

# **BINDING AFFINITY PREDICTION USING KAN**

## **Table of Contents**

<b>1. Abstract .....</b>	<b>6</b>
<b>2. Introduction .....</b>	<b>7</b>
2.1 Motivation .....	7
2.2 Research Focus .....	8
2.3 Kolmogorov-Arnold Networks Function in Predicting Binding Affinity .....	9
2.4 Significance of the Study .....	10
<b>3. Literature Survey .....</b>	<b>11</b>
3.1 Conventional Methods for Predicting Binding Affinity .....	11
3.2 Overview of AI and Machine Learning for Predicting Binding Affinity .....	12
3.3 Binding Affinity Prediction Using Neural Networks .....	13
3.4 Function of ECFP8b Fingerprints in Binding Affinity Prediction .....	14
3.5 Drug Discovery Using Kolmogorov-Arnold Networks .....	14
<b>4. Use of Kolmogorov-Arnold Networks in Binding Affinity Prediction .....</b>	<b>15</b>
4.1 Kolmogorov-Arnold Networks Operational Mechanism .....	16
4.2 Preprocessing and Data Representation .....	17
4.3 Kolmogorov-Arnold Network Architecture .....	18
4.4 Important Distinctions between MLPs and KANs .....	19
4.5 KAN Optimization .....	22
4.6 Binding Affinity Prediction Loss Function .....	23
<b>5. Summary .....</b>	<b>25</b>
<b>6. Conclusion .....</b>	<b>25</b>
<b>7. Scope for Future Work .....</b>	<b>26</b>
<b>8. References .....</b>	<b>27</b>

## **Table of Figures, Tables and Formulas**

<b>2.3 Key Differences between MLPs and KANs .....</b>	<b>10</b>
<b>4.1 Mathematical Formula of KANs Decomposition .....</b>	<b>16</b>
<b>4.2 Table of Preprocessing and Data Representation .....</b>	<b>18</b>
<b>4.3 KAN Architecture .....</b>	<b>19</b>
<b>4.4 Performance Trade-offs .....</b>	<b>21</b>
<b>4.4 Table of Performance Trade-offs .....</b>	<b>21</b>
<b>4.5.1 Optimizer for Gradient-Based Optimization .....</b>	<b>22</b>
<b>4.5.2 Metrics for Evaluation .....</b>	<b>22</b>
<b>4.5.3 Mathematical formula of Early Stopping .....</b>	<b>23</b>
<b>4.5 Table of KAN Optimization .....</b>	<b>23</b>

## **1. Abstract**

Binding affinity prediction is an important step in drug development because it quantifies the interaction strength between chemical compounds and their biological targets, allowing researchers to identify viable therapeutic possibilities. Conventional methodologies, such as molecular docking and free-energy calculations, while widely utilized, are frequently computationally expensive, time-consuming, and unscalable, especially for huge datasets or high-throughput screening. These restrictions need the creation of new, efficient, and scalable prediction frameworks. This study proposes an improved Kolmogorov-Arnold Network (KAN), a pioneering machine learning network designed specifically for binding affinity prediction. The model uses Extended-Connectivity Fingerprints (ECFP8b) to represent molecular structures, offering a strong numerical alternative to classic sequence-based SMILES embedding methods. ECFP8b fingerprints maintain crucial chemical and structural information while simplifying preprocessing by eliminating the requirement for tokenization and embedding layers. The KAN model uses unique Linear Layers and additive composition procedures that follow the Kolmogorov-Arnold representation theorem, reducing the complicated problem of molecule interactions to interpretable and computationally efficient univariate and bivariate functions. The architecture prioritizes transparency, scalability, and precision above the black-box aspect of typical neural networks. To verify model consistency and generalizability, performance was evaluated using robust metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE), which were used throughout a 5-fold cross validation framework. Advanced optimization approaches, such as EarlyStopping and ReduceLROnPlateau, were included into the training pipeline to improve convergence and reduce overfitting. Experimental results support the model's predictive capabilities, demonstrating its capacity to deliver high accuracy with low computing cost. By accurately capturing molecular connections while maintaining interpretability, the KAN model provides a scalable and efficient method for cheminformatics and computational drug development. This study emphasizes KAN's transformative potential in accelerating the exploration of vast chemical spaces and identifying novel therapeutic candidates, providing a thorough account of the model's theoretical foundation, methodology, implementation, and experimental results to advance modern drug discovery pipelines.

## 2. Introduction

A crucial step in **drug development** is **binding affinity prediction**, which gauges how strongly chemical compounds interact with biological targets to find prospective treatment options. Conventional approaches, such as **free-energy calculations** and **molecular docking**, are popular but have limited scalability, high computational costs, and time constraints. Their efficacy for extensive chemical space exploration and high-throughput screening is hampered by these issues. The **Kolmogorov-Arnold Network (KAN)**, one of the potent and scalable alternatives brought about by **machine learning**, makes binding affinity prediction easier by decomposing it into interpretable **univariate** and **bivariate functions**. By combining interpretability and computational efficiency, this architecture overcomes the drawbacks of conventional methods and offers a deeper understanding of **molecular interactions**.

An improved **KAN model** specifically designed for **binding affinity prediction** is created and validated in this project. This study uses **Extended-Connectivity Fingerprints (ECFP8b)**, which directly encode molecular structures as numerical vectors, capturing important substructures and interactions while streamlining preparation, in contrast to sequence-based **SMILES representations** that necessitate intricate preprocessing. In order to guarantee scalability and interpretability, the model uses additive composition procedures and bespoke linear layers while abiding by the Kolmogorov-Arnold theorem. The improved KAN model is a game-changing tool for computational drug discovery since it exhibits great predictive capabilities, scalability, and efficiency when evaluated using metrics like **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)** with rigorous cross-validation.

