration and provides an almost unbiased estimate of generalization error.

In addition to cross-validation, a comprehensive hyperparameter search was conducted to identify the optimal model configuration. The search space included the learning rate, the number of training epochs, the batch size, and the choice of feature subsets. For each configuration, the Q-squared (Q^2) statistic was computed from the cross-validation results. The Q^2 statistic, analogous to the R^2 coefficient in standard regression, measures the proportion of variance in the target variable that is explained by the model. A Q^2 value of 1 indicates perfect prediction, while a value of 0 indicates that the model performs no better than predicting the mean of the training data. The Q^2 statistic is defined as:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the observed values, and n is the number of observations.

The dataset augmentation strategies described in Chapter 4 (interpolation and origin point addition) were applied prior to feature extraction. Different augmentation configurations were tested, and the combination that yielded the best cross-validation performance was selected for the final model.

5.3 Cross-Validation Results and Performance Metrics

The hyperparameter optimization process evaluated a wide range of configurations using Leave-One-Out Cross-Validation. The results provide insight into the model's predictive capability under realistic data constraints.

Among the configurations tested, the optimal hyperparameters were identified based on the highest Q^2 score achieved during cross-validation. The best-performing configuration achieved a Q^2 score of 0.984 using LOOCV evaluation. This result indicates that the model explains approximately 98.4% of the variance in the adsorption capacity.

The distribution of Q^2 scores across different hyperparameter configurations revealed several important findings. Configurations that included the SoftPlus output activation consistently outperformed those using a linear output, confirming the importance of constraining predictions to the physically meaningful non-negative range. The choice of feature subset had a moderate impact on performance, with configurations using the full feature set (including both Polymetrix descriptors and Morgan fingerprints) performing better than those using reduced feature sets.

The interpolation-based data augmentation provided measurable benefits. The configuration using interpolation followed by origin point addition produced the most consistent cross-validation results, suggesting that the augmented data provides a

more complete representation of the capacity-concentration relationship near the boundary conditions.

These results were obtained on the augmented dataset; performance on the original, unaugmented data would be expected to be lower.

Table 5.1: LOOCV results across MLP hyperparameter configurations. Top 10 experiments ranked by Q^2.[1]

Experiment	**Q^2**	**MAE**	**RMSE**
hd_16_8_4_4_4	0.984	1.499	6.485
hd_16_8_4_4_4_only_y_scaler	0.980	1.916	7.089
hd_16_8_4_4_4_mae	0.972	2.757	8.533
hd_16_8_4_4_4_4_mae	0.918	4.558	14.518
hd_16_8_8_8	0.913	2.883	14.907
hd_16_8_4_4_4_4	0.904	3.555	15.655
hd_32_16_8_8	0.900	3.592	15.959
hd_8_8_8_8	0.894	3.328	16.493
hd_16_16_16	0.888	3.387	16.926
hd_32_16_8_4	0.834	4.567	20.583

5.4 Exploratory Clustering Analysis of the PDCC Dataset

To gain insight into the structure of the PDCC dataset and identify natural groupings among polymer-molecule pairs, an Agglomerative Hierarchical Clustering (AHC) analysis was performed. AHC iteratively merges data points based on their feature similarity, producing a hierarchical structure that can be visualized as a dendrogram. Two complementary visualizations of the clustering results are presented:

