energy&fuels

Fossil Fuels

Computer-aided Gasoline Compositional Model Development based on GC-FID Analysis

Chen Cui, Linzhou Zhang, Yongjian Ma, Triveni Billa, Zhen Hou, Quan Shi, Suoqi Zhao, Chunming Xu, and Michael T Klein

Energy Fuels, **Just Accepted Manuscript** • DOI: 10.1021/acs.energyfuels.8b01953 • Publication Date (Web): 24 Jul 2018

Downloaded from http://pubs.acs.org on July 31, 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Computer-aided Gasoline Compositional Model Development based on GC-FID Analysis

Chen Cui[1,2], Linzhou Zhang*[1], Yongjian Ma[1], Triveni Billa[2], Zhen Hou[2], Quan Shi[1], Suoqi Zhao[1], Chunming Xu[1] and Michael T. Klein*[2,3]

1. State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing, 102249, P. R. China

2. Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States and Center for Refining and Petrochemicals

3. King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

ABSTRACT

The demand for improved gasoline product quality has helped make molecular-level models become more and more preferred for the modern refinery. Building the molecular compositional model is an essential first step for this quantitative molecular management of gasoline streams. Gas chromatography equipped with flame ion detection (GC-FID) is commonly used in the gasoline detailed hydrocarbon analysis (DHA). The combination of GC-FID analysis and molecular-level modeling is thus very attractive. In the present study, we developed a gasoline compositional model based solely on GC-FID as input. To suppress the negative influence of peak coelution, we developed a statistics-based peak tuning algorithm to obtain individual compound resolution at higher carbon number range. Using the tuned result as input, the molecular-level gasoline compositional model was built by estimating the quantitative percentages of the species in a predefined molecular library (573 molecules). The molecular-level compositional model has good extensibility and can link to the molecule-based physical properties prediction model. The model has been verified via applications on various gasoline samples. The prediction of research octane number for large-scale gasoline samples was also revealed.

1. INTRODUCTION

Although the pseudo-component-based methodology is widely applied in flowsheet simulation software, it cannot meet the present requirements of gasoline blending and process modeling for the purpose of accurate quality control. The physical and chemical properties of a gasoline are determined by its molecular composition. Thus, it is becoming increasingly important to develop molecular-level models to achieve accurate optimization of gasoline refining processes. In the molecular-level management of gasoline streams, two key problems should be addressed: 1) the acquisition of molecular composition, and 2) the establishment of molecular-based property prediction models.

Gas chromatography coupled with a flame ionization detector (GC-FID) has become a standard method for gasoline detailed hydrocarbon analysis (DHA).[1-3] In GC-FID DHA, the retention index is used for the identity and the peak area or peak height is used for quantification. In a typical GC-FID analysis, around 200-400 components can be detected depending on the gasoline type and measuring method. The compositional result is widely used in refineries for product quality control, novel refining process development and catalyst design.

Although CG-FID provides detailed molecular information of a gasoline, it cannot separate all the compounds in the mixture. The number of isomers increases significantly with the increase in the carbon number. The separation capacity of a gas chromatography column is not enough to resolve all the isomer peaks in the high carbon number region. Many chromatogram peaks at high retention times have coeluting components. In recent years, several new technologies, such as GC coupled with vacuum ultraviolet detector (GC-VUV)[4-6] and comprehensive two-dimensional GC (GC×GC)[7-9], have been developed for molecular composition analysis of gasoline, yet standard methods are still not widely accepted in the field.

Another approach to the specification of gasoline composition is the molecular reconstruction of gasoline via computer-aided modeling. For this purpose, a series of virtual molecules are manually defined on the computer, rather than identified directly from the instrumental analysis. The limited measured data (normally bulk properties) are used to calculate the quantitative abundance information. In the 1990s, Quann et al.[10-12] proposed the Structured Oriented Lumping (SOL) method to describe the molecular-level composition of petroleum. Ghosh et al.[13] successfully applied the SOL method in the simulation of the naphtha hydro-desulfurization process.

Stochastic and predefined methodologies were also mentioned in various publications. Neurock et al.[14] developed a stochastic reconstruction method which randomly sampled different statistical distributions of structural attributes by a Monte Carlo method. Hudebine et al.[15] optimized the method by fine-tuning the result with the entropy maximization method. Generally, the stochastic method was used for the heavy petroleum fraction.[16, 17] Recently, Pan et al.[18] extended the stochastic strategy to the gasoline fraction, and structural attributes were replaced by SOL vectors in random sampling.

Since Monte Carlo sampling was time-consuming, some researchers preferred to establish a predefined molecular library for gasoline fraction. Peng[19] proposed the molecular type homologous series matrix (MTHS) method to represent the composition of petroleum fractions. A series of papers about gasoline composition based on the concept of MTHS were published by the research group from University of Manchester[20-23] who also extended the MTHS method to heavy fraction of petroleum.[24, 25] Albahri et al.[26] selected 68 representative hydrocarbon molecules to establish the naphtha composition. Hudebine et al.[27] selected 230 hydrocarbon molecules to establish the FCC gasoline composition. Cui et al.[28] selected 170 hydrocarbons and heteroatom species to establish a general composition model for different types of gasoline. However, these compositional models faced the uncertainty problem caused by multiple solutions. Cui et al.[28] also evaluated the influence of the uncertainty problem on the prediction of gasoline properties. Moreover, researchers also reconstructed composition by machine learning, such as Bayesian estimation[29, 30] methods and by using artificial neural networks.[31]

Molecular reconstruction methods give an approximate gasoline molecular composition that exhibits predicted bulk property values similar to the experimental data since the composition is deduced from bulk properties. There will be deviations in any specific result. Moreover, the measurement of bulk properties may be time consuming. On the other hand, since GC-FID is capable of giving detailed hydrocarbon information of gasoline samples, some researchers had made some attempts to use GC-FID result as input to build a molecular-based composition model directly. Ghosh et al.[32] proposed a method that fine-tunes the GC analysis data by using bulk properties.

Building on the foregoing research, we have developed a new model strategy named Computer-aided Gasoline Compositional (CGC) modeling for the gasoline characterization. CGC modeling fine-tunes the GC-FID-based DHA results by using a statistics-based algorithm to provide the complete gasoline composition. The reconstructed composition can then be used in molecular-based property prediction.

2. THEORY AND METHOD

2.1 Theoretical Basis and Model Overview

One would expect GC-FID to directly give the complete qualitative and quantitative molecular composition of a gasoline sample. However, the GC column separation capacity is not sufficient for high carbon number molecules, resulting in severe peak overlapping at their corresponding elution region. Thus, the GC-FID result can be categorized into three types: known peaks, coeluting peaks, and unidentifiable peaks, which represent the identification of single component, multiple components, and unknown components, respectively. Figure 1 shows a typical GC-FID spectrum of gasoline. Taking the range of retention time from 41 to 44 min as an example, peak type 1 was an identified peak, in which n-octane was recognized. Peak type 2 was a coeluting peak, in which trans-2-octane and 1,2,3-cyclopentane eluted at the same time. The relative content of the two components could not be determined. Peak type 3 was an unidentified peak. The identity in the peak is unknown.

The goal of the presently described approach is to use solely GC-FID results to obtain the molecular composition and, based on this, to build a compositional model that can predict bulk properties precisely. Thus, an algorithm was developed to infer the full molecular compositional information from the GC-FID result, particularly, from the coeluting and undefined peaks.

The nature of petroleum helped guide the strategy. In most cases, compounds in gasoline form homologous series. It is widely accepted that content of each molecule in different homologous series is not independent and follows a certain statistical distribution.[33] This is the technical basis for our approach that uses the content of some detected molecule to infer the content of a missing component. To infer the identity and determine the relative content of components in coeluting peaks and unidentified peaks, a statistics-based peak tuning algorithm was established. The reconstructed composition which contained more accurate qualitative and quantitative information would be improved by using this tuning algorithm on DHA results. Based on the reconstructed composition, the properties of gasoline can be predicted by combining the properties of each molecule and corresponding mixing rules.

Figure 2 shows the flowchart of CGC model. The scheme could be divided into two parts. One is the statistics-based peak tuning (SPT) algorithm, and the other is a molecular-based properties prediction module. After analyzing gasoline samples using GC-FID, and the preliminary results were pre-processed and served as input for the SPT algorithm. This consisted of three steps: (1) fitting the known peaks; (2) splitting the coeluting peaks; (3) inferring possible molecules from the unidentified peaks. After the three steps, most of the components' identities and quantities in coeluting peaks and the unidentified peaks were clarified.

The reconstructed composition was then sent to the properties prediction module. There were three parts in the properties prediction module: (1) molecular composition module; (2) molecular library and properties library; (3) properties prediction rules. A gasoline model would be created when the molecular composition module received the reconstructed composition. Finally, the gasoline model could make a prediction of a certain property by querying the property of each molecule in the model and the corresponding prediction rules from the second and third part of the properties prediction module, respectively.

