KAPSARC
مركز الملك عبدالله للدراسات والبحوث البترولية
King Abdullah Petroleum Studies and Research Center

Global Crude Oil Refinery Models: Parametric and Nonparametric Methods

Evar Umeozor

September 2023

Doi: 10.30573/KS--2023-MP02

Acknowledgments

The author thanks all KAPSARC colleagues and support staff for their constructive discussions on the Oil Value-Chain Analyzer (KOVA) project under which this research was funded and for their help with the data acquisition procedures, respectively.

About KAPSARC

KAPSARC is an advisory think tank within global energy economics and sustainability providing advisory services to entities and authorities in the Saudi energy sector to advance Saudi Arabia's energy sector and inform global policies through evidence-based advice and applied research.

This publication is also available in Arabic.

Legal Notice

Abstract: There is a dearth of publicly available literature on reliable, evidence-based modeling representations of global crude oil refineries, which has led to a lack of transparency and consistency regarding how business and government policy processes are informed in many energy systems and energy transition modeling scenarios. This paper uses actual field data to identify that 40 unique refinery configurations account for the global refinery landscape. A machine learning algorithm training extremely randomized tree regressor (ETR) predictors is applied to develop a nonparametric production model of refineries. To facilitate computational convenience and suitable deployments in diverse use cases, a multivariate linear regression (MLR) model of refineries is also presented. For all refined products captured, the machine learning model demonstrates superior predictive performance, with coefficients of determination of over 90%, but both the machine learning model and the regression model are useful. Other performance metrics are assessed for both models.

***Keywords**: Crude oil, Refinery, Configuration, Machine learning, Regression*

Acronym	Definition
CDU	Crude distillation unit
CS	Condensate splitter
HDTR	Hydrotreater
LPG	Liquified petroleum gas
RFMR	Reformer
DSLR	Desulfurizer
ISMR	Isomerizer
VDU	Vacuum distillation unit
RCC	Resid catalytic cracker
FCC	Fluid catalytic cracker
DH	Distillate hydrocracker
RH	Resid Hydrocracker
CKR	Coker
NAP	Naphtha
GAS	Gasoline
KJF	Kerosene/jet fuel
DGO	Diesel/gas oil
HFO	Heavy fuel oil
COK	Coke
MLR	Multivariate linear regression
ETR	Extremely randomized trees regressor
HSO	Heavy sour crude oil
HSW	Heavy sweet crude oil
MSO	Medium sour crude oil
MSW	Medium sweet crude oil
LSO	Light sour crude oil
LSW	Light sweet crude oil
CAP	Design capacity of refinery
MMb/d	Million barrels per day
Kb/d	Thousand barrels per day

1. Introduction

Crude oil refineries produce finished petroleum products through a series of processing steps known as refinery unit operations. The varieties and capacities of process units, along with their operational characteristics, differentiate one refinery from another. Over the years, the refining industry has adopted various metrics to categorize refineries based on their capital investment and upgrading capability. These efforts have resulted in refinery complexity indicators such as the equivalent distillation capacity, the Nelson complexity index and the bottom of the barrel index (Kaiser 2017, Ogjresearch 2022). Although these metrics have helped quantify the refinery complexity and suitability for various crude oil feed qualities, they fail to provide relevant information for refinery modeling applications (Herce, Martini et al. 2022).

Refinery models are essential for assessing the economic and energy performance of refineries and potentially exploiting opportunities to optimize profitability, efficiency, and policy compliance. From a process design perspective, these models facilitate the development of optimal refinery configurations under prevailing operating and logistical conditions in local and/or regional markets. Business and government policymakers draw insights from modeling and analysis for investment selection, downstream diversification and regulatory policy design and compliance. Identifying the fundamental refinery configuration is crucial for realizing reliable and deployable refinery models.

Abella, Motazedi et al. (2015) used a predominantly North American database to identify 10 unique refinery configurations for model development. Their configuration-based modeling approach enabled the deployment of models for estimating refinery energy consumption and greenhouse gas (GHG) emissions. Doing so aided the identification and exploration of efficiency improvements and emission reduction opportunities (Motazedi, Abella et al. 2017). Such models are also useful for assessing crude feed blends and refinery configuration options to maximize the netback (Nduagu, Umeozor et al. 2018). They have also been extended to quantify the environmental impact of the refinery life cycle (Young, Hottle et al. 2019). However, for the effective application of configuration-based modeling, the accurate identification of refining process units and their attributes is crucial.

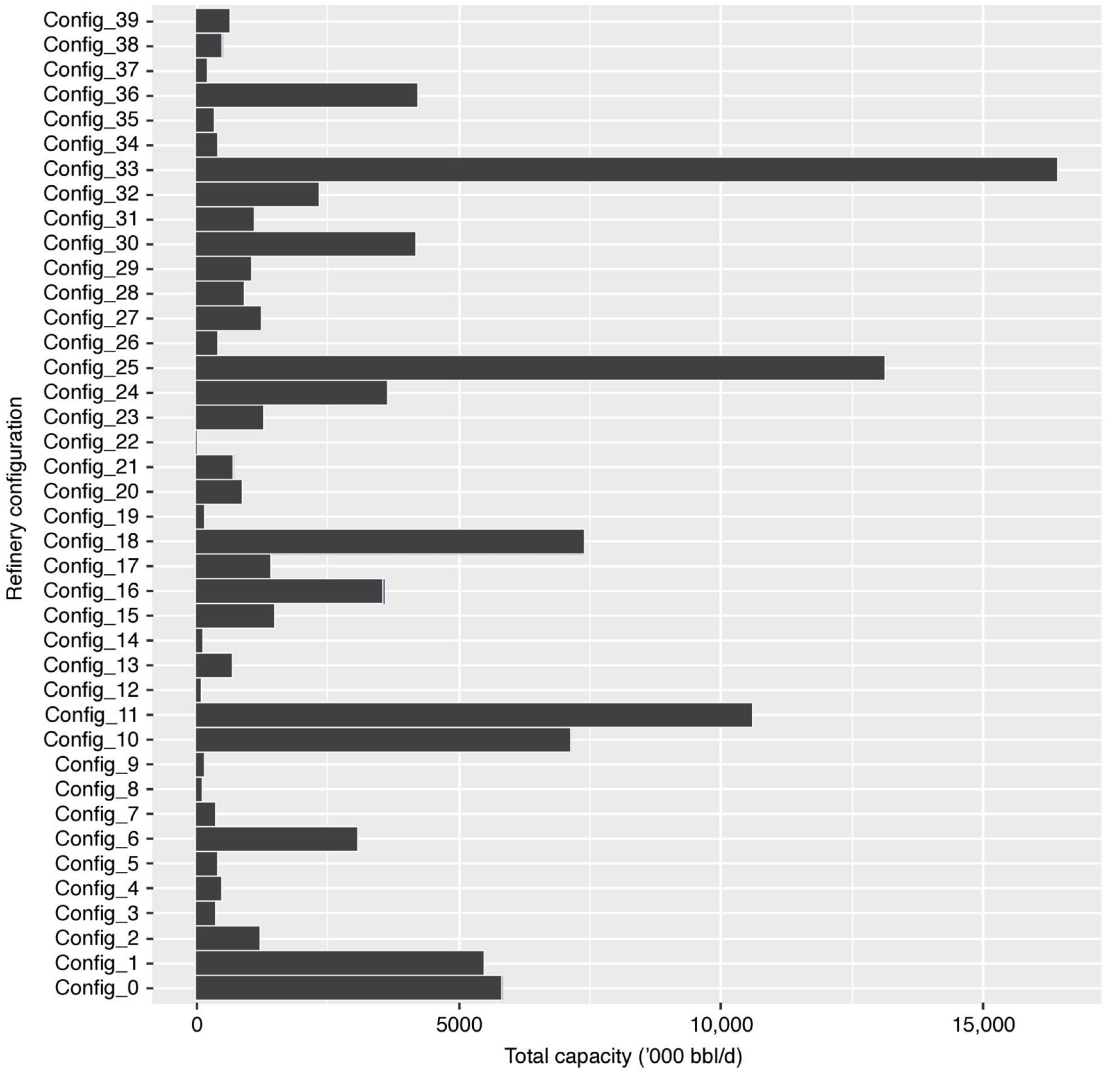
In 2020, there were 867 global crude oil refineries with active capacities of 103.32 million barrels per day (MMb/d), of which approximately 80% were located outside North America (Platts 2022). Identifying the unique configurations across the global crude oil refining landscape would facilitate individual, national, regional, and global refinery modeling. Such modeling could promote various application studies related to, for example, efficiency improvement, emission reduction, policy design, netback optimization, and investment competitiveness analysis. The rest of this paper is organized as follows: global refinery classification into design configurations is presented in the next section. In section 3, modeling methods and the data for both the parametric and nonparametric modeling of refinery configurations are presented. The modeling results are presented in section 4, and the conclusions and further research opportunities are discussed in section 5.

2. Global Refinery Configurations

In this work, the S&P Global Platts World Refinery Database (Platts 2022) is used to investigate global crude oil refinery configurations using process unit-level data for all operating refineries during the 2005-2020 period. A clustering technique is developed based on the availability of major process units and their capacities relative to other contiguous units in each refinery plant. For the entire study period, 40 unique refinery configurations are identified to represent all possible configurations that operated each year over the 16-year period. Essentially, 30 of the identified configurations have not been previously reported and represent those found outside the North American region, particularly in Asia. All identified refinery configurations are presented in supplementary information section SI.A.

Figure 1 shows the global refining capacity in 2020 by the identified refinery configurations. The refinery configurations are classified so that design complexity increases in relative terms from the

Figure 1: Global crude oil refinery capacities by refinery configuration (2020).

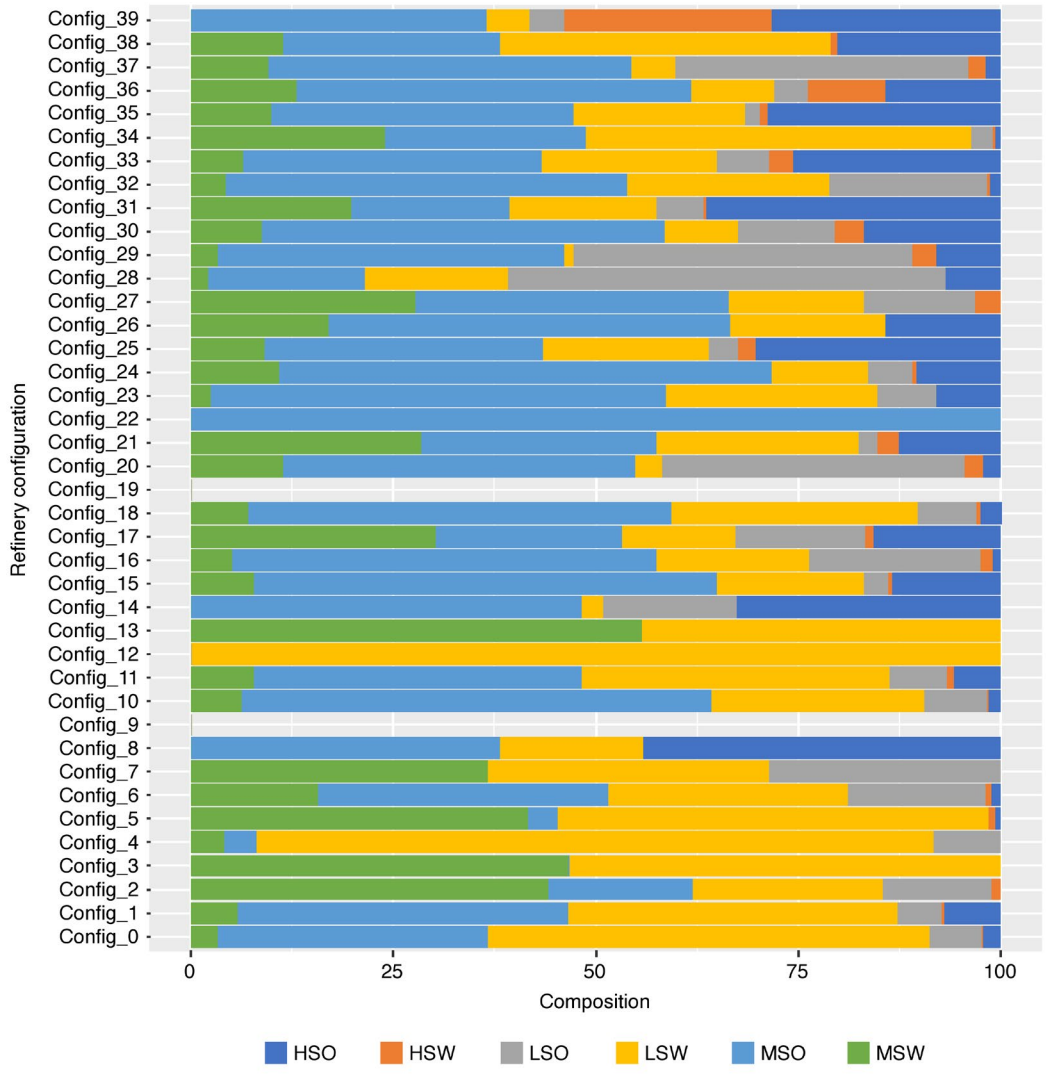


Source: KAPSARC Oil Value-Chain Analyzer

lower-numbered simpler refineries to the higher-numbered complex refineries. For example, the simplest refinery – also known as a hydroskimming refinery – is labeled configuration 0 (Config_0), while the most complex deep-conversion refinery is labeled Config_39. At approximately 16.4 MMb/d, the Config_33 refinery has the highest total capacity in the world, whereas there has been no existing Config_22 refinery capacity since 2016 due to either shutdowns or retrofitting to a different configuration class. The last Config_22 refinery was in operation in 2015, with a capacity of approximately 190.8 Kb/d.

The global refining capacity processes approximately 100 MMb/d of crude oil blends, which are normally categorized into 6 grades based on their sulfur content and density. Figure 2 shows the average global crude oil blend runs by refinery configuration between 2005 and 2020. Heavier (higher density)

Figure 2: Global crude oil blend runs by refinery configuration (2005-2020 average).

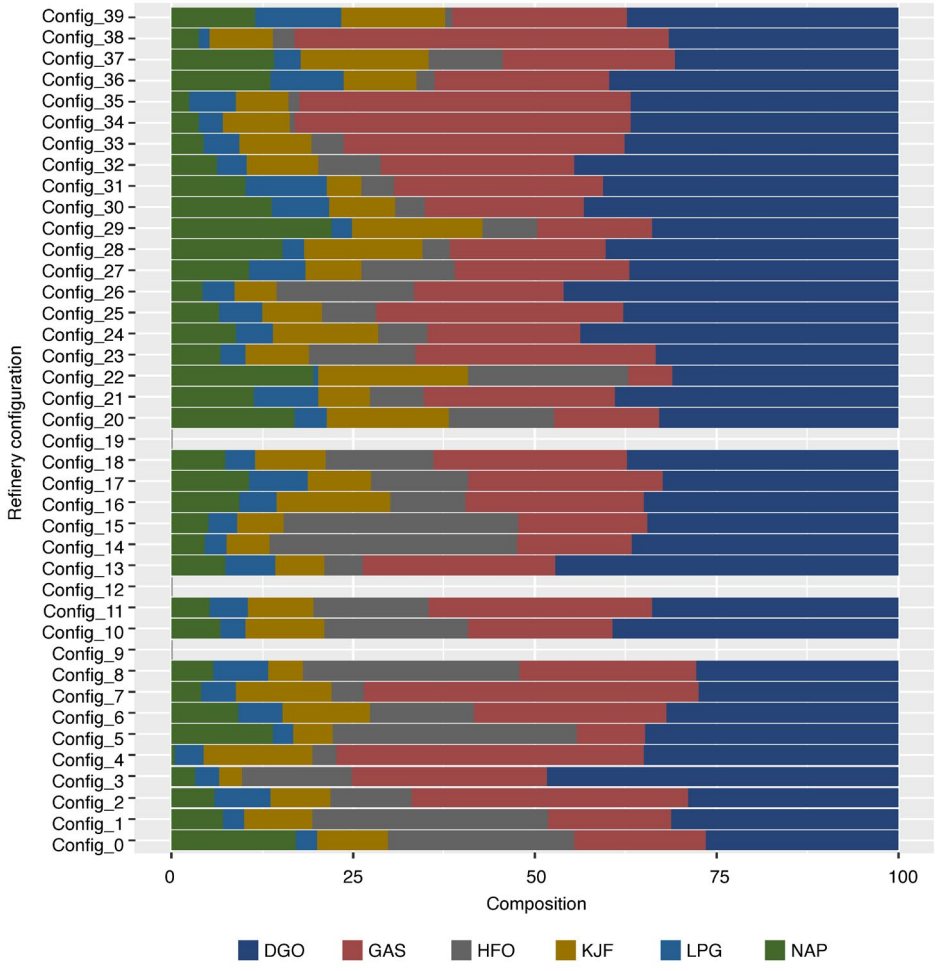


Source: KAPSARC Oil Value-Chain Analyzer

and sour (higher sulfur content) crudes, such as heavy sour crude oil (HSO), heavy sweet crude oil (HSW), and medium sweet crude oil (MSO), are mainly processed by higher complexity refineries, whereas light and medium crude grades (light sweet crude oil (LSW), medium sweet crude oil (MSW) and light sour crude oil (LSO)) are mostly refined at simpler and medium conversion refineries (see Table 1 for a description of crude grades). Furthermore, crude blends are often matched to each refinery configuration to maximize the production of the most desired products at the expense of the least desirable products. Other reasons for crude oil blending include satisfying environmental requirements such as sulfur content limits, exploiting the price differentials among crude grades in the market, utilizing locally available crude oil products, and optimizing the crack spread of refined products. Currently, the global crude run data for Config_9 and Config_19 refineries have yet to be captured in the source database of world refineries.

The combinations of refinery designs and their operations processing various crude oil blends yield various arrays of refined product slates. Figure 3 depicts the global average refined product

Figure 3: Global refined product slates by refinery configuration (2005-2020 average).



Source: KAPSARC Oil Value-Chain Analyzer

slates by refinery type between 2005 and 2020. In contrast to simpler refinery configurations, more complex refineries achieve deeper conversion of the crude feed to yield more lighter to middle distillate products (such as liquefied petroleum gas (LPG), naphtha (NAP), gasoline (GAS), kerosene/jet fuel (KJF) and diesel/gas oil (DGO)) and fewer bottom distillates and residues (such as heavy fuel oil (HFO)). Doing so also translates into higher yields of petrochemical feedstocks (such as LPG and NAP). The composition of products from complex refineries enables them to achieve better crack spreads through higher yields of more valuable products and the processing of heavier and/or sour crude grades, which are normally sold at discounts compared to benchmark lighter and/or sweet crude grades. At the time of this study, the source database had not reported other bottom products, such as resid and coke. Additionally, global production data for Config_9, Config_12, and Config_19 have not yet been captured in the database.

The configuration-based modeling of refineries has several benefits. It reduces the modeling task for large-scale studies encompassing multiregional and multitemporal analyses. It also reduces the number of modeling variables required to accurately specify the appropriate model for a given refinery configuration. The models utilize the larger volume of operational datasets from all refineries of similar configurations, capturing the variabilities in design capacities and uncertainties in operational schemes. Given the proprietary nature of refinery data, information on the applicable refinery configuration can inform the development of process simulator-based designs for further assessment studies.

Moreover, configuration-based modeling can drive best-practice information dissemination among refining plants operating in similar configurations through comparative analysis of performances across operators, countries, and regions. Such information dissemination can support transparent and standard refinery energy efficiency and emission assessments with a view to guiding toward opportunities to optimize relative operational performances focused on efficiency, economy, and emission reduction. Such models are also valuable for developing realistic energy outlook scenarios and exploring future potentials for the co-processing of renewable and hydrocarbon feedstocks in conventional refinery plants.

3. Material and Methods

A global crude oil refinery database is used to assemble design and operational data for the 2005-2020 period, covering over 11,000 observations of process unit-level capacities, crude blend runs and product slate yields (Platts 2022). The dataset is split into two samples for model training and testing at proportions of 70% and 30%, respectively.

