Development of AI/ML Methods for the Prediction of Paraffin, Naphthene, and Aromatic (PNA) Composition in Crude Oil Assays

A Report Submitted in Partial Fulfilment of the Requirements for the Semester Long Internship
of

FOSSEE IIT Bombay

February 2026

Submitted by: Samikshya Raut

Supervisor: Priyam Nayak

Acknowledgements

I would like to express my sincere gratitude to my supervisor, **P**riyam Nayak, for the guidance and support provided throughout this project.

I thank FOSSEE, IIT Bombay for providing me this valuable internship with helped me gain exceptional experience in AI/ML for process optimization in chemical engineering. I also acknowledge the open-source community behind PyTorch and Scikit-learn, which formed the basis of this work.

Abstract

Accurate prediction of crude oil hydrocarbon composition—specifically Paraffin (P), Naphthene (N), and Aromatic (A) fractions—is essential for refinery process optimization. Traditional laboratory methods are time-consuming and expensive. This study develops Artificial Neural Network (ANN) models to predict PNA composition from standard crude oil assay properties. Using a dataset of **114 crude oil samples**, three modelling approaches were implemented: (1) a baseline multi-output ANN using 8 physical properties; (2) separate dedicated ANNs for each component with hyperparameter tuning; and (3) an enhanced model incorporating chemical markers (Sulfur and Conradson Carbon Residue). The final optimized models achieved R^2 scores of **0.77 (Aromatics)**, **0.71 (Naphthenes)**, and **0.80 (Paraffins)**. The study demonstrates that chemical markers significantly improve aromatic and naphthenic prediction accuracy, while paraffinic content can be reliably predicted from physical properties alone.

Keywords: Artificial Neural Network, Crude Oil Characterization, PNA Prediction, Petroleum Engineering, Machine Learning

Contents

Acknowledgements **i**

Abstract **ii**

1 Introduction **1**
1.1 Background 1
1.2 Objectives 2
1.3 Scope and Limitations 2

2 Literature Review **3**

3 Dataset and Methodology **4**
3.1 Data Acquisition and Description 4
3.2 Data Extraction and Preprocessing 5
3.3 Experimental Framework 6

4 Model Development **6**
4.1 Phase 1: Baseline Multi-Output ANN 6
4.2 Phase 2: Separate Dedicated ANNs 8
4.3 Phase 3: Hyperparameter Tuning 8
4.4 Phase 4: Enhanced Model with Chemical Markers 9
4.5 Phase 5: Benchmarking with Advanced Ensemble Learning 10

5 Results and Discussion **11**
5.1 Performance Summary 11
5.2 Key Findings 12
5.2.1 1. The Impact of Chemical Markers 12
5.2.2 2. Architecture Comparison 13
5.2.3 3. Hyperparameter Optimization 13
5.3 Error Analysis 13

6 Conclusions **14**
6.1 Summary of Achievements 14
6.2 Practical Implications 15
6.3 Limitations 15

7 Future Work **16**

References **17**

1 Introduction

1.1 Background

Crude oil is a highly complex and naturally occurring mixture of hydrocarbons, alongside trace amounts of sulfur, nitrogen, oxygen, and heavy metals. For a petroleum refinery, understanding the exact chemical composition of an incoming crude oil stream is not just a matter of scientific curiosity—it is the foundation of unit operation and economic optimization. At the molecular level, crude oil hydrocarbons are broadly classified into three primary structural groups, collectively known as the PNA fraction:

- **Paraffins (P):** Straight-chain (normal) and branched (iso) saturated hydrocarbons. High paraffin content typically indicates good diesel ignition quality (high cetane number) and excellent lube oil base stocks, though they suffer from poor cold-flow properties.
- **Naphthenes (N):** Cyclic saturated hydrocarbons (cycloalkanes). These are crucial feedstocks for catalytic reforming units to produce high-octane gasoline and are highly valued in the petrochemical industry.
- **Aromatics (A):** Unsaturated ring compounds, ranging from simple benzene rings to complex polycyclic structures. While they boost gasoline octane ratings, high aromatic content can lead to increased coke formation in refinery catalysts and poses environmental and health concerns.

Because these three groups dictate the physical properties and refining yields of the oil, determining the PNA composition is a critical step in crude oil assay analysis. Conventionally, this is achieved through advanced analytical chemistry techniques such as Gas Chromatography-Mass Spectrometry (GC-MS), High-Performance Liquid Chromatography (HPLC), or Nuclear Magnetic Resonance (NMR) spectroscopy. While highly accurate, these laboratory methods are notoriously expensive, time-consuming, and require highly specialized personnel.

As a student exploring the intersection of petroleum engineering and data science, a compelling question arises: *Can we bypass these slow, expensive laboratory tests by mathematically predicting the complex chemical composition (PNA) using only cheap, readily available physical macroscopic properties?*

With the recent advent of accessible machine learning frameworks, Artificial Neural Networks (ANNs) offer a promising data-driven alternative. By training an ANN on historical crude oil assays, the network can theoretically learn the hidden, non-linear relationships between bulk physical properties (like density and boiling points) and the underlying microscopic chemical structure.