2.2 Molecular Identities and Properties Library

A predefined molecular library is the basic database for peak tuning and property prediction. We selected 573 common molecules from the GC-FID-based DHA results of various gasoline samples, including FCC gasoline, reformates, straight-run gasoline, coker gasoline, and catalytic pyrolysis gasoline. The diagnostics of the pre-defined molecular library are shown in Figure 3. Molecules in the library can be classified into five categories, which are normal paraffin (NP), isoparaffin (IP), olefin (O), naphthene (NC), and aromatics (A). The number of molecules in each category is shown in Figure 3 (a). The normal paraffins had the smallest number of molecules because there are no isomers. The number of olefins was the highest, occupying half of the molecules in the library, due to the excessive number of possible olefin isomers that vary in double bond position, cis-trans configuration and number of branch groups. Figure 3 (b) shows the variation of the number of molecules with carbon number. Again, the number of normal paraffins was kept at the value of one. For other molecular types, the number of molecules in library first increases and then decreases with increasing carbon number. It should be noted that the theoretical isomer number increases exponentially with carbon numbers. The reason for the drop in molecule number in the predefined library at higher carbon number region is that a lumping strategy was applied. Only one representative molecule was selected for each lump to facilitate the peak tuning and property prediction calculations. For the sake of accuracy, all molecular properties came from experimental values rather than group contribution methods. The experimental data were queried from the NIST database.

2.3 Statistic-based Peak Tuning (SPT) Algorithm

The SPT algorithm is the core of CGC model. Similar approaches have been described. For example, researchers from ExxonMobil used bulk property data to auto-tune the GC field ionization time-of-flight mass spectrometry (GC-FI-ToF) result.[34, 35] SPT is a self-tuning algorithm and does not require information of bulk property. It resolves component identity and quantity in coeluting and undefined peaks based on a statistical characteristic of petroleum molecules and homologous series.

The SPT algorithm is initiated after the preliminary GC-FID results are processed into DHA data. The algorithm tunes identified peaks, coeluting peaks, and unidentified peaks sequentially.

Fitting the known peaks is the first step. The flowchart is shown in Figure 4. A matrix which denoted the data matrix is created by classifying the known peaks by molecular type and carbon number. The elements in the data matrix are the total fraction of the molecules with the same molecular type and carbon number. Based on the notion that the content of each molecular type follows a certain statistical distribution, the three-parameter gamma distribution was adopted in the SPT algorithm to reconstruct the composition. The mathematical expression of the gamma distribution is shown below:

$$p(x) = \frac{(x-\eta)^{\alpha-1}e^{-(x-\eta)/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad (1)$$

where α, β, η were the parameters which need to be tuned during the fitting process. The carbon number and the content data of each molecular type in the data matrix were used as the data for fitting, which corresponded to x and p(x), respectively, in Equation (1). The tuned parameters and root mean square error (RMSE) would be the fitting results. These are denoted as $\text{Parameters}_{1,i}$ and $\text{RMSE}_{1,i}$, where 1 means they resulted from the step 1, and i means that they were belongs to the molecular type i.

The second step was designed to split the coeluting peaks. As shown in Figure 5, the coeluting peaks obtained from the preprocessing would be split one by one. The splitting process consisted of two parts: an assumption and a selection. Because the relative content of each molecule in coeluting peaks was unknown, we had to model the relative content for each molecule. There were so many possibilities of the relative content that a list of modeling assumptions was established. The list of these modeling assumptions is shown in Figure 6 as an example, where each assumption is listed in the sequential trial order. These assumptions of split ratio are totally user-defined. User can set any rational ratio in the program. Every assumption in the list would be tested. Taking the assumption j as an example, the coeluting peak was split based on the assumption j. As a result, the coeluting peak was converted to two or more known peaks. Then, these peaks were classified into the data matrix in terms of molecular type and carbon number. As described in step one, fitting the content data in the matrix and the results would be marked as $\mathrm{Parameters}_{2,i,j}$ and $\mathrm{RMSE}_{2,i,j}$, where 2 means they resulted from the step 2; i means that they belonged to the different molecular type i; and j means that they were belongs to the modeling assumption j. The change of the RMSE from step one to step two would be marked as Change_j, and its mathematical expression was shown below:

$$Change_j = \left(RMSE_{2,i,j} - RMSE_{1,i}\right)/RMSE_{1,i}(2)$$

After testing all the assumptions in the list, the assumption that had the smallest Change would be selected. The data matrix was then updated based on the selected assumption, as well as the $\mathrm{Parameters}_{1,i}$ and $\mathrm{RMSE}_{1,i}$. At this point it was determined that a coeluting peak was split successfully. The foregoing process was repeated until all the coeluting peaks were split. The final state of the parameters and RMSE for each molecular type at the second step was marked as $\mathrm{Parameters}_{2,i}$ and $\mathrm{RMSE}_{2,i}$.

The third step of the SPT algorithm was designed to infer the possible molecules from the unidentified peaks. As shown in Figure 7, the flowchart of this step was similar to that for the second step, as there were once again the two parts of model assumption and selection. The calculation process was similar to that for step two and so only the differences will be pointed out here.

The information of the unidentified peaks was totally unknown. Even the number of components was not known. To simplify the calculations, the first hypothesis we made at the third step was that the unidentified peaks contained just only one component. Based on this hypothesis, the possible carbon number and molecular type of the unidentified peak were assumed. Peaks around the unidentified peak were served as reference for assuming carbon number. Some other basic chemical criterions need to be followed when assume molecular type. For example, there is no need to assume that an unidentified peak was normal paraffin, because all normal paraffins in gasoline can be identified in GC-FID-based DHA system. After processing all the unidentifiable peaks, the final state of the parameters and RMSE for each molecular type at the third step were marked as $\mathrm{Parameters}_{3,i}$ and $\mathrm{RMSE}_{3,i}$.

The execution of the three steps of the SPT algorithm provides an enhanced estimation of the complete composition. It was worth pointing out that molecules in the library could be classified into more detailed molecular types. For example, the isoparaffins could be further divided into mono-branched paraffins (MP), double-branched paraffins (DP), and triple-branched paraffins (TP); the olefins could be further divided into normal olefins and branched olefins. The more detailed molecular types were also suitable for SPT algorithm. Generally, the SPT algorithm is suitable for full-range gasolines without additives. Narrow fractions might not be supported because there is not enough data for fitting.

2.4 Bulk Property Prediction

The molecular composition itself cannot be used alone for gasoline molecular management. The property prediction module that reveals all the key properties of a given molecular composition is also important. In general, the bulk properties can be calculated by combing pure component property data with corresponding mixing rules. The reconstructed gasoline composition was suitable for a variety of molecular-based properties prediction rules.

Table 1 lists all the predictable properties in the CGC model and the corresponding prediction methods. The blending of some properties can be considered as ideal, such as specific gravity and molecular weight. These properties are easy to predict, because they follow linear rules in blending. However, some properties are hard to predict due to their highly non-ideal blending rules, such as octane number and Reid vapor pressure (RVP). Molecular-based prediction models were expected to provide more accurate results for these non-ideal blending properties. Take octane number as an example. The molecule-based prediction of octane number has been widely studied by researchers. Our group developed a new molecular-based octane prediction model, which was called the CUP CNN model. The CUP CNN model is based on convolutional neural network, and the details will be illustrated in a follow-up paper. We also support the EM model, which is a famous molecular-based octane number prediction proposed by Ghosh et al.[32] from ExxonMobil. Both of two models can be used in CGC model since a complete molecular composition was given. Here, we only show the RON prediction result from CUP CNN model.

3. CASE STUDY

3.1 Application of SPT Algorithm on Gasoline Sample

An FCC gasoline was used as an example to illustrate the SPT algorithm. The GC analysis was conducted with Agilent 7890B gas chromatograph equipped with a split injector and flame ion detection. The column was a 50m × 0.2mm × 0.2μm HP-PONA. Nitrogen carrier gas needed to be tuned until the retention time of n-Dodecane in the standard sample was 81±1 min. In this case, it was set at a constant pressure 87 kPa. The oven program was 35 ºC (hold 5 min) to 200 ºC (hold 10 min) at 2 ºC min^{-1}. The inlet was held at 250 ºC, and operated in split mode with 0.5 μL injection at 150 : 1 split ratio. The FID was operated at 250 ºC. Raw chromatographic data were exported from ChemStation (Aglient Technologies) and processed by a commercial GC-based DHA system developed by SINOPEC. After measuring the gasoline sample and processing the raw data, the DHA information was obtained. The DHA data and the peak information were then used as input for SPT algorithm. Figure 8 shows comparison between the composition distribution and the ideal gamma distribution after different steps of the SPT algorithm. In this case, the molecules would be categorized into eight molecular types (namely, NP, MP, DP, TP, NO, BO, NC, and A). The eight molecular types corresponded to the subgraphs (a) to (h) in Figure 8. Each subgraph included three additional subgraphs that represented the results of the three steps in SPT algorithm. The carbon number was served as the X axis, while the mass fraction was served as the Y axis for all these subgraphs.