### 2.1 Motivation

Finding new **medicinal compounds** is a difficult and resource-intensive process that frequently calls for years of study and a substantial financial outlay. Effectively traversing the enormous **chemical space**, which contains every conceivable molecular structure, is a major challenge. Even though they are well-established, traditional techniques like **molecular docking** and **free-energy calculations** are constrained in their scalability, time-consuming procedures, and high computing costs. Because of these limitations, they are not appropriate for **high-throughput drug discovery** or large-scale datasets, which calls for the creation of more effective and scalable alternatives.

The need for a reliable, comprehensible, and **scalable method of binding affinity prediction**—a crucial component of **drug discovery**—is what motivates our study. The study uses **Extended-Connectivity Fingerprints (ECFP8b)** to streamline molecular representation in order to overcome the drawbacks of conventional techniques and current machine learning algorithms. While maintaining important chemical substructures and interactions, ECFP8b fingerprints do away with the preprocessing difficulties that come with SMILES sequences. Furthermore, in accordance with the **Kolmogorov-Arnold representation theorem**, the improved Kolmogorov-Arnold Network (KAN) incorporates additive composition operations and custom Linear Layers. By bridging the gap between traditional computational techniques and contemporary machine learning, this innovative method aims to speed up drug development, lower computational costs, and provide a deeper understanding of molecular interactions.

## **2.2 Research Focus**

The creation of a **Kolmogorov-Arnold Network (KAN)** specifically designed for the precise prediction of **binding affinities** between chemical compounds and biological targets is the main goal of this project. In **drug development**, binding affinity is a crucial metric that directs the identification of compounds with therapeutic potential. This study introduces a **scalable and interpretable machine learning framework** to overcome the drawbacks of conventional techniques like **molecular docking** and **free-energy calculations**. The following are the main research questions that guide this work:

- How can **Extended-Connectivity Fingerprints (ECFP8b)** representations of chemical structures be used to efficiently create the **Kolmogorov-Arnold Network (KAN)** to predict binding affinities?
- How do **ECFP8b fingerprints** make preprocessing easier while maintaining important structural and chemical data that may be entered into the KAN model?
- How can comprehension of **binding affinity predictions** be improved by the interpretability of the KAN design, which breaks down molecular interactions into **univariate** and **bivariate functions**?
- What are the main performance indicators that show the **accuracy, scalability, and generalizability** of the model, such as **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**?

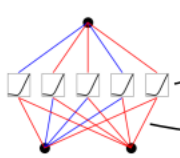
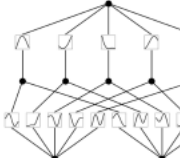
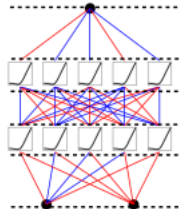
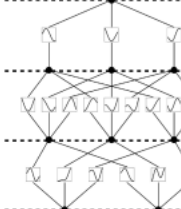
The goal of this research is to leverage the special characteristics of the **KAN architecture** to close the gap between conventional computational methods and contemporary machine learning approaches. The improved model provides a **scalable** and **interpretable solution** by incorporating **unique Linear Layers** and **additive composition processes**, and using **ECFP8b fingerprints** as an effective molecular representation. This work intends to establish KAN as a useful tool for **expediting drug development**, facilitating effective exploration of chemical spaces, and offering insights into molecular interactions through thorough validation and evaluation on diverse datasets.

### **2.3 Kolmogorov-Arnold Networks Function in Predicting Binding Affinity**

By circumventing the drawbacks of conventional approaches, such as **computational inefficiency**, low scalability, and restricted interpretability, **Kolmogorov-Arnold Networks (KANs)** offer a revolutionary approach to **binding affinity prediction**. Through the use of **mathematical decomposition**, the KAN model reduces the complexity of molecular interactions by dividing high-dimensional data into components that can be understood. It works especially well with chemical data represented as **numerical vectors**, like **Extended-Connectivity Fingerprints (ECFP8b)**, which do not require sequence-based processing and capture important substructures and molecular interactions. The following are important aspects of using KAN to predict binding affinity:

- **Mathematical Decomposition:** KAN ensures interpretability without sacrificing accuracy by dividing complex molecular interactions into univariate and bivariate components using the Kolmogorov-Arnold representation theorem.
- **Scalability:** By employing ECFP8b fingerprints to process large datasets effectively, the model is well-suited for high-throughput binding affinity prediction and the investigation of broad chemical spaces.
- **Interpretability:** By demonstrating how particular molecular characteristics and their interactions affect binding affinity, the univariate and bivariate layers of KAN provide insightful information for cheminformatics and facilitate a better comprehension of the prediction process.
- **Efficiency:** The architecture maintains excellent predictive accuracy while reducing computational complexity by concentrating on key structural and chemical properties.

KAN sets a new benchmark for predictive models in **drug development** by combining **ECFP8b fingerprints** as input and utilizing a simplified architecture that uses **additive composition operations** and proprietary **Linear Layers** to balance **computational speed** and scientific clarity.