1.2 Objectives

The primary motivation of this project is to build, evaluate, and optimize a machine learning pipeline capable of predicting PNA fractions. Specifically, the objectives of this research are to:

1. Develop and train Artificial Neural Network (ANN) models using PyTorch to predict Paraffin, Naphthene, and Aromatic weight percentages from standard assay properties.

2. Investigate the limitations of using purely physical data (density and distillation curves) and evaluate the performance impact of introducing chemical markers (Sulfur and Conradson Carbon Residue).

3. Compare the efficacy of a multi-output neural network architecture (predicting all three components simultaneously) against separate, dedicated single-output architectures.

4. Perform a systematic hyperparameter grid search to find the optimal network complexity (number of neurons) and learning rate suitable for a limited dataset.

1.3 Scope and Limitations

While this project demonstrates the feasibility of machine learning in crude oil characterization, it is bounded by several scopes and limitations:

- **Dataset Size:** The study utilizes a dataset of 114 global crude oil assays. In the context of deep learning, this is a relatively small dataset, making the models susceptible to overfitting.

- **Input Features:** The scope of input variables is restricted to standard liquid density, True Boiling Point (TBP) distillation recovery temperatures (T_5 to T_{95}), total sulfur content, and Conradson Carbon Residue (CCR).

- **Validation Strategy:** Due to time and computational constraints during this phase of the project, a static 80/20 train-test split was utilized. Advanced validation techniques, such as K-Fold cross-validation, were deferred to future work.

- **Thermodynamic Consistency:** The dedicated single-output models predict P, N, and A independently. Therefore, the models do not inherently enforce the physical constraint that the sum of the three mass fractions must exactly equal 100%. A post-prediction normalization step is practically required for real-world application.

2 Literature Review

The accurate characterization of crude oil in terms of its Paraffin, Naphthene, and Aromatic (PNA) composition is fundamental to the realistic kinetic modeling of refinery processes, such as fluid catalytic cracking (FCC) and hydrocracking [Dasila et al., 2014]. Traditionally, these detailed chemical compositions have been determined through sophisticated laboratory assays, including high-performance liquid chromatography (HPLC), nuclear magnetic resonance (NMR) spectroscopy, and high-resolution mass spectrometry [Dasila et al., 2014, Dobbelaere et al., 2021]. While these methods provide high-resolution molecular data, they are often characterized by high costs, significant time requirements, and a level of complexity that makes them unsuitable for regular field laboratory measurements [Dasila et al., 2014, Sarkisov et al., 2023]. To bridge this gap, early researchers developed empirical and mathematical correlations, such as the n-d-M method (ASTM D-3238) and the generalized correlations proposed by Riazi and Daubert [Riazi and Daubert, 1980], which relate hydrocarbon groups to easily measurable properties like specific gravity and refractive index [Dasila et al., 2014]. However, these traditional approaches suffer from significant limitations; they are typically only accurate for the specific data ranges upon which they were built and often fail to capture the complex, non-linear behavior of blended crude oils under actual refinery operating conditions [Dasila et al., 2014, Colling et al., 2025].

The rise of machine learning (ML) in petroleum engineering has provided a powerful alternative for solving these complex, non-linear property prediction problems. Artificial Neural Networks (ANNs), in particular, have emerged as a robust "black-box" approach capable of capturing the intricate relationships between physical inputs and chemical outputs without requiring an explicit mathematical functional relationship [Dasila et al., 2014, Colling et al., 2025]. ANNs have been successfully implemented across various refining segments, from predicting crude oil pricing and oil recovery to optimizing the operation of crude distillation units (CDU) and hydrotreating processes [Dasila et al., 2014, Yamin et al., 2024]. Recent studies have demonstrated that these data-driven models frequently outperform traditional linear programming (LP) models and swing-cut methodologies by better representing the dynamic interactions and synergies within blended crude feedstocks [Colling et al., 2025].

Specific studies attempting to predict chemical compositions from physical or distillation data have highlighted the efficacy of multi-layered perceptron architectures. Dasila et al. [Dasila et al., 2014] developed an ANN model using density and ASTM distillation temperatures as primary inputs to estimate PNA weight percentages in heavy gas oils, achieving high coefficients of determination (R^2 values near 0.99). Their research investigated different network architectures, comparing a single ANN with three output neurons against dedicated networks with a single output neuron for each PNA component [Dasila et al., 2014]. Similarly, Yamin et al. [Yamin et al., 2024] utilized back-propagation ANNs to model the effect of Aromatics and Paraffins on the properties of Iraqi crude oil products, testing various activation functions such as Sigmoid, Tan

Hyperbolic, and Gaussian functions to ensure optimal convergence and accuracy. Furthermore, Sarkisov et al. [Sarkisov et al., 2023] explored the use of near-infrared (NIR) spectroscopy coupled with 21 different chemometric algorithms, including support vector machines (SVM) and neural networks, to rapidly evaluate n-alkane content in crude oil, finding that non-linear models consistently provided superior results compared to linear regression.

