Response to the comments of Reviewer #1 (EGUSPHERE-2024-2496)

The mixing states of black carbon are widely measured by the single-particle soot photometer (SP2) instrument. This study employed a machine learning (ML) based method, light gradient boosting machine (LightGBM), to process the scattering and incandescence signals of SP2 and to retrieve the mixing states of particles. ML based method performs more efficiently than the traditional Leading-Edge-Only (LEO) approach with quite consistent retrieval outcomes. The relative importances of the selected signal features in retrieving the particle microphysical properties were studied by SHapley Additive exPlanation (SHAP) method. The authors stated that this ML based method has the potential to be a reliable noise-resistant approach to analyze the SP2 data. My major comments are attached below.

Response: We sincerely thank the reviewer #1 for the insightful suggestions and constructive comments. In response, we have thoroughly revised the manuscript and prepared a detailed, point-by-point response to all comments and questions raised. In the revised manuscript, we have added conceptual definitions for different particle types and clarified their classification criteria. Additionally, we provide a more detailed description of the feature dataset construction process, explaining the intrinsic connection between the feature signals used in the model and the original signals, as well as their scientific significance. Here are our point-to-point responses.

Main Comments:

1) Fig.3, 4, & 5 compare the predicted particle microphysical properties with those named "actual values". How did you define the "actual values", and how they were obtained? Were they the outputs from LEO approach or did you use any other particle sizer instruments to measure the "actual values" of particle size? According to the meaning of "actual", this value should be regarded as the ground-truth, or more reliable measurements of the particle size.

Response: Thank you for pointing this out. These particle microphysical properties named "actual values" are derived through physical inversion methods, as detailed in the "Construction of label dataset" section of our manuscript. Specifically, for internally mixed BC, these values are outputs from the LEO approach. We employed the machine learning model to learn the mapping relationship between the input (SP2 signals) and output (microphysical properties) data. Based on the developed inversion model, when we input SP2 signals, the model can predict the corresponding microphysical properties. These predicted values are then compared with the "actual values" obtained through the physical inversion methods to evaluate the model's performance, as shown in Fig. 3, 4, & 5.

We appreciate your reminder, and we realize that using the "actual values" in the manuscript is not appropriate. In response, we have revised it to “observed values”.

2) Fig. 7 shows the robustness of the ML-based retrievals by comparing the retrieved particle size with the one obtained by LEO approach. LEO method, because it utilizes part of the scattering signal, has the possibility to mischaracterize the particles with different sizes. Comments: I agree that the retrieval accuracy will be improved with more observational constraints or signal feature inputs. However, the reason that LEO method only utilizes the threshold portion but not the entire scattering signals is that the loss or vaporization of particle coatings happens after the particle started to absorb laser energy in SP2. The entire scattering signal function doesn't reflect the scattering properties of the original mixing states of BC-containing particles (coating thickness, BC core size, etc.). Therefore, LEO method utilizes the threshold portion of the signal when the particle properties (size) doesn't change significantly yet in SP2. Though the proposed ML-based method utilizes more signal feature, it doesn't necessarily reflect the true size of the original coated particles.

Response: Thank you for your comment. As you described, the LEO fitting uses the leading edge of the scattering signal, which corresponds to the stage before the coating evaporates, aiming to capture the original characteristics of the particle when it has not yet significantly changed in the SP2. While, as we pointed out in the discussion of Fig. 7 (which corresponds to Fig. 9 in the revised manuscript), the leading edge of the scattering signal is close to the baseline and thus more susceptible to noise interference, which may increase uncertainty in the LEO fitting.

As the BC core absorbs laser energy, the coating evaporates upon reaching its boiling point, the scattering signal gradually deviates from a Gaussian distribution. Although this distorted part of the signal cannot directly reflect the initial physical characteristics of the particle, the changes in the signal are not entirely random. This signal can be considered a complex function of the BC core diameter (D_c) and the entire particle diameter (D_p). In fact, previous studies have tried to use these changes in the scattering signal to depict the variation of particle scattering cross-section in the SP2 or the mixing state of BC (Moteki et al., 2014; Moteki and Kondo, 2008; Sedlacek et al., 2012).

Considering that the leading-edge signal only accounts for a small portion of the entire scattering signal, the information it can provide is limited. Therefore, this study attempts to input the complete scattering signal change into a machine learning model, aiming to parse the information contained in the subsequent signal. This approach can make more comprehensive use of the information expressed by the signal, thereby providing more robust results.