Model	<b>Multi-Layer Perceptron (MLP)</b>	<b>Kolmogorov-Arnold Network (KAN)</b>
Theorem	<b>Universal Approximation Theorem</b>	<b>Kolmogorov-Arnold Representation Theorem</b>
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(e)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left( \sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	(a)  fixed activation functions on nodes learnable weights on edges	(b)  learnable activation functions on edges sum operation on nodes
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	(c)  MLP(x) $\mathbf{W}_3$ $\sigma_2$ $\mathbf{W}_2$ $\sigma_1$ $\mathbf{W}_1$ $\mathbf{x}$ nonlinear, fixed linear, learnable	(d)  KAN(x) $\Phi_3$ $\Phi_2$ $\Phi_1$ $\mathbf{x}$ nonlinear, learnable

**Figure – 2.3 [1]**

## 2.4 Significance of the Study

By overcoming the drawbacks of **conventional techniques** and utilizing the novel capabilities of **Kolmogorov-Arnold Networks (KAN)**, this study presents a revolutionary approach to **computational drug discovery**. A crucial stage in finding potential treatments, **binding affinity prediction** has hitherto depended on computationally costly methods like **molecular docking**. There is a need for **effective** and **interpretable alternatives** because current approaches are ineffective for high-throughput applications and are not scalable.

By breaking down intricate molecular interactions into **univariate** and **bivariate components**, KAN provides a **scalable, effective, and interpretable method** that guarantees precise predictions without compromising transparency. This study uses **Extended-Connectivity Fingerprints (ECFP8b)** as a direct, numerical molecular representation in place of sequence-based SMILES (Simplified Molecular Input Line Entry System)



representations. These fingerprints simplify preprocessing and improve prediction resilience by preserving crucial chemical and structural information.

Metrics like **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)** support the study's high accuracy and scalability, demonstrating KAN's performance. The improved KAN model has the potential to significantly advance novel therapeutic solutions and accelerate the **drug development process** by facilitating the effective exploration of broad chemical regions.

### **3. Literature Survey**

By finding compounds that have the potential to become **potent therapeutic agents**, **binding affinity prediction** is essential to **computational drug discovery**. From conventional **physics-based techniques** like molecular docking and free-energy calculations to contemporary, **data-driven strategies** made possible by artificial intelligence (AI), the discipline has experienced substantial change. The **scalability** of traditional methods is limited by their frequent reliance on computationally demanding simulations and predetermined chemical descriptors. The field of **drug discovery** has changed as a result of the introduction of **effective** and **scalable substitutes** by machine learning and neural networks.

Beginning with the shortcomings of **conventional methods** and moving on to the incorporation of **AI-driven methodologies**, this chapter explores the development of **binding affinity prediction tools**. More sophisticated numerical encodings, such as **Extended-Connectivity Fingerprints (ECFP8b)**, which streamline preprocessing while preserving crucial chemical information, have replaced **SMILES-based representations**. The chapter also examines the revolutionary possibilities of **Kolmogorov-Arnold Networks (KAN)**, highlighting its capacity to offer **scalable** and **interpretable solutions** for managing the intricacy of chemical data in **drug discovery** [2].

#### **3.1 Conventional Methods for Predicting Binding Affinity**

In order to assess the intensity of the interaction between molecules and their **biological targets**, traditional **binding affinity prediction** has depended on computational and **physics-based methods** like **molecular docking** and **free-energy calculations**. Although these

techniques have shown promise in **drug discovery**, their scalability and computational efficiency are frequently limited.

- **Molecular Docking:** This technique estimates binding energies to find interesting candidates by forecasting how a chemical will attach to a protein's active site. Structure-based drug discovery works well, but its scalability is limited by its high computational cost and reliance on high-quality structural data.
- **Free-energy computations:** These computations, which include simulations of molecular dynamics, yield accurate estimations of binding affinities. However, they are not feasible for **high-throughput screening applications** or huge datasets due to their processing expense [4].

### Traditional Method's Drawbacks

Conventional methods are hampered by their **high computing costs**, dependence on **preset chemical characteristics**, and inability to adapt to **new targets**. Because of these limitations, **machine learning** and **deep learning** approaches—which offer **scalable, data-driven solutions** that can effectively capture **intricate chemical interactions**—have become increasingly popular.

### 3.2 An Overview of AI and Machine Learning for Predicting Binding Affinity

Binding affinity prediction has changed as a result of the combination of **artificial intelligence (AI)** and **machine learning (ML)**, which offers **scalable** and **effective substitutes** for conventional methods. By addressing problems like **parameter dependency** and **computational inefficiency**, these sophisticated techniques enable models to depict intricate molecular interactions [3].

### Early Methods of Machine Learning

Molecular descriptors and fingerprints were used by early machine learning models, such as **Support Vector Machines (SVMs)** and **Random Forests (RFs)**, to predict binding affinities. Despite being more effective than conventional techniques, these methods' dependence on **specified features** limited their applicability to a variety of datasets.

## Developments in Deep Learning

Models can now directly extract features from raw data thanks to **representation learning**, which was made possible by **deep learning (DL)**. Neural networks like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used in sequence-based methods that use SMILES strings and protein sequences to predict affinities. Furthermore, in order to capture **atom-level interactions** and offer a more sophisticated understanding of molecular behavior, graph-based techniques, such as graph neural networks (GNNs), model molecules as graphs [6].

## Opportunities and Difficulties

Even while ML and DL techniques increase scalability and adaptability, problems still exist in areas like **interpretability**, reliance on high-quality labeled data, and generalization to unknown datasets. The limits of ML-based binding affinity prediction are being pushed by innovations like attention mechanisms and hybrid architectures, which present promising prospects for future developments.

