Journal of Zhejiang University-SCIENCE A 2026 27(8):825-836
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2500457

Check for updates

Data-driven neural surrogates for ReaxFF molecular dynamics simulations in engine-relevant combustion chemistry

Yuchao YAN[1], Qiao HUANG[2], Tianfang XIE[3], Jinlong LIU[1]✉

[1]*Power Machinery & Vehicular Engineering Institute, Zhejiang University, Hangzhou 310027, China*
[2]*College of Information Engineering, China Jiliang University, Hangzhou 310018, China*
[3]*School of Aeronautics and Astronautics, Purdue University, West Lafayette 47907, USA*

Abstract: Machine learning (ML) has gained significant traction in engine-related research, particularly because of its potential to improve predictive performance while reducing computational costs. However, most current applications rely on feedforward neural networks (FNNs; e.g., conventional artificial neural networks (ANNs)); these are well-suited for modeling static data and capturing nonlinear relationships, but do not explicitly encode temporal dependencies unless sequence context is introduced via feature engineering. Motivated by this limitation, we evaluate sequence-aware neural surrogates for engine-relevant combustion-chemistry time-series data. Specifically, the temporal evolution of an intermediate product group during polycyclic aromatic hydrocarbon (PAH) formation in C_2H_4/NH_3 pyrolysis is modeled using reactive force field (ReaxFF) molecular dynamics (MD) trajectories, comparing an FNN baseline (with explicit time as an input) against a long short-term memory (LSTM)-based recurrent neural network (RNN). The results show that while the FNN baseline benefits from explicit temporal feature engineering, its predictive performance is inferior to that of the LSTM model, even when the network depth is increased. This behavior is consistent with the architectural limitations of feedforward models, which do not maintain an internal memory state, and therefore, tend to generalize poorly when the target dynamics are history dependent. In contrast, the LSTM model leverages gated memory to learn temporal dependencies and consequently improves the predictive accuracy of combustion-chemistry time-series modeling, providing an efficient surrogate once trained. Overall, our findings delineate the conditions under which sequence-aware recurrent architectures offer advantages over feedforward models for ReaxFF MD time-series surrogate modeling.

Key words: Feedforward neural network (FNN); Recurrent neural network (RNN); reactive force field molecular dynamics (ReaxFF MD); Combustion-chemistry surrogate; Internal combustion engine

1 Introduction

Machine learning (ML) has gained significant momentum in recent years, marking a transformative shift in many disciplines. This has led to the increasing integration of ML into fields that were previously considered outside its traditional scope. One such area is internal combustion engines, where ML is gaining traction due to its potential to revolutionize engine design, performance, and emission control (Aliramezani et al., 2022; Huang et al., 2025b). For instance, Huang et al. (2022) employed neural networks to model the spark ignition engine combustion process, using engine control variables (e.g., spark timing, equivalence ratio, and engine speed) as inputs, and engine performance metrics (e.g., power output, engine phasing, and emissions) as outputs, thereby facilitating engine control unit calibration (Liu et al., 2022; Sok et al., 2024). Additionally, ML has been used as a virtual sensor in knock detection (Petrucci et al., 2020; Aramburu et al., 2024) and emission control (Atkinson et al., 1998; Kumar et al., 2024), particularly for reducing nitrogen oxide content in selective catalytic reduction systems (Le Cornec et al., 2020; Kamat et al., 2023). The application of ML has also helped establish relationships between fuel structure and properties, enabling the prediction of cetane and octane numbers for pure fuels and mixtures (Li et al., 2020, 2021). In combustion modeling, ML enables the prediction of combustion parameters under varying operating conditions, marking a transition

✉ Jinlong LIU, ljl199022@zju.edu.cn
Jinlong LIU, https://orcid.org/0000-0002-4820-7460

Received Sept. 17, 2025; Revision accepted Dec. 14, 2025;
Crosschecked Mar. 23, 2026; Online first Apr. 16, 2026

© Zhejiang University Press 2026

from non-predictive Wiebe function models to semi-predictive models (Mishra et al., 2021; Torregrosa et al., 2021). In multi-dimensional engine computational fluid dynamics (CFD) simulations, ML can serve as surrogate models that work in conjunction with optimization algorithms to optimize chamber shape profiles and other performance metrics (Badra et al., 2021; Silva et al., 2023). Additionally, ML has been employed to address sub-grid modeling and closure term issues, thereby enhancing the accuracy and efficiency of simulations (Yao et al., 2020; Liu and Wang, 2022). While ML has been proven effective in these applications, most current research in this area focuses on static data and nonlinear relationships, especially in cases where data points are independent. Conventional ML methods, such as feedforward neural networks (FNNs; e.g., conventional artificial neural networks (ANNs)), support vector machines, and random forests, are predominantly used in existing engine studies due to their proven effectiveness at these tasks (Aghbashlo et al., 2021; Li et al., 2023). However, despite the rapid advancement of deep learning in other fields, its application in engine research remains limited. This underutilization highlights the opportunity to develop more advanced techniques, which could unlock benefits in engine research.

In this paper, we aim to address this gap by evaluating sequence-aware neural surrogates in a combustion-chemistry setting relevant to engines. Specifically, a case study is undertaken to examine whether a sequence-aware recurrent model—represented by a long short-term memory (LSTM) network—provides measurable advantages over FNN baselines for constructing time-series surrogates. Although feedforward models can incorporate time as an explicit input, they do not maintain an internal memory state and may generalize poorly when the target dynamics exhibit history dependence. A key motivation of this work is to reduce the computational costs of reactive force field (ReaxFF) molecular dynamics (MD) simulations, which are widely used to understand aromatic hydrocarbon formation yet remain computationally expensive (Xing et al., 2023; Diao et al., 2024). By training on representative conditions, data-driven surrogates can reduce the need for redundant simulations, and thus, lower the computational demands. Hence, an LSTM-based recurrent neural network (RNN) is applied to build a surrogate for ReaxFF MD trajectories of aromatic hydrocarbon formation during C_2H_4/NH_3 pyrolysis. This is relevant to zero-carbon ammonia fuel strategies for internal combustion engines (Huang and Liu, 2024; Huang et al., 2024), where soot and aromatic chemical reactions in ammonia/hydrocarbon mixtures require appropriate kinetic descriptions. In such systems, capturing aromatic kinetics relevant to soot often requires explicit treatment of carbon–nitrogen interactions. However, resolving carbon–nitrogen interactions in aromatic formation often necessitates the use of ReaxFF MD since existing reaction mechanisms may lack sufficient elementary steps for these pathways (Yan et al., 2025). This is particularly important for ammonia/hydrocarbon combustion applications (e.g., ammonia/diesel dual-fuel engines), where soot-relevant aromatic chemistry must be represented in engine models (Huang et al., 2025a). Moreover, first-principles-based C/H/O/N force-field simulations provide a practical route for probing the detailed reaction pathways underlying aromatic formation (Zhang et al., 2023). Overall, the results of this study highlight conditions under which sequence-aware modeling outperforms FNN baselines for combustion-chemistry time-series prediction, which may help accelerate engine-related simulation workflows.