Figure 8 (c) can serve as an example to elaborate on the results of DP at the end of the three steps of SPT algorithm. For convenience, the three subgraphs of Figure 8 (c) were marked as (c)-1, (c)-2 and (c)-3 from the top to the bottom. The red dots in these three subgraphs were the mass fractions of DP in the data matrix at the end of each step of SPT algorithm. The blue dashed lines are the distributions calculated using $\text{Parameters}_{1,\text{DP}}$, $\text{Parameters}_{2,\text{DP}}$, and $\text{Parameters}_{3,\text{DP}}$, respectively. It can be seen in (c)-1 that there were large deviations between some red dots and the blue dashed line. Especially the red dot at carbon number value of seven was significantly lower than the blue dashed line. This is because some of the DP molecules existed in the form of coeluting peaks or unidentified peaks. The mass fractions of these peaks were not counted into the data matrix at the end of step one. After the step 2, namely, all the coeluting peaks involved DP were split，it can be seen from (c)-2 that the distribution of the red dots was very close to the blue dashed line.

In Figure 8 (c)-3, the step 3 did not show a significant improvement with respect to step 2. One possible reason for this is that the content of unidentified peaks was too low to make an obvious change on step 3. The other was that step three preferred to classify unidentified peak into the molecular type whose composition distribution deviated farther from the ideal gamma distribution. This scenario could be found in Figure 8 (d). Even if the coeluting peak had been split in step 2, the composition distribution of TP was still far from the ideal gamma distribution. Therefore, the SPT algorithm would prefer to categorize the unidentified peaks into TP rather than DP. As a result, improvement of the composition distribution at the end of step three could be found in (d)-3, but not in (c)-3.

As shown in Figure 8 (h), the mass fraction of aromatics was zero at a carbon number value of seven. This is because toluene was always eluted with 2,3,3-trimethylpentane in the GC-FID method used. After processing by the SPT algorithm, a relatively reasonable mass fraction of toluene was obtained. It needs to be pointed out that all the 11 normal paraffins could be identified by GC-FID in most situations. Therefore, all the subgraphs in Figure 8 (a) look the same.

In summary, the molecular composition distribution of gasoline was enriched after processing by the SPT algorithm. In the present work, only hydrocarbons were included into the molecular library. In future work, molecules containing heteroatoms, such as sulfur, nitrogen, and oxygen, will be added, which allows a more informative gasoline composition to be reconstructed.

3.2 Property Prediction Validation for Different Gasolines

With the thus-obtained molecular composition, the bulk properties can be predicted. Since SPT is a self-tuning algorithm and no additional measurement was required, it can be seen as a predictive bulk property prediction process. This is quite different from most current molecular reconstruction methods, in which the bulk property measurement is required and served as model constraints.

The prediction of a gasoline sample is listed in Table 2. The developed model has a capacity of predicting almost all the important properties, including density, distillation profile, octane number, vapor pressure, etc. Although it is difficult to predict the ASTM D86 distillation profile, especially the initial boiling point and the final boiling point, good predictions were given by CGC model. The reconstructed gasoline composition was quite flexible that could be applied to different property prediction rules. The CUP CNN model showed a good performance on octane number prediction. The experimental volume fractions of olefins, naphthenes and aromatics were measured by the fluorescence method. The predicted volume fractions were calculated based on the mass fraction data from GC-FID. In fact, the accuracy and reproducibility of the fluorescence method should be lower than GC-FID. Consequently, the predictions came from CGC model were considered more reliable.

Eight gasoline samples were selected to validate the predictions of CGC model, including straight-run gasoline, FCC gasoline, and reformates. Figure 9 shows the comparisons between the experimental values and predictions of different properties of the eight samples. The predictions show good accordance with experimental values. Some data points of the ASTM D86 distillation profile show greater deviations than other properties. These deviations may be related to the error of the ASTM D86 distillation experiment or the transformation between real boiling point and ASTM D86 distillation curve.

3.3 Octane Number Prediction for Large-scale Gasoline Samples

Given the importance of the octane number, 216 gasoline samples were selected to validate the accuracy of the octane number predictions by the CUP CNN model. Figure 10 shows the comparison between the predictions and experimental values of octane number. The average deviation of the predictions from the CUP CNN model was about 1. The data points distributed uniformly, and the absolute deviation of most data points was smaller than 2. Hence, there will be a good performance on predicting octane number by using the reconstructed composition.

It is expected that molecular-level gasoline assessment and blending models can be established based on CGC model in the future. CGC model can be potentially used as an off-line molecular-level gasoline blending engine or rapid gasoline sample evaluation method in laboratories equipped with GC-FID.

4. CONCLUSION

A computer-aided gasoline composition model based on GC-FID analysis was developed. Based on the assumption that the content data of each molecular type follows a gamma distribution, the preliminary composition from the GC-FID based DHA results was reconstructed by the SPT algorithm in CGC model. The coeluting peaks were split, and then possible molecules were inferred from the unidentified peaks. The reconstructed composition could be used for a variety of molecular-based property prediction rules. The predictions show good accordance with the experimental data for different types of gasoline. With respect to octane number, the accuracy of predictions by using the reconstructed composition was verified by two different models and 216 samples. Both the two models were proven to have good performance. This work has laid the foundation for the development of gasoline assessment and blending models at the molecular level.

AUTHOR INFORMATION

Corresponding Authors

lzz@cup.edu.cn (L. Zhang);

mtk@udel.edu (M. T. Klein);

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 21506254) and Science Foundation of China University of Petroleum, Beijing (No. 2462014YJRC020, 2462018BJC003). Michael T. Klein acknowledges collaborations with and support of colleagues via the Saudi Aramco Chair Program at KFUMP and Saudi Aramco.

Reference

1. ASTM D6729-14, Standard Test Method for Determination of Individual Components in Spark Ignition Engine Fuels by 100 Metre Capillary High Resolution Gas Chromatography. In ASTM International: 2014.

2. ASTM D6730-01(2016), Standard Test Method for Determination of Individual Components in Spark Ignition Engine Fuels by 100-Metre Capillary (with Precolumn) High-Resolution Gas Chromatography. In ASTM International: 2016.

3. ASTM D6733-01(2016), Standard Test Method for Determination of Individual Components in Spark Ignition Engine Fuels by 50-Metre Capillary High Resolution Gas Chromatography. In ASTM International: 2016.

4. Walsh, P.; Garbalena, M.; Schug, K. A., Rapid analysis and time interval deconvolution for comprehensive fuel compound group classification and speciation using gas chromatography–vacuum ultraviolet spectroscopy. *Analytical chemistry* **2016,** *88*, (22), 11130-11138.

5. Weber, B. M.; Walsh, P.; Harynuk, J. J., Determination of hydrocarbon group-type of diesel fuels by gas chromatography with vacuum ultraviolet detection. *Analytical chemistry* **2016,** *88*, (11), 5809-5817.

6. Bai, L.; Smuts, J.; Schenk, J.; Cochran, J.; Schug, K. A., Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC–MS for the characterization of weathered and unweathered diesel fuels. *Fuel* **2018,** *214*, 521-527.

7. Gröger, T.; Gruber, B.; Harrison, D.; Saraji-Bozorgzad, M.; Mthembu, M.; Sutherland, A. e.; Zimmermann, R., A vacuum ultraviolet absorption array spectrometer as a selective detector for comprehensive two-dimensional gas chromatography: concept and first results. *Analytical chemistry* **2016,** *88*, (6), 3031-3039.

8. Egeness, M. J. Comprehensive Two-Dimensional Gas Chromatography: method development and verification by characterisation of petroleum fractions. Institutt for kjemi, 2012.

9. Toussaint, G.; Lorentz, C.; Vrinat, M.; Geantet, C., Comprehensive 2D chromatography with mass spectrometry: a powerful tool for following the hydrotreatment of a Straight Run Gas Oil. *Analytical Methods* **2011,** *3*, (12), 2743-2748.

10. Quann, R.; Jaffe, S., Building useful models of complex reaction systems in petroleum refining. *Chemical Engineering Science* **1996,** *51*, (10), 16151633-16311635.

11. Quann, R. J., Modeling the chemistry of complex petroleum mixtures. *Environmental Health Perspectives* **1998,** *106*, (Suppl 6), 1441.

12. Quann, R. J.; Jaffe, S. B., Structure-oriented lumping: describing the chemistry of complex hydrocarbon mixtures. *Industrial & engineering chemistry research* **1992,** *31*, (11), 2483-2497.

13. Ghosh, P.; Andrews, A. T.; Quann, R. J.; Halbert, T. R., Detailed kinetic model for the hydro-desulfurization of FCC Naphtha. *Energy & Fuels* **2009,** *23*, (12), 5743-5759.

14. Neurock, M.; Nigam, A.; Trauth, D.; Klein, M. T., Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms. *Chemical engineering science* **1994,** *49*, (24), 4153-4177.