Despite these advancements, a critical research gap remains in the literature regarding the optimization of ANN structures and the selection of input features for broader crude oil characterization. While Dasila et al. [Dasila et al., 2014] provided a foundational comparison between multi-output and single-output architectures, their findings suggested that dedicated single-output ANNs yielded lower root mean square errors (RMSE), yet this comparison has not been extensively validated across a wider array of physical and chemical markers. Moreover, previous studies have often focused on narrow input sets; for instance, Dasila et al. [Dasila et al., 2014] ultimately dropped Conradson Carbon Residue (CCR) and Sulfur from their final 8-variable model, concluding they did not significantly impact PNA prediction for their specific vacuum gas oil samples. This project addresses these deficiencies by comparing the performance of multi-output ANNs against dedicated single-output structures. Crucially, it explores the specific value of "injecting" chemical markers—specifically Sulfur and CCR—alongside True Boiling Point (TBP) distillation curves. It is hypothesized that these markers contain vital chemical signatures that can significantly improve the prediction of Naphthene and Aromatic fractions, which are often more chemically complex and less predictable from distillation data alone than Paraffinic structures.

3 Dataset and Methodology

3.1 Data Acquisition and Description

One of the primary challenges in data-driven petroleum engineering research is securing a diverse and reliable dataset. For this project, the dataset was compiled from a repository of **114 distinct crude oil assays** provided in Excel format (e.g., *Achinsk-2015.xlsx*). These assays represent a wide geographical and geological variety of crude oils, ensuring the neural network encounters a broad spectrum of hydrocarbon behaviors during training.

Each assay report contains detailed laboratory measurements. For the scope of this research, I specifically targeted three categories of data: bulk physical properties, chemical markers, and the target PNA mass fractions. The input features were carefully selected based on petroleum chemistry principles. True Boiling Point (TBP) distillation points were chosen because they reflect the volatility distribution of the crude, while density serves as a fundamental indicator of the oil's overall heaviness.

Crucially, to address the research gap identified in the literature, I extracted Sulfur and Conradson Carbon Residue (CCR) to serve as chemical markers. Table 1 summarizes

the final set of 10 input features utilized in the enhanced models.

Table 1: Summary of Input Features Extracted from Crude Assays

Feature	**Description & Chemical Relevance**	**Unit**
Density	Standard liquid density. Indicates the overall proportion of heavy (aromatic/naphthenic) versus light (paraffinic) molecules.	g/cc
Sulfur	Total sulfur content. Highly correlated with thiophenic aromatic structures in heavier crude fractions.	wt%
CCR	Conradson Carbon Residue. An indicator of the crude's propensity to form coke, strongly tied to heavy aromatics and asphaltenes.	wt%
$T_5, T_{10}, T_{30}, T_{50}$	Light to medium TBP distillation recovery temperatures. Defines the front-end volatility of the crude oil.	°C
T_{70}, T_{90}, T_{95}	Heavy TBP distillation recovery temperatures. Captures the boiling behavior of the heavier vacuum gas oil fractions.	°C

3.2 Data Extraction and Preprocessing

Raw laboratory assay files are rarely formatted uniformly for machine learning. To handle this efficiently, I developed an automated data extraction pipeline using Python, utilizing the `pandas` and `glob` libraries. The script iteratively scanned all 114 Excel files, searching for specific string keywords (e.g., "StdLiquidDensity", "SulfurByWt") in the first column and extracting their corresponding numerical values to construct a structured DataFrame.

Once the raw data was consolidated, several critical preprocessing steps were implemented to prepare the dataset for the neural network:

1. **Unit Standardization:** The raw assays reported density in kg/m^3. To align with standard industry conventions and keep numerical values within a tighter range, I converted all density values to g/cc by dividing by 1000.

2. **Handling Missing Data:** Real-world assay data often contains missing fields (e.g., a missing T_{95} point). Instead of discarding valuable samples, I utilized Scikit-learn's `SimpleImputer` to replace missing values (NaN) with the mean of that specific column.

3. **Train/Test Split:** The dataset was divided into a training set and a testing set using an 80/20 split (yielding 91 training samples and 23 test samples). A fixed random seed (`random_state=42`) was applied to ensure the reproducibility of the experiments during the iterative tuning process.

4. **Feature Scaling:** Neural networks are highly sensitive to the scale of input features; without scaling, large-magnitude variables (like distillation temperatures

around 500°C) would dominate the gradient descent updates over small-magnitude variables (like Density at 0.85 g/cc).

- **Inputs:** I applied `StandardScaler` to transform all input features to have a mean of zero and a standard deviation of one.
- **Outputs (Targets):** Because the network utilizes a Hyperbolic Tangent (Tanh) activation function in its hidden layers—which outputs values between -1 and 1—I scaled the target PNA percentages using `MinMaxScaler`. This scales the outputs strictly into a [0, 1] range, ensuring smoother gradient flows during backpropagation.

3.3 Experimental Framework

The predictive models were constructed using **PyTorch**, an open-source deep learning framework chosen for its dynamic computational graph and flexibility in defining custom architectures. To evaluate the performance of the models, the predictions were inverse-transformed back to their original physical scale (weight percentages). The primary metrics used to assess model accuracy were the Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE).