3) Table 2: This table shows the hyperparameters for each particle type. What are the “Internally mixed BC” and “BC-containing particle” by definition? BC-containing particle is a subset of internally mixed BC in my opinion. Then why did you use different hyperparameters for them?

Response: Thank you for your question. We have provided more detailed explanations of the definitions for different particle types in the revised manuscript. Lines 109 to 122 in the revised manuscript reflect the specific revisions:

"In this study, ambient particles measured by SP2 are classified into pure scattering particles and BC-containing particles. ***Pure scattering particles are those that only scatter light without significant absorption, while BC-containing particles, which contain refractory BC (rBC), both scatter and absorb light. BC-containing particles are further subdivided into externally mixed BC and internally mixed BC. Externally mixed BC refers to freshly emitted BC particles that have not yet mixed with other aerosol components, while internally mixed BC describes BC that has undergone atmospheric aging processes and incorporated with other materials (Oshima et al., 2009).*** *Operationally, we differentiate the pure scattering particle and BC-containing particle depending on whether it has the incandescence signal. ...* ***The incandescence signal peak occurs when all non-BC material has evaporated and the BC reaches its incandescence temperature, thus the magnitude of Δt correlates with the thickness of the coating on BC particles: a larger Δt corresponds to a thicker coating that takes longer to evaporate. By examining the distribution of Δt values in the SP2 measurements, as illustrated in Fig. 2d (Sedlacek et al., 2012; Subramanian et al., 2010; Zhang et al., 2016)****, BC-containing particles with $\Delta t < 2\ \mu s$ are classified as externally mixed BC (Fig. 2b), while those with $\Delta t \geq 2\ \mu s$ are categorized as internally mixed BC (Fig. 2c)."*

In addition, the hyperparameter tuning is optimized for the specific tasks of each model, rather than using a universal setting. By adjusting the hyperparameters, we establish the mapping between the model's inputs and outputs. Since each model uses different features and inverses different microphysical properties, the hyperparameters used in model construction also vary accordingly.

Line-by-line comments:

1) Line 73: "quality of the refractory components", what does "quality" mean here?

Response: Thank you for pointing this out. The "quality" here refers to the mass of the refractory BC. For clarity, we have replaced it with the word "mass" in the revised manuscript.

Line 84 in the revised manuscript:

"The intensity of this thermal radiation depends on the composition and ***mass*** *of the refractory components, regardless of the particle morphology and mixing state (Schwarz et al., 2006; Slowik et al., 2007)."*

2) Line 76: Add reference here (Gao, R. S., et al., 2007, Aerosol Science & Technology)

Response: Thank you for your reminder. We have added the reference in the corresponding position.

Line 88 in the revised manuscript:

*"Particle size, therefore, can be measured based on the amount of light they scatter from the laser, which exhibits a Gaussian dependence with time **(Gao et al., 2007)**."*

3) Line 124: "45-dimensional scattering signals". What is the relationship between the 45- dimensional signals and the abovementioned "100-dimensional signals of SP2 data" in line 94, and 90-dimenional feature data in line 137. It would be better to provide additional contexts/descriptions of the principles of signal feature selection for optical retrievals.

Response: Thank you for your comment. The SP2 signal is recorded based on the elapsed time, with each time window corresponding to information about a single particle. For each particle, the corresponding original scattering signal and incandescence signal are both 100-dimensional. The position of particles within the instrument is not known in advance. Among SP2's four detectors, there is a two-element APD (TEAPD) detector. This detector has a gap perpendicular to the particle's direction of motion, resulting in a notch in the TEAPD signal, as shown in Fig. R1a. Given the stability of SP2's optical alignment and constant sample flow rate, this notch provides a precise time reference for a particle's position within the instrument. In practice, the signal from the leading element is inverted, transforming the notch into a zero-crossing point (Fig. R1b) (Gao et al., 2007). Since SP2 simultaneously records data from all four detector channels, this time reference is valid for the signals from the other three detectors as well.

