UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Large Language Models as Feature Selectors for Predicting ALK-5 Inhibition

**Walter Augusto Perez Casas**

Monograph - MBA in Artificial Intelligence and Big Data

ICMC USP
SÃO CARLOS

**Walter Augusto Perez Casas**

# Large Language Models as Feature Selectors for Predicting ALK-5 Inhibition

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence and Big Data

Advisor: Profa. Dra. Patrícia Rufino Oliveira

**Original version**

**São Carlos**

**2025**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Casas, Walter Perez
S856m Large Language Models as Feature Selectors for Predicting ALK-5 Inhibition / Walter Augusto Perez Casas ; orientadora Patrícia Rufino Oliveira. – São Carlos, 2025.
59 p. : il. (algumas color.) ; 30 cm.

Monograph (MBA in Artificial Intelligence and Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Oliveira, Patrícia Rufino, orient. II. Título.

**Walter Augusto Perez Casas**

# Modelos de Linguagem de Grande Escala como Seletores de Características para Predizer a Inibição de ALK-5

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial e Big Data

Orientadora: Profa. Dra. Patrícia Rufino Oliveira

**Versão original**

**São Carlos**

**2025**

*To Matías, for his smile that is my motivation and his hugs that are my hope.*
*To my uncles Tania and Lucho, for their courageous lifelong battle.*

## ACKNOWLEDGEMENTS

First of all, I would like to thank my mother Lourdes for always being the pillar that drives us not to give up, teaching us through example, and because everything I am is thanks to her. To my sister Lu, for the joy with which she has always accompanied me. And to my little Mati, for being my light and the hope that life is always worth living.

I am grateful to my maternal family, the Casas, because they have always been there, because we have shared many moments, and they were key to my development. To my paternal family, the Pérez, because with them I have always found that other side of life, where unity and laughter are never lacking. I also want to thank Gabriela and the Ponce family, because they are a lovely family who welcomed me with great affection, the same affection I hold for them. To my lifelong friends: Aldo, Jesús, and David.

My sincere gratitude goes to Professora Dra. Patricia Rufino, for guiding me, having patience with me, guiding me with her knowledge, and always demonstrating great availability. To Professora Dra. Solange Rezende, for the advice and conversations that have always made me reflect, for always having a space beyond the academic to see the human side of studies.

Finally, to the University of São Paulo (USP), Instituto de Ciências Matemáticas e de Computação (ICMC), and the professors and collaborators who provided the necessary resources to guide my learning.

*"Act only according to that maxim whereby you can at the same time will that it should become a universal law."*

*Immanuel Kant*

# ABSTRACT

Casas, W. P. **Large Language Models as Feature Selectors for Predicting ALK-5 Inhibition**. 2025. 59 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Cancer is one of the leading causes of mortality around the world, mainly due to the uncontrolled proliferation of tumor cells. A promising approach to developing antineoplastic drugs involves inhibiting ALK-5 (Activin-Like Kinase 5), a key molecule regulating cellular processes associated with cancer growth and dissemination. Machine learning methods are commonly employed to predict the inhibitory activity (pIC50) of candidate compounds, and they are trained on molecular descriptors derived from the chemical structure of these compounds. However, the high dimensionality of these chemical representations and the limited sample sizes hinder generalization, often resulting in overfitting. In this work, we propose an approach that leverages the capabilities of large language models (LLMs) to select more representative molecular features prior to applying conventional machine learning algorithms. Our results demonstrate that LLM-assisted feature selection achieves performance comparable to traditional feature selection methods, such as filter, wrapper, or embedded approaches, relying solely on its knowledge, i.e., in a zero-shot manner. This is particularly relevant in this case, where we reduce the number of features from approximately 1400 to just 50, forcing the model to select the most important ones. This highlights its potential for improving the efficiency and effectiveness of ALK-5 inhibitor discovery and guiding efforts toward more practical and scalable methods, thereby facilitating the implementation of solutions in real-world settings.

**Keywords**: LLM. ALK-5 inhibitors. pIC50.

# RESUMO

Casas, W. P. **Modelos de Linguagem de Grande Escala como Seletores de Características para Predizer a Inibição de ALK-5**. 2025. 59 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

O câncer é uma das principais causas de mortalidade em todo o mundo, principalmente devido à proliferação descontrolada de células tumorais. Uma abordagem promissora para o desenvolvimento de fármacos antineoplásicos envolve a inibição da ALK-5 (Activin-Like Kinase 5), uma molécula-chave na regulação de processos celulares associados ao crescimento e à disseminação do câncer. Métodos de aprendizado de máquina são amplamente empregados para prever a atividade inibitória (pIC50) de compostos candidatos, sendo treinados com descritores moleculares extraídos da estrutura química dessas moléculas. No entanto, a alta dimensionalidade dessas representações químicas, aliada ao tamanho limitado das amostras, dificulta a generalização dos modelos, frequentemente resultando em overfitting. Neste trabalho, propomos uma abordagem que aproveita as capacidades dos modelos de linguagem de grande escala (LLMs) para selecionar características moleculares mais representativas antes da aplicação de algoritmos convencionais de aprendizado de máquina. Nossos resultados demonstram que a seleção de características assistida por LLMs alcança um desempenho comparável aos métodos tradicionais de seleção de variáveis, como abordagens de filtro, wrapper ou embutidas, baseando-se exclusivamente no conhecimento do modelo, ou seja, de maneira zero-shot. Esse aspecto é particularmente relevante no presente caso, onde reduzimos o número de características de aproximadamente 1400 para apenas 50, forçando o modelo a selecionar as mais relevantes. Isso evidencia o potencial dessa abordagem para aprimorar a eficiência e a eficácia da descoberta de inibidores da ALK-5, além de direcionar os esforços para métodos mais práticos e escaláveis, facilitando, assim, a implementação de soluções em cenários reais.

**Palavras-chave**: LLM. inhibidores ALK-5. pIC50.

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| ALK-5 | Activin-Like Kinase 5 |
| EDA | Exploratory Data Analysis |
| IC50 | Inhibitory Concentration 50 |
| LLM | Large Language Models |
| MAE | Mean Absolute Error |
| ML | Machine Learning |
| MSE | Mean Squared Error |
| NLP | Natural Language Processing |
| PCA | Principal Component Analysis |
| QSAR | Quantitative Structure-Activity Relationship |
| RMSE | Root Mean Square Error |
| TGF-$\beta$ | Transforming Growth Factor - Beta |
| t-SNE | t-distributed Stochastic Neighbor Embedding |

## CONTENTS

# 1 INTRODUCTION

Cancer represents one of the significant challenges in global public health due to its high mortality rate and the associated social and economic impact (Ferlay *et al.*, 2021; Bray *et al.*, 2024). Therefore, developing novel therapeutic strategies is imperative, particularly those aimed at addressing the uncontrolled proliferation of tumor cells, which is one of the most critical characteristics in the progression of the disease.

Several studies focus on the ALK-5 (Activin-Like Kinase 5) protein due to its essential role in regulating the proliferation and metastasis of tumor cells. These investigations aim to identify novel therapeutic agents capable of inhibiting ALK-5 activity, offering a promising strategy for developing new anticancer therapies (Kargbo, 2022; Zia *et al.*, 2023; Poei *et al.*, 2024). Researchers evaluate multiple chemical compounds to identify potential inhibitors, using the pIC50 value (potency Inhibitory Concentration 50) as a standard metric to quantify their effectiveness. However, obtaining results often requires testing numerous compounds, which is both time-consuming and costly.

Machine learning techniques have been widely applied in drug discovery to optimize this process, reducing the required tests and the associated time and costs (Noviandy *et al.*, 2024; Ion; Nitulescu; Mihai, 2024). One study, in particular, highlights the effective use of machine learning to identify inhibitors of ALK-5 (Espinoza *et al.*, 2021). This underscores the potential of these methods addressing specific targets such as ALK-5.

Given the characteristics of drug discovery problems, such as predicting pIC50 values, machine learning methods often face challenges due to the high dimensionality of features compared to the limited number of samples available (Turzo; Hantz; Lindert, 2022). To address this, many approaches employ feature selection techniques to reduce the feature space while retaining the most relevant information. (Redkar *et al.*, 2020; Labjar; Labjar; Kissi, 2022; Ramapraba *et al.*, 2025). However, these methods require evaluating the available data to assess the relevance of each feature. In our case, this is limited due to the small sample size, which is also costly. This constraint can affect the effectiveness of feature selection, leading to biases in the model and reducing its ability to generalize, as there is not enough data for a robust evaluation.

Large Language Models (LLMs) have gained significant popularity due to their remarkable capabilities in Natural Language Processing (NLP), including tasks such as text generation, sentiment analysis, translation, and question answering (Zhao *et al.*, 2023). Beyond their traditional applications in NLP, recent studies suggest that LLMs can also be utilized in feature selection processes (Li; Tan; Liu, 2025; Jeong; Lipton; Ravikumar, 2024). By leveraging their capacity for contextual learning and the internal knowledge acquired

during pre-training, they can autonomously determine which features are most relevant to the problem without the need for additional data. These results demonstrate their potential to contribute to fields like drug discovery, where selecting the most informative features is crucial for building efficient and accurate predictive models.

In this work, we propose using LLMs as feature selectors to address the challenges of high-dimensional datasets with limited samples in drug discovery, focusing on predicting the pIC50 of ALK-5 inhibitors. Unlike traditional methods that require a minimum number of samples to evaluate features, our approach leverages the models' intrinsic ability to understand context and utilize prior knowledge from pretraining. By employing both zero-shot (Wei *et al.*, 2021) and in-context learning (Dong *et al.*, 2024), we can identify and select the most relevant features without additional task-specific training. This method improves the efficiency and accuracy of predictive models in critical tasks and optimizes the drug discovery process by reducing the number of variables and using all available samples for model training.