### 3.3 Binding Affinity Prediction Using Neural Networks

In **binding affinity prediction**, **neural networks** are now essential because they allow models to learn complex chemical interactions straight from data without the need for intensive feature creation. They are perfect for this field because of their adaptability to molecular representations like **molecular graphs** and **Extended-Connectivity Fingerprints (ECFP8b)**. Various architectures each contribute special advantages to the task:

- **CNNs or convolutional neural networks:** CNNs were first created for image identification, but they have since been modified to extract hierarchical features from fingerprints and chemical graphs. They improve predictions in binding affinity prediction by capturing spatial correlations in molecular data.
- **Long Short-Term Memory Networks (LSTMs) and Recurrent Neural Networks (RNNs):** RNNs are good at sequence-based tasks like protein sequences and SMILES strings, but they suffer from the vanishing gradient issue. This is fixed by **LSTMs**, which are appropriate for sequential molecular representations because they capture long-term relationships.
- **Graph Neural Networks (GNNs):** GNNs capture topological properties and atom-level interactions by representing molecules as graphs. They work exceptionally well

for simulating attributes dependent on connection and structure, notably for three-dimensional molecular conformations.

- **Transformers:** Transformers are excellent at identifying long-range dependencies in molecular data by utilizing **self-attention techniques**. They are becoming more and more useful for predicting binding affinities with intricate chemical structures due to their capacity for parallel processing.

The scalability and adaptability needed for contemporary binding affinity prediction are offered by neural networks. They provide the groundwork for cutting-edge methods to improve interpretability and computational efficiency in this area, like Kolmogorov-Arnold Networks (KAN) [5].

### **3.4 The Function of ECFP8b Fingerprints in Binding Affinity Prediction**

A reliable molecular representation is provided by **Extended-Connectivity Fingerprints (ECFP8b)**, which encode connectivity and substructural information into fixed-length binary vectors. **ECFP8b** immediately captures chemical substructures, which makes it perfect for **binding affinity prediction** in machine learning models, in contrast to SMILES, which depends on sequential data.

#### **Benefits of Models for Machine Learning**

ECFP8b fingerprints do not require preprocessing processes like tokenization or embedding because they give molecules a numerical representation. They preserve important information including **functional group linkages** and **atom-level connectivity**, which helps models effectively spot patterns connected to binding interactions [7].

#### **Strengths and Preprocessing**

Using procedures similar to **Morgan fingerprints**, **ECFP8b fingerprints** are generated by transforming chemical structures into binary vectors. This maintains important **structural relationships** while streamlining preparation. Even while fingerprints are excellent at recording 2D structural information, modeling 3D chemical interactions can be improved by combining them with graph-based representations. **ECFP8b fingerprints** are revolutionizing **binding affinity prediction** and promoting advancements in drug development by providing a scalable, effective, and interpretable input format.

### **3.5 Drug Discovery Using Kolmogorov-Arnold Networks**

**Kolmogorov-Arnold Networks (KAN)** offer a novel framework for **drug discovery** that overcomes the drawbacks of several contemporary machine learning models and conventional computational techniques. By breaking down intricate molecular interactions into **univariate** and **bivariate components** using the **Kolmogorov-Arnold theorem**, KAN provides a special blend of scalability, interpretability, and adaptability for **binding affinity prediction**.

#### **Benefits of KAN**

- **Scalability:** It can handle big datasets with ease, which makes it appropriate for high-throughput screening.
- **Interpretability:** Provides lucid insights on binding affinity determinants by breaking down molecular interactions into simpler components.
- **Adaptability:** Its efficacy and versatility are increased by compatibility with molecular representations such as **ECFP8b fingerprints**.

#### **Uses and Contrasts**

Pairwise interactions and **molecular patterns** are captured by **KAN models**, which are tuned to handle structured data. This enhances **prediction accuracy** and expands our knowledge of **binding mechanisms**. KAN strikes a compromise between interpretability and efficiency when compared to topologies such as transformers and graph neural networks (GNNs). By offering comprehensive but computationally effective insights into **molecular interactions**, KAN is positioned as a potent tool for investigating **chemical space** and improving **binding affinity prediction**.

### **4. The Use of Kolmogorov-Arnold Networks in Binding Affinity Prediction**

In computational drug discovery, **Kolmogorov-Arnold Networks (KAN)** offer a revolutionary approach to binding affinity prediction, addressing important issues including scalability, interpretability, and efficiency. KAN breaks down high-dimensional molecular interactions into **univariate** and **bivariate components** by utilizing the **Kolmogorov-Arnold representation theorem**, guaranteeing clear interpretability and effective processing.

KAN is optimized for molecular data represented as **Extended-Connectivity Fingerprints (ECFP8b)**, which capture comprehensive substructural and connectivity information without requiring sequence-based preprocessing like SMILES tokenization, in contrast to conventional techniques or black-box machine learning models. KAN offers a potent and useful method for increasing cheminformatics and drug development by combining customized loss functions and demythologizing intricate chemical interactions. This helps to close the gap between computational efficiency and scientific understanding.

#### **4.1 Kolmogorov-Arnold Networks Operational Mechanism**

**Kolmogorov-Arnold Networks (KAN)** break down complex multivariate functions into more manageable, interpretable parts by using the Kolmogorov-Arnold representation theorem to solve **high-dimensional issues**. According to the theory, any multivariate function may be written as a sum of **univariate** and **bivariate functions**, which makes KAN especially useful for **predicting binding affinity**.