2 Materials and methods

In this study, we utilize ReaxFF MD simulations of aromatic hydrocarbon formation during C_2H_4/NH_3 pyrolysis to compare the performance of FNN-based models and an LSTM-based RNN in modeling this combustion process, illustrating how sequence-aware architectures can be used as surrogates for ReaxFF MD. The ReaxFF MD simulations, which are conducted using LAMMPS software (large-scale atomic/molecular massively parallel simulator) (Thompson et al., 2022) under NVT conditions (a constant number of particles, volume, and temperature) at 3000 K with periodic boundary conditions, provide the data for modeling. The parameter settings used are from Zhang et al. (2022) and are based on the CHON-2019 force field (Kowalik et al., 2019), and Ar parameters are provided by Yoon et al. (2016). In the combined experimental-simulation study by Zhang et al. (2022), the ReaxFF MD predictions of polycyclic aromatic hydrocarbon (PAH) formation, nitrogen-containing species, and soot propensity as a function of NH_3 ratio were shown

to be consistent with laser-diagnostic measurements (laser-induced incandesence and laser-induced fluorescence) in C_2H_4/NH_3 diffusion flames; as such, this provides experimental support for the force field and simulation parameters adopted in the present work.

The simulations track the formation and concentration of key species—including molecules with varying numbers of carbon atoms—as well as the evolution of reactive radicals. The system density is maintained at 0.1 g/cm^3. Given the short (sub-nanosecond) timescale of ReaxFF MD simulations compared to experimental observations (typically milliseconds to seconds), a higher temperature is applied in order to accelerate atomic collisions, which is common practice in ReaxFF MD studies (Kamat et al., 2010; Zhao et al., 2020). This allows chemical reactions to be observed within an acceptable timeframe and at a lower computational cost (Wang et al., 2023). Accordingly, these accelerated MD trajectories are not intended for one-to-one quantitative comparison with in-cylinder engine measurements. The simulation time step is set to 0.1 fs for accuracy, with a total simulation duration of 500 ps, and recording trajectories every 0.1 ps. To investigate reactions with varying NH_3 ratios, different initial concentration cases (A0, A20, A40, and A60) are initialized, as shown in Table 1. A schematic diagram of the A0 and A60 in the MD simulations is shown in Fig. 1.

Table 1 Initial system compositions for the ReaxFF MD simulations

System	Dataset	Number of molecules		
		C_2H_4	NH_3	Ar
A0	Training	400	0	240
A20	Training	400	80	160
A40	Test	400	160	80
A60	Training	400	240	0

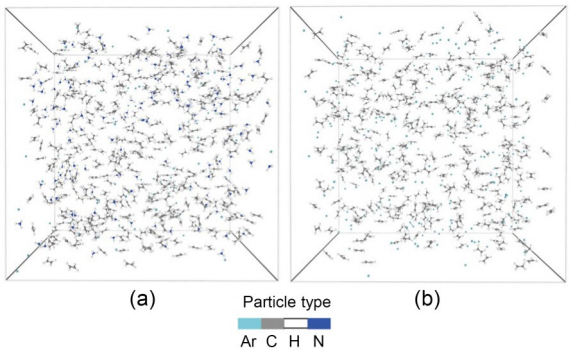


Fig. 1 Initial simulation boxes: (a) A0; (b) A60

This varied NH_3 ratio helps us assess its impact on PAH formation. By holding the number of C_2H_4 molecules constant, consistent comparisons can be made between PAH outcomes for different cases. The NH_3 ratios vary from 0% to 60% relative to C_2H_4, with Ar used as an inert balancing gas to keep the total number of molecules constant across cases (Table 1); hence, the comparisons focus on composition-driven chemical effects under this fixed total molecular number setup. Data from the MD simulations are used to train and test the FNN and LSTM models. Three repetitions are performed to improve reliability and reduce random fluctuations. The evolution of C_6 species (including benzene as a key intermediate in PAH formation) serves as the prediction target, which indicates PAH formation trends. Data from A0, A20, and A60 are used for the training dataset, while data from A40 are used for the test dataset. Note that this train/test split evaluates interpolation within the training composition domain, rather than extrapolation beyond the trained range.

In this work, we compare FNN and LSTM-based RNN architectures for predicting MD time-series data. FNN baselines (implemented as ANNs) are used as feedforward models, while an LSTM-based RNN is used as the sequence-modeling approach. Unlike recurrent models, FNNs process each input independently and do not maintain an intrinsic memory state; therefore, we introduce temporal information explicitly by augmenting the FNN input with time. Within the FNN framework, two models are constructed, namely a single-hidden-layer "shallow" network (input layer→one hidden layer→output layer) and a three-hidden-layer "moderate" network (input layer→three hidden layers→output layer), to explore the impact of network depth on performance. Both FNN models take feature vectors of $[t, r]$ (t is the time; r is the NH_3 ratio) as inputs and predict the quantity of the intermediate product group C_6 species (e.g., benzene and other six-carbon ring structures) as the output; note that min–max scaling is applied for normalization. The FNN models consist of an input layer followed by one or more hidden layers with rectified linear unit (ReLU) activation functions ($f(x)=\max(0, x)$, and x denotes the pre-activation input to a neuron) to capture nonlinear relationships. The shallow FNN includes one hidden layer with 64 neurons, while the moderate FNN contains three hidden layers, each with 64 neurons. Both models are trained using the Adam optimizer, which

adjusts weights iteratively through backpropagation to minimize the loss function. The LSTM network—a variant of RNNs—is designed to capture temporal dependencies in the PAH formation process. To preserve temporal relationships, the dataset is structured sequentially and not shuffled. The LSTM architecture incorporates a primary LSTM layer with 50 neurons and uses gating mechanisms, such as input, forget, and output gates, to regulate information flow, enabling the retention of information over extended sequences while mitigating the vanishing gradient problem. A fully connected output layer generates the final predictions. All models are implemented using Python software (version 3.9) and trained with the Adam optimizer. The training process is conducted with a batch size of 32 over 1000 epochs. Performance is evaluated using the coefficient of determination (R^2) and normalized root mean square error (NRMSE), where higher R^2 values (closer to 1) and lower NRMSE values (closer to 0) indicate better predictive accuracy and model fit.

3 Results and discussion

The goal of this work is to explore the application of sequence-aware neural networks in engine-related research. In this section, we compare the modeling performance of FNN baselines and an LSTM-based RNN in simulating aromatic hydrocarbon formation during C_2H_4/NH_3 pyrolysis; we also evaluate whether sequence-aware modeling provides measurable benefits over feedforward baselines under consistent datasets and evaluation metrics. The effectiveness of the data obtained from ReaxFF MD simulations will first be validated before proceeding with the performance comparison.