## 1.1 Objectives

In this context, the present work aims to explore the use of Large Language Models (LLMs) in drug discovery, providing a scalable and accurate solution for identifying critical features in high-dimensional datasets. Specifically, this study focuses on predicting the pIC50 values of ALK-5 inhibitors, a crucial step in developing effective anticancer therapies.

The objectives of this work are as follows:

- Develop a methodology to leverage LLMs for feature selection in datasets associated with ALK-5 inhibitors.
- Evaluate the effectiveness of the proposed approach in predicting pIC50 values, using metrics such as mean squared error, mean absolute error and determination coefficient.
- Compare performance between the LLM-based feature selection method and traditional feature selection methods, such as filter, wrapper and embedded methods.
- Assess the biological relevance of the selected features, ensuring their interpretability and applicability in the design of new ALK-5 inhibitors.
- Provide insights and recommendations for integrating LLMs into drug discovery pipelines, highlighting their scalability and potential impact on similar problems.

## 1.2 Contributions

The main contributions of this work are summarized as follows:

- Novel application of LLMs: We propose a novel use of Large Language Models (LLMs) as feature selectors to address the challenges of high-dimensional datasets in drug discovery, specifically focusing on predicting the pIC50 of ALK-5 inhibitors.
- Methodological pipeline: We design and implement a scalable pipeline that leverages the contextual understanding capabilities of LLMs to identify the most relevant features in complex datasets.
- Performance evaluation: We conduct an extensive evaluation of the proposed approach, demonstrating its superiority over traditional dimensionality reduction techniques in terms of predictive accuracy and feature interpretability.
- Biological insights: We provide a detailed analysis of the selected features, highlighting their biological relevance and potential applications in the design of new ALK-5 inhibitors.
- Generalizability: We discuss the broader implications of using LLMs in drug discovery, offering insights and recommendations for applying this methodology to other targets and tasks.

## 1.3 Document Organization

The remainder of this text is organized as follows. Chapter 2 outlines the key concepts and reviews the related work. Chapter 3 details the proposed methodology, evaluation metrics, and datasets used in the experiments. Chapter 4 presents the experiments and their results. Finally, Chapter 5 concludes with final considerations and outlines future research directions.

## 1.4 Concluding Remarks

In summary, the introduction, objectives, and contributions presented in this work establish the foundation for exploring the potential of Large Language Models (LLMs) as feature selectors in drug discovery. By addressing the challenges associated with high-dimensional datasets and focusing specifically on predicting the pIC50 of ALK-5 inhibitors, this study aims to advance the state of the art in both computational methods and their biological applications.

The outlined objectives highlight a clear and focused research direction, while the contributions emphasize the novelty and impact of our approach. These elements set the stage for the detailed methodology and experimental results to follow, where the proposed techniques are evaluated and their implications for drug discovery are explored.

# 2 THEORETICAL BACKGROUND

In this chapter, divided into two parts, we first introduce the essential concepts needed to understand and describe the proposed methodology in the next chapter. Afterward, we review existing studies that address the same or similar problems, highlighting their main contributions and limitations.

## 2.1 Background

In this section, we focus on the essential concepts for the development of this work. We begin by presenting the dimensionality problem, which motivates our research, and the most typically employed techniques to address it. Next, we provide a detailed discussion of feature selection, which constitutes our primary technique of study, as well as the various methods associated with this approach. Finally, we review the fundamentals needed to understand Large Language Models (LLMs), covering their Transformer-based architecture and the existing techniques for improving them without retraining their weights, which is especially relevant since we will employ this model in our research.

### 2.1.1 Measuring ALK-5 inhibition with pIC50

In the search for new effective drugs, it is essential to act against the target molecule, the one responsible for triggering the harmful effects. Various inhibitory molecules capable of interacting with the target molecule and mitigating its harmful activity are evaluated to mitigate these effects. In this process, the chemical characteristics of each candidate are examined and its impact on the target molecule is studied through the measurement of the $IC_{50}$, this measurement indicates the concentration necessary to inhibit the activity of the target molecule by 50% (Caldwell *et al.*, 2012); thus, a lower $IC_{50}$ means a higher inhibitory potency. Alternatively, it is possible to employ the $pIC_{50}$, defined in Formula 2.1, which transforms the $IC_{50}$ to a logarithmic scale, thus facilitating the handling of wide ranges of inhibitory concentrations, since the scale of many compounds is exponential. In this context, a higher value of $pIC_{50}$ indicates that the molecule possesses greater efficacy in interacting with the target. The use of $pIC_{50}$ allows the establishment of precise quantitative relationships that accelerate the identification of effective compounds for the development of new drugs.

$$pIC_{50} = -\log_{10}(IC_{50}) \tag{2.1}$$

Therefore, given the importance of pIC50 in evaluating the effectiveness of inhibitory compounds, we have chosen ALK-5 (Activin Receptor-Like Kinase 5) as the target of

this study. ALK-5 is a type I receptor for Transforming Growth Factor-Beta (TGF-$\beta$), which, as a transmembrane protein, receives extracellular signals like TGF-$\beta$ and transmits them into the cell, regulating essential processes, such as tissue proliferation and repair, essential for maintaining homeostasis and facilitating processes such as healing (Mansour *et al.*, 2024). However, in tumor contexts, ALK-5 promotes cell invasion and metastasis. Consequently, modulating ALK-5 activity presents a promising therapeutic strategy for cancer treatment. In this study, we will use $pIC_{50}$ to evaluate the inhibitory efficacy of various molecules in our dataset, which will be discussed in the next chapter. This approach will help us identify the most effective compounds for inhibiting ALK-5 and developing new antitumor drugs.

### 2.1.2 The Curse of Dimensionality

To predict the pIC50 of various molecules against ALK-5, we can employ computational approaches such as Machine Learning (ML) or Quantitative structure-activity relationship (QSAR) models. These methods allow us to avoid experimental testing, which is often costly in time and resources. Such approaches establish a relationship between chemical and structural descriptors of compounds and their biological activity. However, the high number of molecular features generates a high dimensionality space, leading to a phenomenon known as the Curse of Dimensionality (Peng; Gui; Wu, 2023).

This phenomenon occurs because as the number of features (dimensions) increases, the data becomes more and more sparsely distributed in the multidimensional space. Consequently, distance metrics, fundamental to many ML algorithms, lose their ability to discriminate between near and far points. In the context of pIC50 prediction, this is compounded by the complex and highly correlated nature of molecular descriptors, increasing dimensionality and introducing noise and redundancy. In addition, the search space grows exponentially with each new dimension, increasing the computational demand and raising the risk of overfitting, limiting the model's ability to generalize to new molecules.

To address this problem, several techniques are available to reduce the dimensionality of the data and thus improve model performance. Among them are dimensionality reduction methods, such as Principal Component Analysis (PCA) or t-SNE. These techniques transform the data into a lower dimensional space, preserving the most relevant information, but with the counterpart of losing information from the original variables since these are combined or transformed into new dimensions (Cunningham; Ghahramani, 2015). For example, in PCA, the original characteristics are linearly mixed to form principal components, which implies that the new dimensions no longer directly represent the initial variables. This affects the model's interpretability since it is impossible to attribute the result to a specific variable from the original set.

Alternatively, feature selection identifies and retains only the most informative

features, eliminating redundancies and noise but preserving the original variables. This strategy preserves the interpretability of the model since the selected features correspond directly to the initial variables, making it easier to understand how they contribute to the result and their relationship with the study problem. In summary, both approaches simplify the search space and improve the generalization of the model and its computational efficiency. The choice between one or the other will depend on whether interpretability is crucial, such as keeping the original data, in which case feature selection will be the most appropriate option.

### 2.1.3 Feature Selection

Explaining why one component is more relevant than others is essential, as interpretability is as critical as model accuracy in discovering new drugs. In this context, feature selection methods have become increasingly important because they enable the construction of more efficient machine learning models. These methods accelerate the drug discovery process by eliminating redundant or irrelevant information and enhancing the models' interpretability and generalization capabilities during testing.

Feature selection methods are commonly classified into filtering, wrapper, and embedding. **Filtering methods** are techniques that evaluate the relevance of features before training the model, using statistical criteria independent of the learning algorithm. These features are filtered according to a predefined threshold or by selecting a specific number of the most relevant features.

One straightforward filter method uses *Pearson correlation* to measure the linear relationship between features and the target. Given a feature matrix $X \in \mathbb{R}^{n \times d}$, where $n$ is the number of samples and $d$ is the number of features and a target vector $Y \in \mathbb{R}^n$, the Pearson correlation is computed between each column of $X$ and $Y$. The Pearson correlation is defined in Formula 2.2, where $X_j$ represents the $j$-th column (feature) of $X$, $\text{cov}(X_j, Y)$ is the covariance between $X_j$ and $Y$, and $\sigma_{X_j}$ and $\sigma_Y$ are the standard deviations of $X_j$ and $Y$, respectively. Pearson correlation measures the linear relationship between two variables. Given a feature matrix $X \in \mathbb{R}^{n \times d}$, where $n$ is the number of samples and $d$ is the number of features and a target variable vector $Y \in \mathbb{R}^n$, the Pearson correlation is computed between each column of $X$ and $Y$.

$$\rho(X_j, Y) = \frac{\text{cov}(X_j, Y)}{\sigma_{X_j} \sigma_Y} \tag{2.2}$$

On the other hand, a different filtering technique uses *Mutual information* to measure the statistical dependence between a feature and the target variable, specifically evaluating whether knowing the value of a feature $X_j$ provides meaningful information about the target $Y$. Given the feature matrix $X \in \mathbb{R}^{n \times d}$ with $n$ samples and $d$ features,

along with the target vector $Y \in \mathbb{R}^n$, mutual information quantifies the shared information between each feature column $X_j$ and $Y$, as defined in Formula 2.3. The mutual information is calculated as the sum over all pairs $(x, y)$ of the joint probability $p(x, y)$ multiplied by the logarithm of the ratio between the joint probability and the product of the marginal probabilities, $p(x)$ and $p(y)$. This measure reflects how knowledge of one variable reduces uncertainty about the other.

$$I(X_j, Y) = \sum_{x_j \in X_j} \sum_{y \in Y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)\, p(y)} \tag{2.3}$$

The resulting mutual information scores provide a ranking of features by their predictive potential, analogous to but more general than the Pearson correlation. While both methods operate on the same matrix structure and share similar computational efficiency, mutual information extends beyond linear relationships to capture any form of statistical dependence. This advantage comes with increased computational complexity in estimating the probability distributions. As with correlation-based filtering, this approach maintains the limitations of not accounting for feature interactions or downstream model performance, though it remains valuable for initial feature screening where non-linear relationships may be significant.