15. Hudebine, D.; Verstraete, J. J., Molecular reconstruction of LCO gasoils from overall petroleum analyses. *Chemical Engineering Science* **2004,** *59*, (22-23), 4755-4763.

16. de Oliveira, L. P.; Verstraete, J. J.; Kolb, M., Simulating vacuum residue hydroconversion by means of Monte-Carlo techniques. *Catalysis Today* **2014,** *220*, 208-220.

17. Zhang, L.; Hou, Z.; Horton, S. R.; Klein, M. T.; Shi, Q.; Zhao, S.; Xu, C., Molecular representation of petroleum vacuum resid. *Energy & Fuels* **2014,** *28*, (3), 1736-1749.

18. Pan, Y.; Yang, B.; Zhou, X., Feedstock molecular reconstruction for secondary reactions of fluid catalytic cracking gasoline by maximum information entropy method. *Chemical Engineering Journal* **2015,** *281*, 945-952.

19. Peng, B. Molecular modelling of petroleum process. Ph.D. Thesis, UMIST, 1999.

20. Aye, M. M. S.; Zhang, N., A novel methodology in transforming bulk properties of refining streams into molecular information. *Chemical engineering science* **2005,** *60*, (23), 6702-6717.

21. Hu, S.; Towler, G.; Zhu, X., Combine molecular modeling with optimization to stretch refinery operation. *Industrial & engineering chemistry research* **2002,** *41*, (4), 825-841.

22. Wu, Y.; Zhang, N., Molecular characterization of gasoline and diesel streams. *Industrial & Engineering Chemistry Research* **2010,** *49*, (24), 12773-12782.

23. Zhang, Y. A molecular approach for characterisation and property predictions of petroleum mixtures with applications to refinery modelling. . Ph.D. Thesis, UMIST, 1999.

24. Ahmad, M. I.; Zhang, N.; Jobson, M., Molecular components-based representation of petroleum fractions. *Chemical Engineering Research and Design* **2011,** *89*, (4), 410-420.

25. Gomez-Prado, J.; Zhang, N.; Theodoropoulos, C., Characterisation of heavy petroleum fractions using modified molecular-type homologous series (MTHS) representation. *Energy* **2008,** *33*, (6), 974-987.

26. Albahri, T. A., Molecularly explicit characterization model (MECM) for light petroleum fractions. *Industrial & engineering chemistry research* **2005,** *44*, (24), 9286-9298.

27. Hudebine, D.; Verstraete, J. J., Reconstruction of petroleum feedstocks by entropy maximization. Application to FCC gasolines. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles* **2011,** *66*, (3), 437-460.

28. Cui, C.; Billa, T.; Zhang, L.; Shi, Q.; Zhao, S.; Klein, M. T.; Xu, C., Molecular Representation of Petroleum Gasoline Fraction. *Energy & Fuels* **2018**.

29. Conjeevaram Krishnakumar, N. K. A Bayesian approach to feed reconstruction. Massachusetts Institute of Technology, 2013.

30. Mei, H.; Wang, Z.; Huang, B., Molecular-based Bayesian regression model of petroleum fractions. *Industrial & Engineering Chemistry Research* **2017,** *56*, (50), 14865-14872.

31. Pyl, S. P.; Van Geem, K. M.; Reyniers, M. F.; Marin, G. B., Molecular reconstruction of complex hydrocarbon mixtures: An application of principal component analysis. *AIChE journal* **2010,** *56*, (12), 3174-3188.

32. Ghosh, P.; Hickey, K. J.; Jaffe, S. B., Development of a detailed gasoline composition-based octane model. *Industrial & engineering chemistry research* **2006,** *45*, (1), 337-345.

33. Klein, M. T.; Hou, G.; Bertolacini, R.; Broadbelt, L. J.; Kumar, A., *Molecular modeling in heavy hydrocarbon conversions*. CRC Press: 2005.

34. Brown, J. M.; Sundaram, A.; Saeger, R. B.; Wellons, H. S.; Kennedy, C. R.; Jaffe, S. B. Estimating detailed compositional information from limited analytical data. US7598487 B2, 2014.

35. Qian, K.; Olmstead, W. N.; English, J. B.; Green, L. A.; Saeger, R. B.; Jaffe, S. B. Micro-hydrocarbon analysis. US2009/0105966 A1, 2009.

36. Riazi, M., *Characterization and properties of petroleum fractions*. ASTM international: 2005; Vol. 50.

37. Department, A. P. I. R.; Daubert, T. E., *Technical Data Book-Petroleum Refining*. The Department: 1976.

Table 1. Properties prediction rules used in the CGC model

Property	Prediction method
Molecular weight	CPPF Eq 3.45
Specific gravity	CPPF Eq 5.126
Research octane number	CUP CNN Model
Motor octane number	CUP CNN Model
Reid vapor pressure	CPPF Eq 3.103
Refractive index	CPPF Eq 3.45
Critical pressure	API 4B2.1-3
Critical temperature	API 4B2.1-2
Critical volume	API 4B1.1-3
Watson K	API 2-0.9
Acentric factor	CPPF Eq 5.115
ASTM D86	API 3A3.2
Critical compressibility factor	CPPF Eq 6.74
Surface tension	CPPF Eq 6.74
Aniline point	API 2-1.8/2-1.9
Cloud point	CPPF Eq 3.117/3.121
CH weight ratio	Calculated
Heat of vaporization	CPPF Eq 5.126

CPPF: Characterization and Properties of Petroleum Fractions;[36] API: API Technical Data Book;[37] CUP CNN Model: a CNN based octane number model

Table 2. Comparison between the prediction and experimental values of properties for a gasoline sample.

Property	Experiment	Prediction	Property	Experiment	Prediction
Specific gravity	0.7320 ± 0.0012	0.7402	Critical temperature (ºC)	-	268.8
ASTM D86 IBP (ºC)	30.0 ± 7.2	28.5	Critical pressure (kPa)	-	3338.1
ASTM D86 5% (ºC)	46.0 ± 8.9	45.9	Critical volume (m^3/kmol)	-	0.3754
ASTM D86 10% (ºC)	52.7 ± 6.3	48.3	Critical compressibility factor	-	0.2696
ASTM D86 30% (ºC)	75.8 ± 5.3	70.1	Acentric factor	-	0.2926
ASTM D86 50% (ºC)	107.1 ± 5.7	103.7	Refractive index	-	1.4148
ASTM D86 70% (ºC)	138.8 ± 5.7	127.9	Watson K	-	11.77
ASTM D86 90% (ºC)	168.7 ± 3.7	156.8	Surface tension (dyne/cm)	-	22.04
ASTM D86 95% (ºC)	179.6 ± 7.0	170.6	Aniline point (ºC)	-	55.9
ASTM D86 FBP (ºC)	193.3 ± 8.9	178.1	Cloud point (ºC)	-	-77.1
RON	91.1± 0.7	91.9	MON	-	79.3
Olefins (volume%）	29.7 ± 7.8	32.1	CH weight ratio	-	6.60
Naphthenes (volume%）	11.5 ± 4.0	10.8	Heat of vaporization (kJ/kg)	-	325.4
Aromatics (volume%)	27.0 ± 3.0	23.1	Molecular weight (g/mol)	-	94.4
RVP (kPa)	59.75 ± 5.5	62.5			

Figure Captions:

Figure 1. Peak types shown in a typical GC-FID chromatogram

Figure 2. Flowchart of the CGC model

Figure 3. Statistics of molecules in the molecular library: (a) number of molecules in each molecular category; (b) number of molecules as a function of carbon number for each molecular type.

Figure 4. Flowchart of the first step of the SPT algorithm.

Figure 5. Flowchart of the second step of the SPT algorithm.