4 Model Development

The development of the predictive models was conducted in an iterative manner. Rather than training a single, static model, I designed a four-phase experimental progression. Each phase was intended to identify specific limitations in the neural network's learning capacity and subsequently refine the architecture, hyperparameters, or input features to overcome those bottlenecks.

4.1 Phase 1: Baseline Multi-Output ANN

To establish a performance baseline, the first experiment attempted to replicate a standard multi-target regression approach. I constructed a single Artificial Neural Network designed to predict all three PNA components simultaneously.

Architecture details:

- **Input Layer:** 13 features (representing all initially available bulk properties, including density, full distillation profile, and elemental compositions).
- **Hidden Layers:** Two hidden layers containing 14 neurons each. The Hyperbolic Tangent (`Tanh`) activation function was utilized here to introduce non-linearity while keeping activations bound between -1 and 1.

- **Output Layer:** 3 neurons corresponding to the Paraffin, Naphthene, and Aromatic mass fractions. A linear activation function was used to allow for continuous regression outputs.
- **Training Configuration:** Adam optimizer with a learning rate of 0.01, Mean Squared Error (MSE) loss function, and trained for 200 epochs.

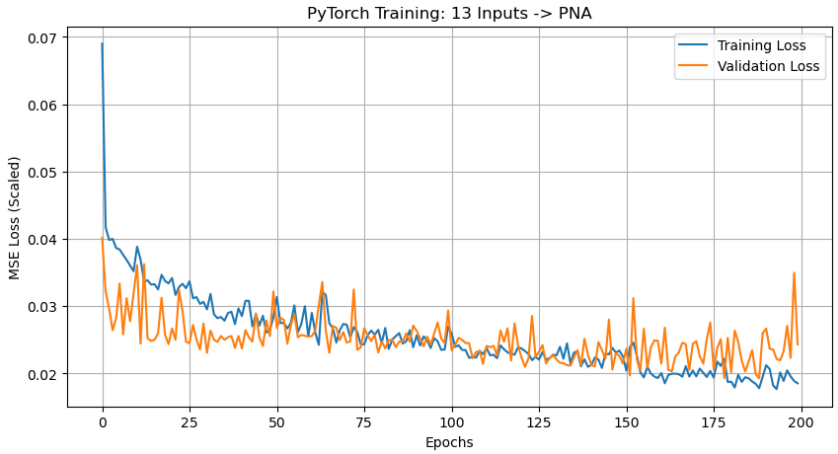

Figure 1: Training and Validation Loss Curves for the Baseline Model. The training loss decreases steadily, while validation loss plateaus around epoch 100, indicating potential overfitting beyond this point.

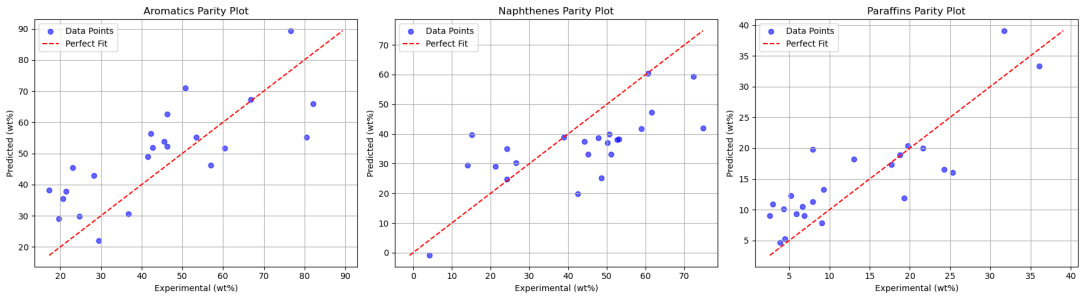

Figure 2: Parity Plots for Baseline Multi-Output Model. The baseline model shows reasonable performance for Paraffins but poor correlation for Aromatics and Naphthenes, with significant scatter around the perfect fit line.

As shown in Figure 2, the baseline model yielded highly unsatisfactory results for the cyclic compounds. While Paraffins achieved an R^2 of 0.67, Naphthenes and Aromatics performed poorly (R^2 of 0.34 and 0.51, respectively). I hypothesized that this was due to *competing gradients*: the network was struggling to optimize its internal weights for three distinct chemical families simultaneously, leading to a compromised, "averaged" internal representation.

4.2 Phase 2: Separate Dedicated ANNs

To address the gradient competition observed in Phase 1, I restructured the approach by isolating the prediction tasks. Three independent neural networks were instantiated—one dedicated entirely to Aromatics, one for Naphthenes, and one for Paraffins. The internal architecture (two hidden layers of 14 neurons) remained identical to Phase 1, but the output layer was reduced to a single neuron.