In **Wrapper methods**, the feature selection process is directly associated with the optimization objective of the model. Consider a predictive model $\mathcal{M}$ with parameters $\theta$ trained on a subset of features $S \subseteq \{X_1, ..., X_j\}$. The goal of the Wrapper approach is to find the optimal feature subset $S^*$ that minimizes the loss function while maintaining or improving model performance, as described in Formula 2.4, where $X_S$ is the matrix containing only the columns from subset $S$, and $\mathcal{L}$ represents the loss function.

$$S^* = \arg \min_{S \subseteq \{X_1, ..., X_j\}} \mathcal{L}(\mathcal{M}_\theta(X_S), Y) \tag{2.4}$$

Wrapper methods commonly employ one of two main search strategies: Forward selection or Backward elimination, which differ in their feature space exploration.

*Forward selection* begins with an empty feature set and iteratively adds the most promising features one at a time. It selects the feature that yields the best model performance at each step. Subsequent iterations add the feature that, when combined with the currently selected subset, provides the greatest improvement in predictive accuracy. This process continues until a stopping criterion is met, such as when additional features no longer significantly enhance performance.

Contrarily, *Backward elimination* follows the opposite approach, starting with the complete set of features and progressively removing the least relevant ones. Initially, the method evaluates the impact of removing each feature individually, discarding the

one whose exclusion minimally affects performance. This process continues until further removals would substantially reduce model accuracy.

Both strategies assess feature subsets by training the model and evaluating performance, typically using cross-validation to mitigate overfitting. Forward selection is generally more computationally efficient for high-dimensional datasets, as it starts with few features. In contrast, Backward elimination can be more suitable when most features are relevant since it accounts for feature interactions from the outset. The decision on which method to use depends on the specific characteristics of the problem and the available computational resources.

Finally, **Embedded methods** perform feature selection by identifying an optimal feature subset $S^* \subseteq \{X_1, \ldots, X_j\}$, evaluating feature importance during model training. These methods integrate model optimization with automatic relevance assessment, learning both the predictive relationship and feature significance through a feature importance score, as shown in Equation 2.5. In this equation, $\psi_j$ represents the importance score derived for the feature $X_j$ through a function $f(\cdot)$ of the learned parameters $\theta_j$.

$$\psi_j = f(\theta_j, X_j, Y) \quad \forall X_j \in S \tag{2.5}$$

The selection mechanism ranks features by their $\psi_j$ scores and selects those whose importance score exceeds a threshold $\tau$ to obtain the final subset $S^*$, as expressed in Equation 2.6. The importance scores $\psi_j$ offer valuable interpretability, reflecting the actual contribution of $X_j$ to predicting $Y$.

$$S^* = \{X_j \mid \psi_j > \tau\} \tag{2.6}$$

Compared to other feature selection approaches, embedded methods are computationally efficient, as importance computation is seamlessly integrated into model training. The selected features capture both individual predictive power and synergistic effects within the feature space, as all importance assessments occur in the context of the complete model. In practice, this approach often combines threshold-based selection with top-$k$ ranking, where features are chosen based on $\psi_j > \tau$ or by selecting the $k$ largest $\psi_j$ values.

This method is particularly advantageous in high-dimensional scenarios, as it preserves the model's discriminative power while reducing dimensionality. Since the importance scores are derived directly from the learning process, they tend to produce more robust feature subsets compared to filter methods while avoiding the computational expense typical of wrapper methods. Moreover, the automatic nature of this selection process makes it adaptable to various data types and scales, provided that the underlying model yields reliable importance metrics.

In summary, the presented feature selection methods offer interpretability and dimensionality reduction, key advantages in drug discovery. However, their reliance on statistical techniques or machine learning models limits scalability in high-dimensional settings due to the curse of dimensionality, as molecular data often presents a significantly higher number of features compared to samples. This challenge motivates the search for a model capable of handling large datasets while providing insights into the selection rationale.

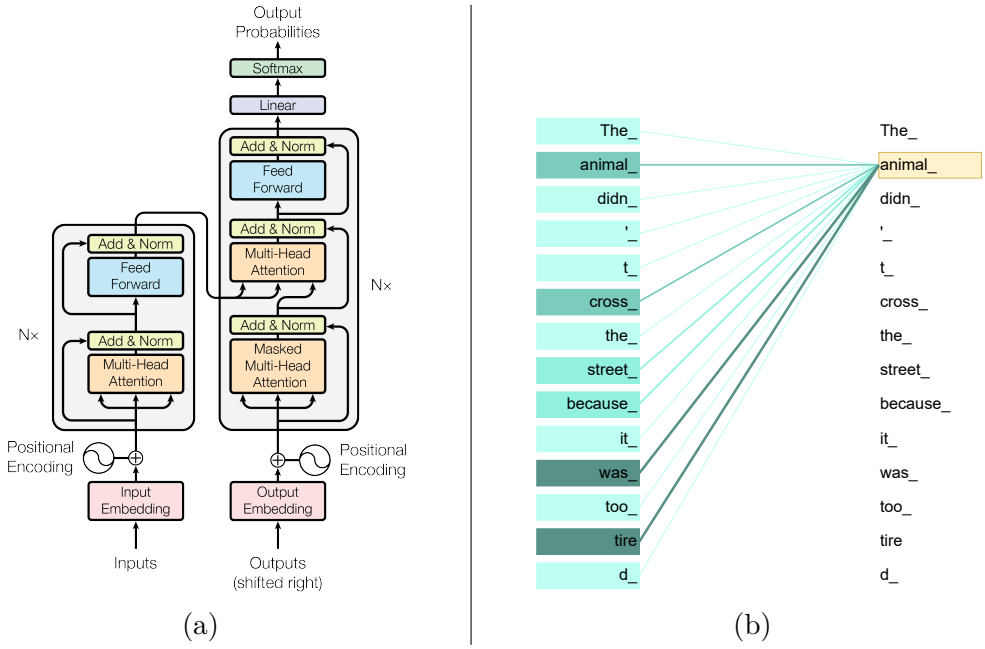
### 2.1.4 Large Language Models

Large Language Models (LLMs) have revolutionized natural language processing (NLP), enabling machines to understand and generate text fluently and coherently (Fan *et al.*, 2023; Minaee *et al.*, 2024). Built upon the Transformer architecture (Vaswani, 2017), these models are characterized by their large scale and extensive training data, allowing them to capture linguistic nuances more effectively (Zhao *et al.*, 2023; Ren *et al.*, 2023; Kaplan *et al.*, 2020).

Transformers, introduced in (Vaswani, 2017), are deep learning models that quickly gained prominence due to their innovative architecture. Unlike earlier models based on Recurrent Neural Networks (RNN), such as LSTM (Graves, 2012) and GRU (Cho, 2014), Transformers utilize attention mechanisms (Niu; Zhong; Yu, 2021). This architectural shift not only enhances generalization and the understanding of complex relationships but also supports parallel training, positioning Transformers as the new state of the art in NLP. Over time, their application has expanded beyond NLP to other areas, such as computer vision, with models like Vision Transformers (Dosovitskiy, 2020).

Inspired by sequence-to-sequence (seq2seq) models (Sutskever, 2014), the Transformer architecture employs an encoder-decoder structure, both leveraging attention mechanisms to efficiently process sequences. As illustrated in Figure 1a, the model transforms each input word into a dense vector (embedding), augmented with positional encodings to preserve the sequence order. Unlike traditional recurrent models, this approach enables efficient parallel processing, accelerating model training and scalability.

The embeddings then pass through a multi-head attention layer, allowing the model to simultaneously focus on various parts of the input sequence. As shown in Figure 1a, the output from this layer is processed through a feed-forward network before being passed to the decoder. The decoder structure mirrors the encoder, with the key distinction of incorporating a masked multi-head attention mechanism. This mechanism ensures that the decoder only attends to preceding positions in the sequence during output generation, crucial for tasks like next-word prediction, where generating a word relies solely on the preceding context. Finally, a linear layer followed by a softmax function at the decoder's output produces the probability distribution over the vocabulary tokens.