Fig. 2 shows the evolution of C_6 species during C_2H_4/NH_3 pyrolysis under varying C_2H_4/NH_3 ratios. In the absence of NH_3 (A0), C_6 species formation is initially slow but increases steadily, reaching a high level at later stages. With a low NH_3 ratio (A20), the trend is similar, although the overall molecular number of C_6 species is slightly lower, which indicates minor inhibition by CN radicals. At a moderate NH_3 ratio (A40), the C_6 species formation accelerates significantly at an early stage, reaching high levels by 200 ps. Between 200 and 400 ps, the molecular number of C_6 species fluctuates, likely due to the competing effects

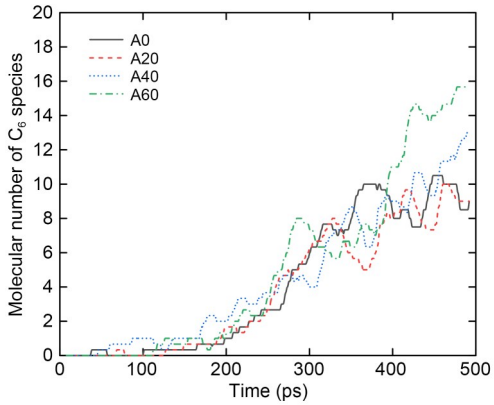


Fig. 2 Time evolution of molecular number of C_6 species during C_2H_4/NH_3 pyrolysis under different NH_3 ratios

of NH_3 inhibiting aromatic hydrocarbon formation while also promoting radical activity. After 400 ps, C_6 species reaches a level slightly higher than A0 and A20. This suggests that while a moderate NH_3 ratio enhances early-stage reactions, it also balances the consumption and generation of C_6 species in subsequent aromatic hydrocarbon growth pathways. At a high NH_3 ratio (A60), C_6 species formation exhibits steady initial growth and does not surpass the early-stage growth rate observed in the moderate NH_3 ratio (A40). After 150 ps, the formation rate of C_6 species under A60 remains steady. After 400 ps, the C_6 species under A60 continues to grow and eventually exceeds all other conditions, reaching the highest level overall. Additionally, the lower C_6 species levels observed at 500 ps under A0 and A20 suggest that a significant portion of C_6 species molecules may be consumed in the hydrogen abstraction acetylene addition (HACA) mechanism, leading to the growth of larger PAHs.

In contrast, under A40 and A60 conditions, NH_3 likely inhibits the HACA mechanism by reducing the availability of C_2H_2 radicals, resulting in less conversion of C_6 species into larger PAHs and allowing more C_6 species to accumulate. As highlighted in prior studies (Bennett et al., 2020; Liu et al., 2023), NH_3 suppresses PAH formation by capturing key species and radicals such as C_2H_2 and CH, which play essential roles in the HACA mechanism. Meanwhile, NH_3 promotes the formation of CN radicals, which can accelerate specific radical-driven pathways, as noted by Zhang et al. (2022). These observations emphasize the complexity of C_6 species kinetics, as it represents a balance between multiple competing mechanisms, including radical-driven formation and consumption

pathways. This inherent complexity makes the intermediate product group of C_6 species (e.g., benzene and other six-carbon ring structures) an ideal target for evaluating the reliability of ML models. Accurately predicting the evolution of C_6 species under varying NH_3 ratios requires the model to capture both the nonlinear dynamics and the interplay of competing reaction pathways, thus providing a rigorous test of its applicability to complex chemical systems.

The evolution of carbon atoms in C_3, N–C_3, C_2+C_4, and N–C_2+N–C_4 species during C_2H_4/NH_3 pyrolysis at varying NH_3 ratios is depicted in Fig. 3. These species represent the dominant hydrocarbons contributing to the aromatic precursor pool and offer insights into how NH_3 addition influences C_2H_4 decomposition and aromatic-precursor formation. In this figure and discussion, C_2 denotes dicarbon species, C_3 refers to tricarbon hydrocarbons, and C_4 represents tetracarbon hydrocarbons. Notably, C_2 excludes the reactant C_2H_4 to better capture the effect of NH_3 during decomposition. As the NH_3 ratio increases, the peak quantities of all hydrocarbon species decrease, which is consistent with previous studies (Bennett et al., 2020). Conversely, the corresponding nitrogen-containing hydrocarbons increase as NH_3 is added. The impact of NH_3 on C_3 species is relatively minor, whereas its influence on C_2 and C_4 species is more pronounced, with N–C_2+N–C_4 exhibiting significant increases. This trend is likely due to C_2H_4 serving as the primary fuel, which predominantly dehydrogenates to form C_2+C_4 species. The differential effects of NH_3 on hydrocarbons with varying carbon numbers suggest that NH_3 may affect benzene ring formation through both odd- and even-carbon species chemistry—a phenomenon which was also reported in previous studies (Zhang et al., 2022). In particular, acetylene (C_2H_2), an even-carbon species,

