CHEMICAL ENGINEERING TRANSACTIONS
VOL. 120, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Bing Shen How, Viknesh Andiappan, Denny K.S. Ng, Hon Loong Lam, Petar S. Varbanov

ISBN 979-12-81206-21-2; **ISSN** 2283-9216
DOI: 10.3303/CET25120067

Improved Molecular Composition Reconstruction Framework for Characterising Petroleum and Bio-oil Fractions

Jianfeng Jiao[a], Xi Gao[b], Jie Li[a,*]

[a]Centre for Process Integration, Department of Chemical Engineering, School of Engineering, The University of Manchester, Manchester M13 9PL, UK
[b]School of Mechanical and Electrical Engineering, Jinggangshan University, Ji'an, Jiangxi, China 343009
jie.li-2@manchester.ac.uk

This study introduces a molecular composition reconstruction framework specifically designed to characterize complex petroleum and bio-oil fractions at the molecular level. By identifying representative core molecular structures and systematically generating homologous series (HS) with varying sidechains, the proposed framework constructs a comprehensive molecular library. Leveraging connectivity matrix-based molecular descriptors combined with artificial neural networks, accurate predictions of molecular properties are achieved. To effectively reduce the high dimensionality, the Gamma probability density function is employed to statistically represent HS distributions, significantly reducing computational complexity. An optimization strategy integrating genetic algorithms and sequential quadratic programming refines the molecular content distributions, ensuring robustness and flexibility across diverse feedstocks. Case studies on petroleum diesel and bio-oil fractions demonstrate superior prediction accuracy of the proposed framework compared to previous methods. For the calculation of diesel properties, the total sum of absolute relative errors (AREs) has decreased from 2.41 % to 1.95 %, for the bio-oil sample, the total ARE has decreased from 17.6 % to 16.01 %. Sensitivity analyses further highlight that the overall HS content significantly outweighs the internal distribution parameters in affecting property predictions, guiding future efforts in molecular-level characterization and process optimization.

1. Introduction

Modern refineries increasingly rely on precise molecular management strategies to enhance efficiency, sustainability, and feedstock flexibility. Traditionally, refining processes have commonly employed lumped methods to simplify simulations by grouping components into a limited number of categories. For instance, feedstock can be represented using wide-boiling-range pseudo-components or classified by chemical types, such as PONA (paraffins, olefins, naphthenes, aromatics) and SARA (saturates, aromatics, resins, asphaltenes). The general idea is that the virtual properties of a sample are determined by summing the properties of individual molecular lumps. Accordingly, mixtures are grouped into a few lumps based on their reaction characteristics. However, traditional lumping methods simplify mixtures into broad pseudo-components, losing detailed molecular-level information. Advanced statistical reconstruction techniques, such as stochastic reconstruction (SR) (Neurock et al., 1990), reconstruction by entropy maximization (REM) (Hudebine et al., 2004), structural-oriented lumping (SOL) (Tong et al., 2016), molecular-type homologous series (MTHS) (Ren et al., 2019), and structural unit and bond electron matrix (SU-BEM) (Feng et al., 2019), have emerged to overcome these limitations by providing more accurate molecular representations.
However, existing petroleum-oriented methods are unsuitable for bio-oil fractions, and integrating property estimation approaches like GC and QSPR into molecular reconstruction is challenging. For example, BEM's low-level matrix representation lacks direct alignment with GC/QSPR's functional groups or SMILES fragments, necessitating extensive and complex manual rule definitions. Furthermore, traditional GC methods neglect proximity and nonlinear effects among groups, and machine learning approaches, although improved, still struggle with unrecognized functional groups.