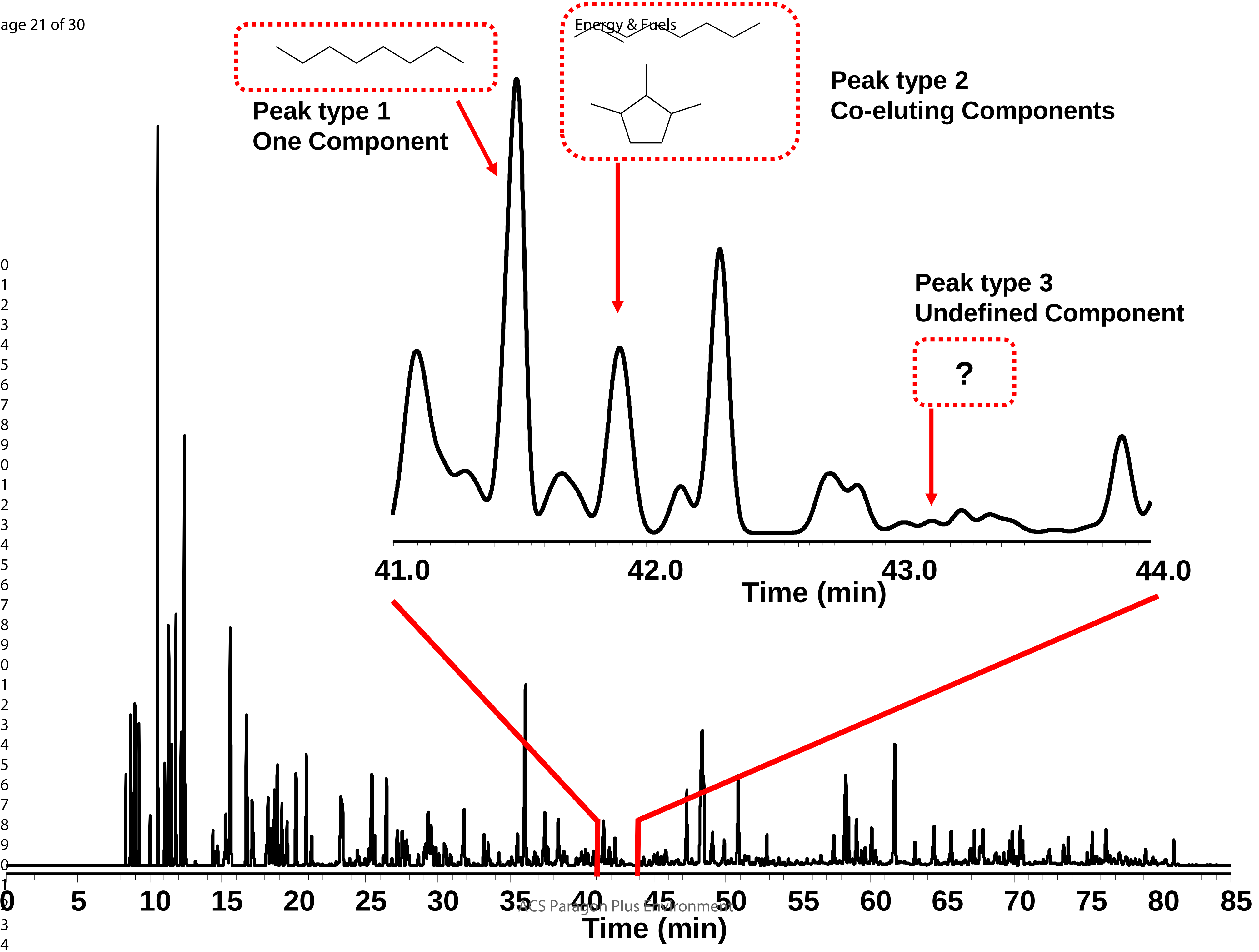
Figure 6. An example of assumption list.

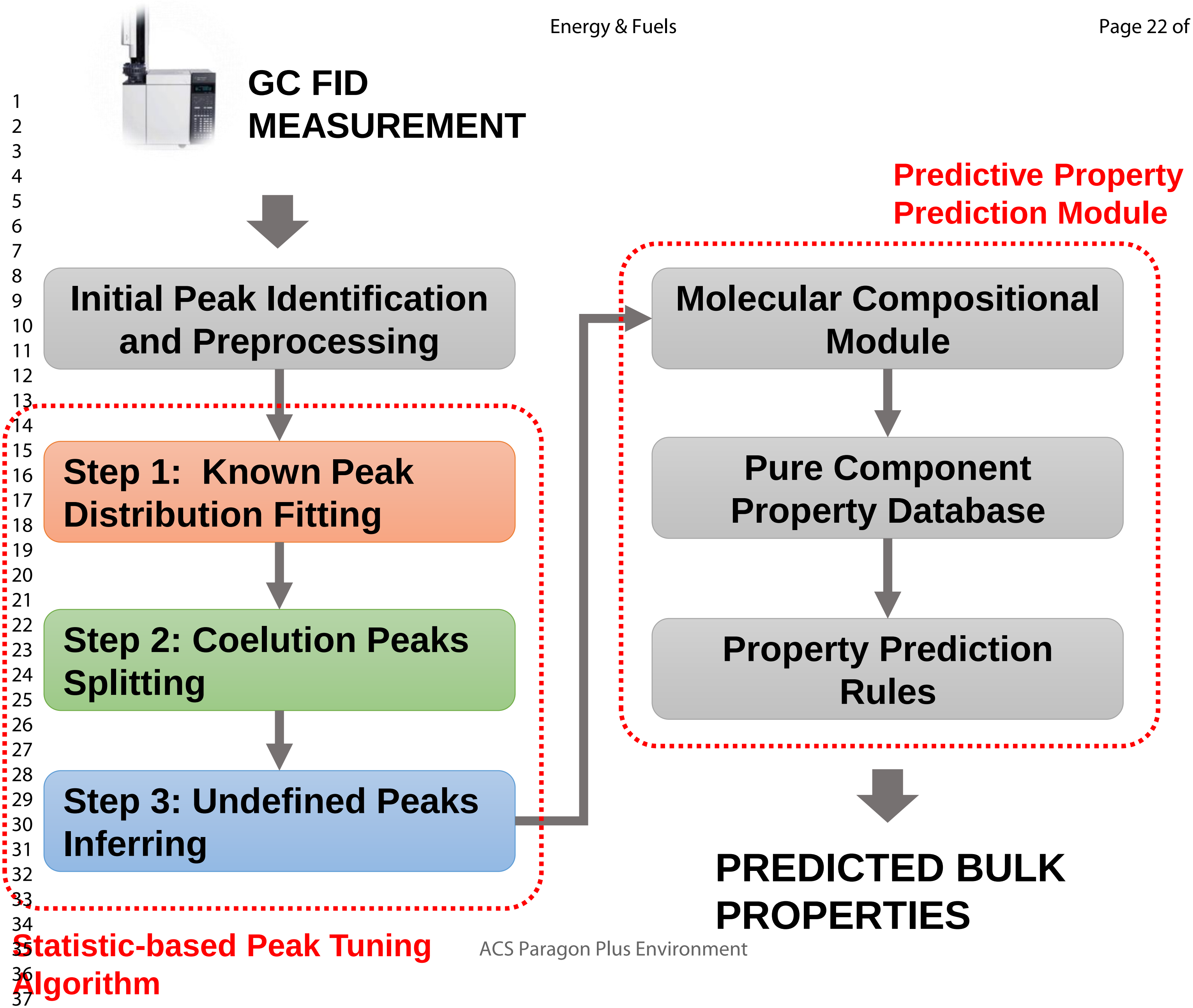
Figure 7. Flowchart of the third step of the SPT algorithm.

Figure 8. Comparison between the composition distribution and the regressed gamma distribution at the end of the three steps of the SPT algorithm for each molecular type: (a)normal paraffins; (b) mono-branched paraffins; (c) double-branched paraffins; (d) triple-branched paraffins; (e) normal olefins; (f) branched olefins; (g) naphthenes; (h) aromatics.

Figure 9. Comparison between the predicted and experimental values of various properties for eight gasoline samples.

Figure 10. Comparison between the predicted and experimental values of research octane numbers for 216 gasoline samples.


Peak type 1
One Component
Peak type 2
Co-eluting Components
Peak type 3
Undefined Component
?
41.0
42.0
43.0
44.0
Time (min)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
Time (min)