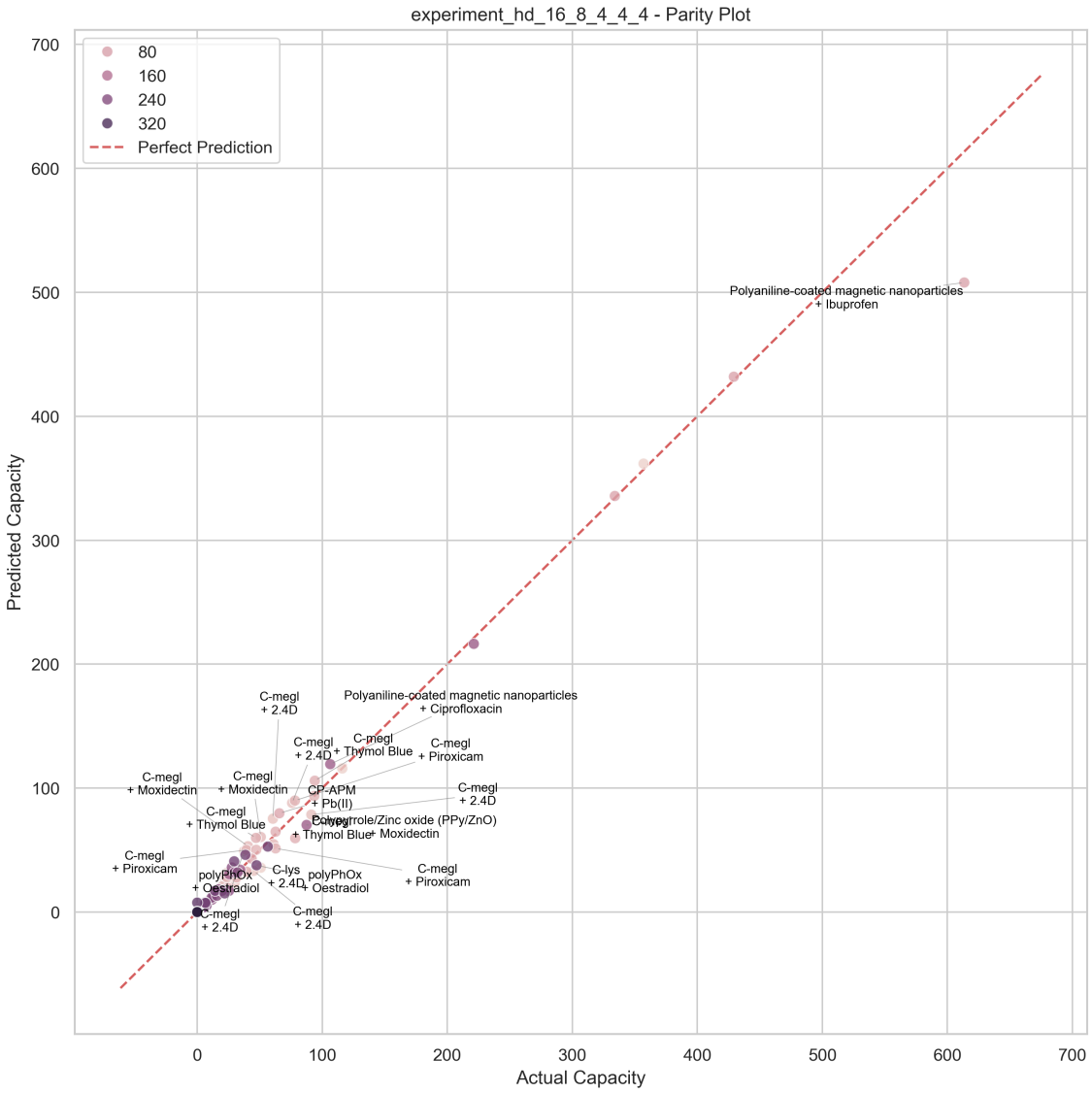

Figure 5.1: Parity plot for the optimal MLP configuration ($Q^2 = 0.984$). Predicted vs. actual adsorption capacity values.