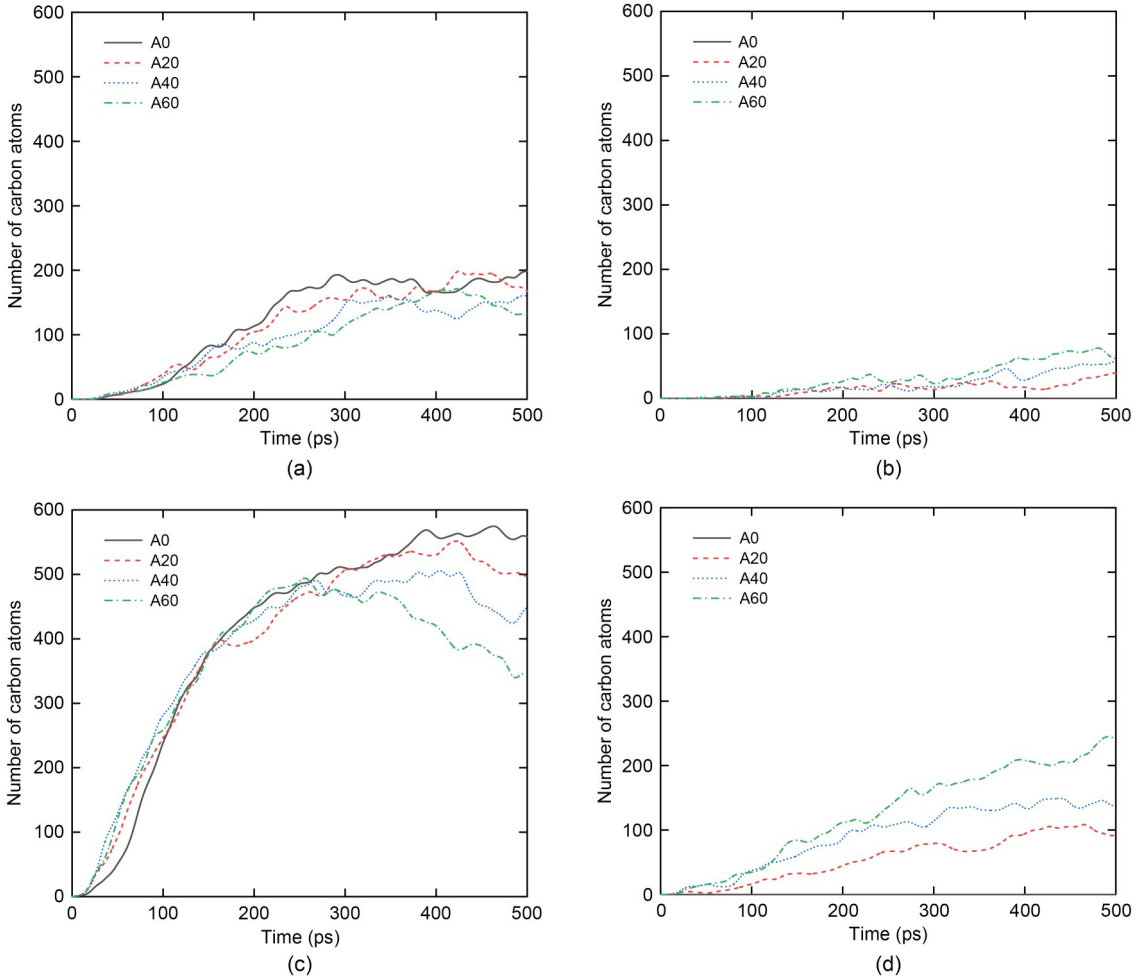


Fig. 3 Time evolution of number of carbon atoms in (a) C_3, (b) N–C_3, (c) C_2+C_4, and (d) N–C_2+N–C_4 species during C_2H_4/NH_3 pyrolysis at different NH_3 ratios

is a key growth reactant in the HACA mechanism and directly contributes to aromatic ring growth. Therefore, changes in the C_2H_2 pool can strongly modulate benzene/PAH formation and soot propensity. In keeping with observations of ammonia–ethylene flames, NH_3 is known to influence soot propensity by reducing CH, CH_2, and CH_3 radicals in C_2H_4 combustion. Soot propensity is governed by multiple factors, including diluents (Kailasanathan et al., 2013) and fuel properties (Tang et al., 2019). These findings suggest that NH_3 addition primarily affects the chemistry of even-carbon species, thereby altering soot propensity pathways.

Previous studies (Bennett et al., 2020; Liu et al., 2023) have suggested that NH_3 addition inhibits soot formation by sequestering carbon atoms from hydrocarbon fuels, consequently reducing the availability of the carbon species necessary for PAH and soot formation and growth. As reported by Bennett et al. (2020), CN species are potential carbon–nitrogen intermediates formed during the co-firing of NH_3 with hydrocarbon fuels, and they suppress soot formation by reacting with PAHs. Hydrogen cyanide (HCN) and hydrogen isocyanide (HNC) are also nitrogen-containing intermediates that may form in C_2H_4/NH_3 systems; however, under the present ReaxFF MD conditions (3000 K), their populations remain low relative to CN species throughout the trajectories and do not exhibit sustained accumulation. Because the following discussion focuses on CN-radical-dominated interactions (e.g., occupancy of active sites on carbon chains and PAHs), the influence of HCN/HNC pathways on these dynamics is considered of secondary importance within the investigated parameter space.

Fig. 4 presents the evolution of molecular number of CN species during C_2H_4/NH_3 pyrolysis at varying NH_3 ratios. As shown, CN species formation increases significantly with higher NH_3 ratios. As a result, when NH_3 ratio rises, the availability of carbon atoms that participate in soot formation decreases. This trend is consistent with previous findings, including the work of Liu et al. (2023), which examined the impact of NH_3 addition on soot formation during n-decane pyrolysis. Their results indicated that adding NH_3 leads to the consumption of CH_3 and C_2H_5 by nitrogen-containing species. Reactions between CH_3 and NH_2, as well as between C_2H_5 and NH_2, appear to be key processes in NH_3-containing fuel systems. For one, the depletion of these carbon precursors directly reduces PAH formation. Moreover, PAH growth is inhibited further due to disruptions in the formation of key reactants such as C_2H_2. These behaviors are consistent with the combined experimental ReaxFF MD study by Zhang et al. (2022), which provides additional support for the physical relevance of the MD trajectories used for surrogate modeling in this work.

Fig. 5 presents the chemical structures of nitrogen-containing polycyclic aromatic hydrocarbons (N-PAHs) formed during the MD simulations of C_2H_4/NH_3 pyrolysis under A40 at 3000 K. This serves to illustrate how nitrogen incorporates into N-PAH structures. The resulting N-PAHs exhibit irregular and multi-ring aromatic configurations. In these representative structures, most nitrogen atoms are located at the periphery of the structure, with few penetrating the inner carbon framework.

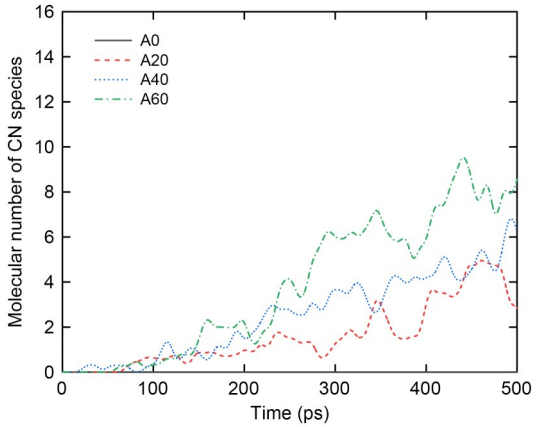


Fig. 4 Time evolution of molecular number of CN species during C_2H_4/NH_3 pyrolysis at different NH_3 ratios

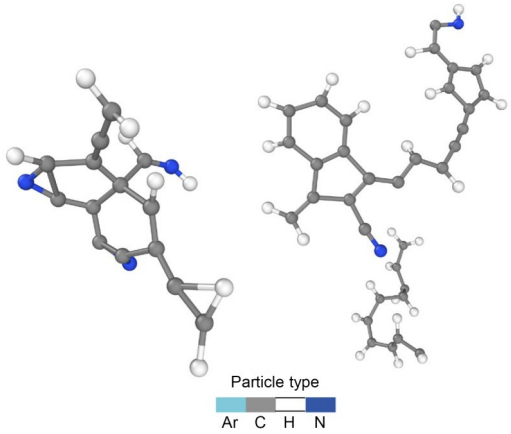


Fig. 5 Representative two snapshots of N-PAHs formed during ReaxFF MD simulations of C_2H_4/NH_3 pyrolysis (A40, 3000 K)

This trend aligns with the expected growth process of N-PAHs, where nitrogen atoms predominantly remain on the branches rather than integrating into the inner rings to form heterocycles. These observations are also consistent with prior reports on nitrogen incorporation and N-PAH growth. Zhang et al. (2024) reported that at temperatures below 2500 K, N-PAHs primarily grow via the HACA mechanism. However, at temperatures exceeding 2500 K, their growth is mainly driven by continuous carbon-chain attachment followed by condensation polymerization. Additionally, it was observed that nitrogen atoms in nitrogen-containing carbon chains are almost exclusively located at the chain ends, rather than in the middle. This is attributed to the relatively high energy barrier associated with the cyclization of nitrogen-containing carbon chains, making the formation of new aromatic rings more difficult (Zaher et al., 2023). In summary, the MD simulation results for C_2H_4/NH_3 pyrolysis and C_6 species formation are consistent with the theoretical predictions and previous studies, confirming the validity of this dataset for subsequent data-driven surrogate modeling.