GC FID MEASUREMENT

Initial Peak Identification and Preprocessing

Step 1: Known Peak Distribution Fitting

Step 2: Coelution Peaks Splitting

Step 3: Undefined Peaks Inferring

Statistic-based Peak Tuning Algorithm

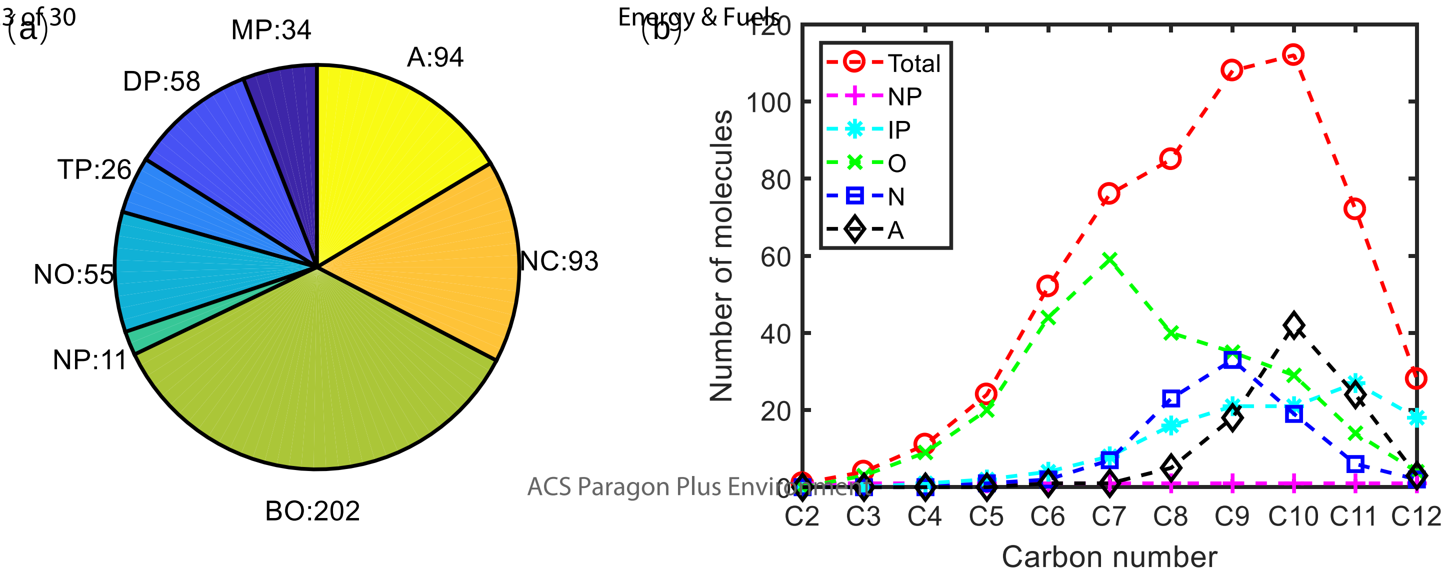
Predictive Property Prediction Module

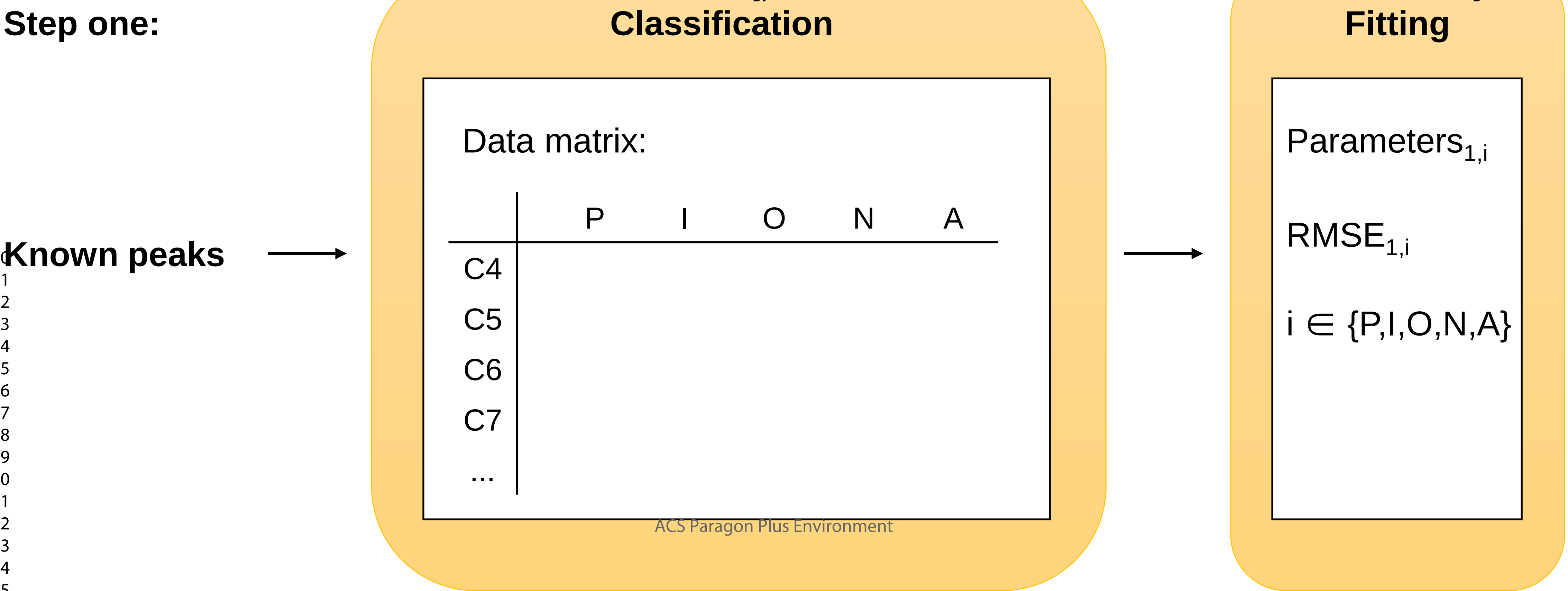
Molecular Compositional Module

Pure Component Property Database

Property Prediction Rules

PREDICTED BULK PROPERTIES


(a)
MP:34
A:94
DP:58
TP:26
NC:93
NO:55
NP:11
BO:202
(b)
Total
NP
IP
O
N
A
Number of molecules
0
20
40
60
80
100
120
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
Carbon number



Step one:

Known peaks

Classification

Data matrix:

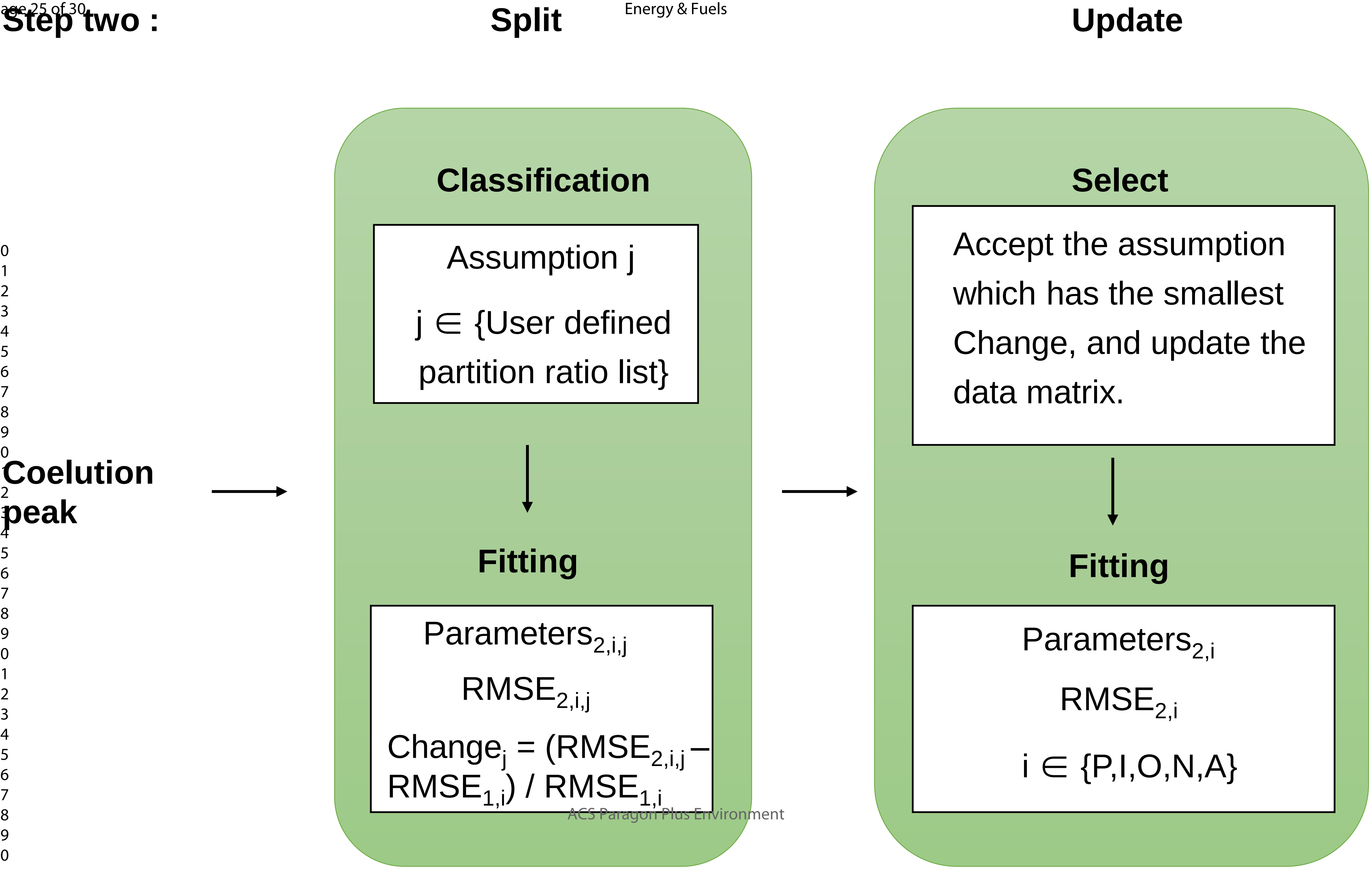
	P	I	O	N	A
C4					
C5					
C6					
C7					
...					