A nonparametric model is developed by implementing the extremely randomized tree regressor (ETR) method in the Python programming language. Details on the ETR algorithm are available in (Geurts, Ernst et al. 2006). However, in this context, for a refinery learning data sample of size *N*, the application of the algorithm is expressed as follows:

$$ls_N = \left\{\left(x^i, y^i\right) : i = 1, \ldots, N\right\} \quad (1)$$

where x^i is the vector of explanatory variables (called features) and y^i is the vector of corresponding output variables (called targets). For the case of *n* features and *p* targets, both can be denoted as follows:

$$x^i = \left(x_1^i, \ldots, x_n^i\right) \quad (2)$$

$$y^i = \left(y_1^i, \ldots, y_p^i\right) \quad (3)$$

The sample values of the j^{th} feature organized in the order of increasing value are $\left(x_j^{(1)}, \ldots, x_j^{(N)}\right)$ for features $(j = 1, \ldots, n)$. An infinite ensemble of extremely random trees estimates the value of the target variable ($\hat{y}_p$) in the following form:

$$\hat{y}_p(x) = \sum_{i_1=0}^{N} \ldots \sum_{i_n=0}^{N} I_{(i_1,\ldots,i_n)}(x) \sum_{X \subset \{x_1,\ldots,x_n\}} \lambda_{(i_1,\ldots,i_n)}^{X} \prod_{x_j \in X} x_j \quad (4)$$

where $I_{(i_1,\ldots,i_n)}(x)$ is the characteristic function of the hyper-interval, as given by the following:

$$[x_1^{(i_1)}, x_1^{(i_1+1)}[\times \cdots \times [x_n^{(i_n)}, x_n^{(i_n+1)}$$

$$[. \qquad \forall (i_1, \ldots, i_n) \in \{0, \ldots, N\}^n \quad (5)$$

$\lambda_{(i_1,\ldots,i_n)}^{X}$ are real-valued parameters that depend on examples x^i and y^i, as well as the parameters for the minimum sample size for splitting a node (n_{min}) and the number of randomly selected features at each node (*K*). For our refinery model, the set of features consists of the plant design capacity (CDU basis), the runs of grades of crude oil feedstock, as categorized using density and sulfur contents (see Table 1) into LSW, LSO, MSW, MSO, HSW and HSO, and the types of refinery configurations. Overall, there are 47 features in the model, including the plant capacities, the 6 crude grades and the 40 refinery configurations. The ETR algorithm is implemented using the PyCaret modeling package in Python (Ali 2020).

Table 1. Categories of crude oil grades based on density and sulfur characteristics.

Crude Grade	Density (API)	Sulfur (%)
LSW	> 34	< 0.6
LSO	> 34	> 0.6
MSW	> 25 and < 34	< 0.6
MSO	> 25 and < 34	> 0.6
HSW	< 25	< 0.6
HSO	< 25	> 0.6

Given the nonparametric nature of the ETR model, deployment in refinery linear programming (LP) applications faces tractability challenges due to the uncertain number of operations needed to obtain the unique optimum solution of the problem. Consequently, a multivariate linear regression (MLR) model that furnishes a model for closed-form mathematical optimization use cases is also presented. The model is also implemented in the Python programming language using the statsmodels library. Using the same dataset and explanatory variables, the MLR model is estimated as follows:

$$\hat{y}_p(x) = \alpha_{0,p} + \sum_{j=1}^{n} \alpha_{j,p}\, x_j \tag{6}$$

For the categorical variables (x_j') representing refinery configurations,

$$x_{j' \subset j} \in \{0,1\} \tag{7}$$

In analyzing the model parameters for Equation 6, heteroscedasticity is observed to be present due to the nature of the dataset. Significant sparsity results from actual realizations of crude runs and product yields for the array of crude grades and product slates. For this reason, a Yeo-Johnson transformation (Yeo and Johnson 2000) is applied to the formulation to achieve heteroscedasticity robustness. This transformation necessitates a reformulation of the model as follows:

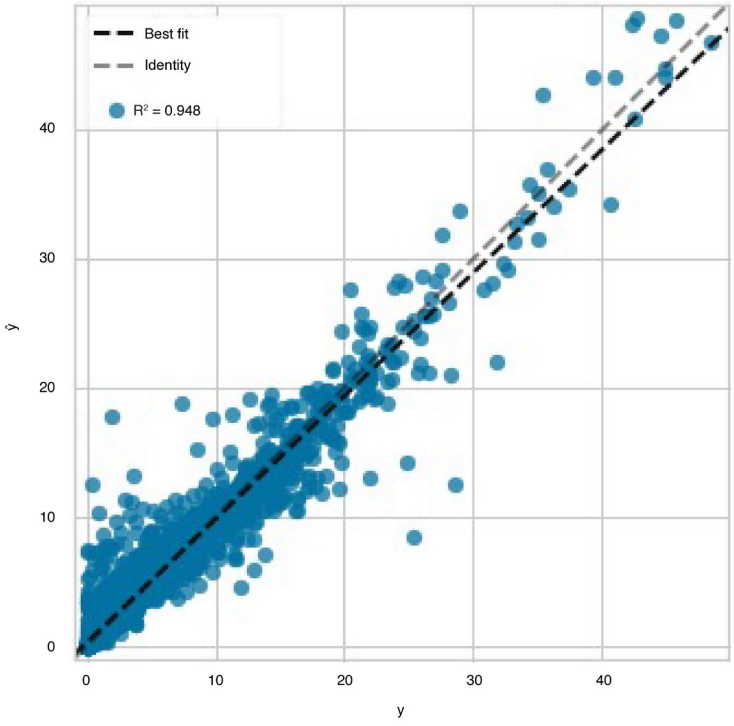
$$\log\left(\hat{y}_p(x) + \omega\right) = \alpha'_{0,p} + \sum_{j'} \alpha'_{j',p}\, x_{j'} + \sum_{j \neq j'} \alpha'_{j,p} \log\left(x_j + \omega\right) \tag{8}$$

For the transformation parameter value, $\omega = 1$, the model is heteroscedasticity robust. In addition to improving homoscedasticity, the transformation supports desirable attributes in the lower bound of the variable space required for mathematical optimization modeling applications. Furthermore, other statistical tests are performed to assess normality, autocorrelation, etc. Details on the statistical tests and results are provided in the supplementary material (Section SI.C).

4. Results and Discussion

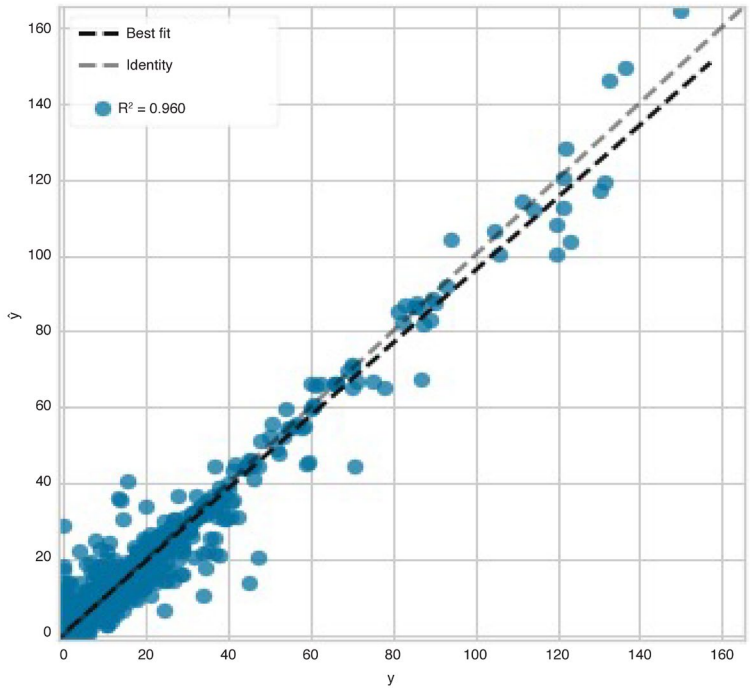
The models are trained with source data from the Platts World Refinery Database. The database has approximately 15,000 observations of crude runs and product slates covering every crude oil refinery plant in the world that operated between 2005 and 2020. Some of the observations had incomplete information about either crude runs or product yields, but all observations reported individual refinery capacities. The database was preprocessed to remove the incomplete entries, bringing the final number of observations for modeling to approximately 11,000. Model training and testing data sample splits of 70% and 30%, respectively, are applied. Figures 4 to 9 show parity plots for all the refined products predicted with the ETR model.

Figure 4: Parity plot for LPG product prediction with the ETR model.



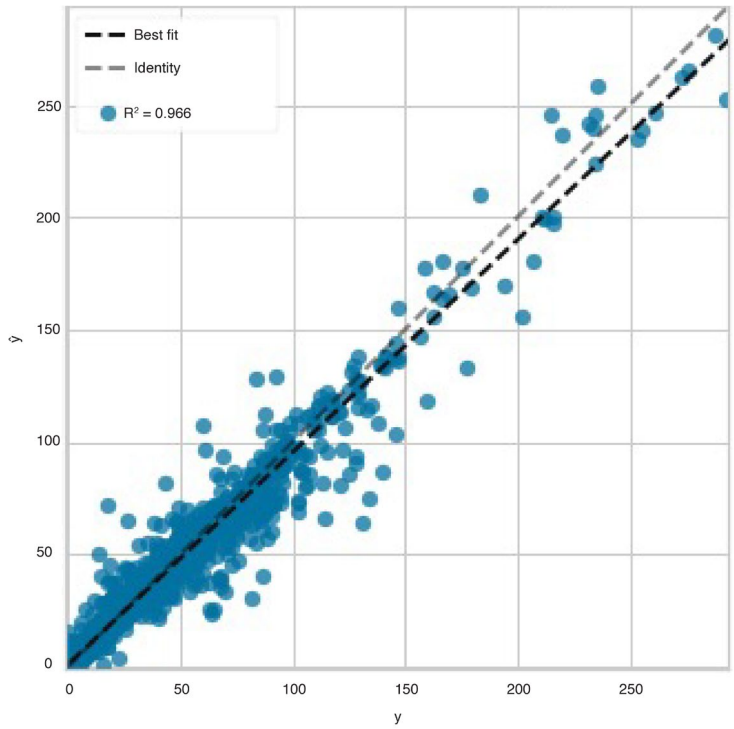
Source: KAPSARC Oil Value-Chain Analyzer

Figure 5: Parity plot for NAP product prediction with the ETR model.



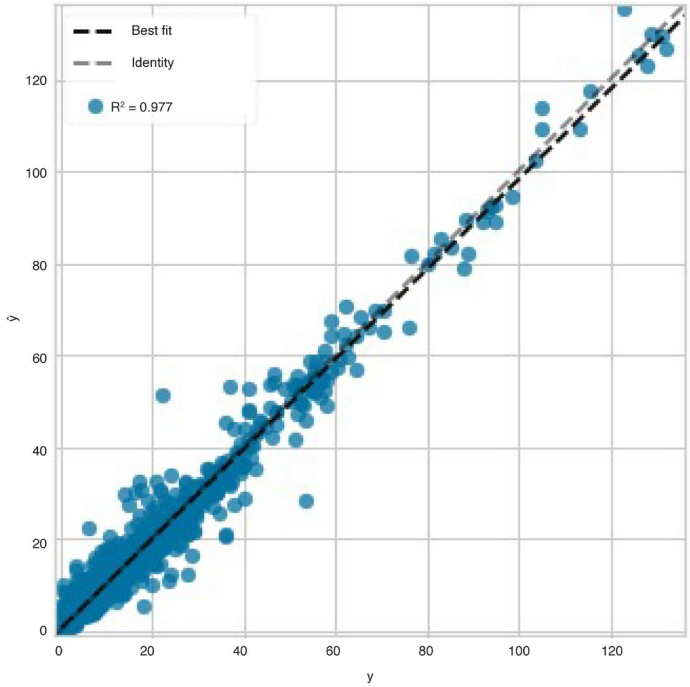
Source: KAPSARC Oil Value-Chain Analyzer

Figure 6: Parity plot for GAS product prediction with the ETR model.



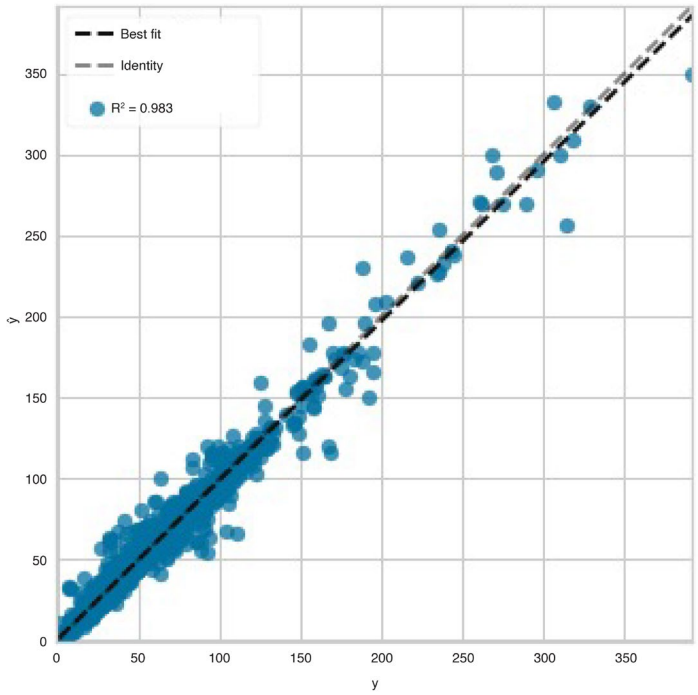
Source: KAPSARC Oil Value-Chain Analyzer

Figure 7: Parity plot for KJF product prediction with the ETR model.



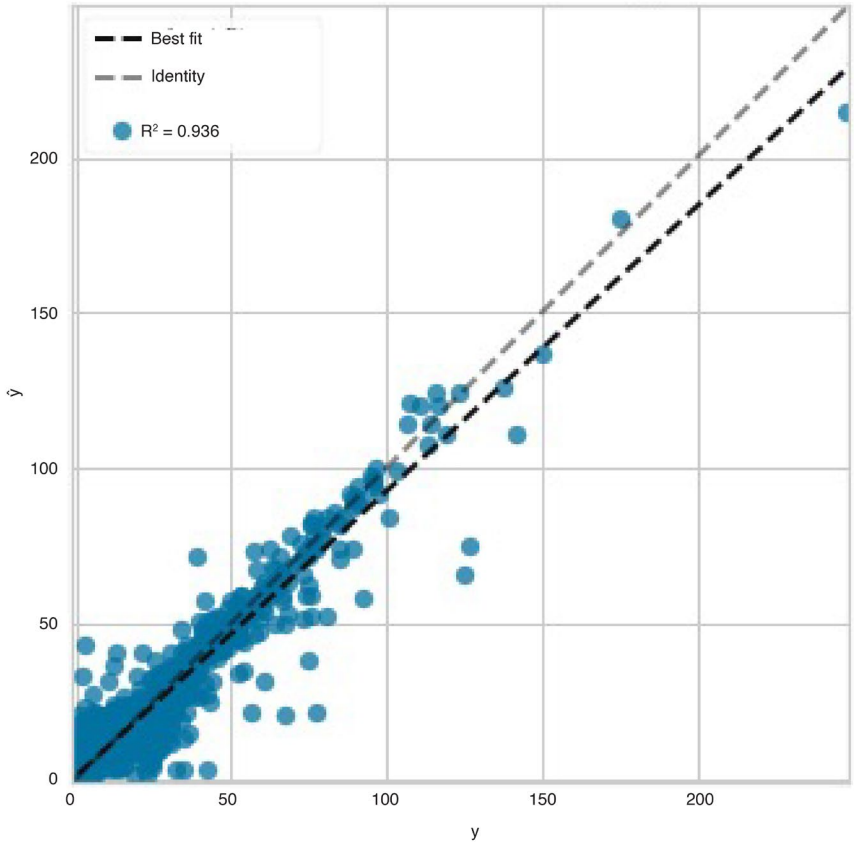
Source: KAPSARC Oil Value-Chain Analyzer

Figure 8: Parity plot for DGO product prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

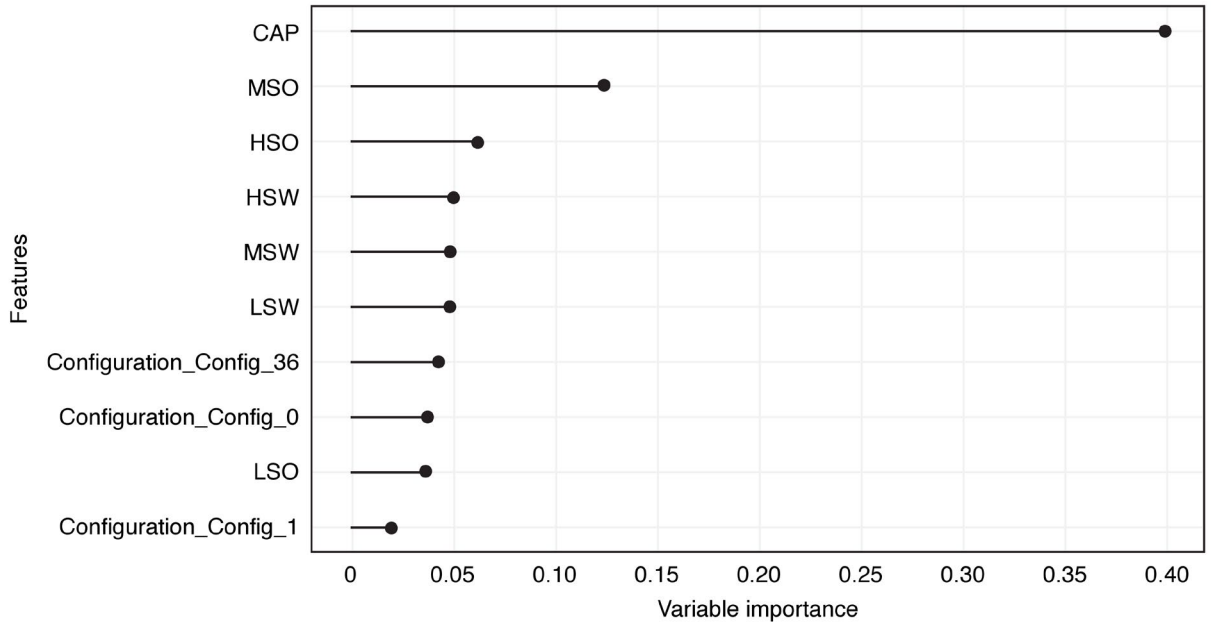
Figure 9: Parity plot for HFO product prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

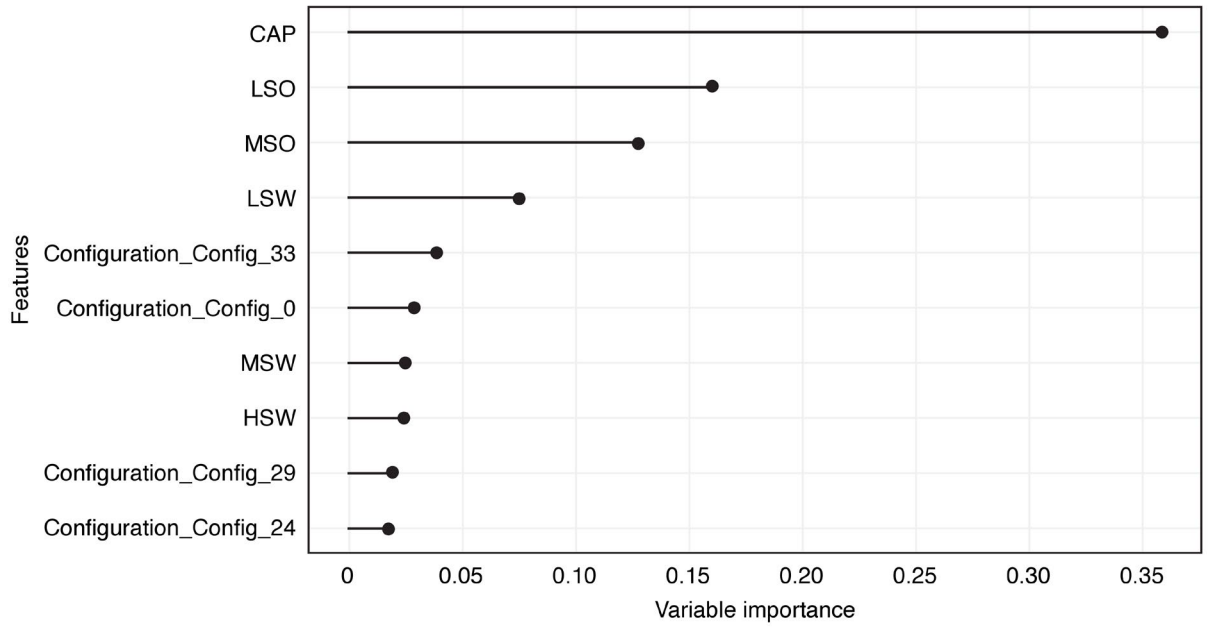
The coefficient of determination (R^2) is above 90% for all refinery products predicted with the ETR model, meaning that over 90% of the variation in the target variables is accounted for by the model. Using the information from the coefficient of determination for each target variable and the values of the coefficients of partial determination, which measures the proportion of total variation in the target variable that is explained by each individual feature (explanatory variable) in the model, it is possible to evaluate and rank all features based on their importance in predicting the target. Figures 10 to 15 present the top ten features accounting for the total variation in the predicted refined products. The figures confirm that the refinery capacity, crude oil blend runs, and refinery configurations are essential variables for quantifying variations in the amounts of the refined products but with different levels of importance for each estimation of the volumes of refined products.