(a) (b)

Figure 1 – (a)Transformer architecture. (b) An example of the attention mechanism: one can see how attention is focused in different parts of the text without necessarily being next to the subject being spoken about. In this case, the 'animal' is 'tired'. Source: Adapted from (Vaswani, 2017)

Both the encoder and the decoder leverage the multi-head attention layer to simultaneously focus on multiple parts of the input sequence. This capability is achieved by dividing the attention mechanism into several "heads". To implement this approach, three fundamental components are used: Query ($Q$), Key ($K$), and Value ($V$). As shown in equation 2.7, these vectors are combined to determine the attention that each word should assign to other words within the sequence. The Query vector represents the word that seeks to establish relationships with others by comparing itself with the Keys of the other words, while the resulting relationships are weighted using a softmax function. Finally, the computed weights are applied to the corresponding Values, generating a numerical representation that captures the attention each word receives based on its context, as illustrated in Figure 1b.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{2.7}$$

In essence, the core functionality of a Transformer-based model, and consequently of an LLM, is to predict the most probable next word given a context. This is achieved through the decoder's masked multi-head attention mechanism, which ensures that the model only considers preceding words when generating the next token. The model generates coherent and contextually appropriate text by analyzing the entire context word by word, even in long sequences. It learns complex patterns, linguistic structures, and contextual relationships through training on vast amounts of data, progressively evolving into a powerful LLM

capable of natural language understanding and generation. This capability has attracted significant interest from both academia and industry. Driven by commercial demand, companies have developed advanced LLMs such as OpenAI's GPT (Brown, 2020; Achiam *et al.*, 2023), Google's PaLM (Chowdhery *et al.*, 2023), and Meta's LLaMA (Touvron *et al.*, 2023).

Despite their ability to generate coherent text, LLMs face challenges when applied to domain-specific tasks. Their knowledge is inherently limited by the scope of their training data, which, while extensive, often lacks depth in specialized areas (Liu *et al.*, 2024). Consequently, they may produce superficial or inaccurate responses in fields requiring technical expertise, such as communication networks (Maatouk *et al.*, 2024) or medicine (Singhal *et al.*, 2023; Thirunavukarasu *et al.*, 2023). Moreover, their significant computational and energy requirements raise sustainability concerns (Faiz *et al.*, 2023).

To overcome these limitations, researchers have developed various methods to enhance LLM performance without requiring retraining or modifying the model. Retrieval-Augmented Generation (RAG) integrates external knowledge sources to supplement the model's responses, allowing it to access updated or domain-specific information. Additionally, few-shot and zero-shot learning techniques enable the model to adapt to new tasks with minimal examples, leveraging its pre-trained knowledge efficiently. Another practical approach is prompt engineering, which involves crafting inputs strategically to elicit more accurate and contextually relevant responses. These methods significantly enhance the versatility and performance of LLMs, enabling them to tackle a broader range of tasks without altering the underlying model.

In conclusion, in the field of drug discovery, LLMs offer significant advantages over traditional feature selection methods, which primarily rely on statistical criteria or model training. Unlike conventional approaches, LLMs can contextualize features by incorporating scientific literature and biomedical data, providing interpretable explanations of their relevance. This capability helps mitigate challenges related to high dimensionality and supports more informed decision-making by justifying feature importance. Therefore, LLMs represent a promising tool in drug discovery, offering a more nuanced and context-aware approach than traditional methods.

## 2.2 Related Works

Feature selection has been extensively studied due to its relevance in building predictive models, especially in contexts involving high-dimensional and complex data. Over the years, various studies have addressed this problem using several approaches, from traditional statistics and machine learning or deep learning techniques to innovative strategies based on LLMs. Although each approach provides valuable perspectives, they also leave certain areas unexplored, suggesting opportunities for future developments.

Firstly, studies like the one presented in (Li; Tan; Liu, 2025) highlight using LLMs to identify relevant features in unstructured datasets. This approach leverages the semantic capabilities of these models to enhance both accuracy and computational efficiency. Similarly, research such as (Jia *et al.*, 2024) and (Yang *et al.*, 2024) combines semantic extraction with traditional techniques; these methods assign initial scores to features using LLMs and then refine the selection with classic algorithms such as clustering, recursive elimination, and regularized regression.

On the other hand, the study published in (Espinoza *et al.*, 2021) adopts a hybrid methodology that combines initial statistical filtering with an iterative wrapper-based process. This approach has proven effective in complex domains, such as genomic data analysis, by reducing dimensionality without compromising model performance.

Another line of research focuses on deep learning-based techniques, such as the approach proposed in (Brown, 2020), which utilizes attention mechanisms in pre-trained models to assess feature relevance. This approach enables the capture of complex nonlinear interactions, thereby increasing the model's predictive capacity, although sometimes at the cost of reduced interpretability and increased computational load.

An innovative approach is presented in (Jeong; Lipton; Ravikumar, 2024), where the authors propose a method that leverages LLMs to perform feature selection using only the feature names and a brief description of the prediction task. Despite the limited input, the models effectively identify the most relevant features, achieving performance comparable to traditional methods. This approach highlights the potential of LLMs in feature engineering by capturing semantic context with minimal information.

A critical perspective on the use of LLMs for feature selection was presented in the study (Küken; Purucker; Hutter, 2024), where the authors examine how LLMs tend to generate simple features, often neglecting more complex operators that are essential for advanced predictive tasks. This bias towards simplicity can result in lower performance when directly using features produced by LLMs. The findings emphasize the need for strategies that integrate the semantic capabilities of LLMs with more robust feature selection techniques, particularly in scenarios where capturing complex relationships between variables is crucial for model performance.

In this context, our work aims to address some of the gaps identified in the literature. Additionally, existing approaches do not validate using LLMs on datasets characterized by very high dimensionality and limited sample size. To bridge this gap, we propose leveraging the semantic capabilities of LLMs to extract relevant information from structured data while capturing complex nonlinear relationships. This approach aims to enhance both the interpretability and scalability of feature selection. Our goal is to provide a robust and adaptable solution that complements and surpasses existing methods' limitations.

# 3 METHODOLOGY

In this chapter, we describe our methodological approach for building a feature selection pipeline using a pre-trained LLM. We detail the prompts designed for the LLM, the process of refining and validating its outputs for coherence, and the comparison of these outputs with state-of-the-art feature selection methods. Additionally, we explain how the selected features were integrated into our evaluation framework, including the metrics and datasets used to assess their effectiveness.

## 3.1 Transformers as Feature Selectors

Our proposed pipeline consists of two main stages: feature selection and subsequent validation using the selected features. This structured approach enables a direct comparison between traditional feature selection methods and a novel method that uses a pre-trained LLM model (the Foundational model) as a feature selector, as illustrated in Figure 2. This process aims to evaluate the effectiveness of the employed methods, establishing a robust framework for comparison and analysis.

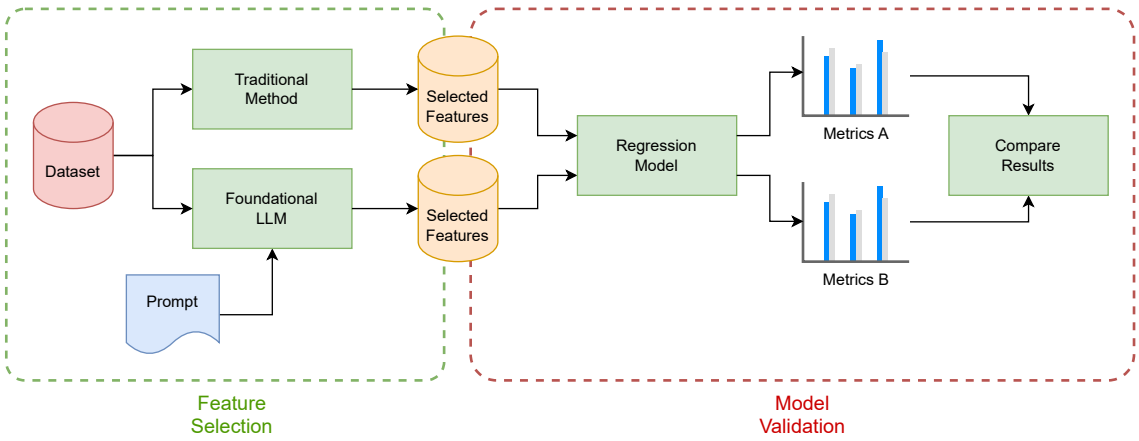



Figure 2 – The proposed pipeline has two stages: feature selection and validation, combining classical methods with our novel pre-trained LLM-based method. Source: Author.

To ensure data quality, we performed comprehensive Preprocessing. First, we validated the dataset by identifying variable types, checking for missing values, and calculating key descriptive statistics. Given the very high dimensionality, we focused on measures that effectively summarize the data without excessive computational cost. These included the mean and standard deviation to assess central tendency and variability, as well as the minimum, maximum, and interquartile range (IQR) to detect potential outliers.

Then, in the first stage, we explore classical feature selection methods and a novel approach based on a pre-trained LLM model. The classical methods considered include

techniques from the literature presented in the previous chapter, such as *filtering*, *wrapper*, and *embedding*. We developed specific functions for each method that receives data and the number of $k$ features to return, identifying the top $k$ features. Depending on the method, it either uses a machine learning model or a statistical metric. For *filtering* methods, we implemented Pearson correlation and mutual information. For *wrapper* methods, we used forward selection and backward elimination, incorporating a Ridge model internally. Finally, we employed a Random Forest regressor for *embedding* methods.

Simultaneously, we implemented an innovative approach for our proposed method using a pre-trained LLM model. To achieve this, we used the most widely adopted and powerful pre-trained LLM models available. Specifically, we utilized OpenAI's API with the GPT-4, GPT-o1, and GPT o3-mini models, as well as the DeepSeek R1 model, a Chinese model known for its impressive capabilities, through its API.

For this approach, the LLM is guided by a carefully designed prompt that shapes the model's output. Figure 3 shows an example of such a prompt. As illustrated, the prompt should include three key components: the context, which informs the model about the task at hand; the instruction or query, specifying the response or action required; and, optionally, example inputs that serve as references. Additionally, in some cases, it may also include the desired output format.

**Context:**
*You are a data scientist and your goal is to assign a relevance score to each feature based on its predictive power or importance within the given dataset.*

**Instruction:**
*I will provide you with a dataset. Your task is to analyze the dataset and calculate a relevance score for each feature. Return a ranked list of features with their corresponding scores.*

**Sample:**
*Input: A dataframe containing numerical and categorical features.*
*Output: A table with two columns: Feature Name and Relevance Score.*

Figure 3 – Sample of prompt. Source: Author.