##### **Broken Down Molecular Interactions**

KAN breaks down the input function  $f(x_1, x_2, \dots, x_n)$  into the following form:

$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \phi_i(x_i) + \sum_{i=1}^n \sum_{j=1}^n \psi_{ij}(x_i, x_j)$$

Here:

- $\phi_i(x_i)$ : Univariate functions that capture individual feature contributions.
- $\psi_{ij}(x_i, x_j)$ : Bivariate functions that model pairwise feature interactions.

**Figure – 4.1**

##### **Important Elements of KAN**

- **Univariate Function Layers:** These layers determine how particular molecular characteristics affect binding affinity by separating the contributions of individual components.
- **Bivariate Function Layers:** By modeling pairwise feature interactions, these layers allow KAN to capture dependencies like conformational changes that are essential for molecular interactions.

- **Pooling Mechanisms:** To create a final prediction, aggregation layers integrate the outputs of univariate and bivariate functions, making sure that both independent and interaction effects are taken into account.

**KAN** is a potent method for **computational drug development** because it breaks down high-dimensional chemical interactions into discrete constituents, offering a **scalable** and interpretable framework for **binding affinity prediction**.

## **4.2 Preprocessing and Data Representation**

The effective use of **Kolmogorov-Arnold Networks (KAN)** in binding affinity prediction depends on effective **data representation** and preprocessing. KAN relies on input data that is organized for computational efficiency while maintaining crucial molecular characteristics. Preparing molecular data, especially **Extended-Connectivity Fingerprints (ECFP8b)**, is the main goal of this section in order to provide precise and understandable predictions.

### **Binary Vectors Known as ECFP8b Representation**

Extended-Connectivity Fingerprints (ECFP8b) use connectivity and substructural data to encode molecular structures. **ECFP8b** offers a direct **numerical representation** in contrast to sequential formats like SMILES, while maintaining important characteristics such as substructural linkages, functional group interactions, and atom-level connectivity.

### **Steps in Preprocessing**

Among the steps involved in preparing molecular data for **KAN** is **fingerprint generation**, which transforms molecules into **binary ECFP8b vectors** and records connection patterns that are essential for forecasts.

- **Normalization:** To ensure uniformity across datasets, binding affinity values are scaled.
- **Data Cleaning:** To enhance model performance, invalid or missing molecular entries are eliminated.
- **Data Splitting:** To ensure robust evaluation, data is separated into training, validation, and test sets using **K-fold cross-validation**.

Step	Description	Example
Fingerprint Generation	Converts molecular structures into binary ECFP8b vectors.	CCO $\rightarrow$ [0, 1, 0, ...]
Normalization	Scales binding affinity values for consistency.	Affinity: 2.5 $\rightarrow$ 0.62
Data Cleaning	Removes invalid or missing data entries.	NaN $\rightarrow$ Exclude
Data Splitting	Divides data for model training and testing.	20% Train, 80% Test

**Table – 4.2**

Preprocessing guarantees that the model accurately captures **univariate** and **bivariate interactions** by aligning molecular data with **KAN's architecture**. The precision and comprehensibility of **binding affinity predictions** are improved by this simplified preparation.

### **4.3 Kolmogorov-Arnold Network Architecture**

For **binding affinity prediction**, **Kolmogorov-Arnold Networks (KAN)** offer a novel paradigm that tackles issues including **accuracy**, **scalability**, and **interpretability**. KAN makes predictions clear and effective by breaking down molecular interactions into univariate and bivariate components.

#### **Essential Elements of the KAN Architecture**

- **Input Layer:** Ensures model compliance by integrating molecular data as **ECFP8b fingerprints**.
- **Univariate Function Layers:** Use  $\phi_i(x_i)$  to convert each  $x_i$  in order to capture the independent contributions of features.
- **Bivariate Function Layers:** Crucial for comprehending molecular dependencies, these layers use  $\psi_{ij}(x_i, x_j)$  to model pairwise interactions between features.
- **Pooling Mechanisms:** Combine independent and interaction effects to create a final prediction by aggregating outputs from univariate and bivariate layers.

#### **Effective Preprocessing and Design**

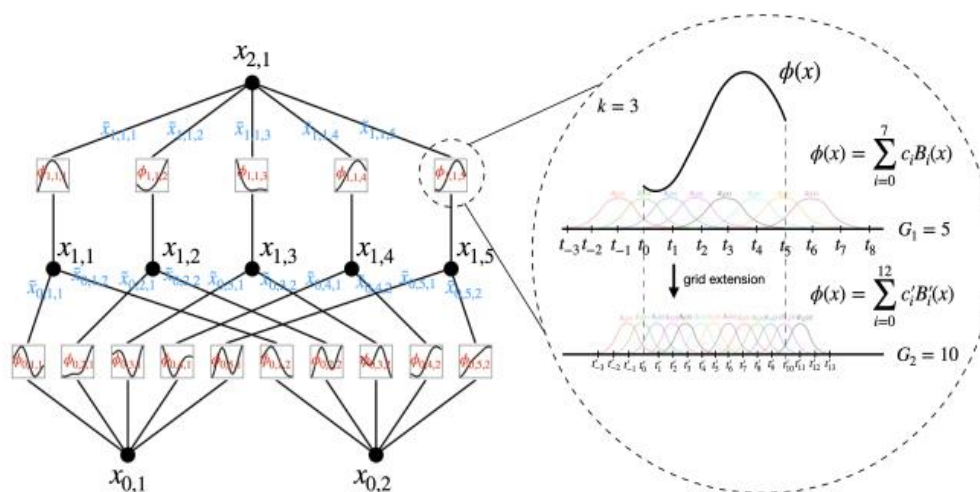
By minimizing overhead in **bivariate layers**, parallel processing in **univariate levels**, and **sparse computation**, **KAN** places a strong emphasis on **computational efficiency**. The creation of ECFP8b fingerprints, which involves transforming chemical structures into binary vectors, is one important function of preprocessing.