Paper Received: 16 May 2025; Revised: 29 August 2025; Accepted: 2 October 2025
Please cite this article as: Jiao J., Gao X., Li J., 2025, Improved Molecular Composition Reconstruction Framework for Characterising Petroleum and Bio-oil Fractions, Chemical Engineering Transactions, 120, 397-402 DOI:10.3303/CET25120067

For the molecular reconstruction of complex hydrocarbon mixtures, each homologous series (HS) typically includes dozens or even hundreds of consecutive carbon-numbered molecules. If each molecule is assigned an individual concentration, the number of variables would increase drastically—potentially exceeding 10^3—leading to the "curse of dimensionality.", However, available experimental constraints are very limited. Therefore, the molecular reconstruction of mixtures represents a typical underdetermined optimization problem characterized by excessive degrees of freedom, insufficient observations, and complex objectives.

To address these issues, this study proposes a molecular composition reconstruction framework using statistical assumptions and a hybrid heuristic–gradient optimization strategy. Initially, a Gamma PDF condenses homologous series (HS) compositions into statistical parameters, reducing problem dimensionality and improving efficiency. A two-stage optimization employing Genetic Algorithm GA and Sequential Quadratic Programming (SQP) is implemented—GA rapidly explores global regions but struggles near precise solutions, while SQP provides accurate local refinement. The improved ANN-CM model (Jiao et al., 2025) is integrated to predict molecular properties accurately. Compared to GC-based methods, this framework reduces test-set RMSE by up to 83.8 %, outperforming conventional models.

2. Improved molecular composition reconstruction framework

The proposed reconstruction framework consists of three steps: (1) Preparation—identifying core structures from experimental data and generating homologous series (HS) to build a qualitative molecular library. (2) Transformation—predicting molecular properties using ANN, assuming initial gamma PDF parameters and core-structure mole fractions. (3) Calculation—defining an objective function to minimize discrepancies between predicted and experimental bulk properties using mixing rules.

2.1 Preparation step: qualification

Using experimentally obtained molecular data, common HS structures—including aromatic/naphthenic rings and heteroatom-substituted units—are identified. Sidechains of varying lengths attach to core structures based on bond positions and steric effects, though only one sidechain per core is considered due to limited NMR data. Longer sidechains correlate with higher boiling points, enabling determination of sidechain length from boiling point ranges. This predefined core-sidechain assembly is termed qualification.

2.2 Transformation step: quantification

Despite using a reduced subset of molecules as a representative model, quantifying the full molecular composition remains a challenge due to the high degree of freedom. To solve this challenge, statistical methods such as the Gamma PDF can be used to reduce the number of variables required to determine the contents of a homologue series. This PDF is widely used to characterize petroleum fraction molecules and other organic molecules as well (Feng et al., 2019). The functional form of this PDF is shown in Eqs(1)-(2)

$$f(x) = \frac{(x-\gamma)^{\alpha-1} e^{-\frac{x-\gamma}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \tag{1}$$

$$\Gamma(\alpha) = \int_0^{\infty} t^{z-1} e^{-t} dt \tag{2}$$

where α, β and γ are shape、scale and location factor; α controls the shape of the distribution, β determines the width, γ defines the starting point of the distribution on the left; $\Gamma(\alpha)$ is the Gamma function; *t* is the dummy variable; e^{-t} is the exponential decay term.

2.3 Calculation of the bulk properties of the mixture

The objective function minimizes differences between estimated and experimental properties using equally weighted factors (coefficients = 1), as shown in Eq(3). Virtual hydrocarbons mimicking real substances are generated, optimized via GA, followed by sensitivity-based uncertainty allocation. The SQP method further refines GA solutions, enhancing accuracy and computational efficiency.

$$Obj = \sum_k abs\left(\frac{C_k^{msd} - C_k^{pred}}{C_k^{msd}}\right) \tag{3}$$

where superscripts *msd* and *pred* denote the measured and predicted values respectively. The subscripts correspond to the different measurable properties，including distillation temperature points，elemental content and specific gravity (SG).