Furthermore, we explored various strategies to obtain a score for the features or columns of the dataset. To achieve this, we followed the methodology described in (Li; Tan; Liu, 2025), structuring the input data to enable the LLM to produce a score, helping us identify the top $k$ most relevant features. These strategies include providing data samples, specifying the entire dataset, or presenting features individually.

The complete process followed with the pre-trained LLM model is depicted in Figure 4. IInitially, the original data goes through a preprocessing stage, after which

samples are selected to include in the prompt that will be read by the LLM. The model then generates a score for each feature and selects the top-$k$ most relevant features, creating the set of selected features. This allows us to extract the relevant columns from the preprocessed data. These selected columns are then used as input for the baseline model, which, in our case, is a ridge regression model. Ridge regression is chosen for its ability to handle multicollinearity and produce robust coefficients, making it well-suited for scenarios where data may exhibit linear relationships with noise or redundancy.

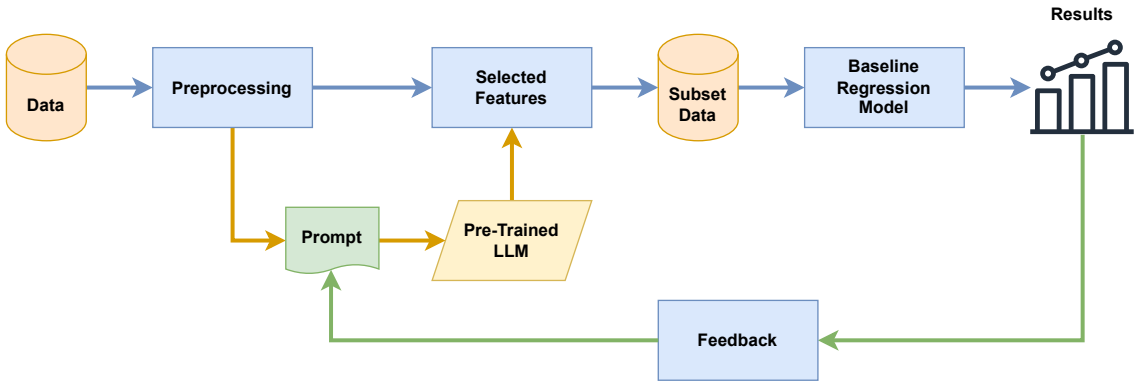



Figure 4 – Pipeline for feature selection using our pre-trained LLM. Data is preprocessed, a sample is extracted for the prompt, the LLM returns a score, and the top-$k$ features are evaluated in the baseline regression model. Metrics provide feedback to refine the prompt iteratively. Source: Author.

Given the limited data samples, we opted to perform cross-validation to ensure the model's reliability. Before this, we conducted a grid search to optimize the model's hyperparameters for the given dataset. Finally, we validated the model's performance using metrics such as Mean Squared Error ($MSE$), Mean Absolute Error ($MAE$), and the coefficient of determination ($R^2$), as explained in the following section. This evaluation allowed us to compare the results obtained with traditional feature selection methods.

When we observed that the LLM's performance was inferior to that of traditional methods, we adjusted the prompt using prompt engineering techniques. This iterative process was repeated several times until we identified a prompt that yielded results comparable to those obtained using conventional methods.

With the finalized pipeline and fine-tuned prompt, a reliable feature selection process is now ready for evaluation using the metrics described in the following section.

## 3.2 Metrics

To evaluate the results obtained from traditional feature selection methods and those generated by the LLM, we will not conduct a direct comparison. Instead, we will employ a baseline regression model, specifically an L2-regularized linear regression (Ridge regression), to validate the metrics described below.

### 3.2.1 Coefficient of Determination ($R^2$)

The coefficient of determination, commonly denoted as $R^2$, measures the proportion of variance in the dependent variable that can be explained by the independent variables. Its value ranges between 0 and 1, where an $R^2$ close to 1 indicates that the model fits the data well, effectively capturing the variability. In contrast, a value near 0 implies that the model poorly explains the observed variance.

The $R^2$ metric is formally defined in Equation 3.1, where $y_i$ denotes the observed value, $\hat{y}_i$ the value predicted by the model and $\bar{y}$ the mean of the observed values. A higher $R^2$ signifies a more significant predictive capability, meaning the selected features accurately reflect the relationship between explanatory variables and the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.1}$$

### 3.2.2 Mean Squared Error (MSE)

Mean Squared Error (MSE) measures the average squared difference between the observed values and the model's predictions, as expressed formally in Equation 3.2. This metric assigns a heavier penalty to larger errors by squaring the deviations, making it particularly sensitive to outliers or significant prediction inaccuracies. Consequently, a lower MSE indicates superior model performance, reflecting a closer alignment between predicted and actual values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

### 3.2.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) quantifies the average absolute differences between observed and predicted values, as defined in Equation 3.3. Unlike MSE, MAE does not disproportionately penalize large errors, making it less sensitive to extreme outliers and thus especially valuable when evaluating model performance based on absolute prediction accuracy. Additionally, MAE offers intuitive interpretability by directly indicating the average magnitude of prediction errors. Lower MAE values reflect more precise and reliable model predictions.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

Together, these metrics offer a comprehensive evaluation of model performance. By applying them simultaneously across different feature subsets, we can objectively compare the datasets obtained from traditional feature selection methods with those produced by the LLM.

## 3.3 Dataset

In this work, we utilized a dataset initially introduced in the study "Evaluating Deep Learning Models for Predicting ALK-5 Inhibition" by (Espinoza *et al.*, 2021). The dataset was explicitly designed to predict the biological activity (pIC50) of ALK-5 inhibitors, potential candidates for cancer treatment. The choice of this dataset is motivated by its relevance to the study of drug discovery models and its suitability for evaluating feature selection techniques.

The dataset was obtained through one of the coauthors of the original study, as it is not publicly available. It was constructed by collecting compounds with known activity against the TGF-beta receptor type I (ALK-5) from the ChEMBL database (target ID CHEMBL4439). After a thorough data curation and preprocessing performed by the original authors, the final dataset consists of 545 unique molecules, each represented by a set of molecular descriptors calculated using the Mordred Python library. These descriptors capture the structural, electronic, and physicochemical properties of the molecules, resulting in over 1,453 calculated features.

Although the dataset was provided in a preprocessed form, we performed additional analysis and standardization to adapt it to our regression model, as detailed in the previous section. This step ensured the numerical stability and compatibility required for the subsequent modeling tasks.

The decision to use this dataset was driven by its comprehensive nature, high dimensionality, and specific design to evaluate models predicting biological activity. Additionally, performing our analysis and standardization allowed us to adapt the dataset for optimal use within our regression framework. By leveraging this well-prepared and further refined dataset, we aimed to evaluate the effectiveness of feature selection techniques using pre-trained LLM models.

# 4 EXPERIMENTAL RESULTS

In this section, we present the results of our proposed methodology. We begin with an Exploratory Data Analysis (EDA) to assess the dataset's quality and structure. Next, we evaluate the baseline regression model (Ridge regressor) using feature subsets obtained through traditional selection methods and our LLM-based approach, highlighting the performance improvements achieved with the LLM-based technique. Finally, we discuss the iterative prompt optimization process and its role in enhancing predictive accuracy.

The results presented here aim to validate the effectiveness of the proposed methodology in accurately predicting the biological inhibition activity of $ALK-5$ inhibitors, while also highlighting the challenges and improvements encountered throughout the process.

## 4.1 Exploratory Data Analysis (EDA)

Before model training, we performed an EDA to understand the structure and distribution of the dataset.

Initial EDA revealed key characteristics of the target variable, $pIC_{50}$. he distribution showed a mean of 7.184 and a median of 7.366, indicating a slightly left-skewed distribution (mean $<$ median). Values spanned from 4.102 to 9.244, with a standard deviation of 0.953, suggesting moderate variability in the biological activity of the compounds. Figure 5 illustrates the distribution of $pIC_{50}$ confirming the left skew, with most data clustered near the median.

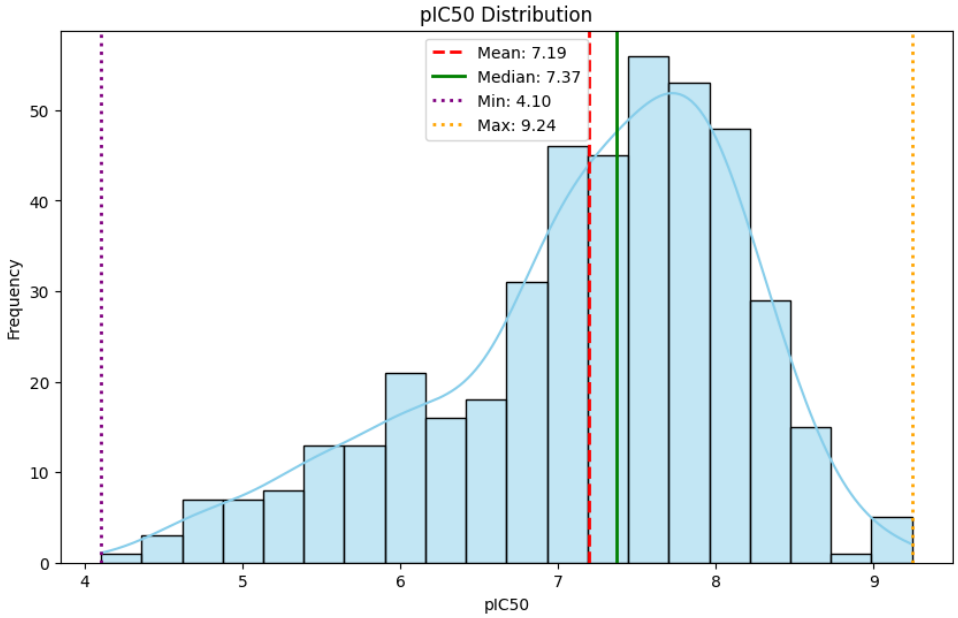



Figure 5 – Distribution of $pIC_{50}$ values showing left-skewed asymmetry. Source: Author.