Since the MD results have been validated against existing literature, this section now presents a comparative analysis of three models, namely the shallow FNN (FNN baseline), moderate FNN (FNN baseline), and an LSTM-based RNN. The objective is to evaluate their effectiveness on both the training and test datasets, so as to assess their feasibility as surrogates for reducing the computational costs of ReaxFF MD simulations of engine-relevant combustion chemistry. Fig. 6 shows a comparison of the C_6 species predictions from the shallow FNN (one hidden layer) and the MD simulations on the training dataset (with [t, r] as inputs). The shallow FNN captures the overall trend of C_6 species evolution, indicating its ability to learn general patterns in the data. However, certain regions show notable deviations from the results of MD simulations, thus highlighting limitations in reproducing finer details of the C_6 species dynamics. The shallow FNNs achieve an R^2 value of approximately 0.980 and an NRMSE of about 0.03–0.04 on the training dataset, indicating strong fitting performance. Nevertheless, the observed discrepancies suggest that further refinement is needed to improve the predictive accuracy, particularly in terms of capturing subtle variations in the trajectories.

Increasing the number of layers in a neural network generally enhances its ability to capture complex features and learn intricate relationships. Therefore, to assess whether greater depth improves model performance for the present task, we compare a shallow FNN (one hidden layer) with a moderate FNN (three hidden layers). The prediction results from the moderate FNN and the MD simulations on the training dataset are compared in Fig. 7. The moderate FNN effectively captures the overall trends observed in the MD simulation data and demonstrates an improved ability to model subtle fluctuations in the molecular number of C_6 species. The model achieves an R^2 value of 0.992 and an NRMSE of 0.028 (on the order of 10^{-2}), demonstrating the benefits of increased network depth for achieving higher accuracy and lower error rates. In general, the moderate FNN outperforms the shallow FNN, although the performance difference on the training dataset remains minimal.

Fig. 8 compares the prediction results from the LSTM-based RNN and the MD simulations on the training dataset. The LSTM-based RNN exhibits closer agreement with the MD simulation data, with reduced deviation and improved reproduction of the C_6 evolution curve. In addition, it achieves higher R^2 and lower NRMSE values than both the shallow and moderate

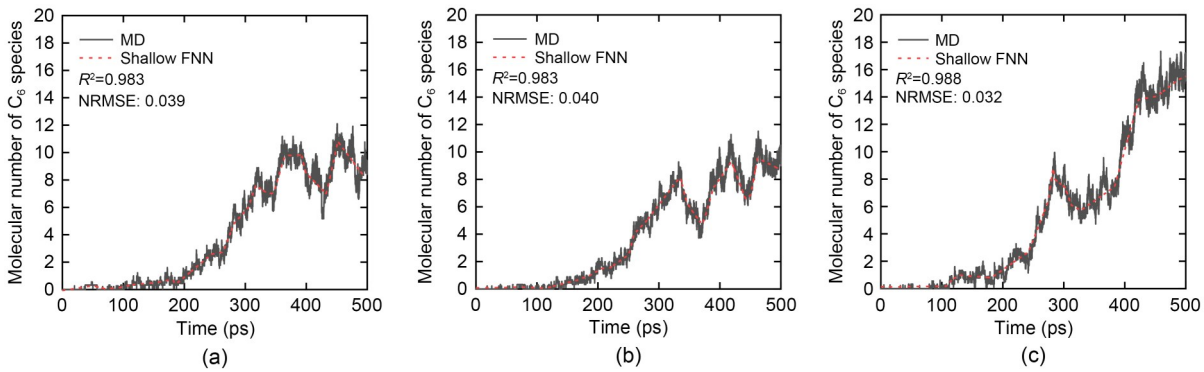


Fig. 6 Prediction results of the shallow FNN versus MD simulations on the training dataset: (a) A0; (b) A20; (c) A60

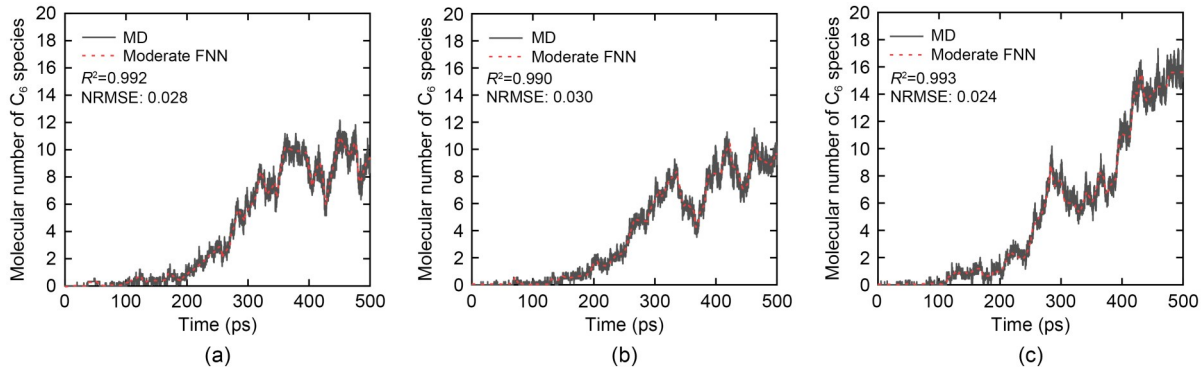


Fig. 7 Prediction results of the moderate FNN versus MD simulations on the training dataset: (a) A0; (b) A20; (c) A60