Figure 10: Feature importance plot showing the top 10 features accounting for the total variation in LPG product prediction.



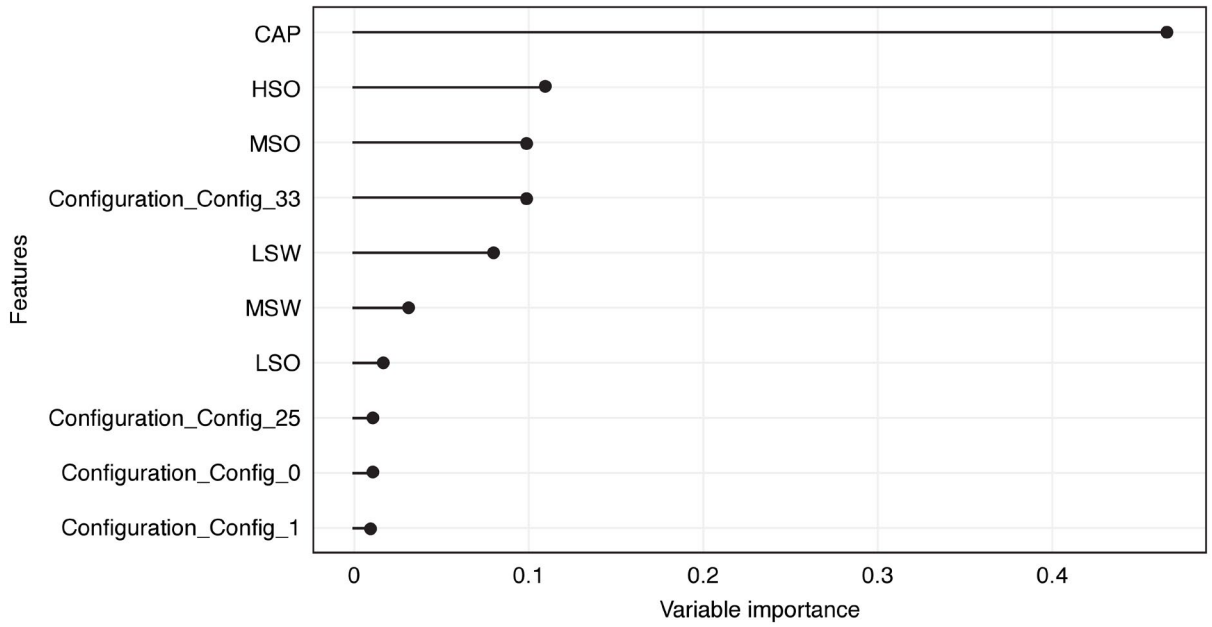
Source: KAPSARC Oil Value-Chain Analyzer

Figure 11: Feature importance plot showing the top 10 variables accounting for the total variation in NAP product prediction.



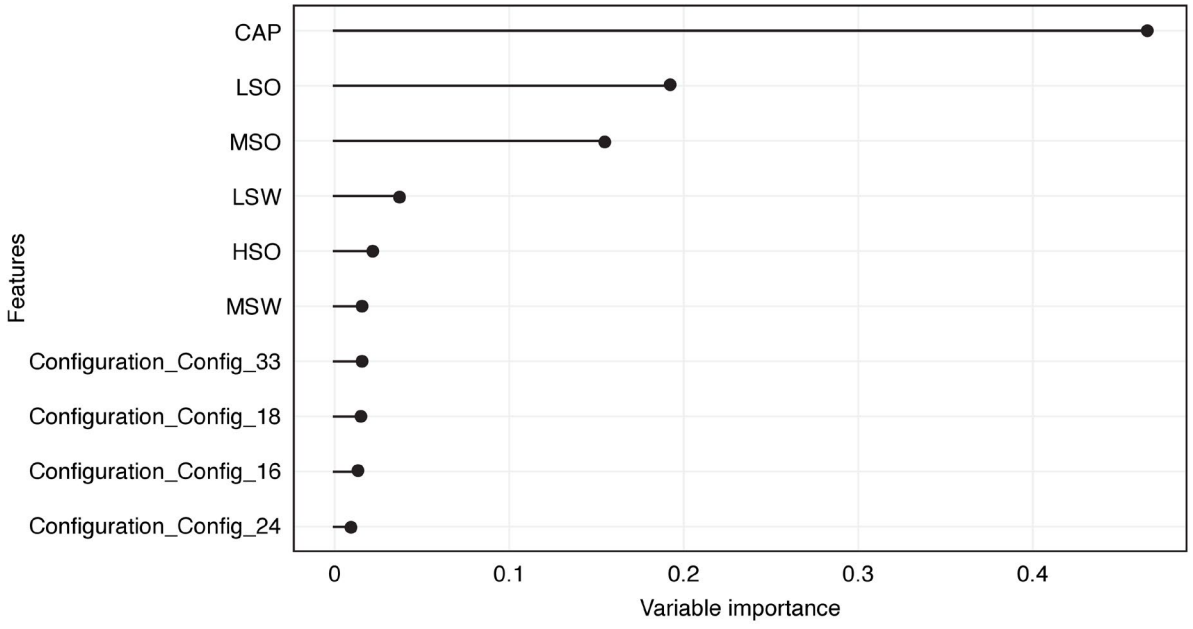
Source: KAPSARC Oil Value-Chain Analyzer

Figure 12: Feature importance plot showing the top 10 variables accounting for the total variation in GAS product prediction.



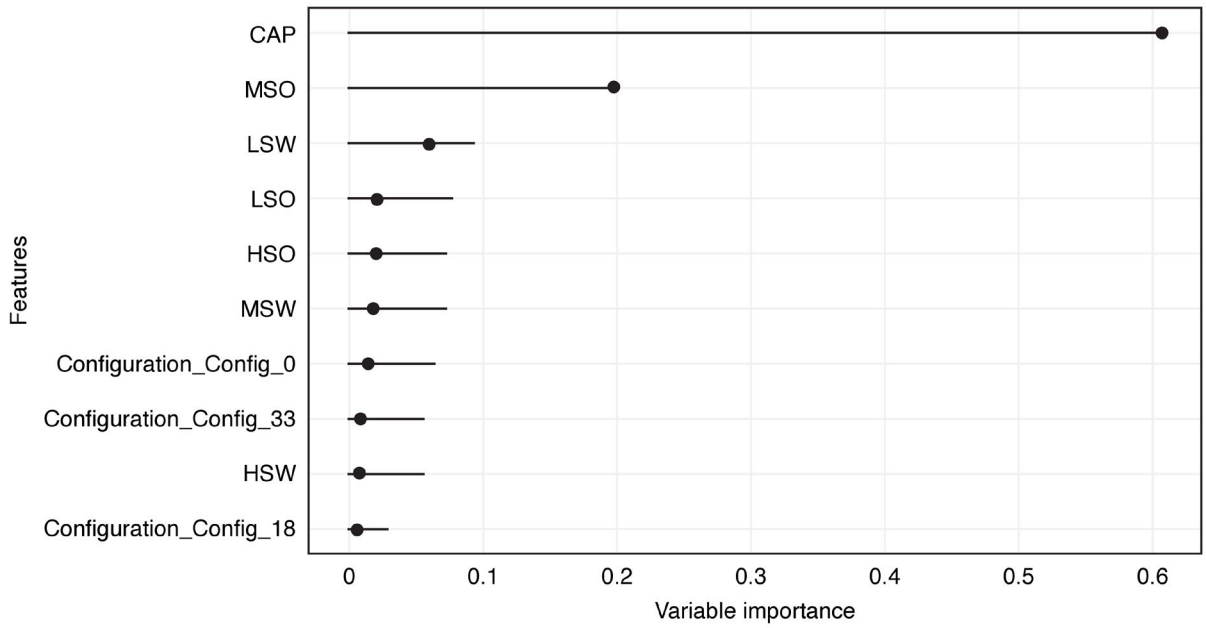
Source: KAPSARC Oil Value-Chain Analyzer

Figure 13: Feature importance plot showing the top 10 variables accounting for the total variation in KJF product prediction.



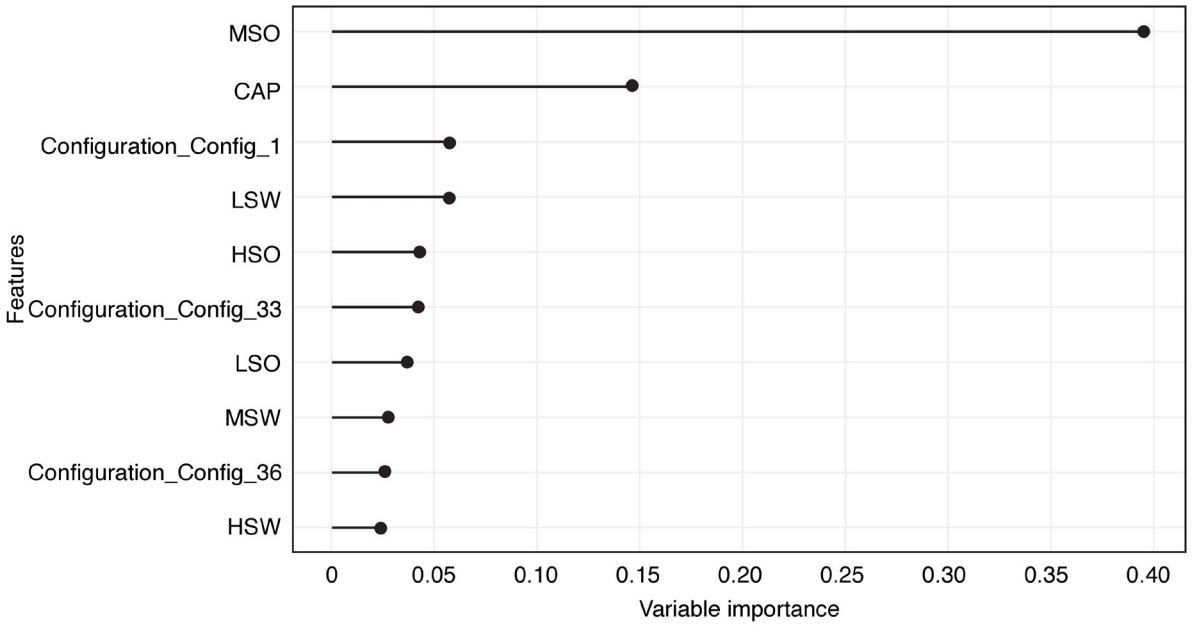
Source: KAPSARC Oil Value-Chain Analyzer

Figure 14: Feature importance plot showing the top 10 variables accounting for the total variation in DGO product prediction.



Source: KAPSARC Oil Value-Chain Analyzer

Figure 15: Feature importance plot showing the top 10 variables accounting for the total variation in HFO product prediction.



Source: KAPSARC Oil Value-Chain Analyzer

In addition to the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE) are two other performance metrics that are evaluated for both the parametric and nonparametric models of refined products. Table 2 presents all performance metrics for the ETR and MLR models. Evidently, the machine learning-based, nonparametric ETR model demonstrates higher predictive performance for the target refined products. For a given refinery configuration, the potential sources of deviations between the predicted and observed values of refined product flows could be from the differences in the process subunit-level designs, process technology, the operational expertise of plant operators (human factors), and model suitability. Consequently, 15 other training algorithms were applied to the datasets, and none outperformed the ETR model on all target refined product predictions. On the R^2 basis, the MLR model shows the lowest performance on NAP prediction, whereas the ETR model has the lowest performance on HFO prediction. Additional statistical analysis results for both models are available in supplementary information sections SI.B and SI.C.

Refinery configuration-based modeling not only is useful for estimating the production of refined products but also can facilitate the quantification of operational energy consumption, GHG emissions and economic performance for each type of refinery plant design or configuration. For a given plant capacity and configuration, variabilities in the properties and economics of crude oil blend feedstocks determine the product yield, product types and the overall refinery economic margin performance. However, the subjects of refinery energy requirements and economics are not the focus of this paper; rather, they constitute the objectives of future research leveraging the refinery configuration framework and models presented in this work. By identifying global refinery configurations, which encapsulates the attributes of each refinery plant design in the form of the configuration variable in the production model, the approach presented in this paper avoids the need to explicitly represent and specify design and operating conditions – such as temperature and pressure – in the presented production model. The use of actual field data on global refineries in the model training ensures that the effects of design and operational differences across the refinery configuration classes are captured.

Table 2. Prediction performance metrics for the test of the ETR and MLR models.

Target	Test Performance – ETR Model			Test Performance – MLR Model		
	MAE	RMSE	R^2	MAE	RMSE	R^2
LPG	0.62	1.43	0.95	2.07	3.43	0.77
NAP	1.12	2.72	0.96	4.64	8.53	0.57
GAS	2.53	6.29	0.96	8.32	14.71	0.84
KJF	0.94	2.11	0.97	3.21	5.82	0.83
DGO	2.32	5.34	0.98	6.41	11.43	0.88
HFO	1.93	4.54	0.93	7.46	12.30	0.65

5. Conclusions

Until now, no comprehensive representation of the global crude oil refining landscape has been modeled or reported in the publicly available literature. In this paper, real refinery design and operating information from over 800 existing crude oil refineries globally is used to establish that there are essentially 40 unique configurations. On this basis, refinery plant data are used to train a machine learning model to predict refined product yields given plant capacities, configurations, and crude blend feedstocks. Due to the nonparametric form of the machine learning algorithm and the computational challenges for refinery LP applications, a parametric model based on MLR is also proposed. Although the machine learning model demonstrates superior predictive performance, the parametric model is convenient for the closed-form representation of refinery models as part of a wider energy system for various purposes or simply for optimal decision-making.

In addition to operational scheduling and crack spread optimization use cases, refinery models are needed in energy outlook and energy transition scenario analysis. They can provide business and government decision makers with information on market opportunities, environmental performance, and regulatory compliance under specified operating environments and conditions. This work addresses the production modeling aspect of refinery models. Extensions of the modeling approach to energy requirements and emissions constitute further research opportunities.

References

Abella, Jessica, Motazedi Kavan, John Guo, and Joule Bergerson. 2015. "Petroleum Refinery Life Cycle Inventory Model (PRELIM) PRELIM v1.5."

Ali, Moez. 2020. "PyCaret: An open source, low-code machine learning library in Python."

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. "Extremely randomized trees." *Machine Learning* 63, no. 1: 3-42. DOI: https://doi.org/10.1007/s10994-006-6226-1

Herce, Carlos, Chiara Martini, Marcello Salvio, and Carlo Toro. 2022. "Energy Performance of Italian Oil Refineries Based on Mandatory Energy Audits." *Energies* 15, no. 2: 532. DOI: https://doi.org/10.3390/en15020532

Kaiser, Mark. 2017. "A review of refinery complexity applications." *Petroleum Science* 14. DOI: https://doi.org/10.1007/s12182-016-0137-y

Motazedi, Kavan, Jessica Abella, and Joule Bergerson. 2017. "Techno–Economic Evaluation of Technologies to Mitigate Greenhouse Gas Emissions at North American Refineries." *Environmental Science & Technology* 51, no. 3: 1918-1928. DOI: https://doi.org/10.1021/acs.est.6b04606

Nduagu, Experience, Evar Umeozor, Alpha Sow, and Dinara Millington. 2018. "An Economic Assessment of the International Maritime Organization Sulphur Regulations on Markets for Canadian Crude Oil." Canadian Energy Research Institute.

Ogjresearch. 2022. "Worldwide Refinery Survey and Complexity Analysis." *Oil and Gas Journal*. Online Research Center.

S&P Global Platts. 2022. "Platts World Refinery Database." Accessed August 15, 2022. https://www.spglobal.com/commodityinsights/en/products-services/oil/platts-world-refinery-database.