2.3.1 Distillation temperature points

Boiling points are predicted by ANN to create a temperature sequence. Volume fraction key points, ranging from 0 to 1 in increments of 0.1, are determined following ASTM D86, thus forming the distillation profile. For volume fraction composition $x_{i,j}^{v}$ is calculated as Eqs(4)-(6).

$$x_{i,j}^{v} = \frac{\frac{x_{i,j}^{w}}{SG_{i,j}}}{\sum_i \sum_j \frac{x_{i,j}^{w}}{SG_{i,j}}} \tag{4}$$

$$x_{i,j}^{w} = \frac{x_{i,j}^{m} \cdot MW_{i,j}}{\sum_i \sum_j x_{i,j}^{m} \cdot MW_{i,j}} \tag{5}$$

$$SG_{i,j} = \frac{\frac{M_{i,j}}{L_{mv_{i,j}}}}{\rho_{H_2O}} \tag{6}$$

where $MW_{i,j}$ is the molecular weight; $x_{i,j}^{w}$ is the mass fraction; $x_{i,j}^{m}$ is the molar fraction; $SG_{i,j}$ is the molecule's specific gravity. $L_{MV\,i,j}$ is the molar volume of the liquid can be predicted using the improved ANN-CM model.

2.3.2 Elemental content and SG

The mass fractions of CHNOS are calculated as Eqs(7)-(8), excluding low-content heteroatoms.

$$w_{E,i,j} = \frac{n_{E,i,j} \cdot A_E}{MW_{i,j}} \qquad \forall E \in \{\mathrm{C}, \mathrm{H}, \mathrm{N}, \mathrm{O}, \mathrm{S}\} \tag{7}$$

$$x_{E,i,j}^{w} = \frac{\sum_j \left(x_{i,j}^{m} \cdot MW_{i,j} \cdot w_{E,i,j}\right)}{\sum_j \left(x_{i,j}^{m} \cdot MW_{i,j}\right)} \tag{8}$$

where $MW_{i,j}$ is the molecular weight of the molecule j of core structure i in the molecular repository; $n_{E,i,j}$ is the number of atoms of element *E* in molecule *j*; A_E is the atomic weight of element *E.*
The calculation of the mixture's SG is calculated as Eq(9).

$$SG_{i,j} = 0.001 \cdot \left\{ \sum_j \left(\frac{x_{i,j}^{w}}{\frac{M_{i,j}}{\mathrm{Lmv}_{i,j}}} \right) \right\}^{-1} \tag{9}$$

2.3.3 Sensitive analysis

Uncertainty is introduced via sensitivity analysis, with ±20 % variation around baseline optimization results reflecting experimental uncertainty. The Saltelli-Sobol method generates 128 (variables + 2) quasi-random samples within defined intervals, with each set analyzed through the above steps. Sobol indices, calculated using Eq.(10), quantify the contribution of each input parameter to overall prediction variance.

$$S_i = \frac{Var(\mathbb{E}[E|X_i])}{Var(E)} \tag{10}$$

where S_i is the Sobol index of the *i*-th variable; E is the calculation error; X_i is the *i*-th input variable in the set of input variables.

3. Case study

The proposed framework is evaluated through the reconstruction of molecular compositions for both petroleum fractions and bio-oil derived from biomass.

3.1 Case study 1: diesel

Petroleum-derived diesel is commonly used for transportation and industrial purposes due to its high energy density. 27 seed molecules present in the petroleum-derived diesel (Guan et al., 2022), including normal and isomeric paraffins, naphthenes, aromatics, as well as thiol, thiophene, pyridine pyrrole. By adding sidechains to

seed molecule of each homologues series, 217 molecules are generated within the boiling point range of 473-623 K. Table 1 presents the calculated bulk properties alongside the experimental data. The calculated specific gravity is 0.84 kg/L, which is in agreement with Speight's observation where the density of petroleum diesel is approximately 0.85 kg/L (Speight, 2011). Compared to previous work (Pan et al., 2022), significant improvements have been achieved in 4 out of the 6 distillation curve points. The absolute relative errors (AREs) at the 50 vol% and 70 vol% points have been reduced by 70 % and 97 %, respectively, while the 10 vol% and 90 vol% points have seen reductions of 33 % and 58 %, respectively. The total sum of AREs has decreased from 2.41 % to 1.95 %.