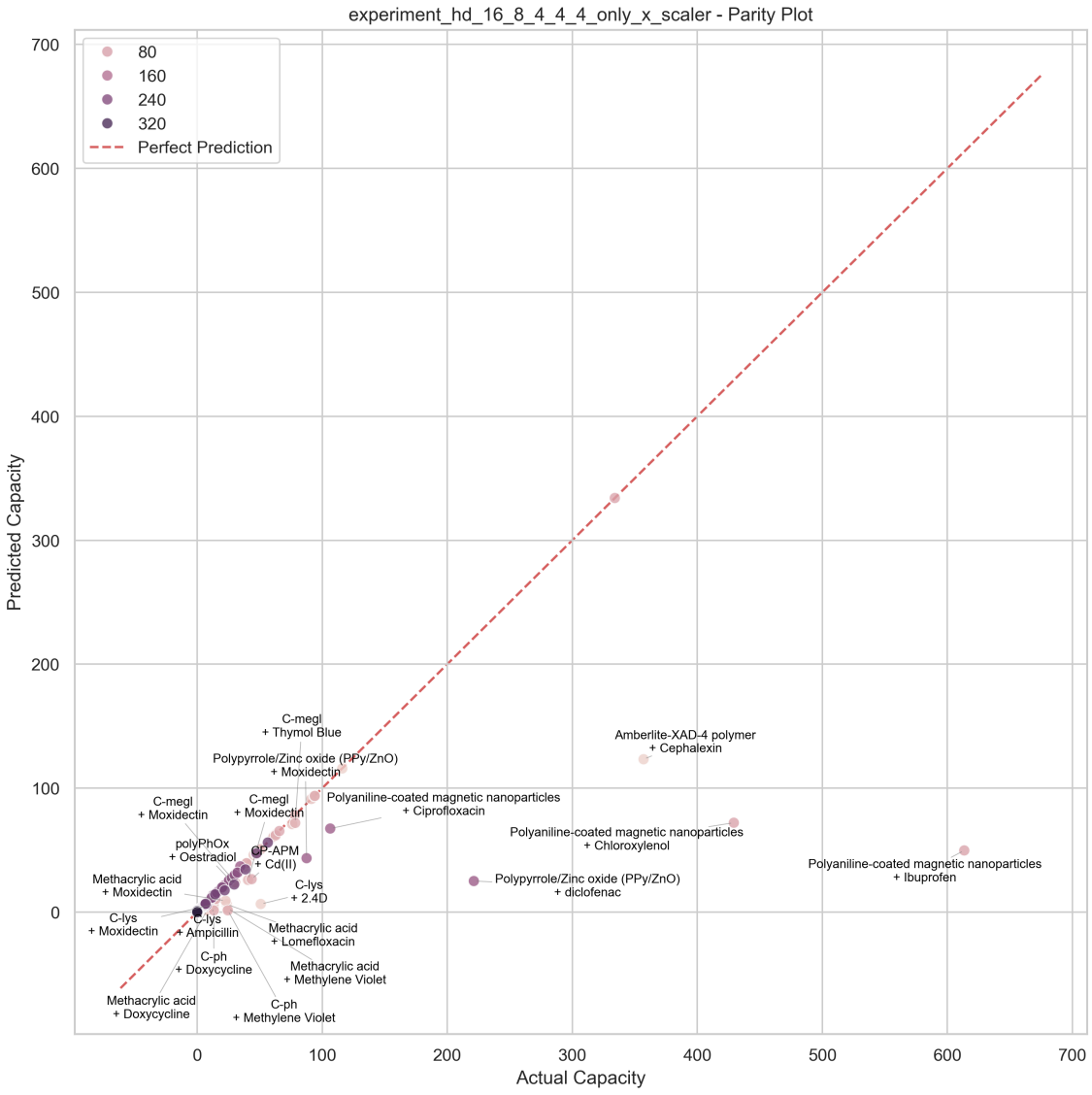

Figure 5.2: Parity plot for the MLP configuration without Y-scaling ($Q^2 = 0.413$). The increased scatter demonstrates the importance of target scaling for prediction quality.

the dendrogram, which shows the full hierarchical merging process, and a two-dimensional projection via Principal Component Analysis (PCA), which reveals the spatial separation of clusters in reduced dimensionality.