Yeo, In-Kwon, and Richard Johnson. 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87, no. 4: 954-959. DOI: https://doi.org/10.1093/biomet/87.4.954

Young, Brett, Troy Hottle, Troy Hawkins, Matt Jamieson, Greg Cooney, Kavan Motazedi and Joule Bergerson. 2019. "Expansion of the Petroleum Refinery Life Cycle Inventory Model to Support Characterization of a Full Suite of Commonly Tracked Impact Potentials." *Environmental Science & Technology* 53(4): 2238-2248. DOI: https://doi.org/10.1021/acs.est.8b05572

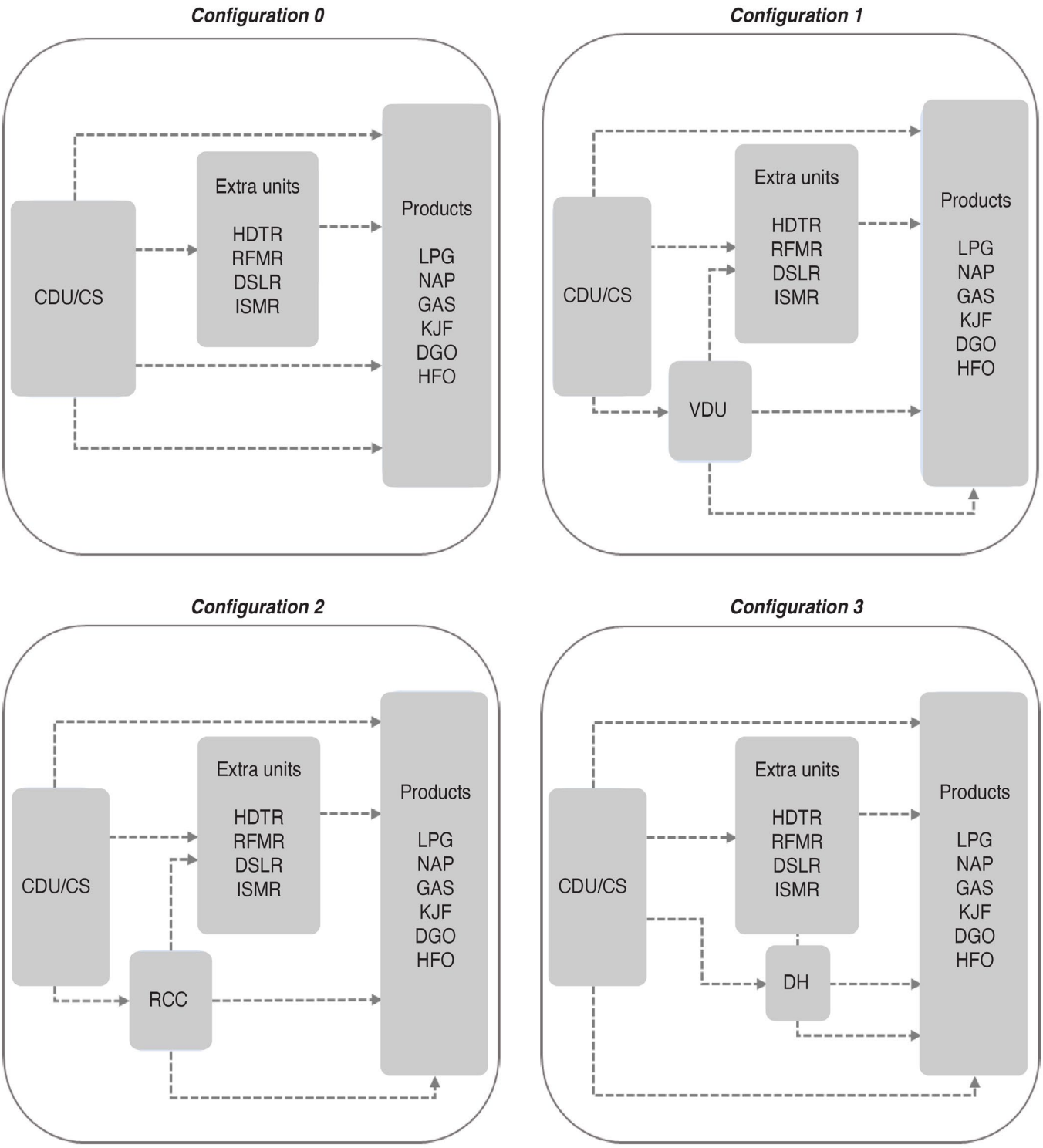
Supplementary Information

Global Crude Oil Refinery Models: Parametric and Nonparametric Methods

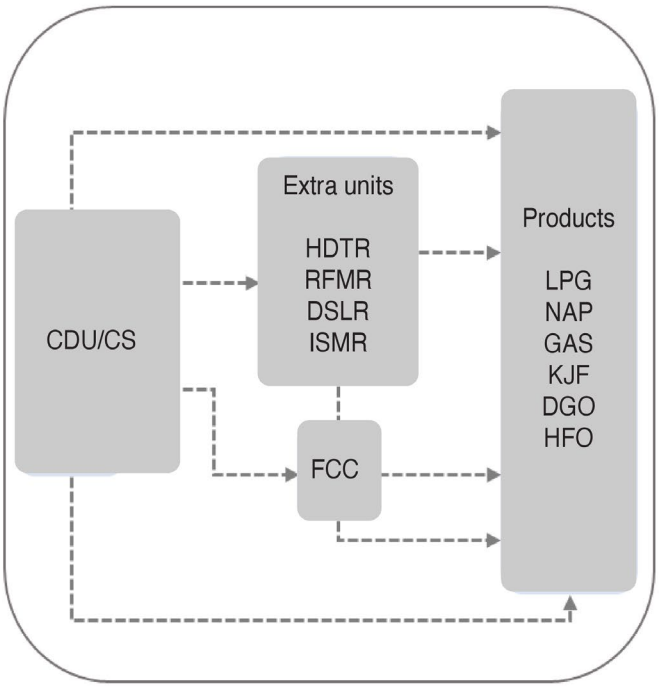
Evar Umeozor[a, †]

[a] King Abdullah Petroleum Studies and Research Center, Riyadh, Saudi Arabia
Email: evar.umeozor@kapsarc.org
Telephone: +966 507 029 871

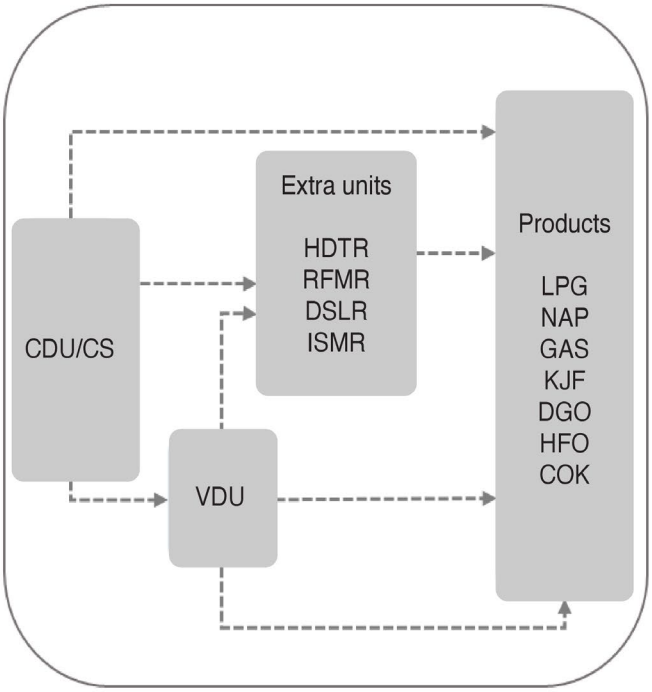
Section SI.A: Summary process diagrams of the world's crude oil refineries