Subsequently, one strategy we considered was to analyze the variables that exhibited the highest correlation to determine whether they showed any relationship with the target variable. We calculated the Pearson correlation coefficient to assess the relationships between variables to achieve this. This method was also employed in the filtering approach, as we aimed to evaluate the potential accuracy of the baseline model when using this technique.

We then generated a table of absolute correlations, sorting the values from highest to lowest, as presented in Table 1. We observed that only three variables reached a correlation of approximately 0.50, indicating a moderate linear relationship between these highly correlated variables and the target variable.

Table 1 – Comparison of the top 5 most correlated (left) and least correlated (right) features with $pIC_{50}$.

| Feature | Correlation |
|---|---|
| AMID_N | 0.512 |
| MID_N | 0.506 |
| nN | 0.503 |
| SMR_VSA3 | 0.469 |
| SssNH | 0.450 |

| Feature | Correlation |
|---|---|
| AATS2dv | 0.000 |
| MATS3c | 0.001 |
| FPSA4 | 0.001 |
| SM1_Dt | 0.001 |
| nBondsA | 0.001 |

Afterward, we plotted the three variables that exhibited a correlation greater than 0.50 to examine their behavior in relation to the target variable, as illustrated in Figure 6. We found that all three variables displayed a positive (direct) correlation. Additionally, the last two plots revealed that the variables were discrete, while the target variable was continuous. This can create particular challenges in the Ridge regression model, as this type of model assumes a continuous relationship between the features and the target. If the feature is discrete, the model may overestimate the importance of some specific values or assign significant coefficients inappropriately, especially if the discrete values are not uniformly distributed. Therefore, it is important to consider this characteristic when selecting and processing the variables.

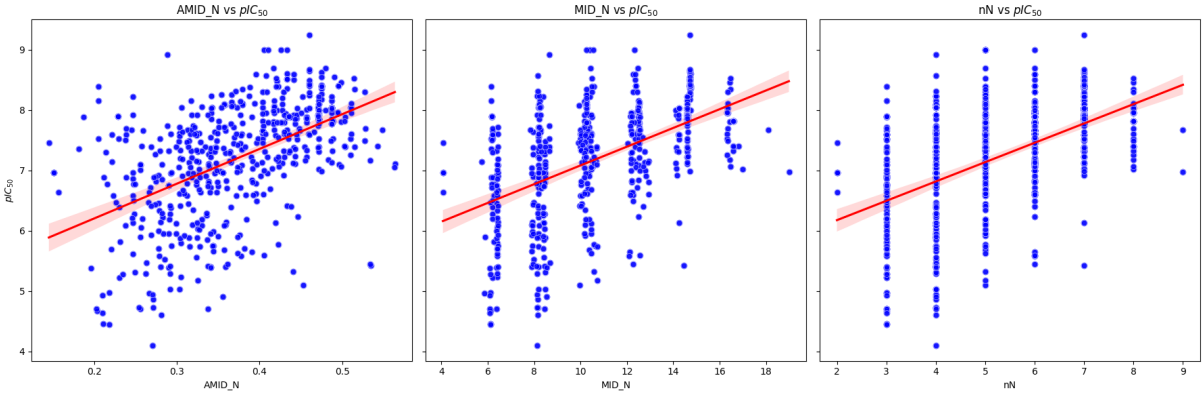
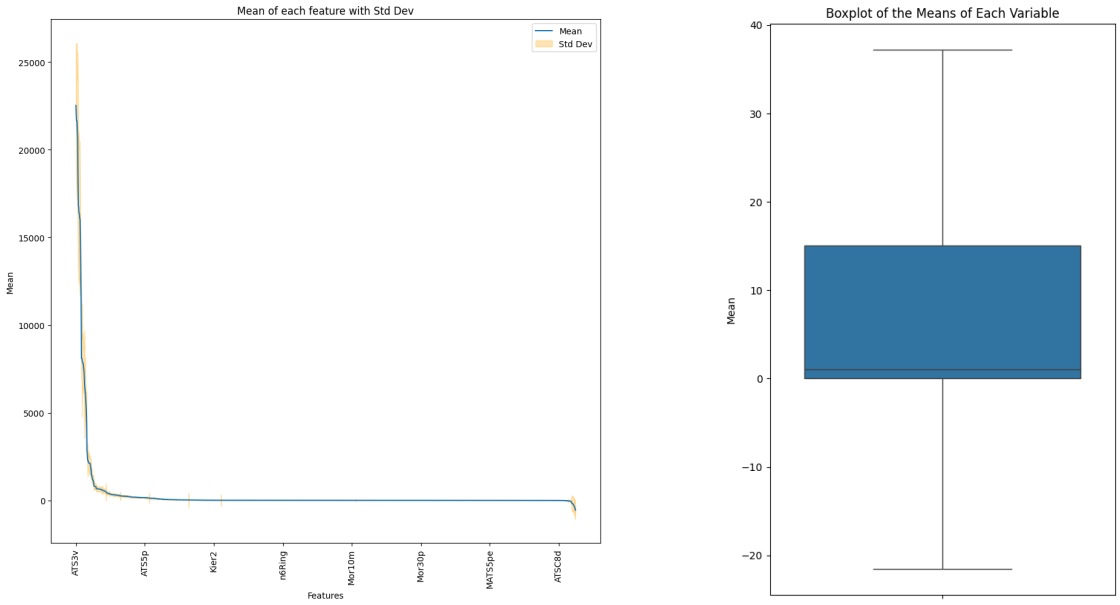



Figure 6 – Top Three Most Correlated Features versus $pIC_{50}$. Source: Author.

Given that we have a large number of variables but very few samples (almost one-third of the total columns), the classical plots obtained, such as individual heatmaps or boxplots, were not helpful, as they did not provide relevant information. Moreover, due to the scarcity of samples, we did not consider removing outliers.

For this reason, we decided to calculate the mean and standard deviation of each variable, as shown in Figure 7a. We observed that the variables have significantly different scales, with some taking small or even negative values while others reach magnitudes exceeding ten thousand. Additionally, the ranges of variation differ significantly between variables. Therefore, we decided to standardize the data before using it in the model. This decision is reinforced by observing the boxplot in Figure 7b, where the means of the variables are generally between $-20$ and 40, while many others fall outside this range. Standardization is particularly important in our case, as we are working on a regression problem and want to prevent scale differences from affecting the model's performance.




(a) (b)

Figure 7 – (a) shows feature means, while (b) shows a boxplot of means without outliers. Source: Author.

Finally, we removed non-numerical data, such as the SMILE representation and the molecule ID. No duplicates or missing values were found since the data originated from a preprocessed source. We split the dataset into training and testing sets with a proportion of 80 and 20 percent, respectively. These steps were performed solely before conducting the experiments and applying the feature extraction methods presented in previous chapters.

## 4.2 Evaluation of Feature Selection Techniques

After implementing all feature selection methods, we experimented with different numbers of features, such as 1000, 500, and 200, but observed no significant difference between methods. Since the previous work presented in (Espinoza *et al.*, 2021) uses only 50 features for training their model, we decided to align with that number. Additionally, given the high dimensionality of the data, reducing it to 50 features would correspond to just 3.45% of the original dimensions, which increases the challenge for all methods and better highlights performance differences. Therefore, in our pipeline, we consistently aimed to extract 50 features using each of the methods.

In our methodology, the LLM models provide a score for each feature based on their internal knowledge and the desired objective. These scores are sorted in descending order, and the top-$k$ elements are extracted. This process is illustrated in Figure 8, where the top 10 most important features identified by the GPT-o1 model are shown, ranked by relevance.

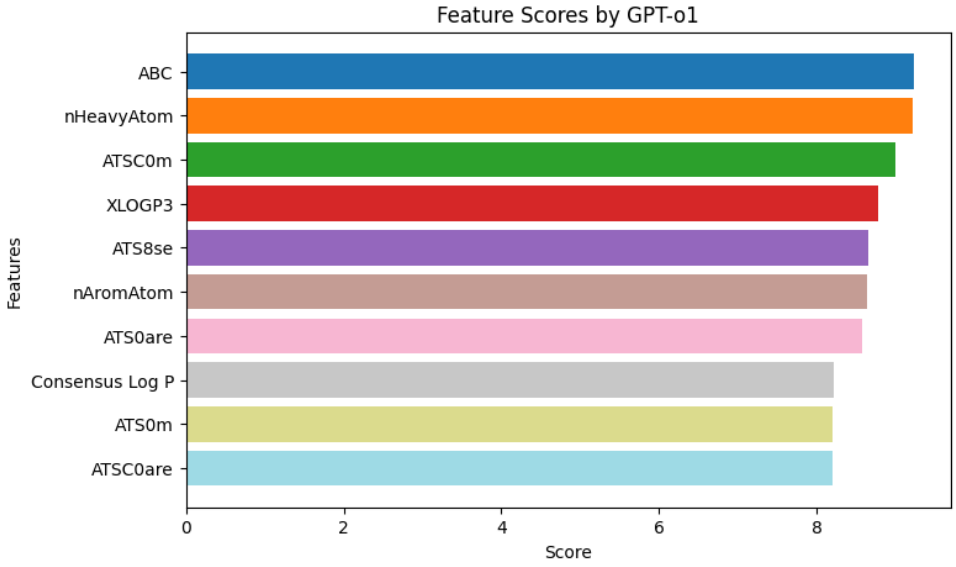



Figure 8 – Top 10 Features by Score (Generated by GPT-01 Model). Source: Author.