Table 1: Experimental and predicted bulk properties of the diesel sample

Properties	Measured	Predicted	ARE (%)	ARE (%) from Pan et al. (2022)
Distillation Curve (ASTM D86, ℃)				
10 vol %	267.96	266.91	0.39	0.58
30 vol %	275.98	276.63	0.23	0.19
50 vol %	287.15	287.34	0.066	0.22
70 vol %	299.49	299.50	0.003	0.11
90 vol %	313.81	314.14	0.105	0.25
100 vol %	325.58	329.22	1.11	1.02
Element (wt.%)				
C	0.8715	0.8715	0.00	0.00
H	0.1270	0.1270	0.05	0.04
N	0.0002	0.0002	0.00	-
S	0.0013	0.0013	0.00	-
Specific gravity	0.89	0.84	5.61	-
PONA (wt. %)				
P	0.14	0.14	0.00	-
N	0.58	0.58	0.00	-
A	0.28	0.28	0.00	-

The case study included sensitivity analysis to assess the impact of each optimized parameter on the result. $x_{76} - x_{100}$ represent the fractions of each HS, with much higher sensitivity than Gamma PDF parameters ($x_1 - x_{75}$), because Gamma PDF adjusts the distribution of homologues in a HS, while the fraction parameters determine the total content of one HS. However, the property variation between HS is much more significant than within one HS as shown in Figure 1. The results indicate that the fractions of each HS contribute significantly more to the variance in the distillation curve and elemental analysis than the shape parameters of the Gamma PDF. This suggests that the evolution of carbon numbers within the same HS has a minimal impact on the macroscopic properties of diesel.

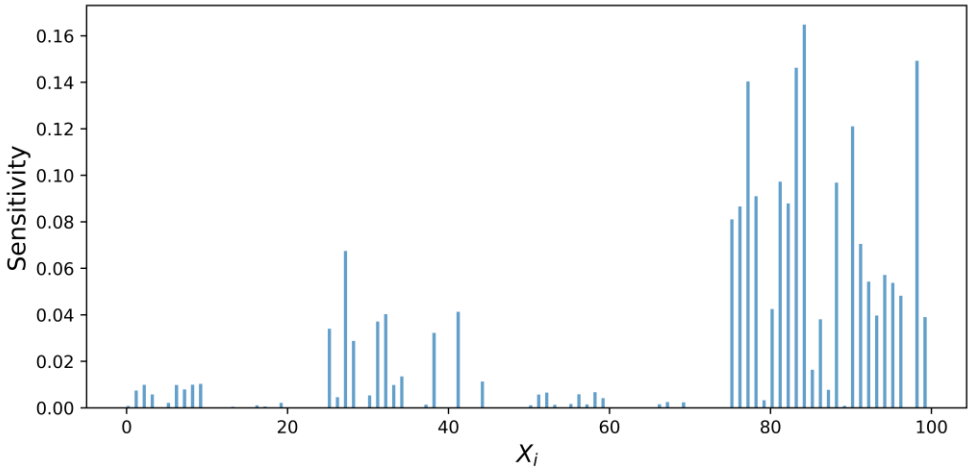

Figure 1: Sensitivity analysis of input parameters in diesel molecular composition reconstruction

3.2 Case study 2: ENSYN bio-oil characterization

Unlike petroleum fractions, most of the bio-oil molecules contain oxygen atoms. Due to the complexity of bio-oil, the analytical data of the samples are often unavailable, and fully characterized samples are rare. Jamri et al. (2020) used the MTHS method to reconstruct the molecular composition of the bio-oil (Al Jamri et al., 2020). Residues (37.7 wt.%) are assumed to follow the same trend as that of the organic distillate fraction according to their work. Bulk properties and the distillation curve of the bio-oil sample are obtained from their work. Based on reported information, there are 37 seed molecules covering the above-mentioned oxygenates and hydrocarbons (Al Jamri et al., 2020). HS of each molecule are generated within the boiling point range between 323.15 K and 873.15 K. Consequently, 952 molecules are generated to represent the sample.
Table 2 compares the measured and predicted bulk properties, which include boiling points and weight fractions of C, H, and O elements. The initial boiling point (IBP) and FBP are not considered objectives in this work because they are generally difficult to measure.