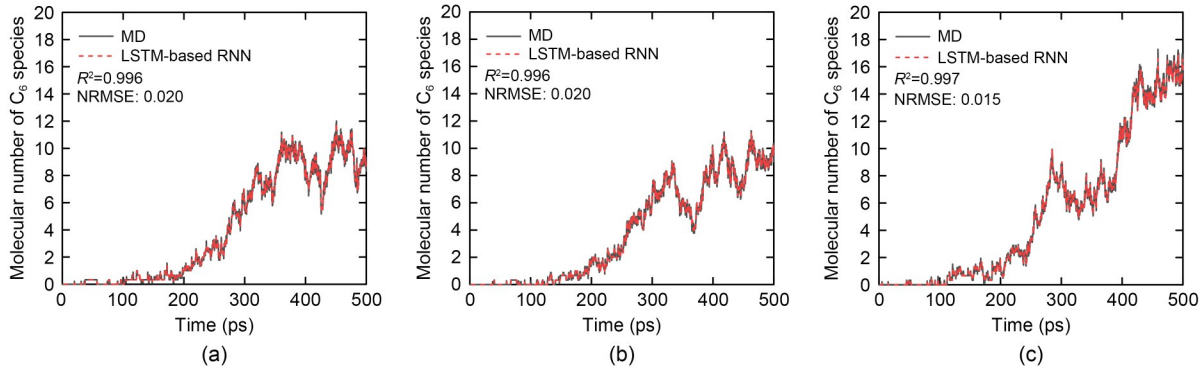


Fig. 8 Prediction results of the LSTM-based RNN versus MD simulations on the training dataset: (a) A0; (b) A20; (c) A60

FNNs on the training dataset. This result suggests that the LSTM-based RNN can accurately model sequential trajectories by leveraging its internal state to represent temporal dependencies, making it well-suited for the present time-series surrogate task. Overall, while all three models perform well on the training dataset, the LSTM-based RNN exhibits a modest advantage in accuracy.

Although all three models perform well on the training dataset, this does not necessarily ensure strong performance on the test dataset, due to potential challenges such as overfitting and limited generalization. Overfitting occurs when a model not only learns the underlying patterns in the training data, but also noise and specific anomalies, thereby reducing its performance on unseen data. Limited generalization refers to the inability of the model to effectively apply learned patterns to different datasets, which results in decreased accuracy when data that deviates from the training dataset is encountered. To assess the generalization ability of the models, in Fig. 9, we compare prediction results from the shallow FNN, moderate FNN, LSTM-based RNN, and MD simulations on the test dataset (A40). The shallow FNN exhibits the largest prediction deviations, followed by the moderate FNN. Additionally, the moderate FNN struggles to capture the overall trend and fails to accurately predict fluctuations in the data. These findings underscore the challenges that FNN baselines may face in maintaining accuracy and reliability when applied to unseen conditions. In contrast, the LSTM-based RNN effectively captures the pattern of C_6 species evolution, demonstrating high alignment with the MD simulations. Moreover, with an R^2 value of 0.996 and an NRMSE of 0.017, the LSTM-based RNN shows strong predictive performance. In comparison, the shallow and moderate FNN baselines achieve R^2 values of 0.738 and 0.867, respectively.

Fig. 9 also indicates that the FNN baselines exhibit a larger gap in train–test generalization compared to the LSTM-based RNN, which is consistent with overfitting under limited training coverage. This behavior reflects an architectural limitation: feedforward models do not maintain an internal memory state to encode reaction-history effects. When the trajectory at a given time depends on prior evolution (i.e., the system is

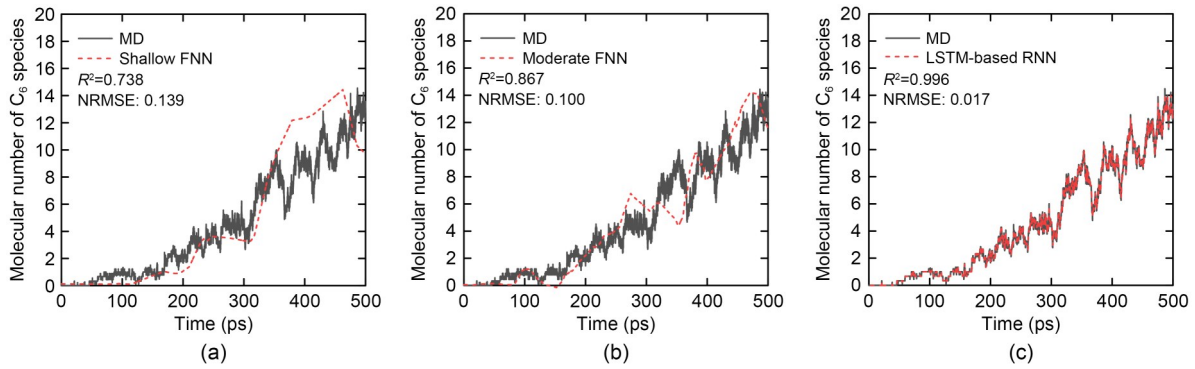


Fig. 9 Prediction results of (a) shallow FNN, (b) moderate FNN, and (c) LSTM-based RNN versus MD simulations on the test dataset

history dependent), sequence models such as LSTM can represent this dependence through their internal memory, whereas feedforward models—which process each input independently—can face difficulties in generalizing. It is worth noting that the preceding evaluation on A40 corresponds to interpolation, since A40 lies within the composition range covered by the training dataset (NH_3 ratio up to 60%). To further probe extrapolation over mixture composition, an additional split is examined by training the LSTM-based RNN on A0, A20, and A40 (NH_3 ratio up to 40%), and then, test it on A60 (r=60%), as defined in Table 1. The resulting prediction shows pronounced deviations from the MD trajectory, indicating that the extrapolation capability of the present data-driven surrogate is limited under the current training coverage. Compositional changes can shift dominant reaction pathways and alter the formation dynamics of C_6 species, leading to a distribution shift that a purely data-driven model may not be able to generalize to without additional training coverage. Therefore, the present model is best suited for predictions within the trained parameter space.

Furthermore, comparing train–test discrepancies reveals clear differences in generalization abilities. The NRMSE values of the shallow and moderate FNNs increase by factors of 3.5 and 3.4, respectively, on the test dataset relative to the training dataset; in contrast, the LSTM-based RNN maintains an NRMSE value of 0.017 with only a negligible increase in error. The larger generalization gap of the feedforward baselines is consistent with overfitting under limited training coverage and reflects a notable architectural constraint: FNNs process each input independently and do not maintain an internal memory state, so temporal dependence can only be incorporated into such models through explicit feature engineering. In the present setting, the trajectory of C_6 species is not fully characterized by the instantaneous inputs provided to the FNNs (t and r), because the same time and NH_3 ratio can correspond to different effective reaction states depending on the preceding evolution (e.g., radical pools and precursor accumulation). The LSTM-based RNN can partially infer such latent state information from the preceding sequence via its gated memory, which helps retain earlier information that is predictive of subsequent trajectory changes; this leads to improved test performance and a smaller train–test gap (Fig. 9). In sum, these results support the usage of sequence-aware surrogates for ReaxFF MD time-series modeling, and indicate a practical pathway for reducing the computational costs associated with repeated reactive MD simulations.