To generate these scores, we had to explicitly specify the range and whether the score should be discrete or continuous, as the models initially returned discrete values, and in some cases, variable intervals depending on the model itself. This inconsistency did not yield satisfactory results when using the baseline regression model.

Once we obtained the features generated by each method, we extracted the samples with the selected features from the training set and calculated the $MSE$, $MAE$, and $R^2$ metrics. As shown in Table 2, the performance values are quite similar across different feature selection methods, and the standard deviation does not vary significantly compared to the mean, indicating consistent and stable results.

Table 2 – Performance of Feature Selection Methods in Training (Mean ± Standard Deviation)

| **Type** | **Method** | $MSE(\pm SD)$ | $MAE(\pm SD)$ | $R^2(\pm SD)$ |
|---|---|---|---|---|
| Filter | Pearson | 0.4653 (0.0845) | 0.5252 (0.0417) | 0.5021 (0.0683) |
| | Mutual Info | 0.4286 (0.0887) | 0.4991 (0.0467) | 0.5427 (0.0717) |
| Wrapper | Forward | **0.3082 (0.0560)** | **0.4261 (0.0371)** | **0.6682 (0.0610)** |
| | Backward | 0.3873 (0.0508) | 0.4729 (0.0242) | 0.5828 (0.0556) |
| Embedded | Random Forest | 0.4305 (0.0804) | 0.4978 (0.0486) | 0.5384 (0.0751) |
| LLM | 4o | 0.4945 (0.0649) | 0.5330 (0.0354) | 0.4664 (0.0791) |
| | o1 | 0.4315 (0.0662) | 0.4989 (0.0368) | 0.5361 (0.0634) |
| | o3-mini | 0.4045 (0.0680) | 0.4867 (0.0419) | 0.5658 (0.0660) |
| | deepseek-r1 | 0.4098 (0.0749) | 0.4914 (0.0407) | 0.5602 (0.0686) |

This consistency in performance across various methods indicates that the extracted features are robust, producing similar predictive results regardless of the selection technique used. This suggests that, despite employing different approaches, the selected features consistently lead to comparable model performance. Among the tested methods, the Forward Selection (Wrapper) method stands out as having the best performance across all three evaluated metrics. This result can be attributed to the iterative nature of wrapper methods, which sequentially build an optimal feature subset that maximizes model quality.

On the other hand, the LLM-based methods, despite their ability to understand the semantic meaning of each variable, achieved results comparable to traditional techniques. Notably, among these approaches, the o3-mini method demonstrated slightly better performance compared to its counterparts and showed results similar to the Random Forest Embedded method. This suggests that LLMs can generate relevant scores for regression performance without requiring additional data, leveraging solely the knowledge acquired during their pre-training. This ability highlights the potential of LLM-based methods to identify meaningful features even when working with structured numerical data, demonstrating their versatility and adaptability in diverse contexts.

The relatively small standard deviations observed across all methods indicate consistent performance. LLM-based feature selection achieves comparable results to conventional techniques in this specific setting; it demonstrates the ability to generate relevant scores solely based on pre-trained knowledge, showcasing its potential to identify meaningful features without requiring additional domain-specific data.

To validate the consistency of the previous results, we evaluated the selected features with the baseline model on the test set, and the obtained metrics are presented in Table 3. The performance on the test set shows a similar pattern to the training results, with relatively small differences among the various feature selection methods.

Table 3 – Test Set Performance of Feature Selection Methods

| **Type** | **Method** | $MSE$ | $MAE$ | $R^2$ |
|---|---|---|---|---|
| Filter | Pearson | 0.4077 | 0.5095 | 0.4562 |
| | Mutual Info | 0.3661 | 0.4955 | 0.5116 |
| Wrapper | Forward | 0.3476 | 0.4458 | 0.5363 |
| | Backward | 0.3201 | 0.4242 | 0.5730 |
| Embedded | Random Forest | 0.3950 | 0.4937 | 0.4731 |
| LLM | 4o | 0.3897 | 0.4855 | 0.4802 |
| | o1 | 0.3449 | 0.4638 | 0.5399 |
| | o3-mini | 0.3541 | 0.4775 | 0.5276 |
| | deepseek-r1 | **0.3176** | **0.4389** | **0.5763** |

The most remarkable finding is the performance of the LLM-based methods, particularly deepseek-r1, which achieves the best results among all evaluated methods. This represents a notable improvement compared to other LLM variants and even outperforms the traditionally strong Backward Selection (Wrapper) method, which previously demonstrated the best performance during training. This outcome indicates that LLM-based feature selection can generalize well when evaluated on unseen data, highlighting the potential of using pre-trained knowledge to capture relevant features that may not be immediately apparent through conventional methods.

Among traditional methods, Backward Selection (Wrapper) continues to show robust performance, maintaining one of the lowest error metrics. This supports the earlier observation that iterative wrapper methods effectively select feature subsets that enhance model performance.

In summary, although LLM-based methods did not consistently outperform traditional approaches during training, their competitive performance on the test set demonstrates their ability to generalize effectively. This reinforces the idea that leveraging LLMs for feature scoring can be beneficial, prioritizing features for a specific problem.

As an additional advantage, the computational efficiency of the LLM-based methods stands out when compared to other feature selection approaches. As shown in Table 4, LLM-based methods not only demonstrate competitive performance but also require significantly less time to select features compared to more computationally intensive methods like Backward Elimination (Wrapper) and Random Forest Regressor (Embedded).

This notable reduction in processing time is particularly valuable when working with high-dimensional data or when feature selection needs to be performed repeatedly as part of model tuning and validation. The ability of LLM methods to efficiently generate relevant features based solely on their pre-trained knowledge makes them an attractive alternative, especially in scenarios where computational resources or time are limited.

Table 4 – Computation Time for Selecting Features

| **Type** | **Method** | **Processing Time** |
|---|---|---|
| Filter | Pearson Correlation | <1 min |
| | Mutual Information | <1 min |
| Wrapper | Forward Selection | ∼7 min |
| | Backward Elimination | ∼61 min |
| Embedded | Random Forest Regressor | ∼126 min |
| LLM | 4o | ∼2 min |
| | o1 | ∼6 min |
| | o3-mini | ∼9 min |
| | deepseek-r1 | ∼3 min |

Therefore, beyond their comparable predictive performance, the efficiency of LLM-based feature selection methods adds another practical advantage, reinforcing their potential utility in real-world applications where time efficiency is crucial.

## 4.3 Prompt Optimization and Iterative Improvement

During the initial experiments, the LLM-based methods did not outperform like traditional techniques. This distance was primarily attributed to the initial prompt design. The models did not generate a reliable feature selection, and even after multiple iterations, the outputs remained unstable. Moreover, we experimented with prompts in Spanish and Portuguese, but these did not yield the same results as those formulated in English. Only when using English did we observe an improvement, although some issues persisted.

To enhance the performance of LLM-based feature selection, we made several adjustments to the prompt. Initially, we tried providing the data context directly, but this approach did not significantly improve the results. We also experimented with giving the model diverse examples of feature selection (commonly known as few-shot learning), but it did not result in consistent or accurate outputs.

The most significant improvements came after systematically refining the prompt. We specified the expected output format, explicitly instructing the model to generate a continuous score within a defined interval, as the models initially returned discrete values or, in some cases, refused to generate a score. This issue was particularly pronounced with the more reasoning-capable models such as o1 and o3-mini, which claimed insufficient information to provide a score. To overcome this, we explicitly instructed the models to always generate a score, regardless of data limitations. Additionally, we required the output to follow a structured JSON format to facilitate downstream processing. However, this formatting request sometimes led to errors, especially with models like deepseek-r1,

GPT-4o, and GPT-o1, where the JSON structure was not consistently adhered to.

Further refinements involved clarifying the problem context and defining expectations more explicitly. We framed the model as a PhD-level expert within the prompt, which unexpectedly improved the consistency of the scores. Additionally, we prompted the model to explain why it selected a particular feature. Although it provided explanations, this did not necessarily improve the quality of the feature ranking itself. A sample of the final prompt is shown in Figure 9, illustrating these incremental changes.

**Context:**
*You are a PhD-level expert in medicinal chemistry and drug development with extensive experience in QSAR modeling and kinase inhibitor design. You specialize in predictive modeling of biological activity (pIC50) for small molecules.*

**Problem:**
*Regression task to predict pIC50 of ALK-5 inhibitors from molecular features.*

**Input Data:**

- Features: *{features}*
- Sample X values: *{X_samples}*
- Sample y values (pIC50): *{y_samples}*

**Task:**
*Assign a continuous relevance score (0-10 scale, max 3 decimals) to each feature based on its predictive power for pIC50. Follow these rules:*

- Score must reflect how strongly the feature relates to target (pIC50)
- Use 0 for unknown/unrelated features
- For doubtful cases, assign values near zero (0.1-1.0)

**Output Format:**
*Return a list of tuples: [(feature_name, score), ...] ordered by descending score. Example:*
`[('MolLogP', 8.215), ('NumHAcceptors', 6.732), ...]`

**Note:**
*You MUST provide a score for every feature. Never omit features.*

Figure 9 – Feature scoring prompt for $pIC_{50}$ prediction problem. Source: Author.

During iterative testing, we observed that the generated scores varied significantly when running the same prompt multiple times with a small sample of features. To address this variability, we continuously monitored how the scores changed when repeating the prompt with a set of 10 variables. Although some fluctuations persisted, the refined prompt

structure led to more stable outputs compared to earlier versions. We also attempted to guide the model by specifying which aspects to focus on when scoring features, but this adjustment did not result in notable improvements.

In summary, the prompt optimization process was crucial to achieving consistent and reliable feature selection using LLM-based methods. Through iterative refinement, we significantly reduced variability and improved the quality of the generated scores. Despite these improvements, some inherent challenges remained, particularly regarding maintaining consistency across multiple runs and managing the model's reasoning when faced with ambiguous or incomplete data.