Table 2: Experimental and predicted bulk properties of the bio-oil sample

Properties		Measured value	Predicted value	ARE (%)	ARE (%) from Pan et al. (2022)
TBP Curve (℃)	10 vol%	96.59	96.55	0.041	0.26
	30 vol %	115.82	118.36	2.19	3.00
	50 vol %	207.79	207.28	0.24	0.67
	70 vol %	292.24	293.08	0.28	0.17
	90 vol %	426.86	427.06	0.047	0.05
C (wt.%)		0.71	0.71	0.00	0.01
H (wt.%)		0.086	0.086	0.00	0.21
O (wt.%)		0.19	0.19	0.00	0.00
Specific gravity		1.21	1.05	13.22	13.23

One limiting factor is that the generated library only contains a few molecules with boiling points under 115.82 ℃, which limits the adaptability of the model (Wang et al., 2020). However, adding more molecules to address the issue would also bring difficulties to model optimization. The largest ARE, which is 13.22 %, is from specific gravity. Compared to a reported error of 19.5 % (Al Jamri et al., 2020), these closer results suggest that the proposed method in this work is more accurate. Moreover, compared to previous work results (Pan et al., 2022), the total ARE has decreased from 17.6 % to 16.01 %. This discrepancy between the measured and calculated specific gravity could be attributed to the 37.7 wt.% of residual substances in the bio-oil, which usually have a higher boiling point and specific gravity than organic distillate fractions. However, experimental data on the residues, particularly their specific gravity and structures are not readily available. Overall, the performance of the proposed methodology is better than the reported result. With more reliable experimental data, the accuracy can be further improved.
A comprehensive sensitivity analysis is conducted, and the findings are illustrated in Figure 2**.** A total of 148 parameters are optimized, with three Gamma PDF parameters and one HS fraction parameter being optimized for each of the 37 seed molecules.

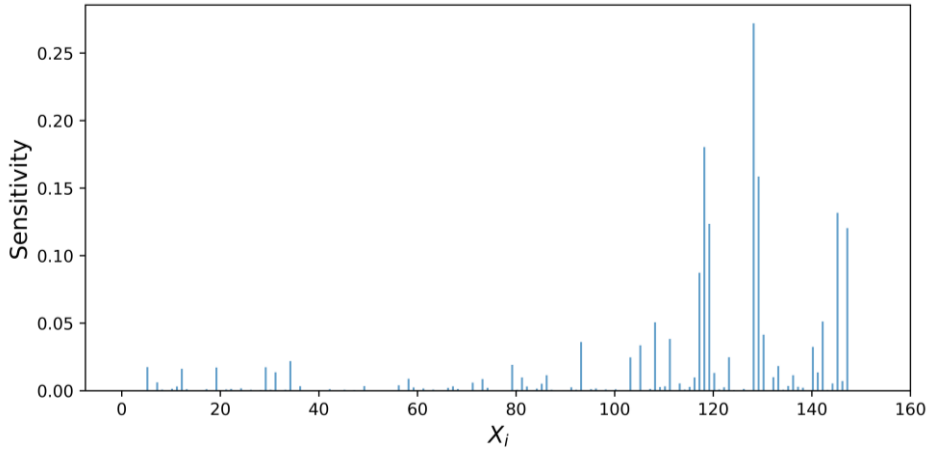

Figure 2: Sensitivity of input parameters in bio-oil molecular composition reconstruction

The results indicate that, the vast majority of the sensitivity indices for the Gamma PDF shape parameters are extremely low, indicating that the internal shape differences of the carbon number distribution have a limited impact on macroscopic properties. In contrast, the content parameters of homologous series exhibit relatively high sensitivity indices. This is attributed to the high oxygen content and structural complexity of bio-oil, where different functional groups exhibit large variations in boiling points and polarity. As a result, the sensitivity of the HS fraction parameters is significantly higher than that observed in the diesel systems.