- **Normalization:** The process of consistently scaling binding affinity values.
- **Data Cleaning:** To ensure reliable training, eliminate missing or erroneous entries.

### Difficulties and Assessment

KAN uses **sparse computation** and **feature selection** to tackle issues like high-dimensional data. **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, and **K-fold cross-validation** are used to assess its efficacy, guaranteeing reliable and comprehensible forecasts. KAN advances drug discovery by bridging theoretical innovation with real-world application through the integration of effective preprocessing, univariate and bivariate components, and pooling techniques.



**Figure – 4.3**

## 4.4 Important Distinctions between MLPs and KANs

Compared to **Multi-Layer Perceptrons (MLPs)**, **Kolmogorov-Arnold Networks (KANs)** offer a substantial gain in terms of **interpretability**, **scalability**, and applicability for intricate applications like binding affinity prediction. The differences between both architectures are described in this subchapter, along with the reasons why cheminformatics applications are a better fit for KANs.

### 1. Representation of Features

- **MLPs:** Learn feature interactions implicitly by processing input data as a single high-dimensional vector with fully connected layers. There is little information available on individual or paired contributions using this "black-box" method.

- **KANs**: Clearly model the contributions of individual and paired features by breaking down molecular data into **univariate** and **bivariate components**. By capturing smooth, continuous fluctuations in molecular characteristics, **B-splines** significantly improve feature representation and allow for finer-grained modeling of interactions.

## 2. Interpretability

- **MLPs**: Add computational complexity by requiring post hoc techniques like **SHAP** or **LIME** to explain predictions, and they lack inherent interpretability.
- **KANs**: Designed for transparency, these models offer direct insights into individual features and their interactions through **univariate** and **bivariate layers**. Furthermore, **B-splines** help with interpretability by providing more lucid insights on molecular trends through the structured, smooth, and interpretable representation of characteristics.

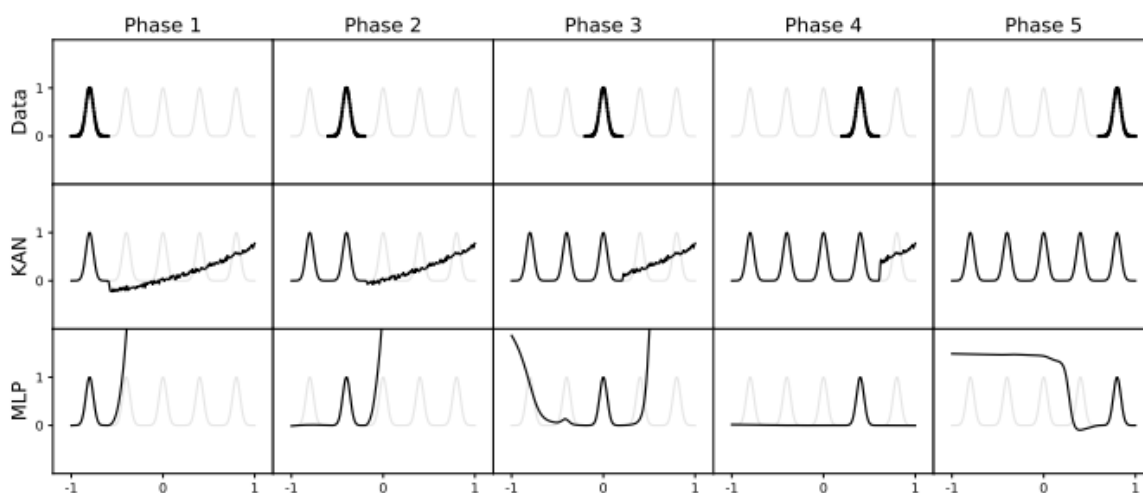
## 3. Efficiency and Scalability

- **MLPs**: High-dimensional data is difficult for **MLPs** to handle because of the dense calculations in fully connected layers, which causes scalability problems as the number of features rises.
- **KANs**: Reduce duplication by separating **univariate** and **bivariate components** and use **sparse calculations** in bivariate layers to handle huge datasets effectively. Additionally, by structuring molecular data into grids, **Grid Extension** approaches allow **KAN** to expand effectively to bigger datasets while optimizing memory utilization and simplifying computations.

## 4. Applicability for Predicting Binding Affinity

- **MLPs**: These general-purpose architectures are less useful for **cheminformatics activities** that call for detailed molecular interaction modeling.
- **KANs**: Specifically built to capture both structural and chemical dependence in molecular interactions. **KAN** can adjust to new datasets or chemical settings without having to retrain from scratch thanks to the integration of **Continual Learning**, guaranteeing constant performance in **binding affinity prediction** across a variety of dynamic datasets.

## 5. Performance Trade-offs



**Figure – 4.4**

Aspect	MLPs	KANs
Feature Representation	Implicit learning of interactions.	Explicit decomposition with B-splines.
Interpretability	Limited; requires post hoc tools.	High; directly interpretable outputs.
Scalability	Computationally intensive for large datasets.	Efficient with sparse computations and Grid Extension.
Suitability for Binding Affinity Prediction	General-purpose, lacks specialization.	Tailored for molecular modeling with Continual Learning.