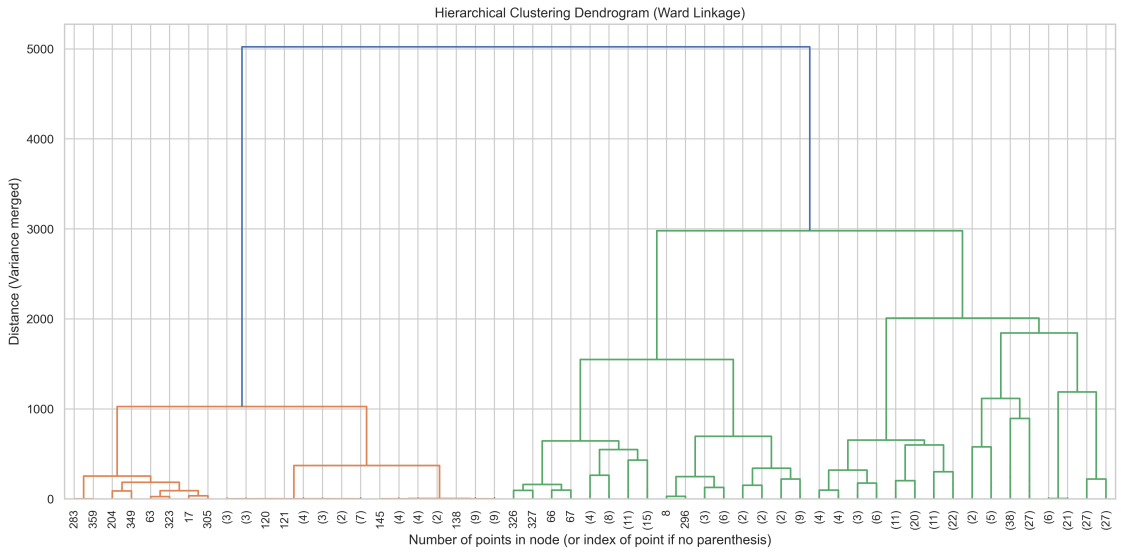

Figure 5.3: Agglomerative Hierarchical Clustering dendrogram of the PDCC dataset. The y-axis represents the linkage distance; lower merges indicate more similar data points.

5.5 Discussion of Model Limitations

Despite the encouraging results, several limitations must be acknowledged. The most fundamental limitation is data scarcity. The PDCC dataset contains orders of magnitude fewer observations than typical machine learning benchmarks. Training deep neural networks on such limited data risks overfitting, where the model memorizes specific training examples rather than learning generalizable patterns. While cross-validation provides a measure of generalization performance, the confidence intervals on these estimates are wide due to the small sample size.

The second limitation concerns the representativeness of the training data. The scraped literature reports successful adsorption experiments, potentially introducing

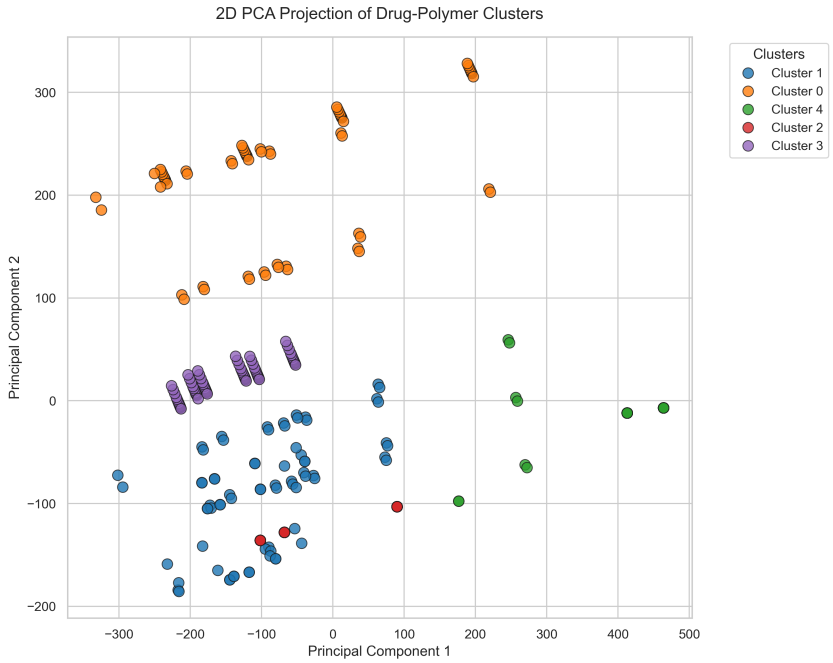

Figure 5.4: PDCC data projected onto the first two principal components, colored by cluster assignment. Cluster membership reveals groupings driven primarily by polymer molecular weight (PC2) and drug molecular properties (PC1).

publication bias. Polymer-molecule pairs that did not exhibit significant adsorption may be underrepresented or absent from the dataset. A robust predictive model requires data on both successes and failures to learn the boundary between effective and ineffective adsorbents.

The third limitation relates to feature quality. As discussed in Chapter 4, polymer features are computed from capped monomers rather than full polymer chains. This approximation neglects higher-order polymer properties such as chain length distribution, cross-linking density, and three-dimensional conformation, all of which influence adsorption behavior.

5.6 Chemical Filtering of Candidate Polymers

Complementing the quantitative predictions of the MLP, a set of heuristic filters based on established chemical principles is applied to eliminate polymers that are unlikely to be effective adsorbents. These filters encode domain knowledge about the requirements for successful adsorption in wastewater treatment.

The first filter concerns the partition coefficient (logP) of the polymer. For a polymer to function as an insoluble adsorbent in aqueous solution, it must have a sufficiently high logP value to remain in a solid phase. A minimum logP threshold of 1.5 is applied, ensuring that the polymer remains hydrophobic enough to avoid dissolution while not being so lipophilic that it loses affinity for the aqueous pollutant.