4. Conclusion

The molecular reconstruction of complex hydrocarbon mixtures is a typical high-dimensional, underdetermined optimization problem. To achieve accurate molecular reconstruction and property estimation, this study proposed a framework for molecular composition reconstruction based on statistical assumptions and a hybrid heuristic–gradient optimization strategy, which integrates molecular representation and bulk property transformation for composition modelling of bio-oil and diesel samples. For the calculation of diesel properties, the total sum of AREs has decreased from 2.41 % to 1.95 %, for the bio-oil sample, the total ARE has decreased from 17.6 % to 16.01 %, the accurate results of the two cases demonstrate the effectiveness of this framework. In conclusion, this framework presents a significant step forward in the development of strategies for understanding complex hydrocarbon mixtures.

Acknowledgments

Jianfeng acknowledges the financial support from China Scholarship Council (CSC) (No. 202406440073).

References

Al Jamri M., Li J., Smith R., 2020, Molecular characterisation of biomass pyrolysis oil and petroleum fraction blends. Computers & Chemical Engineering, 140, 106906.

Feng S., Cui C., Li K., Zhang L., Shi Q., Zhao S., Xu C., 2019, Molecular composition modelling of petroleum fractions based on a hybrid structural unit and bond-electron matrix (SU-BEM) framework. Chemical Engineering Science, 201, 145–156.

Guan Y.M., Guan D., Zhang C., Yuan S.H., Cai G.Q., Zhang L.Z., 2022, Diesel molecular composition and blending modeling based on SU-BEM framework. Petroleum Science, 19, 839–847.

Hudebine D., Verstraete J.J., 2004, Molecular reconstruction of LCO gasoils from overall petroleum analyses. Chemical Engineering Science, 59, 4755–4763.

Jiao J., Gao X., Li J., 2025a, Pure Component Property Estimation Framework Using Explainable Machine Learning Methods. Chinese Journal of Chemical Engineering, DOI:10.1016/j.cjche.2025.05.011.

Neurock M., Libanati C., Nigam A., Klein M.T., 1990, Monte carlo simulation of complex reaction systems: molecular structure and reactivity in modelling heavy oils. Chemical Engineering Science, 45, 2083–2088.

Pan Q., Fan X., Li J., 2022, Unified Characterisation and Property Estimation Framework for Composition Reconstruction of Biomass Pyrolysis Oil and Petroleum Fractions. Chemical Engineering Transactions, 94, 973–978.

Ren Y., Liao Z., Sun J., Jiang B., Wang J., Yang Y., Wu Q., 2019, Molecular reconstruction: Recent progress toward composition modeling of petroleum fractions. Chemical Engineering Journal, 357, 761–775.

Speight J.G., 2011, Production, properties and environmental impact of hydrocarbon fuel conversion. Advances in clean hydrocarbon fuel processing. Chapter In: M Rashid Khan (Ed.), Advances in Clean Hydrocarbon Fuel Processing: Science and Technology. Vol 19, Woodhead Publishing Limited, Cambridge, UK, 54–82.

Tong Q.I.U., Jincai C., Zhou F., 2016, Molecular reconstruction model for petroleum fractions based on structure oriented lumping. Journal of Tsinghua University Science and Technology, 56, 424–429.

Wang Y., Han Y., Hu W., Fu D., Wang G., 2020, Analytical strategies for chemical characterization of bio-oil. Journal of Separation Science, 43, 360–371.