**Table – 4.4 [1]**

Despite their adaptability and broad use, **MLPs** are less appropriate for applications such as **binding affinity prediction** due to their interpretability and scalability issues. By explicitly modeling **individual** and **paired feature contributions**, **KANs**, on the other hand, are excellent at capturing **molecular interactions**, guaranteeing transparency, computational efficiency, and improved performance. Because of these characteristics, **KANs** are positioned as a cutting-edge method for **computational drug discovery** and **molecular interaction analysis**.

## 4.5 KAN Optimization

For **Kolmogorov-Arnold Networks (KAN)** to predict **binding affinity** accurately and efficiently, **optimization** is essential. The **Adam optimizer**, evaluation metrics, and early stopping are among the main KAN approaches highlighted in this section.

### 1. Using Adam Optimizer for Gradient-Based Optimization

The **Adam optimizer** combines momentum and learning rates to adaptively update model parameters:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

Where:

- $\theta_t$ : Model parameters at iteration  $t$ .
- $\eta$ : Learning rate.
- $\hat{m}_t$ : Bias-corrected first moment estimate (gradient).
- $\hat{v}_t$ : Bias-corrected second moment estimate (squared gradient).
- $\epsilon$ : Small constant for numerical stability.

**Figure – 4.5.1**

In contrast to conventional gradient descent, this optimizer guarantees reliable and quicker convergence.

### 2. Metrics for Evaluation

#### Mean Squared Error (MSE)

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

- MSE measures the average squared difference between actual ( $y_i$ ) and predicted ( $\hat{y}_i$ ) values, penalizing larger errors more heavily. It is used as the **loss function**.

#### Mean Absolute Error (MAE)

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

- MAE computes the average absolute error, offering a more interpretable metric for **model evaluation**.

**Figure – 4.5.2**

### **3. Early Termination**

**Early stopping** tracks validation loss and stops training when it plateaus or rises in order to avoid overfitting. The validation loss is computed as follows:

$$L_{\text{val}} = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} (y_i - \hat{y}_i)^2$$

**Figure – 4.5.3**

This ensures the model maintains generalization without overfitting the training data.

Optimization Technique	Purpose	Implementation in KAN
Adam Optimizer	Minimizes the loss function	Combines momentum and adaptive learning rates
Early Stopping	Prevents overfitting	Monitors validation loss
Mean Squared Error (MSE)	Loss function	Penalizes larger errors for accurate fitting
Mean Absolute Error (MAE)	Evaluation metric	Measures average absolute error

**Table – 4.5**

For **binding affinity prediction**, **KAN** achieves reliable performance and effective optimization by combining the **Adam optimizer**, evaluation metrics (**MSE** and **MAE**), and early halting. These methods demonstrate how **KAN** is useful for **computational drug development** and how well it adapts to **high-dimensional molecular datasets**.

### **4.6 Binding Affinity Prediction Loss Function**

A crucial part of **Kolmogorov-Arnold Networks (KAN)**, the **loss function** measures the discrepancy between expected and actual binding affinities and directs the optimization process. Robustness to data variability is maintained while efficient learning is ensured via a suitable loss function.

#### **1. Binding Affinity Loss Functions**

In regression tasks, **loss functions** quantify the discrepancy between actual values and forecasts. Typical loss functions are as follows:

- **MSE (Mean Squared Error):**
  - Ideal for accurate predictions on clean datasets because it penalizes larger errors more severely.
- **Mean Absolute Error (MAE):**
  - Provides robustness to noisy data and outliers by penalizing errors linearly.
- **Huber Loss:**
  - Combines the robustness of MAE with the sensitivity of MSE, making it appropriate for datasets with variable noise levels or mixed distributions.

## 2. Loss Choice in KAN

Because it can penalize huge deviations, **KAN** predominantly uses **MSE** as its loss function, which aligns with the goal of accurately modeling molecular interactions. However, to lessen vulnerability to extreme values, alternatives like **Huber Loss** or **MAE** may be selected for datasets with substantial outliers.

## 3. Loss Function's Role in Optimization

An essential component of **KAN's optimization procedure** is the **loss function**:

- **Guiding Gradient Updates:** To minimize prediction errors, weight updates are guided by the gradient of the loss function.
- **Balancing Performance:** The loss function ensures that **KAN** efficiently captures both individual molecule properties and their interactions by penalizing errors.

## 4. Handling Outlier Difficulties

**Optimization** can be disproportionately affected by **outliers**, particularly when **MSE** is used. Among the tactics to deal with this are:

- **Robust Loss Functions:** Replace **MSE** with **MAE** or **Huber Loss** to lessen sensitivity to extreme values.
- **Preprocessing of the Data:** Outlier removal and **normalization strategies** reduce the impact of extreme values on model performance.

The choice of **loss function** is essential to **KAN's ability to predict binding affinity**. While robust alternatives like **Huber Loss** or **MAE** can handle noisy datasets, **MSE** ensures accurate predictions. **KAN** advances computational drug discovery by achieving accurate and

reliable performance through the integration of an appropriate **loss function** with the optimization methodologies discussed in **Subchapter 4.5**.

## **5. Summary**

The **Kolmogorov-Arnold Networks (KAN)** project tackles important issues in computational drug discovery like scalability, accuracy, and transparency by introducing a novel and interpretable framework for **binding affinity prediction**. **KAN** efficiently handles high-dimensional information and produces interpretable predictions by breaking down molecular interactions into **univariate** and **bivariate components** using the **Kolmogorov-Arnold representation theorem**.