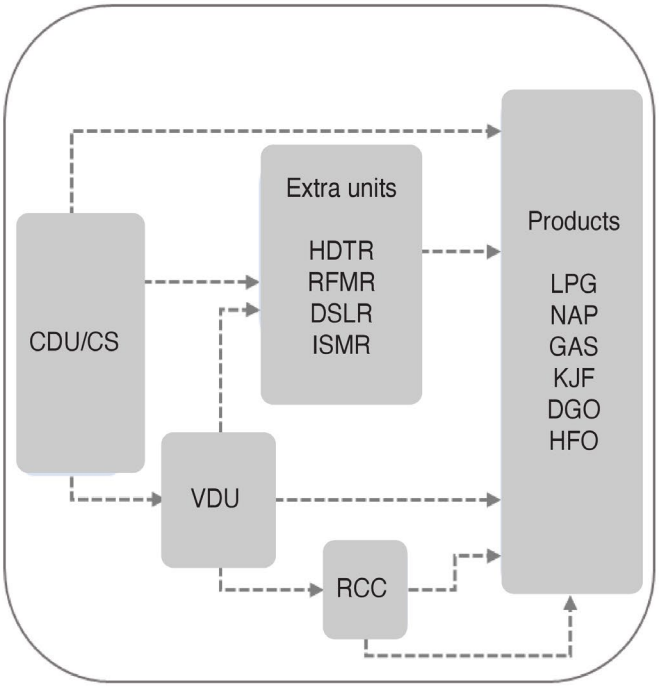
Configuration 4



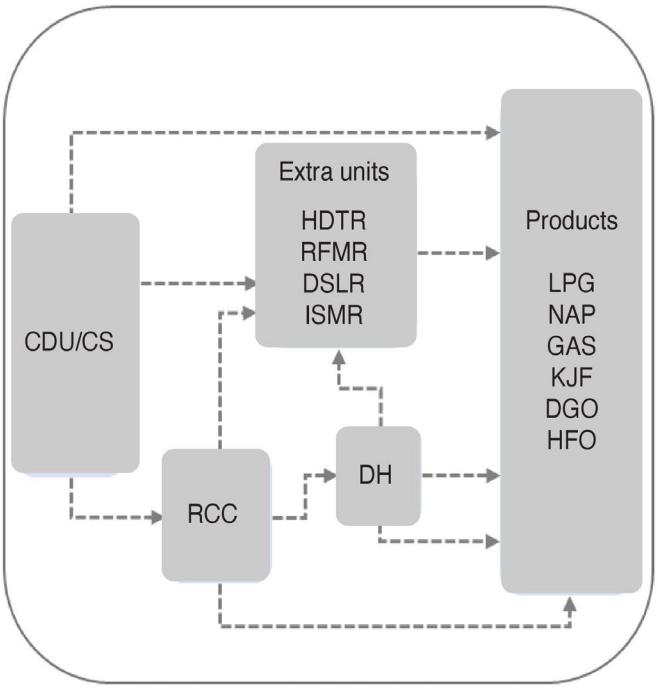
Configuration 5



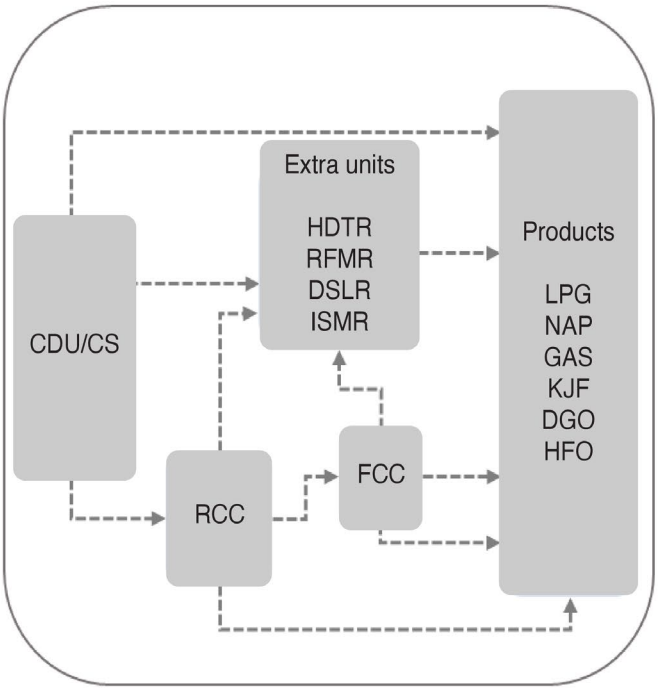
Configuration 6



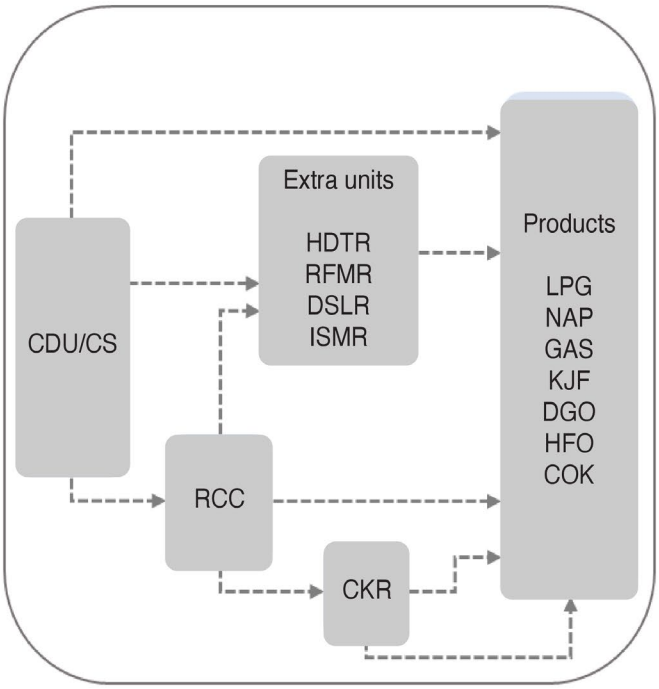
Configuration 7



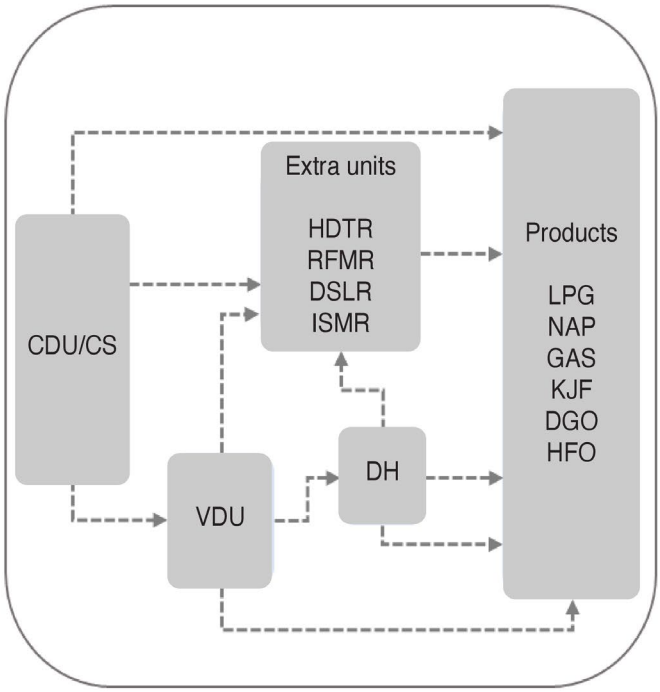
Configuration 8



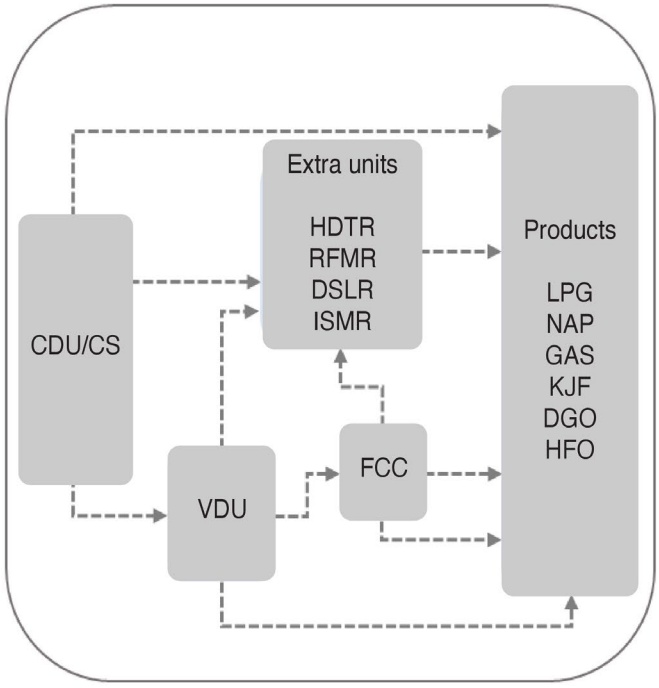
Configuration 9



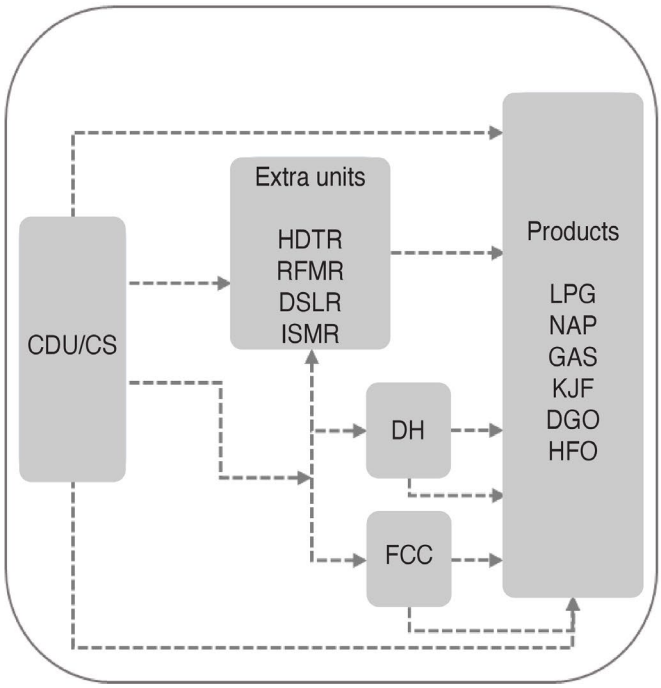
Configuration 10



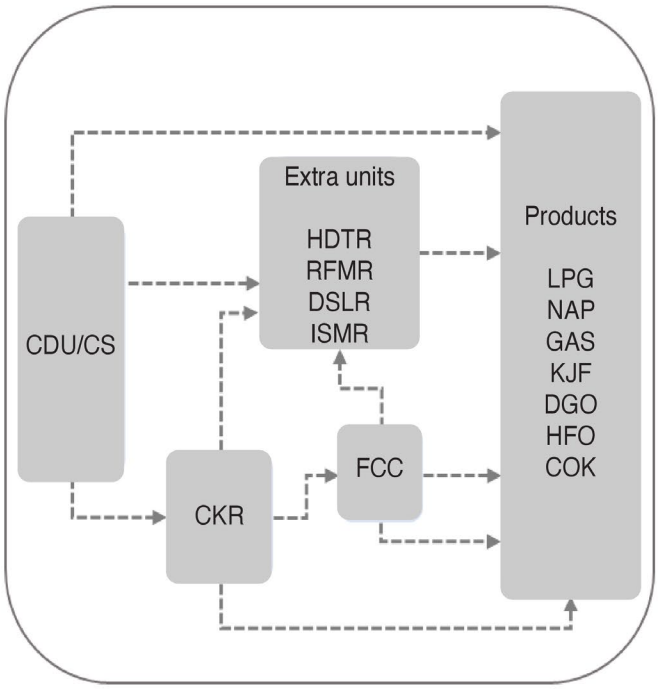
Configuration 11



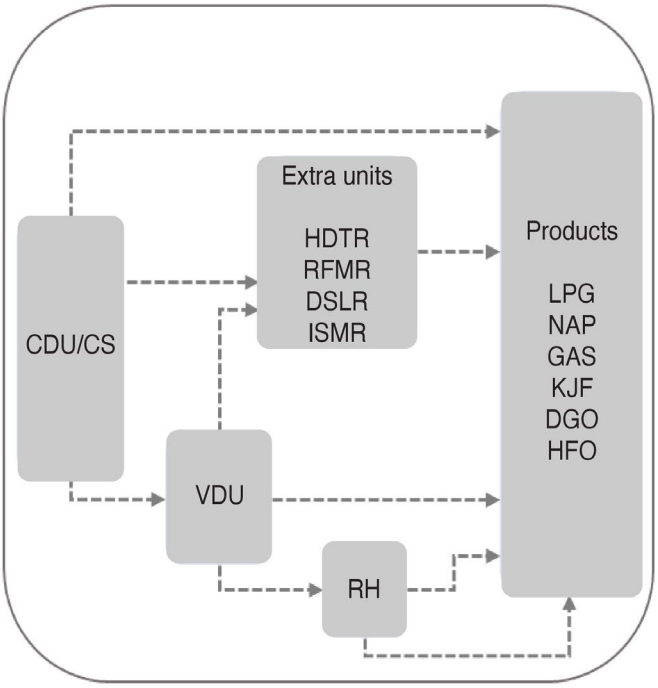
Configuration 12



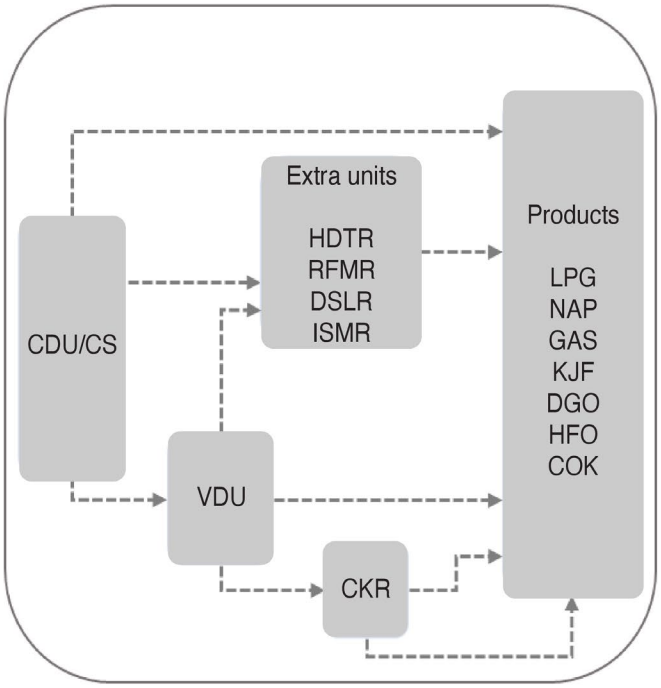
Configuration 13



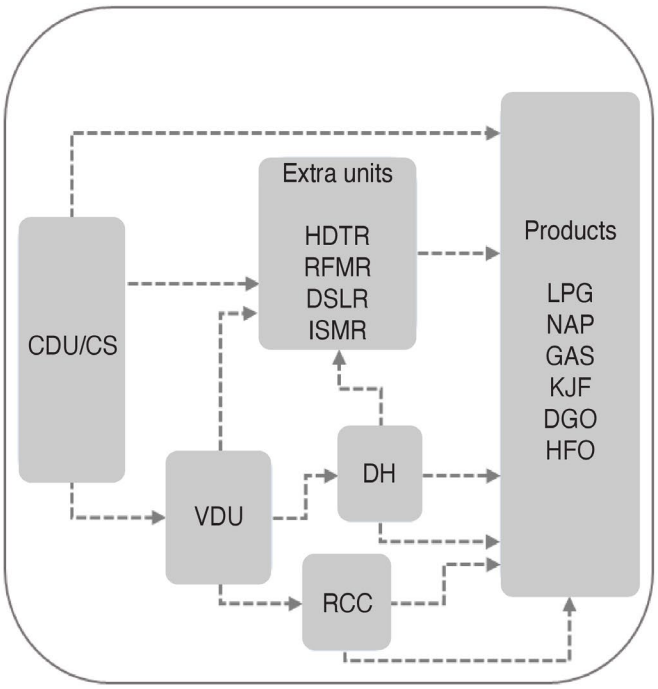
Configuration 14



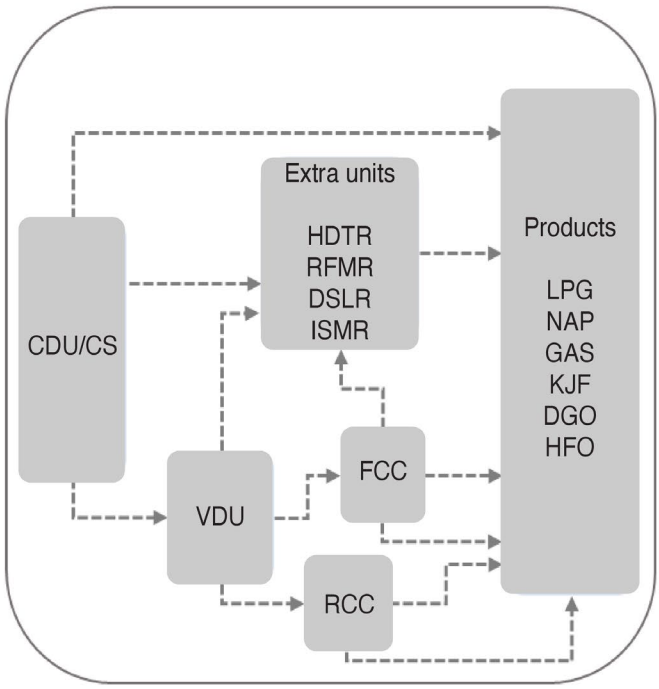
Configuration 15



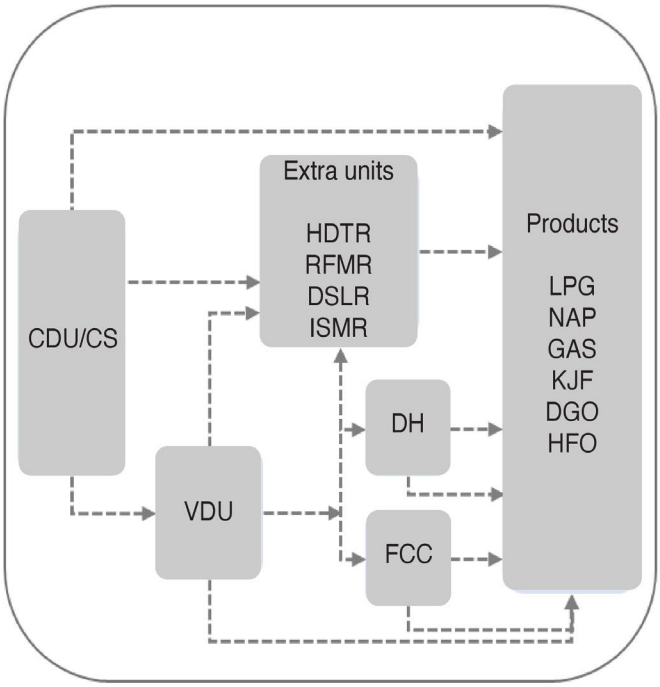
Configuration 16



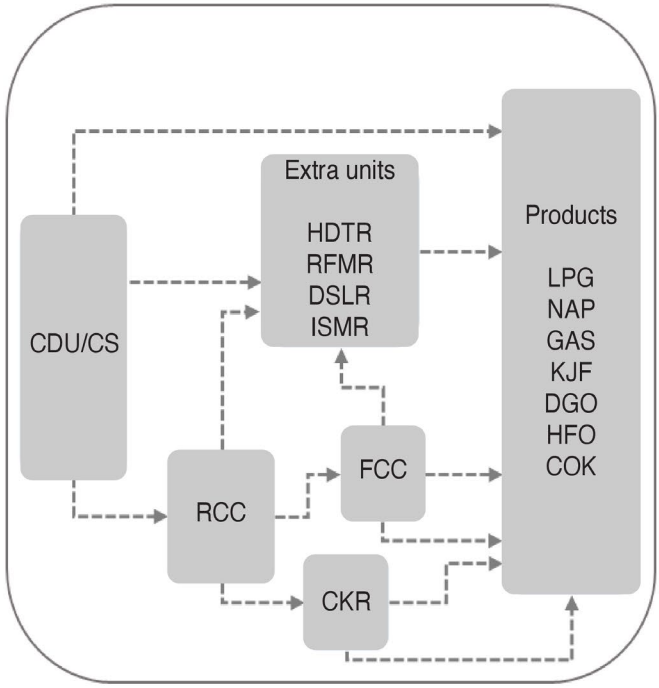
Configuration 17



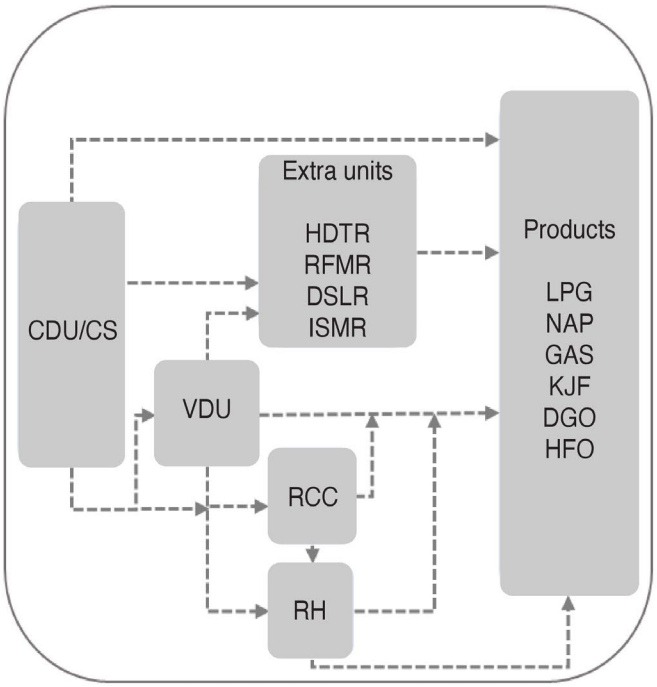
Configuration 18



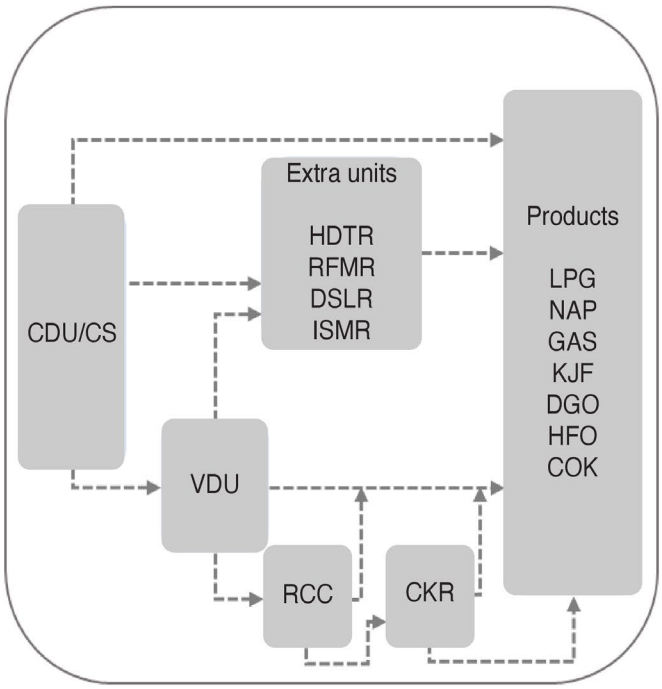
Configuration 19



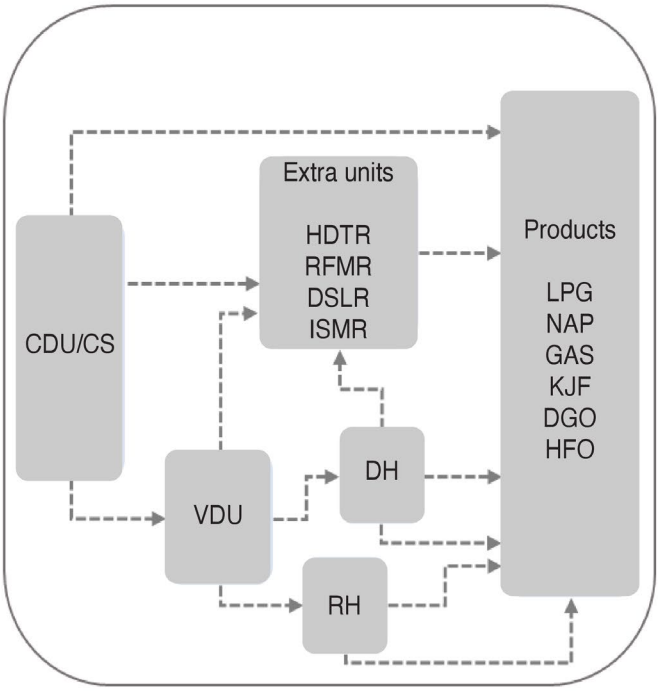
Configuration 20



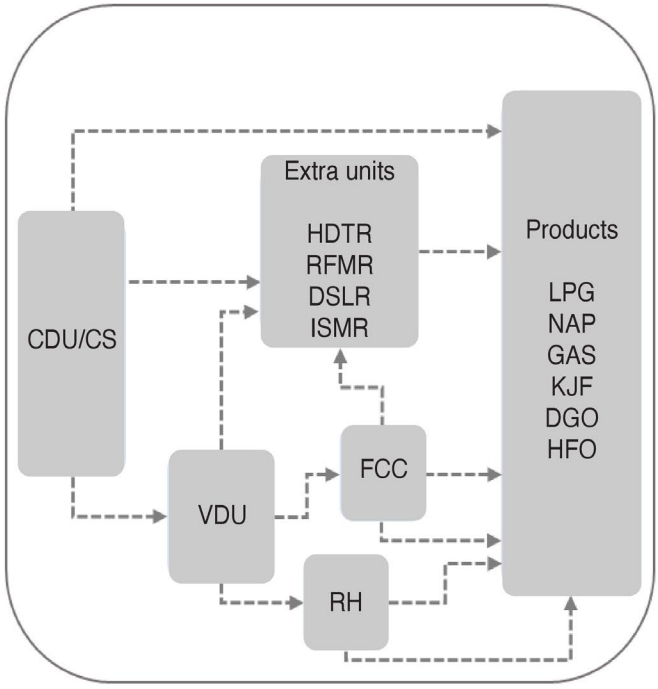
Configuration 21



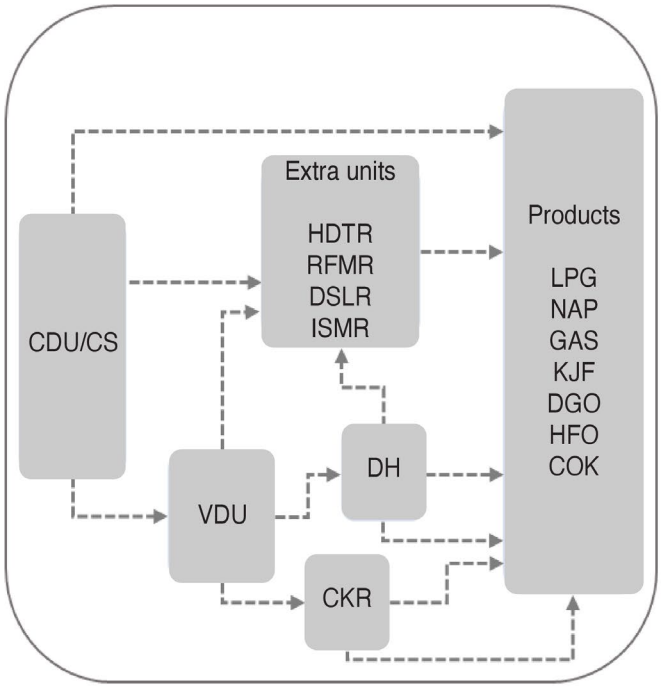
Configuration 22



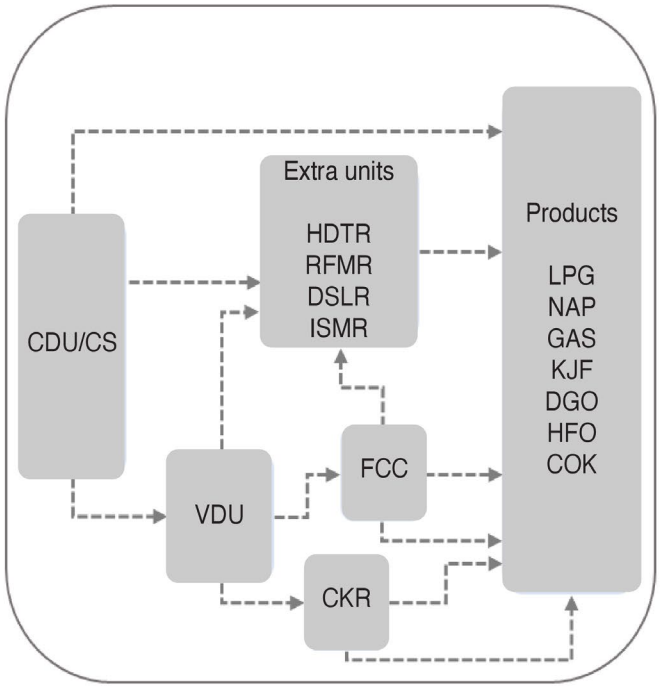
Configuration 23



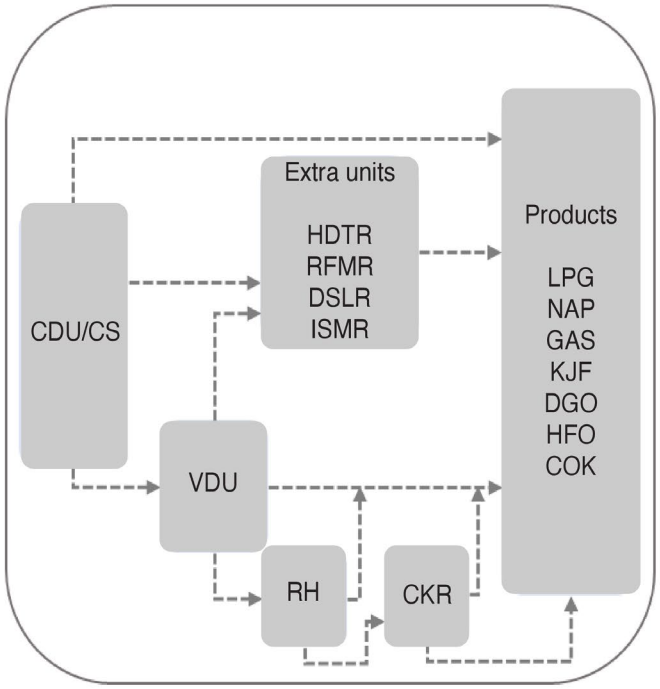