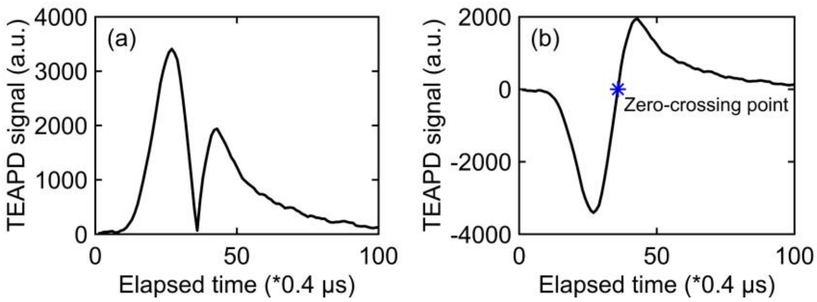

Figure R1. (a) The original scattering signal measured by TEAPD before the signal from the leading element is inverted. (b) The TEAPD signal obtained by SP2, with the blue asterisk indicating the position of the zero-crossing point.

For pure scattering particles, we locate the zero-crossing point in the scattering signal and then extract 22 data points both before and after it, creating a 45-dimensional feature dataset (Fig. R2a). In this newly constructed 45-dimensional feature signal, particles are positioned at consistent locations within the instrument for each corresponding dimension, eliminating the influence of laser intensity distribution on the scattering signal.

When inverting the D_c of BC-containing particles, since the peak intensity of the incandescence signal is positively correlated with the mass of the refractory BC component in the particle, the peak of the incandescence signal is selected as a reference point, from which 22 data points are extracted both preceding and following this point (Fig. R2b), yielding a 45-dimensional feature dataset. This method ensures the incandescence signal peaks from different BC-containing particles are positioned at the same dimension within the feature dataset, facilitating direct comparisons between particles. The optical properties of externally mixed BC are also determined by the refractory BC component, so the same feature selection method is adopted.

For internally mixed BC, its size and optical cross-section characteristics are reflected by both the scattering and incandescence signals. We select scattering features for internally mixed BC using the same approach as for pure scattering particles. Considering that the relative relationship between the incandescence and scattering signals in the original signal can reflect particle characteristics, the incandescence signal is selected with 22 data points before and after the zero-crossing point in a similar way (Fig. R2c). The 90-dimensional feature signal composed of the selected 45-dimensional scattering signal and 45-dimensional incandescence signal is used as the input feature for the internally mixed BC inversion model.

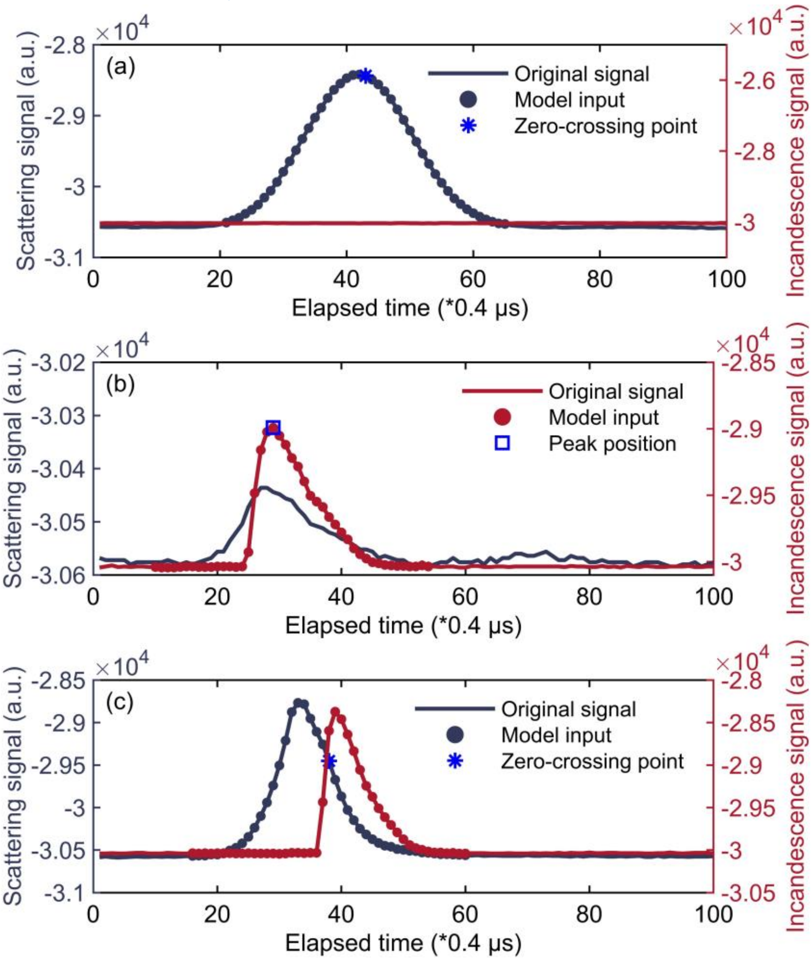

Figure R2. Relationship between the original SP2 signals (line plots) and the feature signals used in machine learning model construction (scatter plots) for different particle types: (a) pure scattering particles; (b) externally mixed BC; (c) internally mixed BC. The method for selecting feature signals used in inverting the core diameter (D_c) of BC-containing particles is identical to that used for externally mixed BC.