Because of its strong **computing efficiency** and capacity to capture intricate **chemical interactions**, **KAN** performs better than conventional techniques like **molecular docking** and general-purpose **machine learning models**. Its value is demonstrated in applications such as **virtual screening**, **protein-ligand interaction modeling**, and **drug discovery pipelines**, where the integration of sophisticated preprocessing techniques, optimization methodologies, and evaluation metrics guarantees accurate predictions across a variety of datasets. **KAN** becomes a potent tool in **cheminformatics** by fusing **computational rigor** with useful design, providing a revolutionary method to speed up the prediction of molecular properties and the development of therapeutics.

## **6. Conclusion**

By using the **Kolmogorov-Arnold representation theorem** to break down complex molecular interactions into univariate and bivariate components, the Kolmogorov-Arnold Networks (KAN) research has developed a unique method for **binding affinity prediction**. By improving prediction **accuracy**, **scalability**, and **interpretability**, this deconstruction overcomes major drawbacks of conventional techniques like molecular docking and general-purpose machine learning models.

The project retains important **structural** and **chemical information** while **removing** the preprocessing hassles associated with sequence-based formats like **SMILES** by using

**ECFP8b fingerprints** as molecular representations. Robust and dependable model performance across a variety of datasets is ensured by combining the **Adam optimizer**, **K-fold cross-validation**, and sophisticated evaluation measures like **MSE** and **MAE**.

**KAN** has demonstrated its effectiveness as a **computational drug discovery method**, allowing for precise **binding affinity prediction** and smooth scaling to **high-dimensional molecular datasets**. It is a formidable **cheminformatics framework** that can expedite the creation of **novel therapeutic solutions** and optimize **drug discovery pipelines** due to its transparent forecasts and versatility.

## **7. Scope for Future Work**

There is a lot of room to grow the **Kolmogorov-Arnold Networks (KAN)** framework in order to handle new problems in **computational drug development**. Extending **KAN** to manage **multi-target binding predictions**, which would allow it to simulate affinities for several protein targets at once, is one encouraging avenue. This would be very helpful in **multi-target drug design** and **polypharmacology**, where it's crucial to interact with numerous targets at once. Furthermore, by **combining KAN with generative models**, it may be possible to predict **binding affinities** and **create new compounds** tailored to particular therapeutic objectives. Improving KAN's ability to handle big, noisy, and unbalanced datasets would make it even more useful for tackling cheminformatics problems in the real world.

Future research could also focus on using **KAN's interpretability** and **scalability** to forecast other molecular characteristics such as **ADMET profiles**, **toxicity**, and **solubility**. **Drug development pipelines** would be streamlined by these extensions, offering a thorough assessment of candidate compounds. Investigating **hybrid strategies** that blend **KAN** with **transformer architectures** or **graph-based representations** may enhance its performance and adaptability for intricate molecular information. By tackling these issues, KAN has the potential to transform the way molecular interactions and characteristics are modeled in computational biology and develop into a flexible and essential tool in cheminformatics.



## **8. References**

1. Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov-Arnold Networks," [Preprint], 2024. DOI: [10.48550/arXiv.2404.19756](https://arxiv.org/abs/2404.19756). [Online]. Available: <https://arxiv.org/abs/2404.19756>
2. R. Gorantla, A. Kubincová, A. Y. Weiße, and A. S. J. S. Mey, "From Proteins to Ligands: Decoding Deep Learning Methods for Binding Affinity Prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 7, pp. 2496–2507, 2024. DOI: [10.1021/acs.jcim.3c01208](https://pubs.acs.org/doi/10.1021/acs.jcim.3c01208). [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.jcim.3c01208>
3. R. Gorantla, A. Kubincová, B. Suutari, B. P. Cossins, and A. S. J. S. Mey, "Benchmarking Active Learning Protocols for Ligand-Binding Affinity Prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 6, pp. 1955–1965, 2024. DOI: [10.1021/acs.jcim.4c00220](https://pubs.acs.org/doi/10.1021/acs.jcim.4c00220). [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.jcim.4c00220>
4. X. Liu, S. Jiang, X. Duan, A. Vasan, C. Liu, C.-C. Tien, H. Ma, T. Brettin, F. Xia, I. T. Foster, and R. L. Stevens, "Binding Affinity Prediction: From Conventional to Machine Learning-Based Approaches," [Preprint], Oct. 2024. DOI: [10.48550/arXiv.2410.00709](https://arxiv.org/abs/2410.00709). [Online]. Available: <https://arxiv.org/abs/2410.00709>.
5. Z. Wang, L. Zheng, Y. Liu, Y. Qu, Y.-Q. Li, M. Zhao, Y. Mu, and W. Li, "OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells," [Preprint], Mar. 2021. DOI: [10.48550/arXiv.2103.11664](https://arxiv.org/abs/2103.11664). [Online]. Available: <https://arxiv.org/abs/2103.11664>.
6. R. Meli, G. M. Morris, and P. C. Biggin, "Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-Based Deep Learning: A Review," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022. DOI: [10.3389/fbinf.2022.885983](https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.885983). [Online]. Available: <https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.885983/full>
7. Y. Yamaguchi, Y. Hasegawa, and K. Nakai, "Machine Learning Methods for Protein-Protein Binding Affinity Prediction," *Frontiers in Bioinformatics*, vol. 2, Dec. 2022. DOI: [10.3389/fbinf.2022.1065703](https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.1065703). [Online]. Available: <https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2022.1065703/full>