Configuration 24



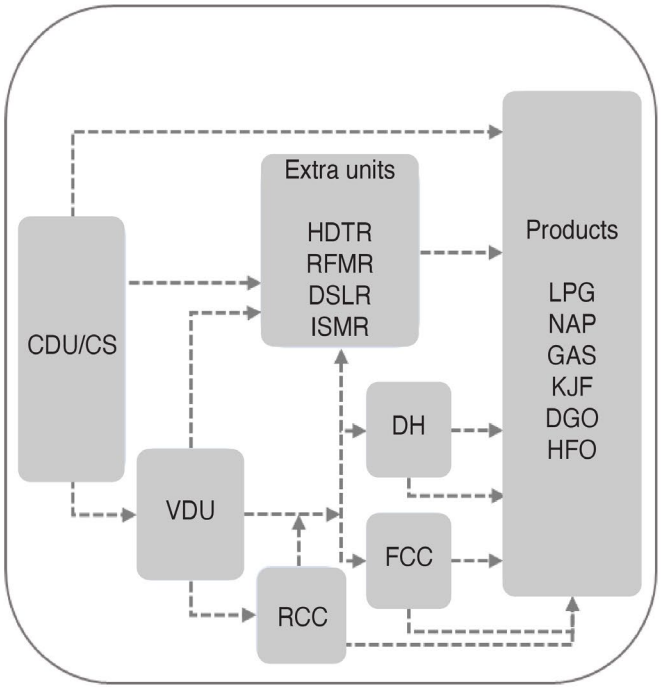
Configuration 25



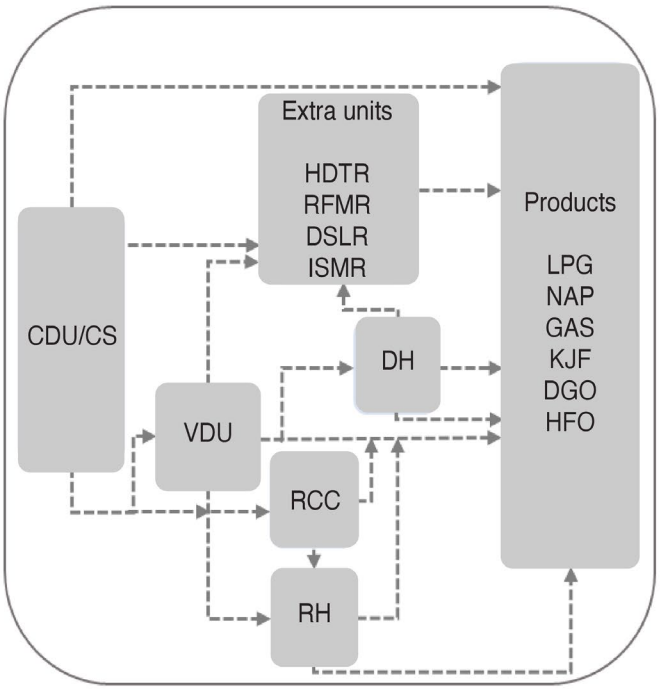
Configuration 26



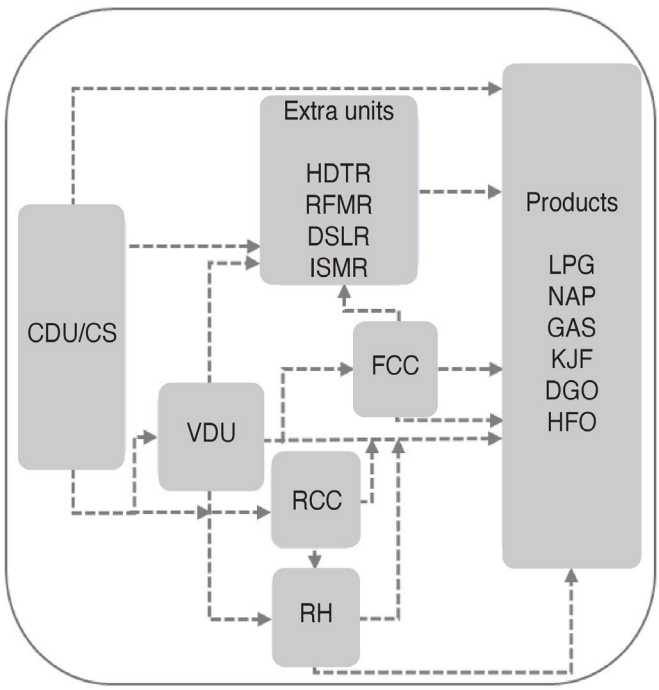
Configuration 27



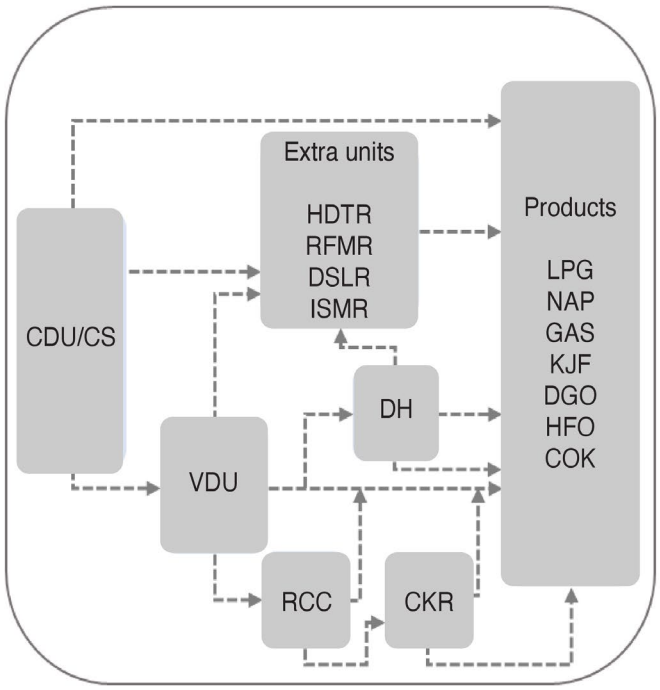
Configuration 28



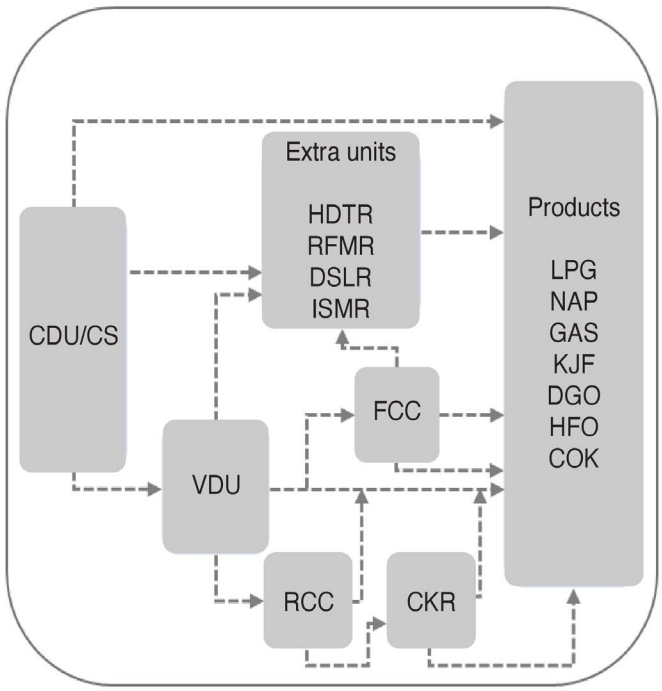
Configuration 29



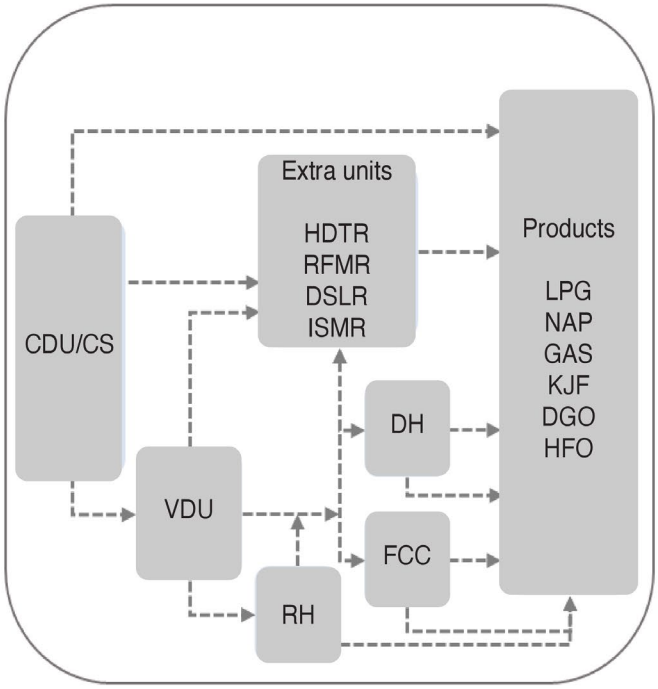
Configuration 30



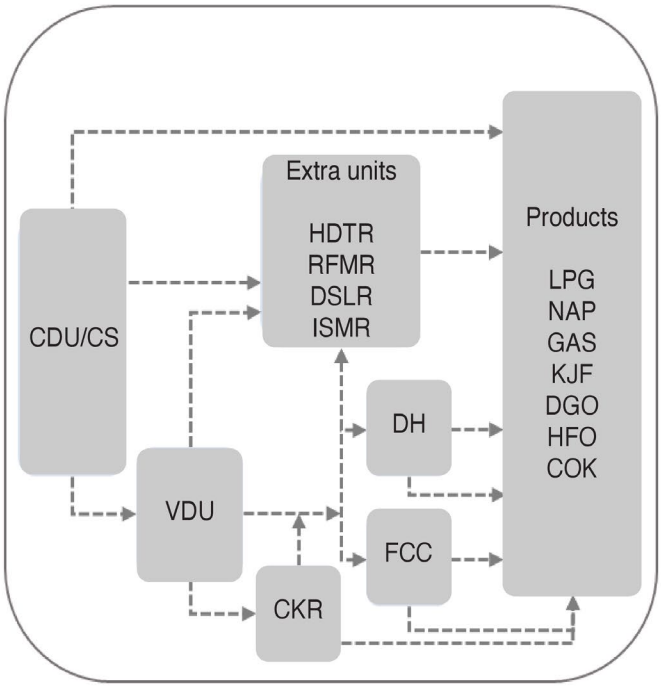
Configuration 31



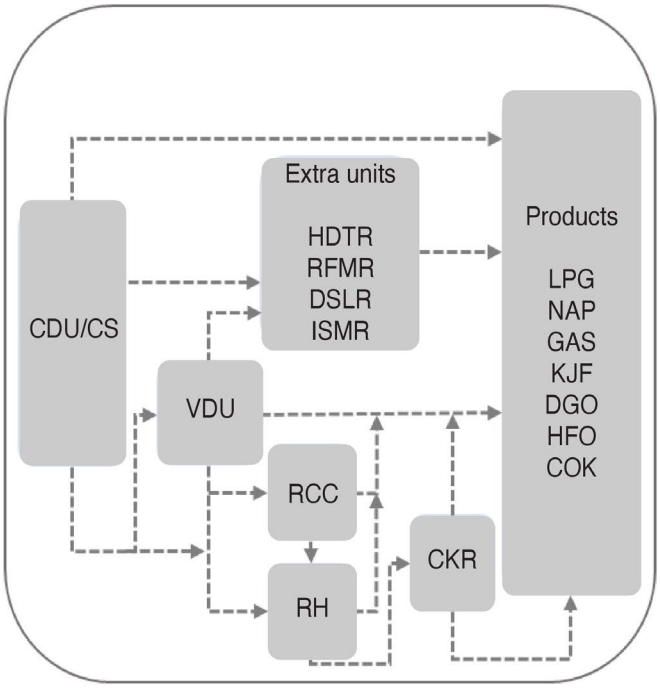
Configuration 32



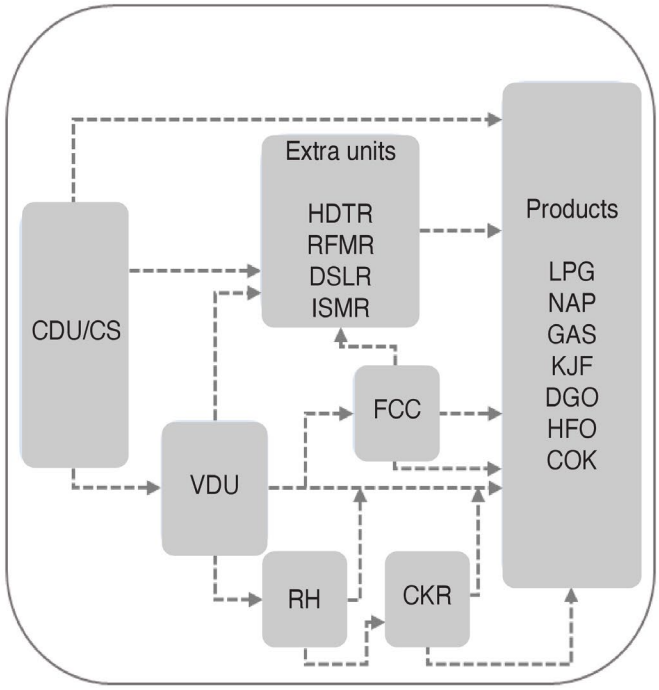
Configuration 33



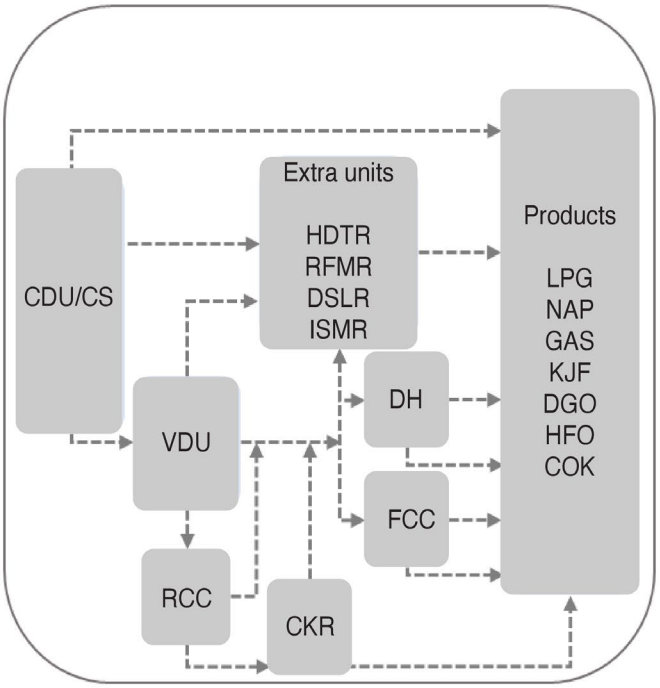
Configuration 34



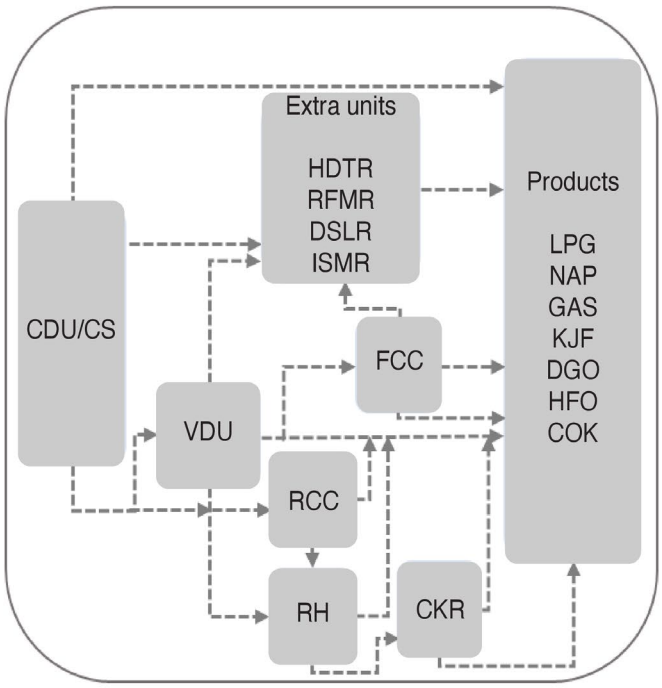
Configuration 35



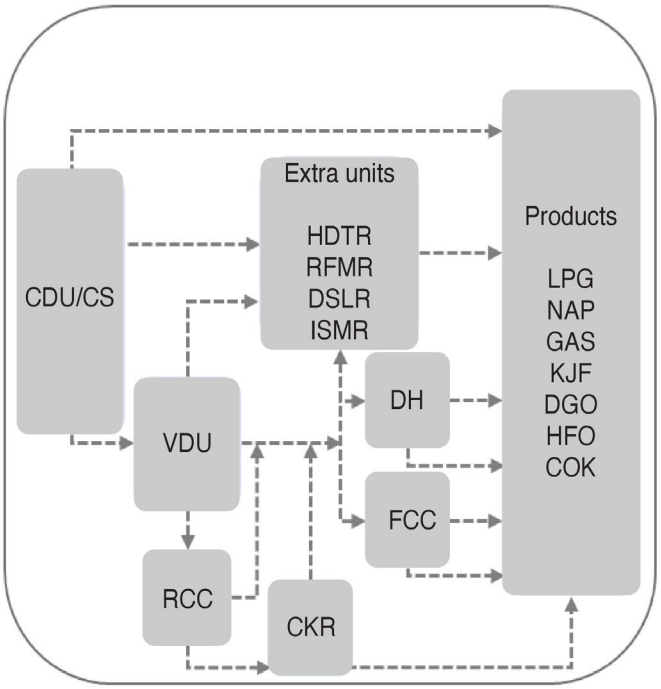
Configuration 36



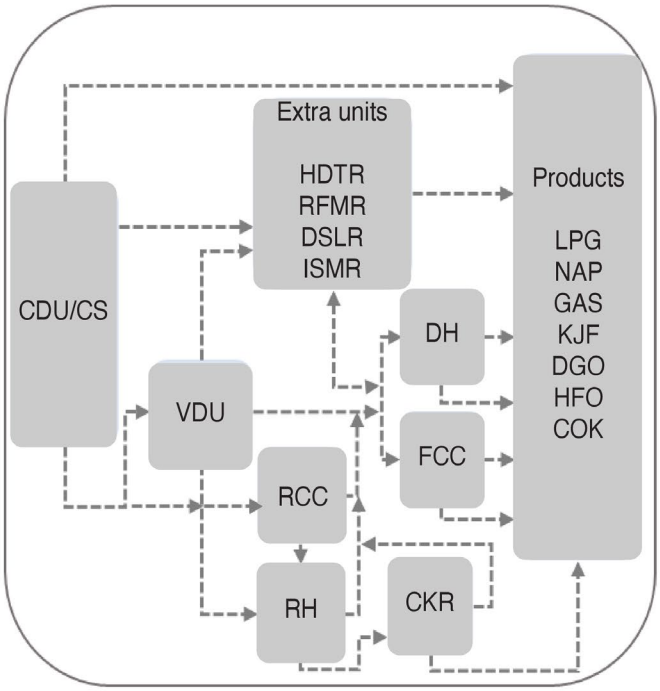
Configuration 37



Configuration 38



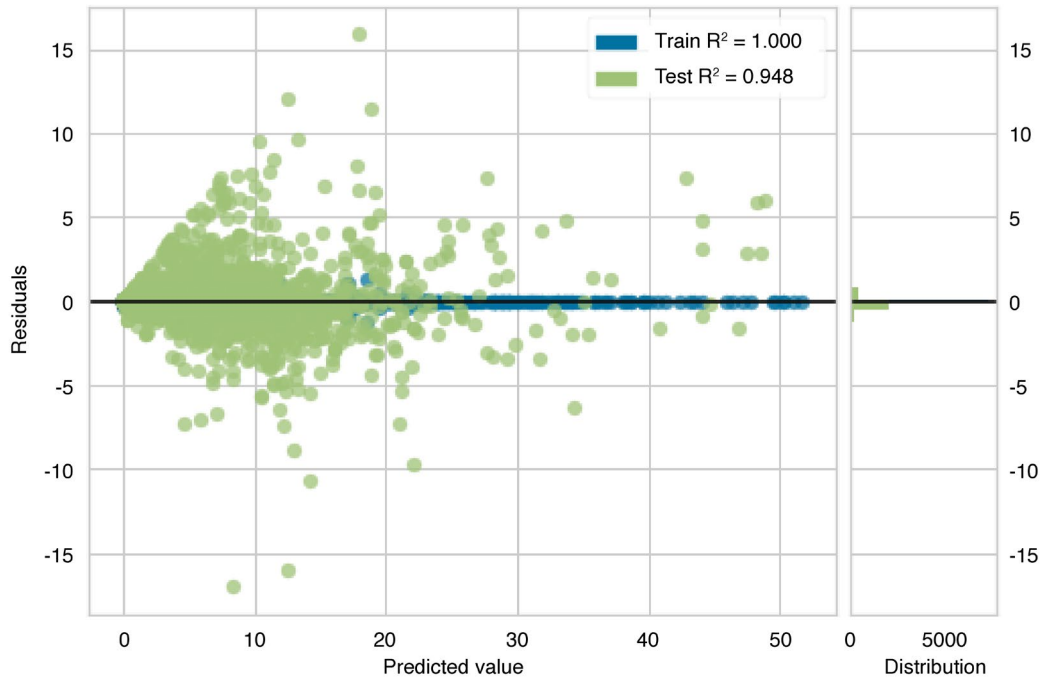
Configuration 39



Section SI. B: Error plots for the extremely randomized tree regressor (ETR) model

In this section, residual analysis is presented in the following figures for the target refinery products, including liquefied petroleum gas (LPG), naphtha (NAP), gasoline (GAS), kerosene/jet fuel (KJF), gas oil diesel (DGO), and heavy fuel oil (HFO).