The second filter addresses polar surface area (TPSA). Polymers with higher TPSA values possess more polar functional groups that can engage in hydrogen bonding and electrostatic interactions with pollutant molecules. A minimum TPSA of 60 Å^2 is required to ensure sufficient polar character for effective binding.

The third and most important filter is based on Frontier Molecular Orbital (FMO) theory. The HOMO and LUMO energies of both the polymer and the target drug determine the nature of their electronic interactions. Two interaction pathways are possible: the polymer acts as an electron donor (using its HOMO) to the drug (acting as an acceptor via its LUMO), or vice versa. The energy gaps for these two pathways are computed as:

$$\Delta E_1 = |E_{\text{LUMO, drug}} - E_{\text{HOMO, polymer}}| \tag{5.2}$$

$$\Delta E_2 = |E_{\text{LUMO, polymer}} - E_{\text{HOMO, drug}}| \tag{5.3}$$

The smaller of these two gaps represents the dominant interaction pathway. A

smaller intermolecular FMO gap indicates stronger electronic coupling and thus stronger binding. A maximum gap threshold of 4.0 eV is applied, retaining only those polymer-molecule pairs with favorable electronic complementarity.

The fourth filter addresses synthetic accessibility (SA score). A practical wastewater adsorbent must be synthesizable at scale using reasonable chemical methods. The SA score, computed from structural features that affect synthetic difficulty, is constrained to a maximum value of 4.5 on a 1-10 scale.

These filters are applied sequentially after generation and validation, dramatically reducing the number of candidate polymers before they are passed to the predictive model for capacity estimation. The combination of chemical filtering and machine learning prediction provides a balanced approach that leverages both domain knowledge and data-driven pattern recognition.

5.7 The Integrated Discovery Pipeline

The complete computational pipeline integrates generation, validation, filtering, and prediction into an end-to-end workflow for polymer discovery. Given a target pharmaceutical molecule (e.g., aspirin, ibuprofen, or metformin), the pipeline proceeds as follows.

First, the target molecule is converted to SMILES notation using the PubChem database. Second, the generative model produces a batch of candidate polymer P-SMILES strings. Third, invalid and duplicate structures are eliminated through validation. Fourth, the heuristic filters based on logP, TPSA, FMO gaps, and SA score are applied. Fifth, the MLP predicts the adsorption capacity for each surviving candidate. Finally, candidates are ranked by predicted capacity, and those exceeding a specified threshold are returned as promising discoveries.

This pipeline was tested with multiple target molecules, demonstrating its ability

to identify candidate polymers with predicted capacities in excess of 1.0 mg/g. Visualization of capacity versus concentration and pH provides additional insight into how the predicted performance varies with experimental conditions, guiding the design of laboratory validation experiments.

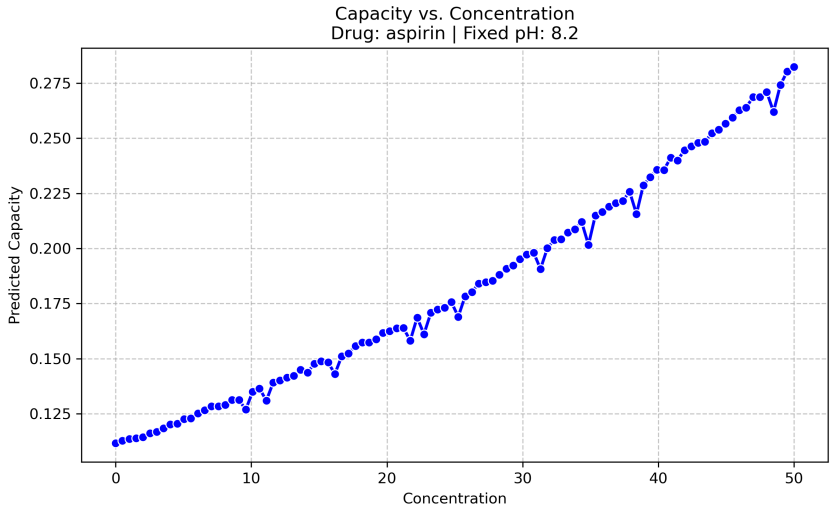

Figure 5.5: Predicted adsorption capacity vs. initial concentration for the top-ranked polymer against aspirin.