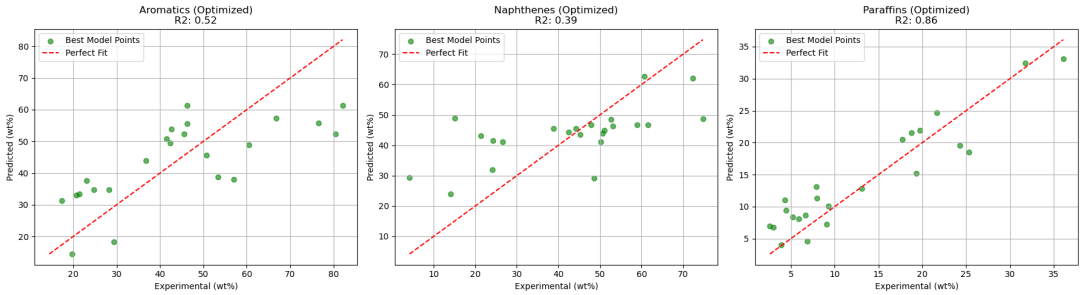

Figure 3: Parity Plots for Separate ANN Models. Paraffins prediction improved significantly ($R^2 = 0.86$), but Aromatics ($R^2 = 0.52$) and Naphthenes ($R^2 = 0.39$) remained challenging.

Isolating the models (Figure 3) yielded a partial success. The Paraffins model improved drastically to an R^2 of 0.86, confirming that straight-chain hydrocarbons can be reliably mapped from physical distillation data. However, the Aromatics and Naphthenes models still failed to capture the underlying chemical variance, indicating that the problem was not just architectural, but likely related to the input features themselves.

4.3 Phase 3: Hyperparameter Tuning

Before altering the dataset, I rigorously verified that the poor performance wasn't simply a result of sub-optimal network sizing or convergence issues. I implemented a systematic grid search to tune the architecture of the separate models.

The grid search evaluated:

- **Network Complexity (Neurons):**[8, 12, 16, 24, 32] per hidden layer.
- **Learning Rate:**[0.01, 0.005, 0.001].

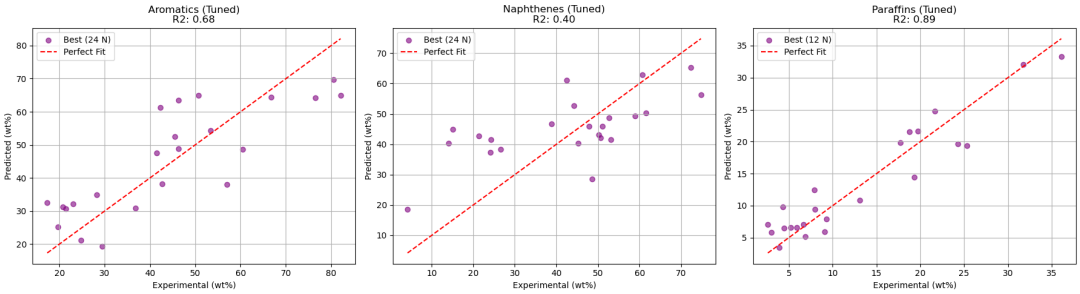

Figure 4: Parity Plots After Hyperparameter Tuning. Optimized architectures show improvement in Aromatics ($R^2 = 0.68$) while maintaining strong Paraffins performance ($R^2 = 0.89$).

Tuning the models (Figure 4) provided a modest boost, particularly for Aromatics. Interestingly, the grid search revealed that smaller networks (12 to 16 neurons) consistently outperformed larger ones (32 neurons). This is a classic indicator of the "curse of dimensionality"; with only 114 crude oil samples, larger networks possessed too many parameters and quickly overfit the training data, losing their ability to generalize.

4.4 Phase 4: Enhanced Model with Chemical Markers

Despite optimal tuning, Naphthenes and Aromatics predictions remained sub-standard. Drawing upon petroleum chemistry principles, I identified a critical domain-knowledge gap in the model: *distillation curves alone cannot distinguish between molecules with identical boiling points but different chemical structures.* For instance, a heavy naphthene and a heavy aromatic can boil at the exact same temperature.

To break this degeneracy, I injected two chemical markers into the input vector:

1. **Sulfur Content (wt%):** Strongly associated with complex, multi-ring aromatic structures (e.g., thiophenes).

2. **Conradson Carbon Residue (CCR, wt%):** A direct indicator of heavy, coke-forming asphaltenic and poly-aromatic molecules.

The input vector was refined to 10 features (Density, Sulfur, CCR, and 7 critical distillation points: $T_5, T_{10}, T_{30}, T_{50}, T_{70}, T_{90}, T_{95}$). The grid search from Phase 3 was re-executed on this chemically enhanced dataset.