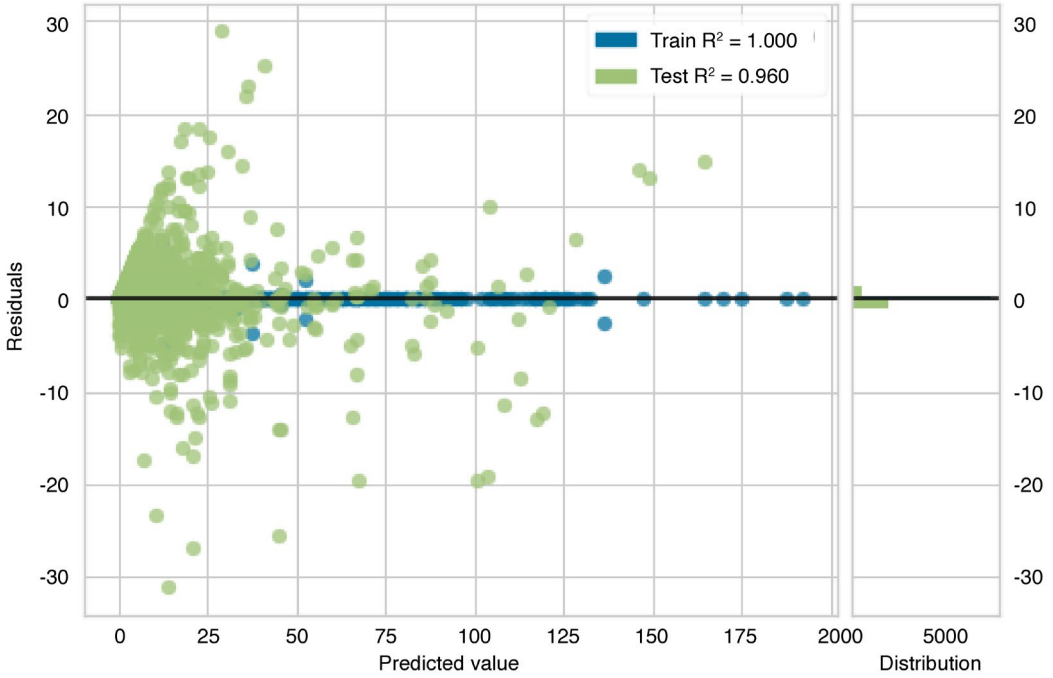
Figure SI.B1: Error analysis for LPG prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

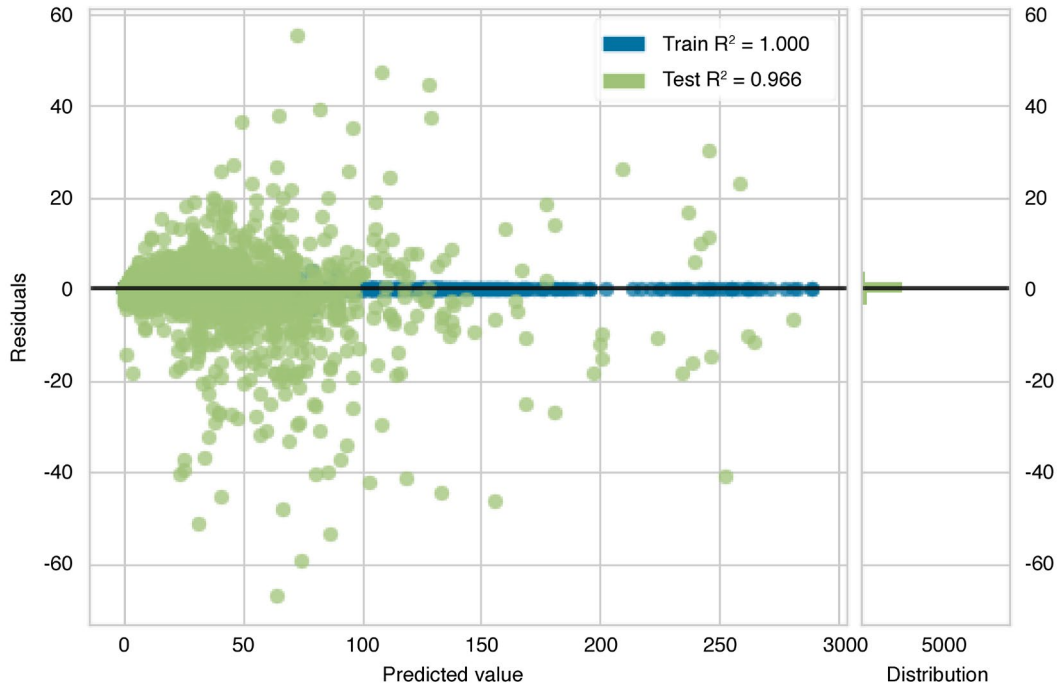
Section SI. B: Error plots for the extremely randomized tree regressor (ETR) model

Figure SI.B2: Error analysis for NAP prediction with the ETR model.



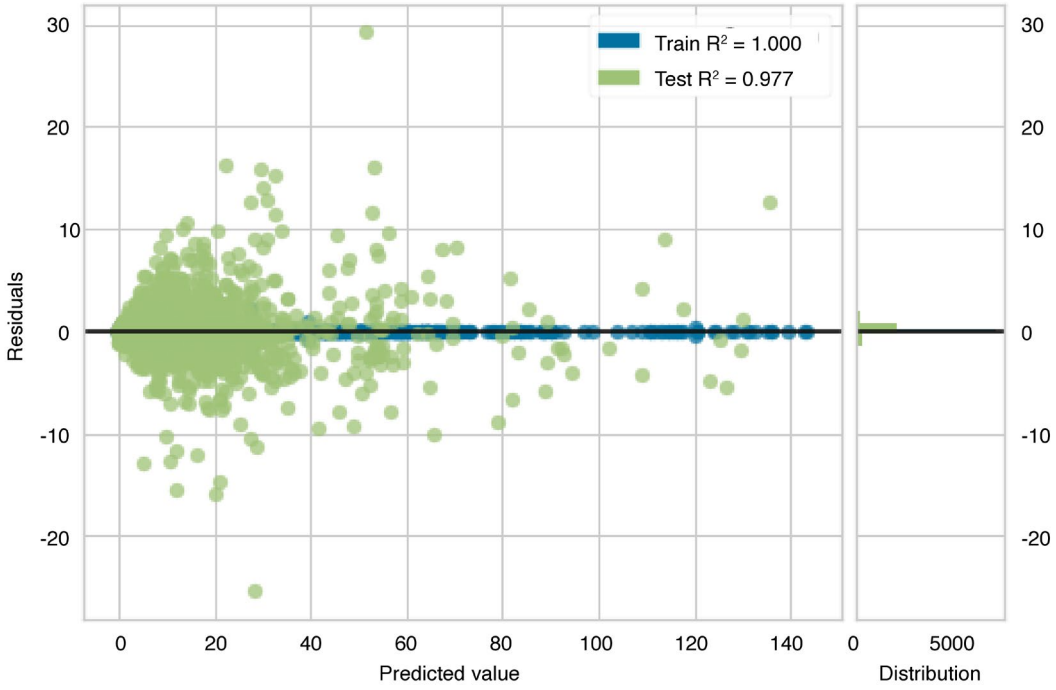
Source: KAPSARC Oil Value-Chain Analyzer

Figure SI.B3: Error analysis for GAS prediction with the ETR model.



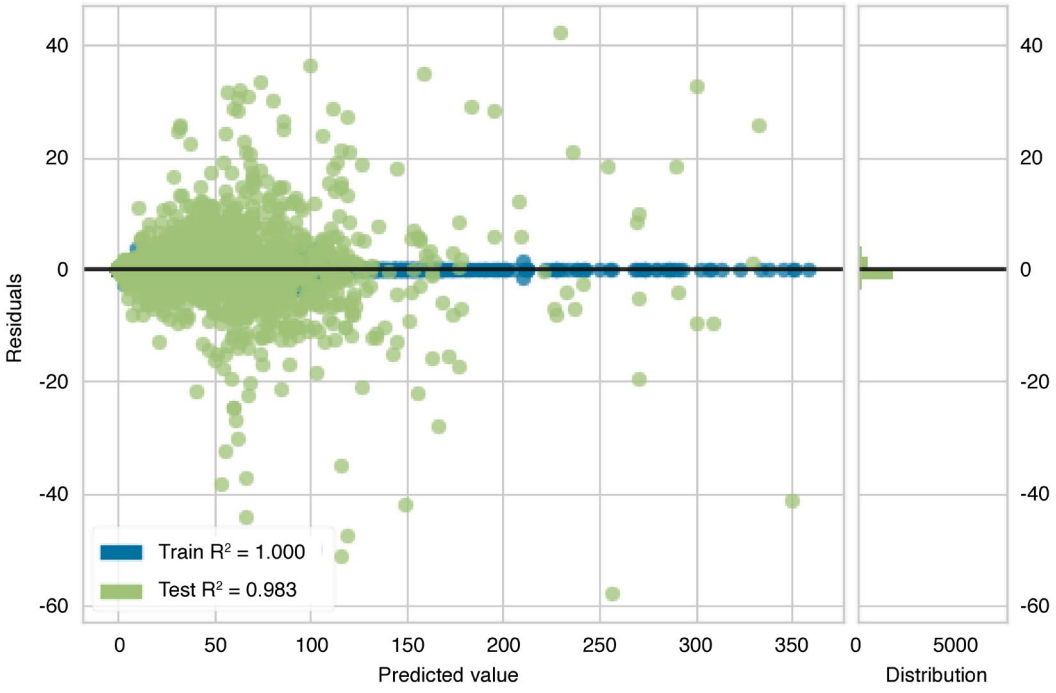
Source: KAPSARC Oil Value-Chain Analyzer

Figure SI.B4: Error analysis for KJF prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

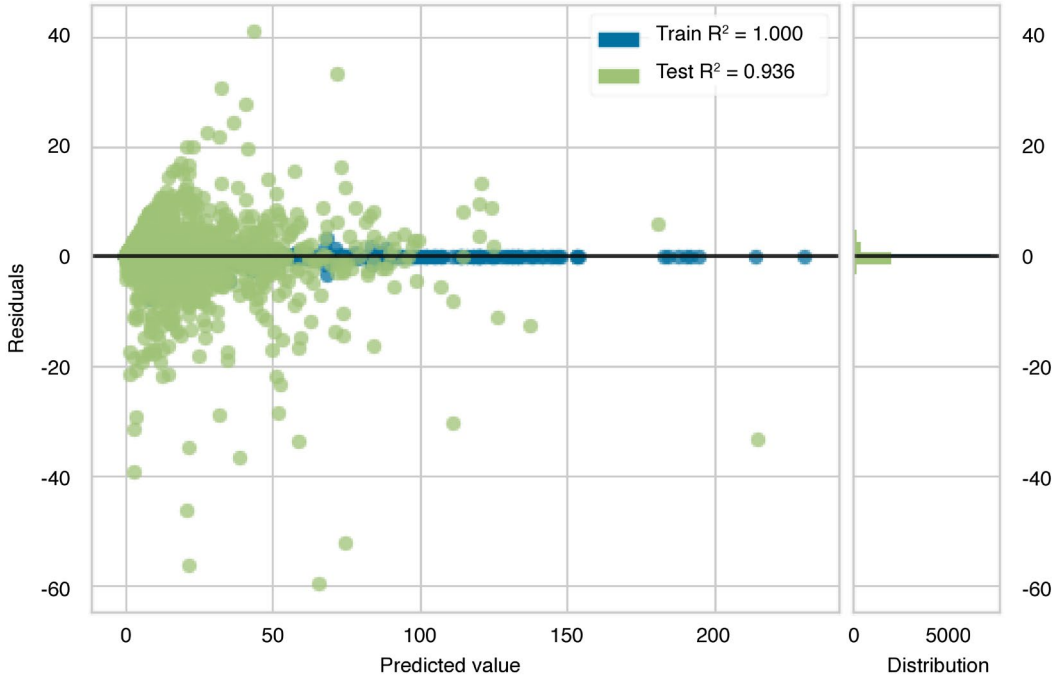
Figure SI.B5: Error analysis for DGO prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

Section SI. B: Error plots for the extremely randomized tree regressor (ETR) model

Figure SI.B6: Error analysis for HFO prediction with the ETR model.



Source: KAPSARC Oil Value-Chain Analyzer

Section SI.C: Results and statistical analysis of the multivariate linear regression (MLR) model

Regression results and statistical analyses are presented in the following tables for refinery products including liquefied petroleum gas (LPG), naphtha (NAP), gasoline (GAS), kerosene/jet fuel (KJF), gas oil diesel (DGO), and heavy fuel oil (HFO).

Table SI.C1. Regression results for LPG prediction with the MLR model.

Dep. Variable:	log(LPG + 1)	R-squared:	0.769
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	1141.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-8128.6
No. Observations:	10922	AIC:	1.635e+04
Df Residuals:	10878	BIC:	1.667e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.4534	0.014	-33.264	0.000	-0.480	-0.427
Configuration [T.Config_1]	-0.0744	0.011	-6.850	0.000	-0.096	-0.053
Configuration [T.Config_10]	0.1112	0.024	4.636	0.000	0.064	0.158
Configuration [T.Config_11]	0.1840	0.018	10.023	0.000	0.148	0.220
Configuration [T.Config_13]	-0.0244	0.038	-0.643	0.520	-0.099	0.050
Configuration [T.Config_14]	0.0171	0.071	0.240	0.810	-0.122	0.157
Configuration [T.Config_15]	0.0788	0.026	3.027	0.002	0.028	0.130
Configuration [T.Config_16]	0.2024	0.060	3.365	0.001	0.085	0.320
Configuration [T.Config_17]	0.6174	0.065	9.481	0.000	0.490	0.745
Configuration [T.Config_18]	0.1192	0.033	3.651	0.000	0.055	0.183
Configuration [T.Config_2]	0.3015	0.044	6.823	0.000	0.215	0.388
Configuration [T.Config_20]	0.8479	0.097	8.767	0.000	0.658	1.037
Configuration [T.Config_21]	0.7114	0.067	10.647	0.000	0.580	0.842
Configuration [T.Config_22]	-0.5430	0.075	-7.208	0.000	-0.691	-0.395
Configuration [T.Config_23]	0.2038	0.079	2.579	0.010	0.049	0.359
Configuration [T.Config_24]	0.5067	0.034	15.124	0.000	0.441	0.572
Configuration [T.Config_25]	0.2188	0.023	9.671	0.000	0.174	0.263
Configuration [T.Config_26]	-0.0711	0.084	-0.849	0.396	-0.235	0.093
Configuration [T.Config_27]	1.0192	0.047	21.635	0.000	0.927	1.112
Configuration [T.Config_28]	0.3244	0.056	5.826	0.000	0.215	0.434
Configuration [T.Config_29]	0.3810	0.102	3.747	0.000	0.182	0.580
Configuration [T.Config_3]	0.1799	0.038	4.689	0.000	0.105	0.255

Configuration [T.Config_30]	0.3533	0.045	7.800	0.000	0.265	0.442
Configuration [T.Config_31]	0.6775	0.069	9.802	0.000	0.542	0.813
Configuration [T.Config_32]	0.0159	0.069	0.232	0.817	-0.119	0.151
Configuration [T.Config_33]	0.3600	0.032	11.225	0.000	0.297	0.423
Configuration [T.Config_34]	-0.1714	0.193	-0.887	0.375	-0.550	0.207
Configuration [T.Config_35]	0.4239	0.115	3.698	0.000	0.199	0.649
Configuration [T.Config_36]	1.2426	0.046	26.772	0.000	1.152	1.334
Configuration [T.Config_37]	-0.1259	0.061	-2.053	0.040	-0.246	-0.006
Configuration [T.Config_38]	-0.4856	0.553	-0.878	0.380	-1.570	0.598
Configuration [T.Config_39]	1.3457	0.108	12.497	0.000	1.135	1.557
Configuration [T.Config_4]	-0.0339	0.020	-1.738	0.082	-0.072	0.004
Configuration [T.Config_5]	-0.1077	0.028	-3.835	0.000	-0.163	-0.053
Configuration [T.Config_6]	0.2761	0.043	6.456	0.000	0.192	0.360
Configuration [T.Config_7]	0.4456	0.136	3.285	0.001	0.180	0.712
Configuration [T.Config_8]	0.0296	0.052	0.571	0.568	-0.072	0.131
log(CAP + 1)	0.1762	0.006	27.996	0.000	0.164	0.189
log(LSO + 1)	0.0580	0.006	9.821	0.000	0.046	0.070
log(LSW + 1)	0.1305	0.004	34.701	0.000	0.123	0.138
log(MSO + 1)	0.1542	0.004	39.055	0.000	0.146	0.162
log(MSW + 1)	0.1364	0.005	24.891	0.000	0.126	0.147
log(HSO + 1)	0.0689	0.005	14.163	0.000	0.059	0.078
log(HSW + 1)	0.0636	0.013	4.948	0.000	0.038	0.089

Omnibus:	367.519	Durbin-Watson:	2.012
Prob(Omnibus):	0.000	Jarque-Bera (JB):	831.602
Skew:	-0.193	Prob(JB):	2.63e-181
Kurtosis:	4.296	Cond. No.	258.

Table SI.C2. Regression results for NAP prediction with the MLR model.

Dep. Variable:	log(NAP + 1)	R-squared:	0.570
Model:	OLS	Adj. R-squared:	0.568
Method:	Least Squares	F-statistic:	1641.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-13259.
No. Observations:	10922	AIC:	2.661e+04
Df Residuals:	10878	BIC:	2.693e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.4614	0.022	-21.350	0.000	-0.504	-0.419
Configuration [T.Config_1]	-0.2452	0.024	-10.357	0.000	-0.292	-0.199
Configuration [T.Config_10]	-0.4393	0.051	-8.575	0.000	-0.540	-0.339
Configuration [T.Config_11]	-0.5424	0.033	-16.447	0.000	-0.607	-0.478

Configuration [T.Config_13]	-0.4970	0.047	-10.511	0.000	-0.590	-0.404
Configuration [T.Config_14]	-0.4682	0.188	-2.488	0.013	-0.837	-0.099
Configuration [T.Config_15]	-0.4474	0.046	-9.636	0.000	-0.538	-0.356
Configuration [T.Config_16]	0.0241	0.060	0.403	0.687	-0.093	0.141
Configuration [T.Config_17]	0.2823	0.067	4.235	0.000	0.152	0.413
Configuration [T.Config_18]	-0.4789	0.053	-9.073	0.000	-0.582	-0.375
Configuration [T.Config_2]	-0.5542	0.049	-11.325	0.000	-0.650	-0.458
Configuration [T.Config_20]	1.1288	0.067	16.775	0.000	0.997	1.261
Configuration [T.Config_21]	0.0852	0.087	0.973	0.330	-0.086	0.257
Configuration [T.Config_22]	1.6348	0.043	37.698	0.000	1.550	1.720
Configuration [T.Config_23]	-0.2702	0.085	-3.172	0.002	-0.437	-0.103
Configuration [T.Config_24]	-0.3665	0.109	-3.355	0.001	-0.581	-0.152
Configuration [T.Config_25]	-0.4261	0.036	-11.886	0.000	-0.496	-0.356
Configuration [T.Config_26]	-0.4739	0.096	-4.919	0.000	-0.663	-0.285
Configuration [T.Config_27]	0.5446	0.065	8.434	0.000	0.418	0.671
Configuration [T.Config_28]	0.8107	0.107	7.585	0.000	0.601	1.020
Configuration [T.Config_29]	1.5174	0.076	19.994	0.000	1.369	1.666
Configuration [T.Config_3]	-0.3403	0.087	-3.894	0.000	-0.512	-0.169
Configuration [T.Config_30]	0.0462	0.059	0.788	0.430	-0.069	0.161
Configuration [T.Config_31]	0.2047	0.083	2.465	0.014	0.042	0.368
Configuration [T.Config_32]	-0.3676	0.076	-4.829	0.000	-0.517	-0.218
Configuration [T.Config_33]	-0.5652	0.053	-10.755	0.000	-0.668	-0.462
Configuration [T.Config_34]	-0.9039	0.287	-3.144	0.002	-1.467	-0.340
Configuration [T.Config_35]	-0.9288	0.184	-5.050	0.000	-1.289	-0.568
Configuration [T.Config_36]	0.9793	0.067	14.663	0.000	0.848	1.110
Configuration [T.Config_37]	0.3372	0.058	5.785	0.000	0.223	0.451
Configuration [T.Config_38]	-0.5872	0.101	-5.841	0.000	-0.784	-0.390
Configuration [T.Config_39]	0.9134	0.118	7.770	0.000	0.683	1.144
Configuration [T.Config_4]	-0.5908	0.043	-13.696	0.000	-0.675	-0.506
Configuration [T.Config_5]	-0.2148	0.056	-3.808	0.000	-0.325	-0.104
Configuration [T.Config_6]	-0.4308	0.079	-5.441	0.000	-0.586	-0.276
Configuration [T.Config_7]	-0.2761	0.094	-2.926	0.003	-0.461	-0.091
Configuration [T.Config_8]	-0.4166	0.064	-6.537	0.000	-0.542	-0.292
log(CAP + 1)	0.3003	0.011	27.165	0.000	0.279	0.322
log(LSO + 1)	0.1199	0.008	14.172	0.000	0.103	0.136
log(LSW + 1)	0.1165	0.006	18.818	0.000	0.104	0.129
log(MSO + 1)	0.1695	0.006	26.372	0.000	0.157	0.182
log(MSW + 1)	0.1144	0.008	14.391	0.000	0.099	0.130
log(HSO + 1)	-0.0206	0.007	-2.751	0.006	-0.035	-0.006
log(HSW + 1)	0.0546	0.020	2.672	0.008	0.015	0.095

Omnibus:	117.178	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	200.182
Skew:	0.025	Prob(JB):	3.40e-44
Kurtosis:	3.661	Cond. No.	258.