The pipeline is configurable, allowing adjustment of generation parameters (temperature, batch size), filter thresholds, and the target capacity cutoff. This flexibility enables adaptation to different application contexts, such as targeting different pollutant classes or optimizing for specific environmental conditions.

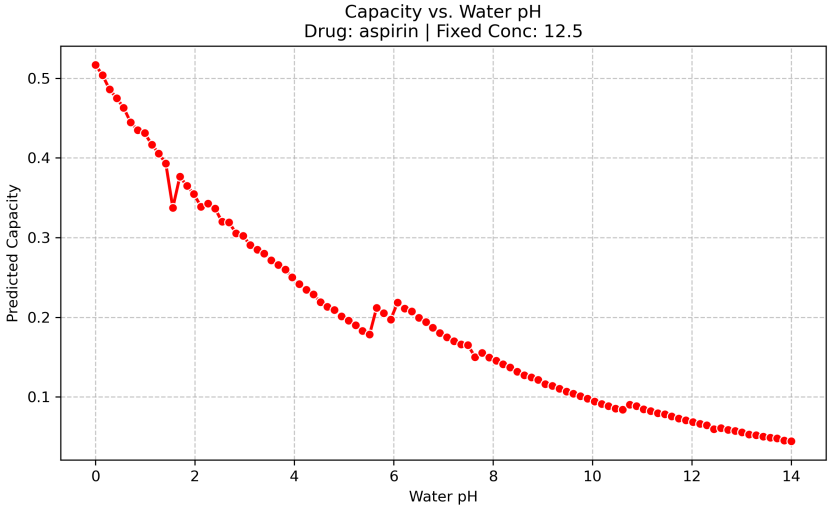

Figure 5.6: Predicted adsorption capacity vs. water pH for the top-ranked polymer against aspirin.

Chapter 6

Conclusions and Future Work

This thesis has explored the application of machine learning techniques to the challenge of discovering novel polymer materials for pharmaceutical adsorption in wastewater. The work addressed both the generative task of creating new polymer structures and the predictive task of estimating their effectiveness, culminating in an integrated computational pipeline for polymer discovery.

6.1 Summary of Contributions

The primary contributions of this thesis are as follows.

First, a systematic data acquisition pipeline was developed to extract experimental adsorption data from the scientific literature. Despite the inherent scarcity of standardized datasets in this domain, the resulting PDCC dataset provides a foundation for machine learning experiments. Data augmentation through linear interpolation and origin point addition was shown to improve model performance, enabling more effective utilization of the limited available data.

Second, a generative model based on the minGPT architecture was successfully adapted for polymer structure generation. The model learns the grammatical rules

of P-SMILES notation and produces novel, chemically valid polymer structures. Evaluation showed that 57.6% of generated P-SMILES strings were valid, and 100% of valid generations were novel relative to the training set. This generative capability enables the exploration of a vast chemical space that would be impractical to sample through laboratory synthesis alone.

Third, a predictive model based on the Multi-Layer Perceptron was developed to estimate adsorption capacity from chemical features. Systematic hyperparameter optimization using Leave-One-Out Cross-Validation identified configurations achieving a Q^2 score of approximately 0.984. While data scarcity imposes fundamental limitations, this result demonstrates that meaningful predictive patterns can be extracted from the available data.

Fourth, a heuristic filtering framework based on established chemical principles was implemented. Filters based on logP, TPSA, FMO theory, and synthetic accessibility effectively eliminate candidates that are unlikely to be effective or practical adsorbents. The integration of these filters with the predictive model provides a balanced approach that combines domain knowledge with data-driven prediction.

Fifth, the individual components were integrated into a complete computational pipeline that, given a target pharmaceutical molecule, generates candidate polymers, filters them based on chemical viability, and prioritizes them by predicted adsorption capacity. This pipeline represents a significant step toward computational acceleration of polymer discovery for environmental remediation.

6.2 Limitations and Challenges

Several limitations of the current work must be acknowledged and addressed in future research. Data scarcity remains the most critical challenge, as the PDCC dataset contains only a few hundred experimental observations, limiting both model

complexity and generalization capability. Publication bias in the training data may skew predictions toward positive outcomes, and the approximation of polymer features from capped monomers neglects higher-order structural properties that influence adsorption behavior. Additionally, the focus on homopolymers and the P-SMILES representation may restrict the applicability of the framework to more complex real-world scenarios. A more detailed discussion of these limitations is provided in Section 5.5.