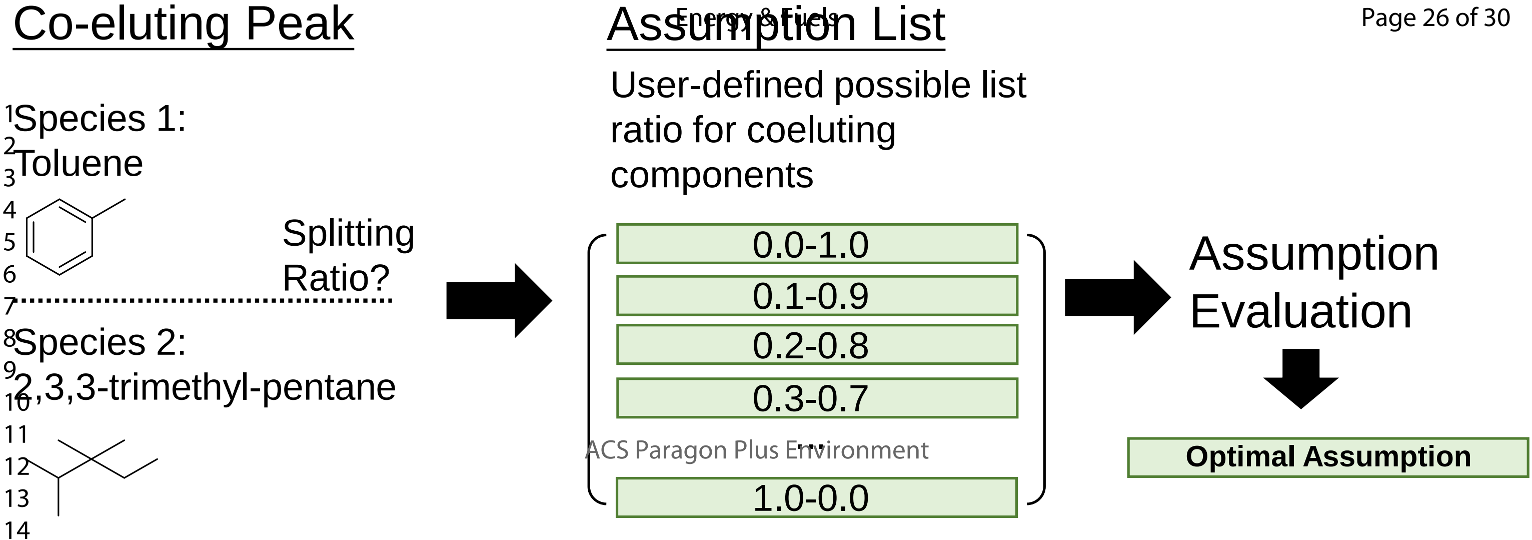
Fitting

Parameters$_{1,i}$

RMSE$_{1,i}$

$i \in \{P,I,O,N,A\}$


Step two :
Split
Update
Coelution peak
Classification
Assumption j
j ∈ {User defined partition ratio list}
Fitting
Parameters$_{2,i,j}$
RMSE$_{2,i,j}$
Change$_j$ = (RMSE$_{2,i,j}$ – RMSE$_{1,i}$) / RMSE$_{1,i}$
Select
Accept the assumption which has the smallest Change, and update the data matrix.
Fitting
Parameters$_{2,i}$
RMSE$_{2,i}$
i ∈ {P,I,O,N,A}



Co-eluting Peak

Species 1:
Toluene

Splitting
Ratio?

Species 2:
2,3,3-trimethyl-pentane

Assumption List

User-defined possible list ratio for coeluting components

0.0-1.0

0.1-0.9

0.2-0.8

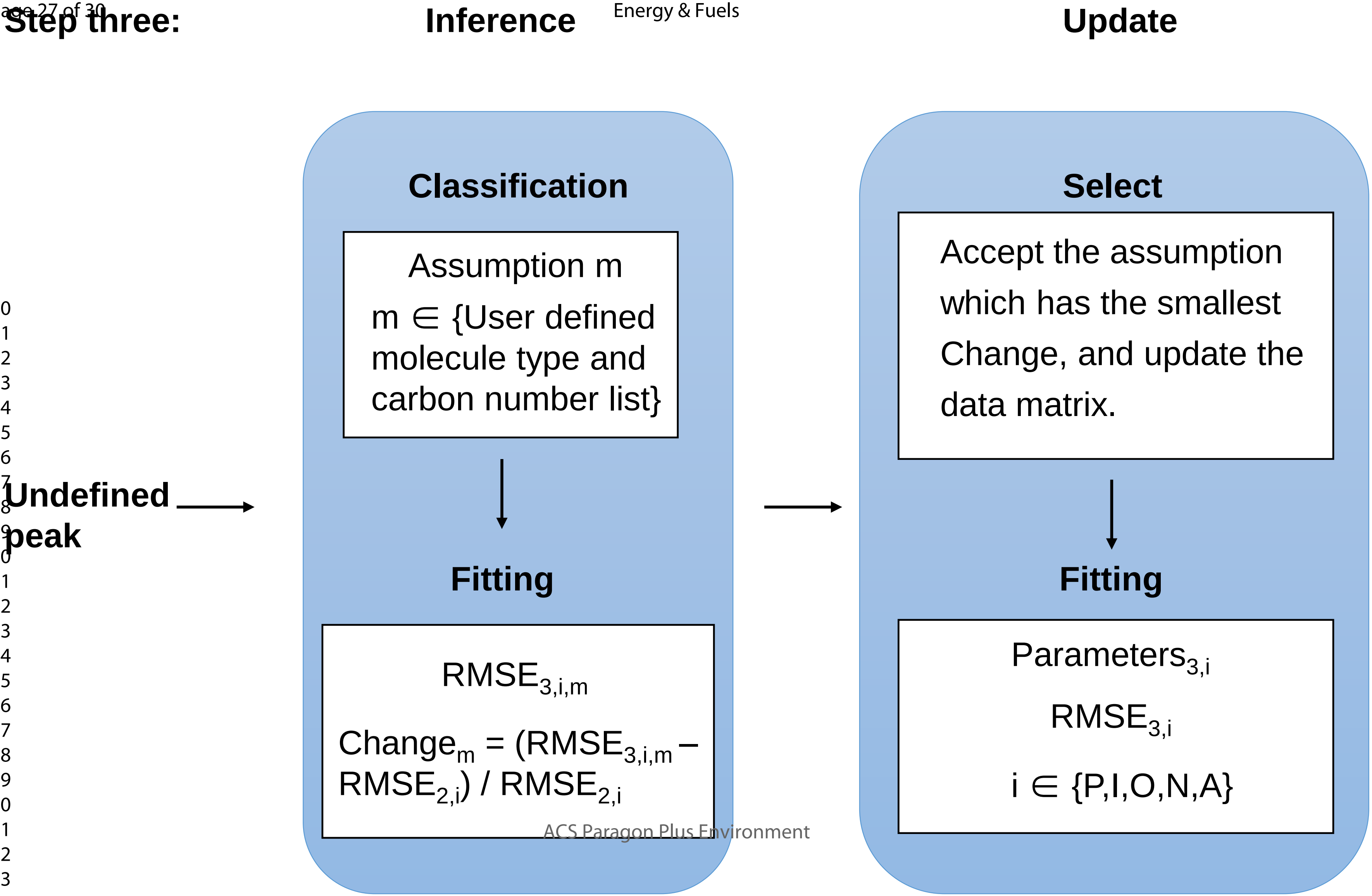
0.3-0.7

…

1.0-0.0

Assumption
Evaluation

Optimal Assumption



Step three:

Inference

Update

Undefined peak

Classification

Assumption m

m ∈ {User defined molecule type and carbon number list}

Fitting

$\mathrm{RMSE}_{3,i,m}$

$\mathrm{Change}_m = (\mathrm{RMSE}_{3,i,m} - \mathrm{RMSE}_{2,i}) / \mathrm{RMSE}_{2,i}$