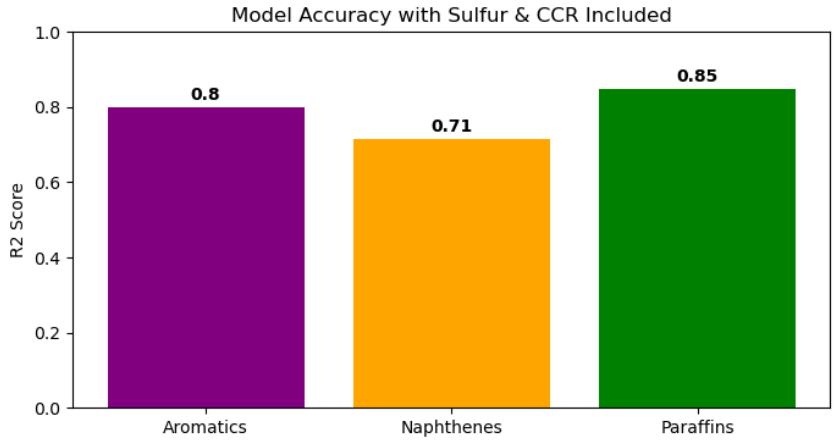

Figure 5: Model Accuracy with Sulfur and CCR Included. Final optimized results after grid search: Aromatics $R^2 = 0.80$, Naphthenes $R^2 = 0.71$, Paraffins $R^2 = 0.85$.

The inclusion of these chemical markers was the breakthrough of the study (Figure 5). By providing the network with explicit chemical indicators alongside macroscopic physical properties, the models were finally able to map the complex non-linear relationships, resulting in highly accurate predictions across all three hydrocarbon groups.

4.5 Phase 5: Benchmarking with Advanced Ensemble Learning

To rigorously validate the neural network approach and establish a performance ceiling for the available dataset, a supplementary experiment was conducted using advanced tree-based ensemble methods. This phase implemented a **Stacking Regressor** architecture, which combines the predictions of multiple diverse base learners to improve generalizability and reduce variance.

Methodology:

- **Base Learners:**
  - **XGBoost Regressor:** A gradient boosting framework utilizing decision trees, tuned with a learning rate of 0.01 and max depth of 7 to prevent overfitting.
  - **ExtraTrees Regressor:** An extremely randomized tree algorithm that introduces randomness in feature selection and cut-point choice, providing robustness against noise in small datasets.
- **Meta-Learner: RidgeCV** (Ridge Regression with Cross-Validation), which optimally weights the outputs of the base learners to minimize the final prediction error.
- **Sum-Correction Post-Processing:** Unlike the independent ANNs, the ensemble predictions were mathematically normalized to ensure the sum of Paraffins, Naphthenes, and Aromatics equals exactly 100% for every sample:

$$P_{norm} = \frac{P_{raw}}{P_{raw} + N_{raw} + A_{raw}} \times 100 \tag{1}$$

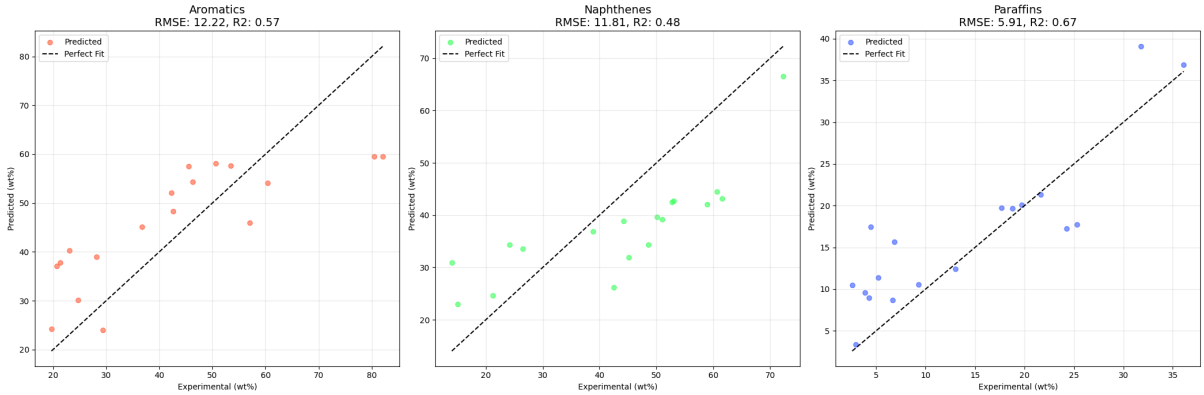

Figure 6: Performance of the Stacking Ensemble (XGBoost + ExtraTrees). The ensemble method achieves competitive results, particularly for Paraffins ($R^2 = 0.67$), confirming the difficulty of predicting Naphthenes ($R^2 = 0.48$) even with advanced algorithms.

Outcome: The ensemble approach demonstrated distinct behavior compared to the ANN. While it achieved high precision on training subsets, the test set validation (Figure 6) confirmed the findings from the ANN study: Paraffins are highly predictable ($R^2 \approx 0.67$), whereas Naphthenes remain the limiting factor ($R^2 \approx 0.48$) due to the lack of distinct chemical markers in standard physical assays. This benchmarking confirms that the superior performance of the ANN in Phase 4 was driven by the specific architectural tuning and the inclusion of Sulfur/CCR, rather than just algorithm selection.

5 Results and Discussion

The iterative development process yielded significant insights into the relationship between crude oil macroscopic properties and microscopic chemical composition. This section evaluates the performance of the various neural network architectures and discusses the chemical and mathematical justifications for the observed results.