Section SI.C: Results and statistical analysis of the multivariate linear regression (MLR) model

Table SI.C3. Regression results for GAS prediction with the MLR model.

Dep. Variable:	log(GAS + 1)	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	2052.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-11304.
No. Observations:	10922	AIC:	2.270e+04
Df Residuals:	10878	BIC:	2.302e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.7067	0.020	-34.841	0.000	-0.747	-0.667
Configuration [T.Config_1]	-0.1880	0.021	-8.933	0.000	-0.229	-0.147
Configuration [T.Config_10]	0.1719	0.036	4.824	0.000	0.102	0.242
Configuration [T.Config_11]	0.2608	0.027	9.535	0.000	0.207	0.314
Configuration [T.Config_13]	-0.5483	0.053	-10.327	0.000	-0.652	-0.444
Configuration [T.Config_14]	0.1048	0.089	1.178	0.239	-0.070	0.279
Configuration [T.Config_15]	-0.0433	0.043	-1.013	0.311	-0.127	0.040
Configuration [T.Config_16]	0.1352	0.059	2.289	0.022	0.019	0.251
Configuration [T.Config_17]	0.1373	0.068	2.031	0.042	0.005	0.270
Configuration [T.Config_18]	0.3600	0.039	9.184	0.000	0.283	0.437
Configuration [T.Config_2]	0.1088	0.068	1.606	0.108	-0.024	0.241
Configuration [T.Config_20]	0.5307	0.139	3.816	0.000	0.258	0.803
Configuration [T.Config_21]	0.2245	0.071	3.170	0.002	0.086	0.363
Configuration [T.Config_22]	0.0083	0.058	0.144	0.885	-0.104	0.121
Configuration [T.Config_23]	0.6167	0.109	5.675	0.000	0.404	0.830
Configuration [T.Config_24]	0.1251	0.081	1.554	0.120	-0.033	0.283
Configuration [T.Config_25]	0.0968	0.032	2.991	0.003	0.033	0.160
Configuration [T.Config_26]	-0.5959	0.180	-3.307	0.001	-0.949	-0.243
Configuration [T.Config_27]	0.6320	0.092	6.849	0.000	0.451	0.813
Configuration [T.Config_28]	0.5921	0.058	10.259	0.000	0.479	0.705
Configuration [T.Config_29]	0.3425	0.154	2.231	0.026	0.042	0.643
Configuration [T.Config_3]	0.2824	0.089	3.170	0.002	0.108	0.457
Configuration [T.Config_30]	-0.2888	0.053	-5.405	0.000	-0.393	-0.184
Configuration [T.Config_31]	-0.0669	0.088	-0.760	0.447	-0.239	0.106
Configuration [T.Config_32]	0.2802	0.063	4.424	0.000	0.156	0.404
Configuration [T.Config_33]	0.5112	0.041	12.613	0.000	0.432	0.591
Configuration [T.Config_34]	0.7608	0.106	7.157	0.000	0.552	0.969
Configuration [T.Config_35]	0.5604	0.080	6.985	0.000	0.403	0.718
Configuration [T.Config_36]	0.5282	0.057	9.192	0.000	0.416	0.641
Configuration [T.Config_37]	-0.0518	0.118	-0.437	0.662	-0.284	0.180

Configuration [T.Config_38]	0.5113	0.366	1.395	0.163	-0.207	1.229
Configuration [T.Config_39]	0.4866	0.127	3.821	0.000	0.237	0.736
Configuration [T.Config_4]	-0.2546	0.058	-4.381	0.000	-0.368	-0.141
Configuration [T.Config_5]	-0.4323	0.050	-8.563	0.000	-0.531	-0.333
Configuration [T.Config_6]	0.1675	0.064	2.614	0.009	0.042	0.293
Configuration [T.Config_7]	0.9885	0.179	5.520	0.000	0.637	1.339
Configuration [T.Config_8]	-0.4431	0.073	-6.069	0.000	-0.586	-0.300
log(CAP + 1)	0.3920	0.010	40.373	0.000	0.373	0.411
log(LSO + 1)	0.0314	0.007	4.794	0.000	0.019	0.044
log(LSW + 1)	0.2607	0.005	51.910	0.000	0.251	0.271
log(MSO + 1)	0.2232	0.005	40.704	0.000	0.212	0.234
log(MSW + 1)	0.2102	0.006	32.852	0.000	0.198	0.223
log(HSO + 1)	0.1588	0.006	26.461	0.000	0.147	0.171
log(HSW + 1)	-0.0048	0.012	-0.406	0.684	-0.028	0.018

Omnibus:	280.999	Durbin-Watson:	2.017
Prob(Omnibus):	0.000	Jarque-Bera (JB):	350.634
Skew:	-0.326	Prob(JB):	7.26e-77
Kurtosis:	3.587	Cond. No.	258.

Table SI.C4. Regression results for KJF prediction with the MLR model.

Dep. Variable:	log(KJF + 1)	R-squared:	0.837
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	3704.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-8168.0
No. Observations:	10922	AIC:	1.642e+04
Df Residuals:	10878	BIC:	1.675e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6000	0.017	-35.501	0.000	-0.633	-0.567
Configuration [T.Config_1]	-0.1208	0.015	-7.935	0.000	-0.151	-0.091
Configuration [T.Config_10]	0.0284	0.032	0.898	0.369	-0.034	0.090
Configuration [T.Config_11]	-0.2395	0.020	-11.698	0.000	-0.280	-0.199
Configuration [T.Config_13]	-0.4837	0.036	-13.474	0.000	-0.554	-0.413
Configuration [T.Config_14]	-0.5217	0.159	-3.281	0.001	-0.833	-0.210
Configuration [T.Config_15]	-0.3123	0.033	-9.453	0.000	-0.377	-0.248
Configuration [T.Config_16]	0.0513	0.057	0.893	0.372	-0.061	0.164

Configuration [T.Config_17]	-0.4568	0.046	-9.965	0.000	-0.547	-0.367
Configuration [T.Config_18]	-0.1369	0.033	-4.140	0.000	-0.202	-0.072
Configuration [T.Config_2]	-0.2581	0.035	-7.435	0.000	-0.326	-0.190
Configuration [T.Config_20]	0.6336	0.063	10.086	0.000	0.510	0.757
Configuration [T.Config_21]	-0.4329	0.047	-9.263	0.000	-0.524	-0.341
Configuration [T.Config_22]	1.5565	0.026	59.640	0.000	1.505	1.608
Configuration [T.Config_23]	-0.1560	0.073	-2.147	0.032	-0.298	-0.014
Configuration [T.Config_24]	0.4401	0.053	8.300	0.000	0.336	0.544
Configuration [T.Config_25]	-0.3571	0.022	-16.551	0.000	-0.399	-0.315
Configuration [T.Config_26]	-0.7439	0.107	-6.940	0.000	-0.954	-0.534
Configuration [T.Config_27]	-0.0287	0.110	-0.262	0.793	-0.244	0.186
Configuration [T.Config_28]	0.2485	0.098	2.540	0.011	0.057	0.440
Configuration [T.Config_29]	0.5427	0.079	6.872	0.000	0.388	0.697
Configuration [T.Config_3]	-0.2217	0.065	-3.429	0.001	-0.348	-0.095
Configuration [T.Config_30]	-0.4222	0.033	-12.628	0.000	-0.488	-0.357
Configuration [T.Config_31]	-0.7316	0.039	-18.824	0.000	-0.808	-0.655
Configuration [T.Config_32]	-0.4531	0.061	-7.388	0.000	-0.573	-0.333
Configuration [T.Config_33]	-0.0496	0.029	-1.728	0.084	-0.106	0.007
Configuration [T.Config_34]	0.0261	0.105	0.250	0.803	-0.179	0.231
Configuration [T.Config_35]	-0.3327	0.061	-5.481	0.000	-0.452	-0.214
Configuration [T.Config_36]	0.1344	0.060	2.227	0.026	0.016	0.253
Configuration [T.Config_37]	0.0599	0.067	0.891	0.373	-0.072	0.192
Configuration [T.Config_38]	-0.2419	0.172	-1.407	0.159	-0.579	0.095
Configuration [T.Config_39]	0.3052	0.182	1.675	0.094	-0.052	0.662
Configuration [T.Config_4]	-0.2591	0.039	-6.696	0.000	-0.335	-0.183
Configuration [T.Config_5]	-0.3023	0.042	-7.268	0.000	-0.384	-0.221
Configuration [T.Config_6]	-0.1405	0.041	-3.427	0.001	-0.221	-0.060
Configuration [T.Config_7]	0.4759	0.113	4.211	0.000	0.254	0.697
Configuration [T.Config_8]	-0.6683	0.025	-27.166	0.000	-0.717	-0.620
log(CAP + 1)	0.3226	0.008	42.426	0.000	0.308	0.338
log(LSO + 1)	0.1250	0.005	22.802	0.000	0.114	0.136
log(LSW + 1)	0.1609	0.004	41.100	0.000	0.153	0.169
log(MSO + 1)	0.1998	0.004	49.404	0.000	0.192	0.208
log(MSW + 1)	0.1417	0.005	30.277	0.000	0.133	0.151
log(HSO + 1)	0.0812	0.004	18.182	0.000	0.072	0.090
log(HSW + 1)	0.0746	0.009	8.487	0.000	0.057	0.092

Omnibus:	307.127	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	480.099
Skew:	0.276	Prob(JB):	5.60e-105
Kurtosis:	3.867	Cond. No.	258.

Table SI.C5. Regression results for DGO prediction with the MLR model.

Dep. Variable:	log(DGO + 1)	R-squared:	0.878
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	3245.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-10081.
No. Observations:	10922	AIC:	2.025e+04
Df Residuals:	10878	BIC:	2.057e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6353	0.019	-33.446	0.000	-0.672	-0.598
Configuration [T.Config_1]	-0.0532	0.017	-3.217	0.001	-0.086	-0.021
Configuration [T.Config_10]	0.2934	0.031	9.452	0.000	0.233	0.354
Configuration [T.Config_11]	-0.1741	0.023	-7.692	0.000	-0.218	-0.130
Configuration [T.Config_13]	-0.7593	0.066	-11.462	0.000	-0.889	-0.629
Configuration [T.Config_14]	0.4459	0.084	5.332	0.000	0.282	0.610
Configuration [T.Config_15]	-0.0560	0.049	-1.135	0.256	-0.153	0.041
Configuration [T.Config_16]	-0.1889	0.055	-3.419	0.001	-0.297	-0.081
Configuration [T.Config_17]	-0.2232	0.065	-3.425	0.001	-0.351	-0.095
Configuration [T.Config_18]	0.0003	0.034	0.009	0.993	-0.066	0.066
Configuration [T.Config_2]	-0.3399	0.057	-5.978	0.000	-0.451	-0.228
Configuration [T.Config_20]	0.5564	0.089	6.237	0.000	0.382	0.731
Configuration [T.Config_21]	0.1503	0.072	2.078	0.038	0.009	0.292
Configuration [T.Config_22]	0.8556	0.029	29.228	0.000	0.798	0.913
Configuration [T.Config_23]	0.0013	0.077	0.017	0.987	-0.150	0.152
Configuration [T.Config_24]	0.4963	0.059	8.471	0.000	0.381	0.611
Configuration [T.Config_25]	-0.2988	0.029	-10.319	0.000	-0.356	-0.242
Configuration [T.Config_26]	-0.3169	0.143	-2.214	0.027	-0.597	-0.036
Configuration [T.Config_27]	0.3061	0.096	3.191	0.001	0.118	0.494
Configuration [T.Config_28]	0.4641	0.040	11.471	0.000	0.385	0.543
Configuration [T.Config_29]	0.2236	0.129	1.740	0.082	-0.028	0.476
Configuration [T.Config_3]	0.2136	0.109	1.955	0.051	-0.000	0.428
Configuration [T.Config_30]	-0.4128	0.053	-7.734	0.000	-0.517	-0.308
Configuration [T.Config_31]	-0.3430	0.087	-3.950	0.000	-0.513	-0.173
Configuration [T.Config_32]	0.0850	0.055	1.555	0.120	-0.022	0.192
Configuration [T.Config_33]	-0.0026	0.038	-0.068	0.946	-0.077	0.072
Configuration [T.Config_34]	0.0806	0.108	0.743	0.458	-0.132	0.293
Configuration [T.Config_35]	-0.2029	0.059	-3.425	0.001	-0.319	-0.087
Configuration [T.Config_36]	0.3791	0.057	6.625	0.000	0.267	0.491
Configuration [T.Config_37]	-0.5920	0.115	-5.144	0.000	-0.818	-0.366

Configuration [T.Config_38]	-0.3032	0.685	-0.443	0.658	-1.645	1.038
Configuration [T.Config_39]	0.3242	0.077	4.229	0.000	0.174	0.474
Configuration [T.Config_4]	-0.5547	0.050	-11.011	0.000	-0.653	-0.456
Configuration [T.Config_5]	-0.3474	0.061	-5.729	0.000	-0.466	-0.229
Configuration [T.Config_6]	-0.1927	0.045	-4.238	0.000	-0.282	-0.104
Configuration [T.Config_7]	0.2681	0.139	1.923	0.054	-0.005	0.541
Configuration [T.Config_8]	-0.7317	0.072	-10.215	0.000	-0.872	-0.591
log(CAP + 1)	0.4509	0.009	48.762	0.000	0.433	0.469
log(LSO + 1)	0.0329	0.006	5.069	0.000	0.020	0.046
log(LSW + 1)	0.2578	0.005	54.622	0.000	0.249	0.267
log(MSO + 1)	0.2877	0.005	55.049	0.000	0.277	0.298
log(MSW + 1)	0.2390	0.006	36.777	0.000	0.226	0.252
log(HSO + 1)	0.1360	0.005	24.757	0.000	0.125	0.147
log(HSW + 1)	0.0225	0.015	1.453	0.146	-0.008	0.053

Omnibus:	33.414	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36.390
Skew:	0.099	Prob(JB):	1.25e-08
Kurtosis:	3.202	Cond. No.	258.

Table SI.C6. Regression results for HFO prediction with the MLR model.

Dep. Variable:	log(HFO + 1)	R-squared:	0.656
Model:	OLS	Adj. R-squared:	0.655
Method:	Least Squares	F-statistic:	1064.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-13035.
No. Observations:	10922	AIC:	2.616e+04
Df Residuals:	10878	BIC:	2.648e+04
Df Model:	43		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.5816	0.023	-25.292	0.000	-0.627	-0.537
Configuration [T.Config_1]	0.1263	0.025	5.134	0.000	0.078	0.175
Configuration [T.Config_10]	0.0021	0.042	0.049	0.961	-0.080	0.084
Configuration [T.Config_11]	-0.4291	0.033	-13.083	0.000	-0.493	-0.365
Configuration [T.Config_13]	-1.1194	0.042	-26.744	0.000	-1.201	-1.037
Configuration [T.Config_14]	0.6554	0.093	7.026	0.000	0.473	0.838
Configuration [T.Config_15]	-0.0817	0.067	-1.213	0.225	-0.214	0.050
Configuration [T.Config_16]	-0.8020	0.075	-10.643	0.000	-0.950	-0.654
Configuration [T.Config_17]	-0.7005	0.089	-7.894	0.000	-0.874	-0.527

Configuration [T.Config_18]	-0.4245	0.045	-9.373	0.000	-0.513	-0.336
Configuration [T.Config_2]	-0.7453	0.048	-15.496	0.000	-0.840	-0.651
Configuration [T.Config_20]	-0.2467	0.162	-1.527	0.127	-0.563	0.070
Configuration [T.Config_21]	-1.1556	0.090	-12.792	0.000	-1.333	-0.979
Configuration [T.Config_22]	0.4615	0.042	11.078	0.000	0.380	0.543
Configuration [T.Config_23]	-0.6237	0.094	-6.640	0.000	-0.808	-0.440
Configuration [T.Config_24]	-1.4601	0.080	-18.342	0.000	-1.616	-1.304
Configuration [T.Config_25]	-1.1318	0.034	-32.902	0.000	-1.199	-1.064
Configuration [T.Config_26]	-0.2048	0.106	-1.926	0.054	-0.413	0.004
Configuration [T.Config_27]	-0.3174	0.159	-2.002	0.045	-0.628	-0.007
Configuration [T.Config_28]	-1.6573	0.111	-14.935	0.000	-1.875	-1.440
Configuration [T.Config_29]	-0.4312	0.177	-2.443	0.015	-0.777	-0.085
Configuration [T.Config_3]	-0.4578	0.141	-3.235	0.001	-0.735	-0.180
Configuration [T.Config_30]	-1.4906	0.054	-27.715	0.000	-1.596	-1.385
Configuration [T.Config_31]	-1.6854	0.068	-24.830	0.000	-1.818	-1.552
Configuration [T.Config_32]	-0.7913	0.089	-8.887	0.000	-0.966	-0.617
Configuration [T.Config_33]	-1.5271	0.053	-28.836	0.000	-1.631	-1.423
Configuration [T.Config_34]	-2.6699	0.170	-15.670	0.000	-3.004	-2.336
Configuration [T.Config_35]	-2.2686	0.067	-33.981	0.000	-2.399	-2.138
Configuration [T.Config_36]	-1.8854	0.084	-22.388	0.000	-2.050	-1.720
Configuration [T.Config_37]	-0.6738	0.106	-6.341	0.000	-0.882	-0.466
Configuration [T.Config_38]	-1.3440	0.718	-1.872	0.061	-2.751	0.063
Configuration [T.Config_39]	-2.2781	0.123	-18.594	0.000	-2.518	-2.038
Configuration [T.Config_4]	-0.8567	0.038	-22.757	0.000	-0.931	-0.783
Configuration [T.Config_5]	-0.2387	0.084	-2.857	0.004	-0.402	-0.075
Configuration [T.Config_6]	-0.4608	0.052	-8.828	0.000	-0.563	-0.359
Configuration [T.Config_7]	-0.4753	0.101	-4.686	0.000	-0.674	-0.277
Configuration [T.Config_8]	-0.4922	0.149	-3.304	0.001	-0.784	-0.200
log(CAP + 1)	0.4748	0.011	42.908	0.000	0.453	0.497
log(LSO + 1)	0.0188	0.008	2.286	0.022	0.003	0.035
log(LSW + 1)	0.0459	0.006	7.660	0.000	0.034	0.058
log(MSO + 1)	0.2631	0.006	42.592	0.000	0.251	0.275
log(MSW + 1)	0.1317	0.007	17.746	0.000	0.117	0.146
log(HSO + 1)	0.0691	0.007	9.304	0.000	0.055	0.084
log(HSW + 1)	-0.0416	0.016	-2.667	0.008	-0.072	-0.011

Omnibus:	123.340	Durbin-Watson:	1.964
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.852
Skew:	-0.212	Prob(JB):	2.60e-31
Kurtosis:	3.361	Cond. No.	258.

Notes

Notes

Notes

Notes

About the Author

Dr. Evar Umeozor is a Research Fellow at KAPSARC leading integrated oil and gas value-chain research, with a focus on the downstream segment. Dr. Umeozor is a process and energy systems engineering expert with multidisciplinary research and industry experience in technical assessments, life cycle cost and emission analyses, and the application of the systems approach to optimal policy design considering multiple objectives and constraints.

He holds a master's degree in chemical engineering from Imperial College London, United Kingdom, and a doctoral degree in energy and environment systems engineering from the University of Calgary, Canada.

About the Project

This study was conducted under the KAPSARC Oil Value-Chain Analyzer (KOVA) project, which develops data, models, and analytics tools to gain insights into the impacts of business and government policies on the economic, environmental and energy efficiencies of the downstream oil and gas sector, including the integral effects on the midstream and upstream sectors in Saudi Arabia and beyond. The objective of the KOVA project is to deploy integrated systems analysis approaches for identifying optimal policy design options that satisfy the performance metrics and targets of stakeholders in the energy ecosystem.

مركـــزالملـــك عبـــدالله للدراســـات والبحوث البترولية
King Abdullah Petroleum Studies and Research Center

www.kapsarc.org