4 Conclusions

The goal of this study is to explore sequence-aware neural networks in engine-related research, offering new insights and expanding application potential. Specifically, we leveraged data-driven surrogates to emulate ReaxFF MD simulations of C_6 species variation during the pyrolysis of C_2H_4/NH_3. The modeling performance of FNN baselines with shallow and moderate network depths was compared with an LSTM-based RNN. There are some meaningful observations in this study, which are presented as follows:

1. The ReaxFF MD simulations of C_2H_4/NH_3 pyrolysis are consistent with prior reports, particularly regarding the inhibitory effect of NH_3 on PAH formation and the formation of N-PAHs. For instance, the

concentration of CN radicals increases with the addition of NH_3, aligning with the observed suppression of N-PAH formation. Furthermore, the presence of NH_3 significantly influences the formation of initial aromatic rings through both odd- and even-carbon species chemistry. Overall, the MD simulation data are consistent with previous research, confirming their suitability for surrogate model development.

2. Feedforward baselines can assist in simulating C_2H_4/NH_3 pyrolysis when temporal information is introduced via explicit inputs/feature engineering; however, they face challenges in capturing detailed time-series variations of C_6 species. In the tests of FNNs, we found that increasing the network depth (i.e., the number of hidden layers) improved the fitting performance, but the difference in the training dataset remained limited. Although the test performance improved (R^2 increased from 0.738 to 0.867), the moderate FNN still struggled to reproduce the trajectory-level dynamics present in the MD trajectories. When evaluated on hold-out trajectories, these models show poorer generalization, indicating their limited ability to represent complex temporal dependencies in cases where the trajectory depends on preceding reaction history (i.e., an implicit state); therefore, such history must be encoded explicitly through additional feature engineering.

3. When comparing the LSTM-based RNN with the feedforward baselines, the LSTM model achieved higher predictive accuracy than both the shallow and moderate FNNs. While all models exhibited acceptable performance on the training dataset, the feedforward baselines generalized less effectively to the test dataset. In contrast, the LSTM-based RNN provided more robust predictions for long time-series trajectories, showcasing its suitability for combustion-chemistry surrogate modeling. This advantage became even more evident when the target trajectory exhibited history dependence or pronounced fluctuations that are difficult to characterize solely by instantaneous input–output mapping. Moreover, the LSTM-based RNN enables accurate emulation of C_2H_4/NH_3 pyrolysis trajectories with reduced computational cost. The improved performance of the LSTM-based RNN is attributed to its recurrent architecture with memory cells and gating mechanisms, which can capture temporal dependencies and complex patterns in sequential data.

Overall, in the context of engine combustion-chemistry modeling, we demonstrate how an LSTM-based RNN can achieve high predictive accuracy of time- and history-dependent trajectories with low computational cost. The proposed approach outperforms feedforward baselines, which rely on explicit time inputs. This case study underscores the potential of sequence-aware surrogate modeling in engine-related analysis and paves the way for broader applications in engine research.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 52306169) and the Zhejiang Provincial Natural Science Foundation of China (No. LQ23E060005).

Author contributions

Yuchao YAN developed the methodology, performed the analysis and validation, prepared the visualizations, and drafted the manuscript. Qiao HUANG and Tianfang XIE developed the software, curated the data, and contributed to the analysis, validation, and manuscript revision. Jinlong LIU conceived and designed the study, acquired the funding, supervised the project, and provided resources. All authors reviewed and approved the final manuscript.

Conflict of interest

Yuchao YAN, Qiao HUANG, Tianfang XIE, and Jinlong LIU declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aghbashlo M, Peng WX, Tabatabaei M, et al., 2021. Machine learning technology in biodiesel research: a review. *Progress in Energy and Combustion Science*, 85:100904. https://doi.org/10.1016/j.pecs.2021.100904

Aliramezani M, Koch CR, Shahbakhti M, 2022. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: a review and future directions. *Progress in Energy and Combustion Science*, 88:100967. https://doi.org/10.1016/j.pecs.2021.100967

Aramburu A, Guido C, Bares P, et al., 2024. Knock detection in spark ignited heavy duty engines: an application of machine learning techniques with various knock sensor locations. *Measurement*, 224:113860. https://doi.org/10.1016/j.measurement.2023.113860

Atkinson CM, Long TW, Hanzevack EL, 1998. Virtual Sensing: a Neural Network-Based Intelligent Performance and Emissions Prediction System for On-Board Diagnostics and Engine Control. SAE International, Warrendale, USA. https://doi.org/10.4271/980516

Badra JA, Khaled F, Tang M, et al., 2021. Engine combustion system optimization using computational fluid dynamics and machine learning: a methodological approach. *Journal of Energy Resources Technology*, 143(2):022306.

https://doi.org/10.1115/1.4047978

Bennett AM, Liu P, Li ZP, et al., 2020. Soot formation in laminar flames of ethylene/ammonia. *Combustion and Flame*, 220:210-218. https://doi.org/10.1016/j.combustflame.2020.06.042

Diao ST, Li HT, Yu MG, 2024. Atomic insights into the combustion mechanism of DME/NH_3 mixtures: a combined ReaxFF-MD and DFT study. *International Journal of Hydrogen Energy*, 80:743-753. https://doi.org/10.1016/j.ijhydene.2024.07.189

Huang Q, Liu JL, 2024. Preliminary assessment of the potential for rapid combustion of pure ammonia in engine cylinders using the multiple spark ignition strategy. *International Journal of Hydrogen Energy*, 55:375-385. https://doi.org/10.1016/j.ijhydene.2023.11.136

Huang Q, Liu JL, Ulishney C, et al., 2022. On the use of artificial neural networks to model the performance and emissions of a heavy-duty natural gas spark ignition engine. *International Journal of Engine Research*, 23(11):1879-1898. https://doi.org/10.1177/14680874211034409

Huang Q, Yang RM, Liu JH, et al., 2024. Investigation of the mechanism behind the surge in nitrogen dioxide emissions in engines transitioning from pure diesel operation to methanol/diesel dual-fuel operation. *Fuel Processing Technology*, 264:108131. https://doi.org/10.1016/j.fuproc.2024.108131

Huang Q, Yang RM, Liu JH, et al., 2025a. CFD-based investigation of ammonia combustion and slip behavior in an ammonia-diesel dual-fuel engine. *Journal of the Energy Institute*, 122:102217. https://doi.org/10.1016/j.joei.2025.102217

Huang Q, Xie TF, Liu JL, 2025b. Machine learning-assisted reconstruction of in-cylinder pressure in internal combustion engines under unmeasured operating conditions. *Energies*, 18(19):5235. https://doi.org/10.3390/en18195235

Kailasanathan RKA, Yelverton TLB, Fang TG, et al., 2013. Effect of diluents on soot precursor formation and temperature in ethylene laminar diffusion flames. *Combustion and Flame*, 160(3):656-670. https://doi.org/10.1016/j.combustflame.2012.11.004

Kamat AM, van Duin ACT, Yakovlev A, 2010. Molecular dynamics simulations of laser-induced incandescence of soot using an extended ReaxFF reactive force field. *The Journal of Physical Chemistry A*, 114(48):12561-12572. https://doi.org/10.1021/jp1080302