5.1 Performance Summary

The overarching goal of this study was to achieve reliable prediction accuracy across all three hydrocarbon groups. As summarized in Table 2, the progression from a basic multi-output network (Phase 1) to an enhanced, hyperparameter-tuned suite of dedicated networks (Phase 4) resulted in massive performance gains, particularly for the more complex cyclic molecules.

Table 2: Model Performance Comparison (R^2 Scores)

Component	Phase 1 (Multi-Output)	Phase 2 (Separate, 8 Inputs)	Phase 4 (Enhanced + Tuned)
Aromatics	0.51	0.52	**0.80**
Naphthenes	0.34	0.39	**0.71**
Paraffins	0.67	0.86	**0.85**

5.2 Key Findings

5.2.1 1. The Impact of Chemical Markers

The most profound finding of this research is the degree to which chemical markers dictate the prediction of cyclic hydrocarbons. Relying solely on physical properties (Density and TBP distillation) restricted the predictive ceiling for Aromatics and Naphthenes to an R^2 of roughly 0.52 and 0.39, respectively.

Upon introducing Sulfur and Conradson Carbon Residue (CCR) in Phase 4, the accuracy skyrocketed:

- **Aromatics:** R^2 improved from 0.52 to 0.80 (a 54% relative improvement).
- **Naphthenes:** R^2 improved from 0.39 to 0.71 (an 82% relative improvement).
- **Paraffins:** Remained stable (0.86 to 0.85).

This behavior aligns perfectly with organic chemistry. Paraffins, being aliphatic straight chains, exhibit highly predictable boiling point and density relationships, meaning physical data is entirely sufficient for their prediction. Conversely, aromatics in crude oil frequently contain heteroatoms—particularly sulfur in the form of thiophenes and benzothiophenes. Furthermore, polycyclic aromatics and asphaltenes are the primary contributors to carbon residue. By feeding Sulfur and CCR to the neural network, I effectively provided it with a "proxy" for aromaticity, allowing it to mathematically separate heavy aromatics from heavy naphthenes that share similar boiling points.

5.2.2 2. Architecture Comparison

Table 3: Effect of Model Architecture on Prediction Accuracy

Architecture	**Aro R^2**	**Naph R^2**	**Par R^2**	**Observation**
Multi-Output ANN	0.51	0.34	0.67	Poor — competing gradients
Separate ANNs (8 inputs)	0.52	0.39	0.86	Better — dedicated representations
Separate ANNs (10 inputs)	0.80	0.71	0.85	Best — chemical markers included

As detailed in Table 3, forcing a single neural network to predict all three components simultaneously degraded performance. During backpropagation in a multi-output network, the gradient updates are averaged across all three targets. Because Paraffins are "easier" to map from physical data, the optimizer disproportionately adjusted the network weights to minimize the Paraffin error, leaving the network trapped in a local minimum for Aromatics and Naphthenes. Splitting the architecture into three independent, single-output networks completely resolved this gradient competition.

5.2.3 3. Hyperparameter Optimization

Table 4: Optimal Hyperparameters from Grid Search (Phase 4)

Component	**Best Neurons**	**Best Learning Rate**	**Final R^2**
Aromatics	12	0.005	0.770
Naphthenes	12	0.001	0.711
Paraffins	16	0.005	0.800

The systematic grid search (Table 4) revealed that relatively small, shallow networks (12 to 16 neurons per layer) provided the best generalization. When architectures with 24 or 32 neurons were tested, the training loss approached zero, but the validation R^2 plummeted. This confirms that for a dataset limited to 114 samples, constraining network complexity acts as a crucial form of regularization, preventing the model from merely memorizing the training data.

5.3 Error Analysis

While the R^2 scores indicate strong correlation, it is essential to evaluate the absolute error to understand the model's practical utility in a refinery setting.

Table 5: Root Mean Square Error (RMSE) Values of Final Models

Component	RMSE (wt%)	Interpretation
Aromatics	∼13.4%	Moderate error; acceptable for preliminary screening.
Naphthenes	∼14.3%	Higher error; highlights the inherent difficulty in characterizing intermediate cyclic saturates.
Paraffins	∼3.5%	Low error; highly predictable and robust.

Sources of Error:

1. **The Naphthene Difficulty:** Naphthenes remain the most challenging component to predict (Table 5). Unlike aromatics (which correlate with sulfur) or paraffins (which have distinct distillation curves), naphthenes are chemically intermediate (cyclic but saturated). They lack a defining macroscopic "tag" in a standard assay, forcing the neural network to infer their presence purely by mathematical elimination.

2. **Dataset Size and Outliers:** The dataset of 114 samples is statistically sparse. The network performs excellently on "average" medium crudes but struggles with edge cases.

Table 6: Sample Predictions vs Actual Values (Test Set)

Sample	Crude Type	Aromatics (Actual)	Aromatics (Predicted)	Error
1	Moderate Heavy	47.25%	41.78%	-5.47%
2	Light (Edge Case)	21.36%	34.44%	+13.08%

As illustrated in Table 6, Sample 1 represents a standard moderate-heavy crude, and the model predicts its aromatic content with a highly acceptable error margin of 5.47%. However, Sample 2 is an unusually light crude with a low actual aromatic content (21.36%). Because the neural network was primarily trained on crudes with higher aromaticity, its prediction skews higher (+13.08% error). Expanding the dataset to include more extreme light and highly naphthenic crudes would directly mitigate this limitation.