Select

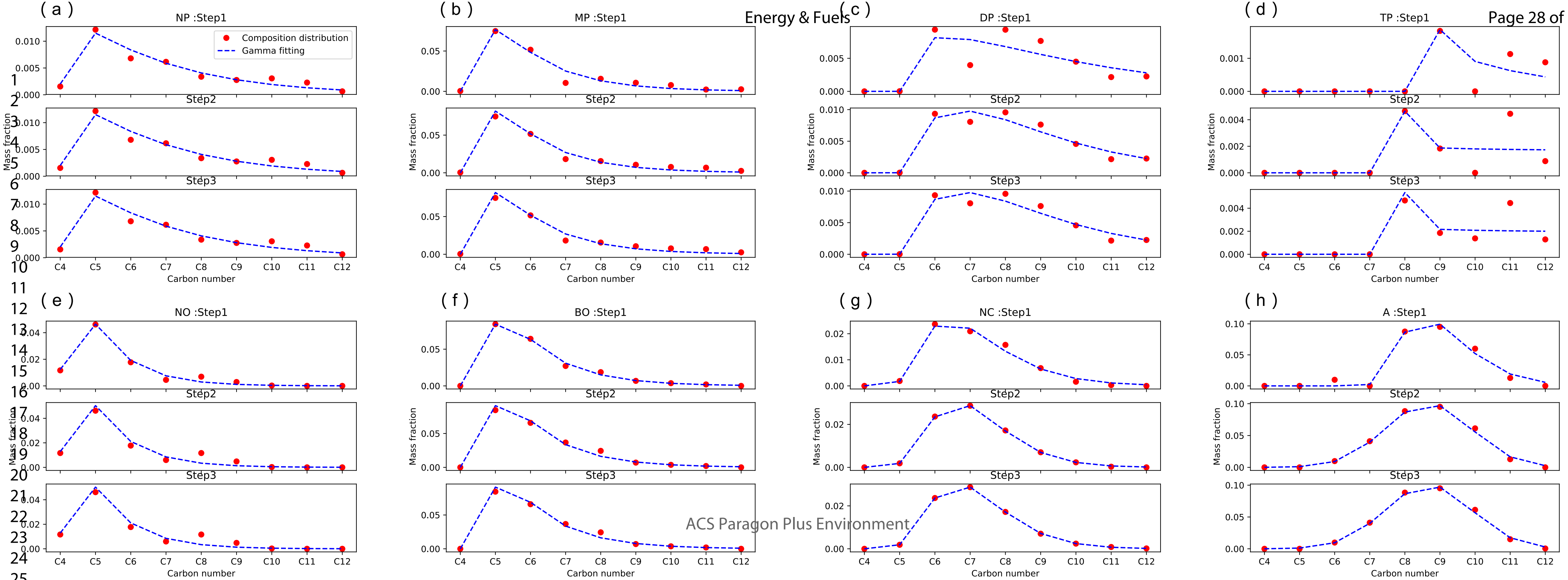
Accept the assumption which has the smallest Change, and update the data matrix.

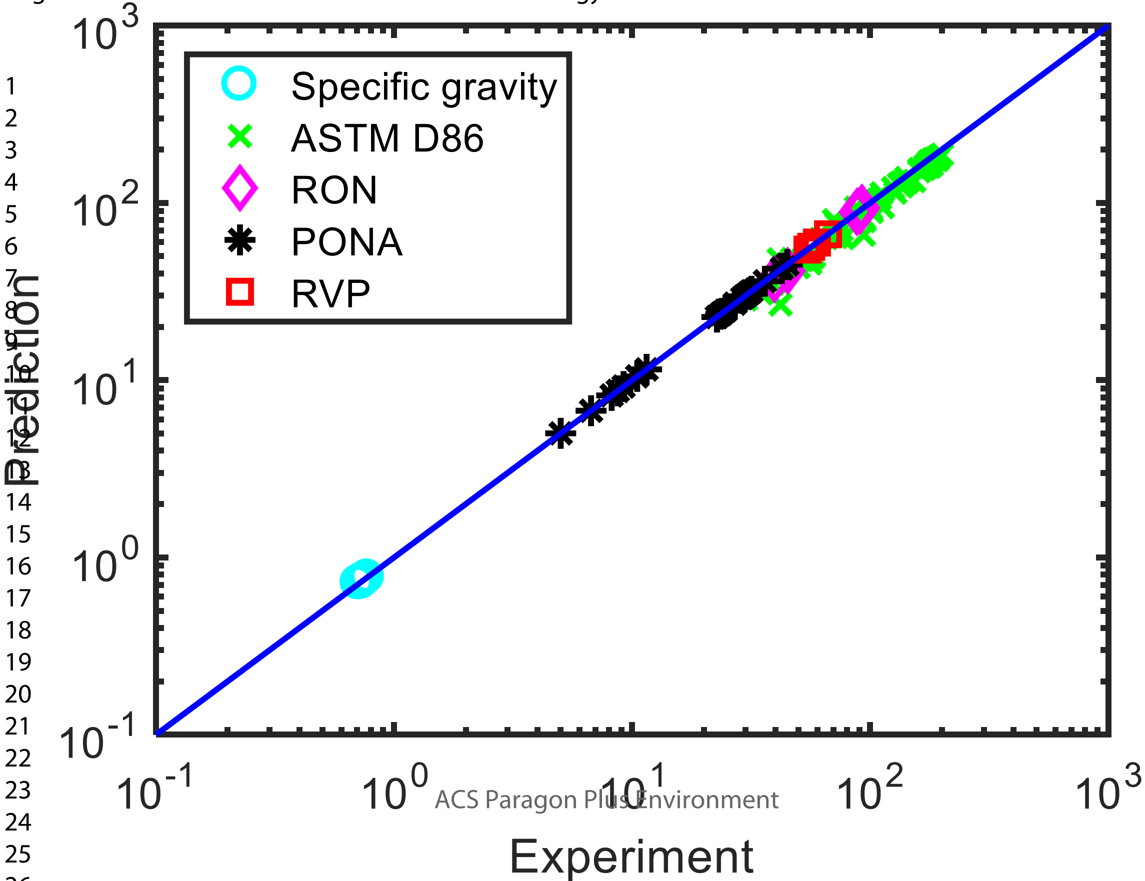
Fitting

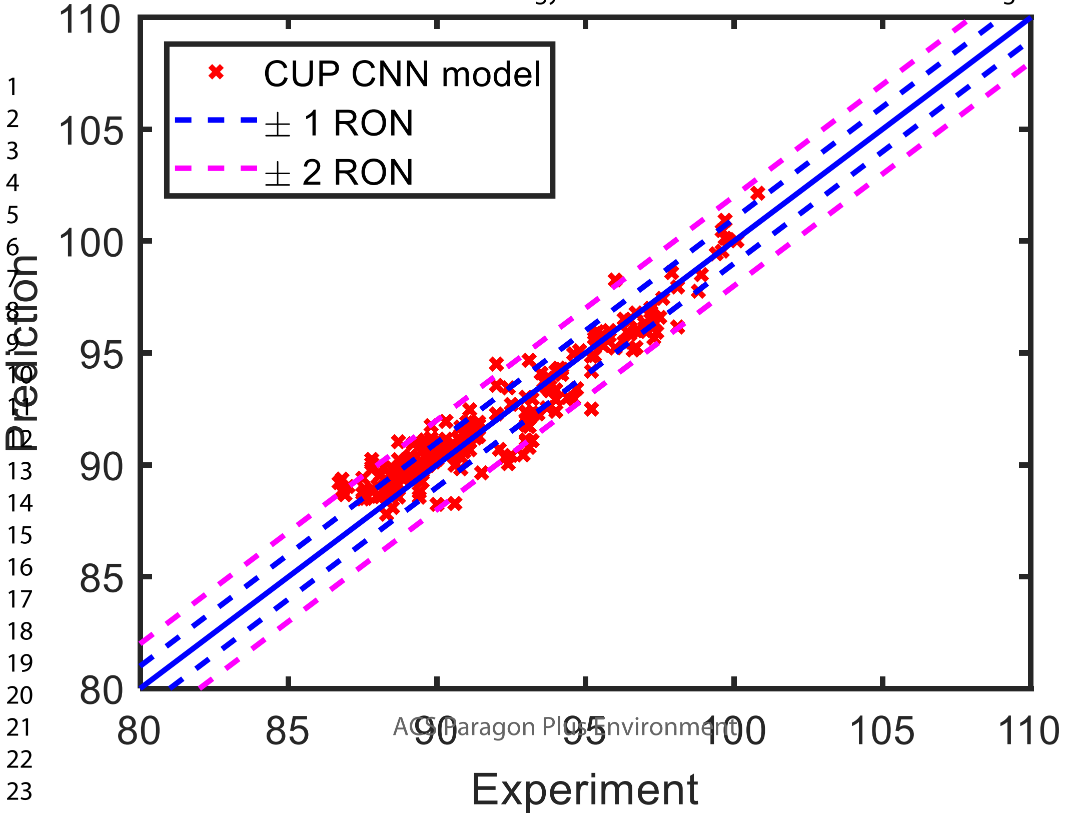
$\mathrm{Parameters}_{3,i}$

$\mathrm{RMSE}_{3,i}$

$i \in \{P,I,O,N,A\}$


(a)
NP :Step1
Composition distribution
Gamma fitting
Step2
Step3
(b)
MP :Step1
Step2
Step3
(c)
DP :Step1
Step2
Step3
(d)
TP :Step1
Step2
Step3
(e)
NO :Step1
Step2
Step3
(f)
BO :Step1
Step2
Step3
(g)
NC :Step1
Step2
Step3
(h)
A :Step1
Step2
Step3
Mass fraction
Carbon number
C4
C5
C6
C7
C8
C9
C10
C11
C12
0.000
0.005
0.010
0.00
0.01
0.02
0.04
0.05
0.10
0.001
0.002
0.004


Specific gravity
ASTM D86
RON
PONA
RVP
Prediction
Experiment
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}


CUP CNN model
± 1 RON
± 2 RON
Prediction
Experiment
80
85
90
95
100
105
110
80
85
90
95
100
105
110