Kamat S, Kapase P, Jain P, et al., 2023. Modeling Virtual Sensor for Engine Nitrogen Oxides Using Variants of Artificial Neural Networks. SAE International, Warrendale, USA. https://doi.org/10.4271/2023-01-5042

Kowalik M, Ashraf C, Damirchi B, et al., 2019. Atomistic scale analysis of the carbonization process for C/H/O/N-based polymers with the ReaxFF reactive force field. *The Journal of Physical Chemistry B*, 123(25):5357-5367. https://doi.org/10.1021/acs.jpcb.9b04298

Kumar A, Yashwanth VH, Kumar R, et al., 2024. Machine Learning Approach to Control Thermal Strategies and Mitigate Sensor Failure Penalty on Emissions. SAE International, Warrendale, USA. https://doi.org/10.4271/2024-28-0170

Le Cornec CMA, Molden N, van Reeuwijk M, et al., 2020. Modelling of instantaneous emissions from diesel vehicles with a special focus on NO_x: insights from machine learning techniques. *Science of the Total Environment*, 737:139625. https://doi.org/10.1016/j.scitotenv.2020.139625

Li J, Zhou Q, He X, et al., 2023. Data-driven enabling technologies in soft sensors of modern internal combustion engines: perspectives. *Energy*, 272:127067. https://doi.org/10.1016/j.energy.2023.127067

Li RZ, Herreros JM, Tsolakis A, et al., 2020. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel*, 280:118589. https://doi.org/10.1016/j.fuel.2020.118589

Li RZ, Herreros JM, Tsolakis A, et al., 2021. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel*, 304:121437. https://doi.org/10.1016/j.fuel.2021.121437

Liu JL, Wang HF, 2022. Machine learning assisted modeling of mixing timescale for LES/PDF of high-Karlovitz turbulent premixed combustion. *Combustion and Flame*, 238:111895. https://doi.org/10.1016/j.combustflame.2021.111895

Liu JL, Huang Q, Ulishney C, et al., 2022. Comparison of random forest and neural network in modeling the performance and emissions of a natural gas spark ignition engine. *Journal of Energy Resources Technology*, 144(3):032310. https://doi.org/10.1115/1.4053301

Liu L, Chen WL, Zhu Q, et al., 2023. Inhibitory mechanisms of ammonia addition on soot formation during n-decane pyrolysis. *Fuel*, 350:128695. https://doi.org/10.1016/j.fuel.2023.128695

Mishra C, Subbarao PMV, 2021. Design, development and testing a hybrid control model for RCCI engine using double Wiebe function and random forest machine learning. *Control Engineering Practice*, 113:104857. https://doi.org/10.1016/j.conengprac.2021.104857

Petrucci L, Ricci F, Mariani F, et al., 2020. Engine Knock Evaluation Using a Machine Learning Approach. SAE International, Warrendale, USA. https://doi.org/10.4271/2020-24-0005

Silva M, Mohan B, Badra J, et al., 2023. DoE-ML guided optimization of an active pre-chamber geometry using CFD. *International Journal of Engine Research*, 24(7):2936-2948. https://doi.org/10.1177/14680874221135278

Sok R, Jeyamoorthy A, Kusaka J, 2024. Novel virtual sensors development based on machine learning combined with convolutional neural-network image processing-translation for feedback control systems of internal combustion engines. *Applied Energy*, 365:123224. https://doi.org/10.1016/j.apenergy.2024.123224

Tang QX, Wang MD, You XQ, 2019. Effects of fuel structure on structural characteristics of soot aggregates. *Combustion and Flame*, 199:301-308. https://doi.org/10.1016/j.combustflame.2018.10.033

Thompson AP, Aktulga HM, Berger R, et al., 2022. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171. https://doi.org/10.1016/j.cpc.2021.108171

Torregrosa AJ, Broatch A, Olmeda P, et al., 2021. Numerical Estimation of Wiebe Function Parameters Using Artificial Neural Networks in SI Engine. SAE International, Warrendale, USA. https://doi.org/10.4271/2021-01-0379

Wang Y, Mao Q, Wang ZY, et al., 2023. A ReaxFF molecular dynamics study of polycyclic aromatic hydrocarbon oxidation assisted by nitrogen oxides. *Combustion and Flame*, 248:112571. https://doi.org/10.1016/j.combustflame.2022.112571

Xing ZH, Chen C, Jiang X, 2023. A molecular investigation on the mechanism of co-pyrolysis of ammonia and biodiesel surrogates. *Energy Conversion and Management*, 289:117164. https://doi.org/10.1016/j.enconman.2023.117164

Yan YC, Xie TF, Liu JL, 2025. Rapid and accurate prediction of molecular dynamics simulations using physics-informed LSTM networks in engine emission analysis: a case study of C_3H_6/NH_3 pyrolysis for PAH formation. *Journal of the Energy Institute*, 120:102090. https://doi.org/10.1016/j.joei.2025.102090

Yao S, Wang B, Kronenburg A, et al., 2020. Modeling of sub-grid conditional mixing statistics in turbulent sprays using machine learning methods. *Physics of Fluids*, 32(11):115124. https://doi.org/10.1063/5.0027524

Yoon K, Rahnamoun A, Swett JL, et al., 2016. Atomistic-scale simulations of defect formation in graphene under noble gas ion irradiation. *ACS Nano*, 10(9):8376-8384. https://doi.org/10.1021/acsnano.6b03036

Zaher MH, Chu C, Dadsetan M, et al., 2023. Experimental and numerical investigation of soot growth and inception in an ammonia-ethylene flame. *Proceedings of the Combustion Institute*, 39(1):929-937. https://doi.org/10.1016/j.proci.2022.07.175

Zhang K, Xu YS, Yu RH, et al., 2024. ReaxFF molecular dynamics study of N-containing PAHs formation in the pyrolysis of C_2H_4/NH_3 mixtures. *Combustion and Flame*, 270:113774. https://doi.org/10.1016/j.combustflame.2024.113774

Zhang P, Zhang K, Cheng XB, et al., 2022. Analysis of inhibitory mechanisms of ammonia addition on soot formation: a combined ReaxFF MD simulations and experimental study. *Energy & Fuels*, 36(19):12350-12364. https://doi.org/10.1021/acs.energyfuels.2c02206

Zhang P, Wu H, Zhang K, et al., 2023. Decoupling effects of C_3H_3/C_4H_5/i-C_4H_5/CN radicals on the formation and growth of aromatics: a ReaxFF molecular dynamics study. *Journal of Aerosol Science*, 171:106185. https://doi.org/10.1016/j.jaerosci.2023.106185

Zhao J, Lin YY, Huang K, et al., 2020. Study on soot evolution under different hydrogen addition conditions at high temperature by ReaxFF molecular dynamics. *Fuel*, 262:116677. https://doi.org/10.1016/j.fuel.2019.116677