We have provided additional descriptions of the principles of signal feature selection for different particle types as suggested. The Figs. R1 and R2 presented here have also been incorporated into the revised manuscript and Supplement Information to further illustrate these principles. Specifically, Fig. R1 has been added as Fig. 3 in the revised manuscript, while Fig. R2 has been included as Fig. S1 in the Supplement Information.

Lines 141 to 143:

*"**The SP2 signal is recorded based on the elapsed time, with each time window corresponding to information about a single particle. For each particle, the original scattering signal and incandescence signal are both 100-dimensional. The position of particles within the instrument is not known in advance.**"*

Lines 154 to 162:

*" **As mentioned in Sect. 2.2, one of the four detectors in the SP2 is a split APD detector. This detector has a gap perpendicular to the particle's direction of motion, resulting in a notch in the TEAPD signal, as shown in Fig. 3a. Given the stability of SP2's optical alignment and constant sample flow rate, this notch provides a precise time reference for a particle's position within the instrument. In practice, the signal from leading element is inverted, transforming the notch into a zero-crossing point (Fig. 3b) (Gao et al., 2007).** Since SP2 simultaneously records data from all four detector channels, this time reference is valid for the signals from the other three detectors as well. **We locate the zero-crossing point in the scattering signal and then extract 22 data points both before and after it, creating a 45-dimensional feature dataset (Fig. S1a). Through this standardization, the differences in signal intensity can be accurately attributed to the inherent physical properties of the particles.***

Lines 168 to 173:

Based on this characteristic, the peak of the incandescence signal is selected as a reference point, from which 22 data points are extracted both preceding and following this point (Fig. S1b), yielding a 45-dimensional feature dataset used for inverting the D_c of BC-containing particles. This method ensures the incandescence signal peaks from different BC-containing particles are positioned at the same dimension within the feature dataset, facilitating direct comparisons between particles while preserving comprehensive information about the incandescence process.

Lines 184 to 189:

*"Compared with other particle types, the internally mixed BC has a more complex structure. ... **Simultaneously, considering that the relative relationship between the original incandescence and scattering signals can reflect particle characteristics, such as coating thickness (Moteki and Kondo, 2007; Schwarz et al., 2006; Subramanian et al., 2010), the incandescence signal is selected with 22 data points before and after the zero-crossing point in a similar way (Fig. S1c). The feature extraction process yields a 90-dimensional feature dataset, comprising 45-dimensional scattering signal and 45-dimensional incandescence signal, ensuring that we can***

comprehensively capture the key characteristics of internally mixed BC."

4) Line 191: What is "GridSearchCV function".

Response: Thank you for your question. GridSearchCV is an automated hyperparameter tuning technique that combines grid search with cross-validation (Ahmad et al., 2022). In this process, a range of values is defined for each hyperparameter, and then all possible combinations are systematically tried. It uses K-fold cross-validation to evaluate the performance of each hyperparameter combination, eliminating the need for manual adjustments by automatically executing tests for all possible combinations. GridSearchCV assesses performance based on the average results of cross-validation and ultimately selects the hyperparameter combination that produces the best performance. The main advantage of this method lies in its ability to comprehensively and systematically explore the hyperparameter space while providing reliable estimates of the model's generalization capability. We have added a brief explanation of the GridSearchCV function in the revised manuscript.

Lines 251 to 254:

"The GridSearchCV with 5-fold cross-validation is employed to optimize the hyperparameter configuration of the LightGBM inversion model. ***This comprehensive approach facilitates an exhaustive search for the optimal hyperparameter combination within a predefined parameter space (Ahmad et al., 2022). By utilizing cross-validation, the methodology effectively mitigates the risk of overfitting and provides robust estimates of the model's generalization performance.***"

5) Line 191: Please check the grammar of the sentence starting with "Based on". What is the subject of this sentence?