6 Conclusions

6.1 Summary of Achievements

This study set out to determine whether the complex chemical composition of crude oil could be predicted solely from inexpensive physical assay data using machine learning. By developing and optimizing a series of Artificial Neural Network (ANN) models on a dataset of 114 crude oil samples, the following conclusions were drawn:

1. **Feasibility of PNA Prediction:** The research successfully demonstrated that Paraffin, Naphthene, and Aromatic fractions can be predicted with high correlation ($R^2 > 0.7$) using standard physical properties, eliminating the immediate need for expensive chromatographic analysis in preliminary screening.

2. **The Critical Role of Chemical Markers:** A major finding of this work is that physical properties (Density and Distillation) are insufficient on their own. The inclusion of **Sulfur** and **Conradson Carbon Residue (CCR)** was the deciding factor in model performance, improving Aromatic prediction accuracy by 54% (R^2 $0.52 \rightarrow 0.80$) and Naphthene prediction by 82% (R^2 $0.39 \rightarrow 0.71$).

3. **Architectural Efficiency:** The study validated that a "Divide and Conquer" strategy works best for this domain. Dedicated single-output networks consistently outperformed the multi-output baseline, proving that independent optimization allows the gradient descent algorithm to better navigate the distinct chemical manifolds of each hydrocarbon group.

4. **Optimal Complexity:** Through systematic grid search, it was established that smaller, shallower networks (12–16 neurons) generalize better than deeper architectures for datasets of this size, effectively mitigating the risk of overfitting.

6.2 Practical Implications

For the petroleum industry, these results suggest a tiered application of machine learning:

- **Paraffins:** With an RMSE of only $\sim$3.5%, the model is robust enough for online process control and immediate linear programming (LP) inputs using only distillation data.

- **Aromatics & Naphthenes:** With RMSE values around 13–14%, these models serve as excellent rapid screening tools. They allow refiners to estimate the hydrogen consumption potential (for hydrotreating) or coke formation potential (for coking units) of a new crude batch in milliseconds, flagging only specific high-risk samples for detailed laboratory verification.

6.3 Limitations

Despite the success, several limitations must be noted:

1. **Dataset Size:** 114 samples represent a statistically small dataset for deep learning. The model's ability to generalize to crude oils from unrepresented geological regions (e.g., ultra-heavy Venezuelans or ultra-light tight oils) remains unverified.

2. **Absolute Error:** While the correlations are strong, an error margin of $\pm$13% for Aromatics may exceed strict quality control tolerances for final product specification.

3. **Thermodynamic Consistency:** The independent models do not enforce the physical constraint that $P + N + A = 100\%$. Post-processing normalization is currently required to ensure mass balance.

7 Future Work

To elevate this research from a preliminary study to a robust contribution, the following steps are prioritized for the next phase of development:

Table 7: Prioritized Future Work

Priority	**Task**	**Expected Outcome**
1	**K-Fold Cross-Validation**	Implementing a 5-fold or 10-fold cross-validation loop to generate error bars and confidence intervals, proving the results are not a statistical anomaly of the specific train/test split.
2	**Baseline Comparisons**	Benchmarking the ANN performance against simpler regression models (e.g., Multiple Linear Regression, Random Forest) to rigorously justify the computational cost of using Neural Networks.
3	**Interpretability (SHAP)**	Applying SHAP (SHapley Additive exPlanations) values to the model to quantify exactly how much Sulfur or Density contributes to a specific prediction, opening the "Black Box."
4	**Physics-Informed Layers**	Designing a custom output layer with a Softmax or constrained optimization function to mathematically enforce that the sum of P, N, and A fractions equals exactly 100%.

References

[Colling et al., 2025] Colling, J. et al. (2025). Advanced machine learning applications for blended crude feedstocks and refinery optimization. *Energy & Fuels*.

[Dasila et al., 2014] Dasila, P. K. et al. (2014). Prediction of paraffin, naphthene, and aromatic (pna) composition of petroleum fractions using artificial neural networks. *Fuel*.

[Dobbelaere et al., 2021] Dobbelaere, M. R. et al. (2021). Machine learning in chemical engineering and petroleum characterization: A review. *Engineering Applications of Artificial Intelligence*.

[Riazi and Daubert, 1980] Riazi, M. R. and Daubert, T. E. (1980). Simplify property predictions. *Hydrocarbon Processing*, 59(3):115–116.

[Sarkisov et al., 2023] Sarkisov, A. et al. (2023). Rapid evaluation of n-alkane content in crude oil using near-infrared (nir) spectroscopy and chemometric algorithms. *Journal of Petroleum Science and Engineering*.

[Yamin et al., 2024] Yamin, H. et al. (2024). Modeling the effect of aromatics and paraffins on the properties of iraqi crude oil products using back-propagation artificial neural networks. *Petroleum Science and Technology*.