6.3 Future Research Directions

Based on the findings and limitations of this work, several promising directions for future research are identified.

The most immediate priority is data expansion. The creation of a comprehensive, publicly available database of polymer-molecule adsorption experiments would benefit not only this work but the broader research community. Automated literature mining, combined with standardized data formatting and quality control, could accelerate this process significantly.

Model architecture exploration represents another productive direction. While this thesis focused on MLPs for prediction, Graph Neural Networks (GNNs) are naturally suited for molecular property prediction because they operate directly on molecular graphs without the need for hand-crafted feature engineering. GNNs have shown state-of-the-art performance in molecular property prediction benchmarks and could potentially improve predictive accuracy for adsorption capacity.

The generative model could be conditioned on target properties to enable directed generation. Instead of generating polymers unconditionally and then filtering, a conditional generative model could be trained to produce polymers with specific property ranges (e.g., high logP, low SA score, specific FMO gap). This approach

would improve the efficiency of the discovery pipeline by focusing generation effort on chemically promising regions of the space.

The predictive model could be extended to multi-task learning, simultaneously predicting multiple adsorption metrics such as capacity, kinetics, and selectivity. Multi-task learning leverages shared representations across related tasks and can improve generalization on each individual task, particularly when data are scarce for each task individually.

Finally, experimental validation of the computationally predicted candidates is essential to close the loop between simulation and reality. The highest-ranked candidates from the discovery pipeline should be synthesized and tested under controlled laboratory conditions to verify the accuracy of the predictions and to identify systematic discrepancies that could guide model improvement.

6.4 Closing Remarks

This thesis has demonstrated the feasibility of applying machine learning techniques to the discovery of polymer adsorbents for pharmaceutical removal from wastewater. The integration of generative modeling, chemical filtering, and predictive analytics provides a computational framework that can accelerate the identification of promising candidates and guide experimental efforts. While data scarcity remains a fundamental constraint, the results obtained show that meaningful patterns can be learned from the available data.

The broader implication of this work is the demonstration that computational approaches can complement traditional laboratory methods in materials discovery. As datasets grow and models improve, the vision of computationally designing materials for specific environmental applications becomes increasingly achievable. The framework developed here provides a foundation upon which future advances in

data, models, and experimental validation can be built, ultimately contributing to the goal of sustainable water treatment through advanced materials science.

Bibliography

[1] Bai X, Zhang X. Artificial Intelligence-Powered Materials Science. Nanomicro Lett. 2025;17(1):135. Published 2025 Feb 6. doi:10.1007/s40820-024-01634-8 https://pmc.ncbi.nlm.nih.gov/articles/PMC11803041/

[2] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." J. Chem. Inf. Comput. Sci. 28 (1988): 31-36.

[3] Lin TS, Coley CW, Mochigase H, et al. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. ACS Cent Sci. 2019;5(9):1523-1531. doi:10.1021/acscentsci.9b00476

[4] RDKit: Open-source cheminformatics. https://www.rdkit.org

[5] Aulifa, D. L., Al Shofwan, A. A., Megantara, S., Fakih, T. M., & Budiman, A. (2024). Elucidation of Molecular Interactions Between Drug-Polymer in Amorphous Solid Dispersion by a Computational Approach Using Molecular Dynamics Simulations. Advances and applications in bioinformatics and chemistry : AABC, 17, 1–19. https://doi.org/10.2147/AABC.S441628

[6] Wei, Z.; Ning, N.; Zhang, L.; Tian, M.; Mi, J. Density Functional Theory of Polymer Structure and Conformations. Polymers 2016, 8, 121. https://doi.org/10.3390/polym8040121

[7] Kunchapu, S., Jablonka, K.M. PolyMetriX: an ecosystem for digital polymer chemistry. npj Comput Mater 11, 312 (2025). https://doi.org/10.1038/s41524-025-01823-y

[8] Kingma, Diederik P. and Jimmy Ba. "Adam: A Method for Stochastic Optimization." CoRR abs/1412.6980 (2014): n. pag.

[9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017), "Attention is All you Need," Neural Information Processing Systems, 30, 5998–6008.

[10] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

Acknowledgements

I would like to thank my parents who raised me since I was born, and still they are doing.