Response: Thank you for your careful review and for pointing out this grammatical issue. The sentence beginning with "Based on" lacks a clear subject, which affects its grammatical structure. We have revised the sentence to "The GridSearchCV with 5-fold cross-validation is employed to optimize the hyperparameter configuration of the LightGBM inversion model." [Page 10, Line 251]

Reference

Ahmad, G. N., Fatima, H., and Saidi, A. S.: Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV, 10, 2022.

Gao, R. S., Schwarz, J. P., Kelly, K. K., Fahey, D. W., Watts, L. A., Thompson, T. L., Spackman, J. R., Slowik, J. G., Cross, E. S., Han, J.-H., Davidovits, P., Onasch, T. B., and Worsnop, D. R.: A Novel Method for Estimating Light-Scattering Properties of Soot Aerosols Using a Modified Single-Particle Soot Photometer, Aerosol Science and Technology, 41, 125–135, https://doi.org/10.1080/02786820601118398, 2007.

Moteki, N. and Kondo, Y.: Effects of Mixing State on Black Carbon Measurements by Laser-Induced Incandescence, Aerosol Science and Technology, 41, 398–417, https://doi.org/10.1080/02786820701199728, 2007.

Moteki, N. and Kondo, Y.: Method to measure time-dependent scattering cross sections of particles evaporating in a laser beam, Journal of Aerosol Science, 39, 348–364, https://doi.org/10.1016/j.jaerosci.2007.12.002, 2008.

Moteki, N., Kondo, Y., and Adachi, K.: Identification by single-particle soot photometer of black carbon particles attached to other particles: Laboratory experiments and ground observations in Tokyo, JGR Atmospheres, 119, 1031–1043, https://doi.org/10.1002/2013JD020655, 2014.

Oshima, N., Koike, M., Zhang, Y., Kondo, Y., Moteki, N., Takegawa, N., and Miyazaki, Y.: Aging of black carbon in outflow from anthropogenic sources using a mixing state resolved model: Model development and evaluation, J. Geophys. Res., 114, 2008JD010680, https://doi.org/10.1029/2008JD010680, 2009.

Schwarz, J. P., Gao, R. S., Fahey, D. W., Thomson, D. S., Watts, L. A., Wilson, J. C., Reeves, J. M., Darbeheshti, M., Baumgardner, D. G., Kok, G. L., Chung, S. H., Schulz, M., Hendricks, J., Lauer, A., Kärcher, B., Slowik, J. G., Rosenlof, K. H., Thompson, T. L., Langford, A. O., Loewenstein, M., and Aikin, K. C.: Single-particle measurements of midlatitude black carbon and light-scattering aerosols from the boundary layer to the lower stratosphere, J. Geophys. Res., 111, 2006JD007076, https://doi.org/10.1029/2006JD007076, 2006.

Sedlacek, A. J., Lewis, E. R., Kleinman, L., Xu, J., and Zhang, Q.: Determination of and evidence for non-core-shell structure of particles containing black carbon using the Single-Particle Soot Photometer (SP2), Geophysical Research Letters, 39, 2012GL050905, https://doi.org/10.1029/2012GL050905, 2012.

Slowik, J. G., Cross, E. S., Han, J.-H., Davidovits, P., Onasch, T. B., Jayne, J. T., Williams, L. R., Canagaratna, M. R., Worsnop, D. R., Chakrabarty, R. K., Moosmüller, H., Arnott, W. P., Schwarz, J. P., Gao, R.-S., Fahey, D. W., Kok, G. L., and Petzold, A.:

An Inter-Comparison of Instruments Measuring Black Carbon Content of Soot Particles, Aerosol Science and Technology, 41, 295–314, https://doi.org/10.1080/02786820701197078, 2007.

Subramanian, R., Kok, G. L., Baumgardner, D., Clarke, A., Shinozuka, Y., Campos, T. L., Heizer, C. G., Stephens, B. B., de Foy, B., Voss, P. B., and Zaveri, R. A.: Black carbon over Mexico: the effect of atmospheric transport on mixing state, mass absorption cross-section, and BC/CO ratios, Atmos. Chem. Phys., 2010.

Zhang, Y., Zhang, Q., Cheng, Y., Su, H., Kecorius, S., Wang, Z., Wu, Z., Hu, M., Zhu, T., and Wiedensohler, A.: Measuring the morphology and density of internally mixed black carbon with SP2 and VTDMA: new insight into the absorption enhancement of black carbon in the atmosphere, Atmospheric Measurement Techniques, 9, 1833–1843, 2016.