## 4.4 Discussion and Interpretation

The results demonstrate that LLM-based feature selection methods can effectively compete with traditional techniques when the prompt is carefully designed. The iterative refinement of the prompt, including specifying output format, framing the model as a PhD-level expert, and using clear contextual instructions, significantly improved the consistency and accuracy of the generated feature scores.

One of the main advantages of LLM-based methods is their ability to generate relevant features based solely on pre-trained knowledge, capturing complex interactions that conventional methods might overlook. Additionally, their computational efficiency makes them an attractive option compared to more time-consuming approaches like backward elimination or embedded techniques.

# 5 CONCLUSIONS

The results obtained in this study highlight the potential of LLM-based methods for feature selection compared to traditional approaches. Our findings show that LLM-based methods can present competitive results when the prompt is carefully optimized.

The most promising LLM-based method was deepseek-r1, which outperformed other LLM variants and demonstrated comparable results to traditional methods. This performance suggests that, LLM-based methods are capable of identifying relevant features when the prompt design is adequate, effectively leveraging pre-trained knowledge to generate relevant feature scores without requiring additional information or retraining the model.

The computational efficiency of LLM-based methods is also noteworthy. Compared to time-consuming approaches like Backward Elimination or Random Forest, LLMs significantly reduce processing time, making them suitable for scenarios requiring rapid feature selection. This advantage, combined with their ability to intuitively rank features based on internal knowledge, positions LLM-based methods as a valuable alternative.

## 5.1 Main Limitations

Despite the promising potential of LLM-based feature selection, several limitations emerged. The most significant challenge was the high dependency on prompt engineering. The initial prompts often led to inconsistent or incorrect feature ranking, particularly when the model did not understand the format or scoring requirements. Additionally, attempts to use languages other than English, such as Spanish or Portuguese, did not yield satisfactory results, indicating a strong language dependency.

Another limitation was the variability in generated scores when the same prompt was executed multiple times, especially in models with advanced reasoning capabilities like o1 and o3-mini. Even after iterative prompt adjustments, complete stability was not achieved. Furthermore, specifying the output format in JSON was prone to errors, which did not consistently adhere to the structured output, leading to data processing challenges.

## 5.2 Future Work

Future research should focus on developing more robust and adaptive prompt engineering strategies for LLM-based feature selection. One promising direction is to create automatic prompt optimization frameworks that dynamically adjust the prompt structure based on model feedback, reducing the reliance on manual adjustments. Additionally,

integrating LLM-derived features with traditional statistical methods could combine the strengths of both approaches, leading to improved performance and robustness.

To address current limitations, it is also important to explore models with enhanced multilingual capabilities, as language understanding proved to be a key challenge. Further work should consider fine-tuning LLMs specifically for feature selection tasks, as well as employing techniques like RAG (Retrieval-Augmented Generation) to enrich the model with contextual information, enabling more accurate score generation.

Finally, applying LLM-based methods to other regression scenarios and real-world datasets would help assess their generalizability and practical value.

# REFERENCES


ACHIAM, J. *et al.* Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

BRAY, F. *et al.* Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 74, n. 3, p. 229–263, 2024.

BROWN, T. B. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

CALDWELL, G. W. *et al.* The ic50 concept revisited. **Current topics in medicinal chemistry**, Bentham Science Publishers, v. 12, n. 11, p. 1282–1290, 2012.

CHO, K. Learning phrase representations using rnn encoder-decoder for statistical machine translation. **arXiv preprint arXiv:1406.1078**, 2014.

CHOWDHERY, A. *et al.* Palm: Scaling language modeling with pathways. **Journal of Machine Learning Research**, v. 24, n. 240, p. 1–113, 2023.

CUNNINGHAM, J. P.; GHAHRAMANI, Z. Linear dimensionality reduction: Survey, insights, and generalizations. **The Journal of Machine Learning Research**, JMLR. org, v. 16, n. 1, p. 2859–2900, 2015.

DONG, Q. *et al.* A survey on in-context learning. *In*: **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**. [*S.l.: s.n.*], 2024. p. 1107–1128.

DOSOVITSKIY, A. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

ESPINOZA, G. Z. *et al.* Evaluating deep learning models for predicting alk-5 inhibition. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 1, p. e0246126, 2021.

FAIZ, A. *et al.* Llmcarbon: Modeling the end-to-end carbon footprint of large language models. **arXiv preprint arXiv:2309.14393**, 2023.

FAN, L. *et al.* A bibliometric review of large language models research from 2017 to 2023. **ACM Transactions on Intelligent Systems and Technology**, 2023.

FERLAY, J. *et al.* Cancer statistics for the year 2020: An overview. **International journal of cancer**, Wiley Online Library, v. 149, n. 4, p. 778–789, 2021.

GRAVES, A. Long short-term memory. **Supervised sequence labelling with recurrent neural networks**, Springer, p. 37–45, 2012.

ION, G. N. D.; NITULESCU, G. M.; MIHAI, D. P. Machine learning-assisted drug repurposing framework for discovery of aurora kinase b inhibitors. **Pharmaceuticals**, MDPI, v. 18, n. 1, p. 13, 2024.

JEONG, D. P.; LIPTON, Z. C.; RAVIKUMAR, P. Llm-select: Feature selection with large language models. **arXiv preprint arXiv:2407.02694**, 2024.

JIA, P. *et al.* Altfs: Agency-light feature selection with large language models in deep recommender systems. **arXiv preprint arXiv:2412.08516**, 2024.

KAPLAN, J. *et al.* Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

KARGBO, R. B. **ALK Inhibitors for Treating Cancer, Blood, and Kidney Diseases**. [*S.l.: s.n.*]: ACS Publications, 2022. 1539–1541 p.

KÜKEN, J.; PURUCKER, L.; HUTTER, F. Large language models engineer too many simple features for tabular data. **arXiv preprint arXiv:2410.17787**, 2024.

LABJAR, H.; LABJAR, N.; KISSI, M. Qsar anti-hiv feature selection and prediction for drug discovery using genetic algorithm and machine learning algorithms. *In*: **Computational Intelligence in Recent Communication Networks**. [*S.l.: s.n.*]: Springer, 2022. p. 191–204.

LI, D.; TAN, Z.; LIU, H. Exploring large language models for feature selection: A data-centric perspective. **ACM SIGKDD Explorations Newsletter**, ACM New York, NY, USA, v. 26, n. 2, p. 44–53, 2025.

LIU, Y. *et al.* Datasets for large language models: A comprehensive survey. **arXiv preprint arXiv:2402.18041**, 2024.

MAATOUK, A. *et al.* Large language models for telecom: Forthcoming impact on the industry. **IEEE Communications Magazine**, IEEE, 2024.

MANSOUR, M. A. *et al.* Advances in the discovery of activin receptor-like kinase 5 (alk5) inhibitors. **Bioorganic Chemistry**, Elsevier, p. 107332, 2024.

MINAEE, S. *et al.* Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024.

NIU, Z.; ZHONG, G.; YU, H. A review on the attention mechanism of deep learning. **Neurocomputing**, Elsevier, v. 452, p. 48–62, 2021.

NOVIANDY, T. R. *et al.* Machine learning for antiviral drug discovery: Application of xgboost-tpe for predicting bioactivity of hepatitis c inhibitors. *In*: IEEE. **2024 International Conference on Electrical Engineering and Informatics (ICELTICs)**. [*S.l.: s.n.*], 2024. p. 62–67.

PENG, D.; GUI, Z.; WU, H. Interpreting the curse of dimensionality from distance concentration and manifold effect. **arXiv preprint arXiv:2401.00422**, 2023.

POEI, D. *et al.* Alk inhibitors in cancer: mechanisms of resistance and therapeutic management strategies. **Cancer Drug Resistance**, OAE Publishing Inc, v. 7, 2024.

RAMAPRABA, P. S. *et al.* Implementing cloud computing in drug discovery and telemedicine for quantitative structure-activity relationship analysis. **International Journal of Electrical and Computer Engineering (IJECE)**, v. 15, n. 1, p. 1132–1141, 2025.

REDKAR, S. *et al.* A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing. **Molecular informatics**, Wiley Online Library, v. 39, n. 5, p. 1900062, 2020.

REN, X. *et al.* Pangu-{\Sigma}: Towards trillion parameter language model with sparse heterogeneous computing. **arXiv preprint arXiv:2303.10845**, 2023.

SINGHAL, K. *et al.* Large language models encode clinical knowledge. **Nature**, Nature Publishing Group, v. 620, n. 7972, p. 172–180, 2023.

SUTSKEVER, I. Sequence to sequence learning with neural networks. **arXiv preprint arXiv:1409.3215**, 2014.

THIRUNAVUKARASU, A. J. *et al.* Large language models in medicine. **Nature medicine**, Nature Publishing Group US New York, v. 29, n. 8, p. 1930–1940, 2023.

TOUVRON, H. *et al.* Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

TURZO, S. B. A.; HANTZ, E. R.; LINDERT, S. Applications of machine learning in computer-aided drug discovery. **QRB discovery**, Cambridge University Press, v. 3, p. e14, 2022.

VASWANI, A. Attention is all you need. **arXiv preprint arXiv:1706.03762**, 2017.

WEI, J. *et al.* Finetuned language models are zero-shot learners. **arXiv preprint arXiv:2109.01652**, 2021.

YANG, T. *et al.* Ice-search: A language model-driven feature selection approach. **arXiv preprint arXiv:2402.18609**, 2024.

ZHAO, W. X. *et al.* A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

ZIA, V. *et al.* Advancements of alk inhibition of non-small cell lung cancer: a literature review. **Translational Lung Cancer Research**, AME Publications, v. 12, n. 7, p. 1563, 2023.