Multi-objective Bayesian Optimization for the Retrieval of Aggregated Aerosol Structures from Microscopic Images

Abhishek Singh, Smruti Ranjan Jena, Abhishek Gupta, Thaseem Thajudeen

PII: S0021-8502(25)00033-3

DOI: https://doi.org/10.1016/j.jaerosci.2025.106556

Reference: AS 106556

To appear in: *Journal of Aerosol Science*

Received Date: 15 November 2024

Revised Date: 5 February 2025

Accepted Date: 28 February 2025

Please cite this article as: Singh A, Jena SR, Gupta A, Thajudeen T, Multi-objective Bayesian Optimization for the Retrieval of Aggregated Aerosol Structures from Microscopic Images, *Journal of Aerosol Science*, https://doi.org/10.1016/j.jaerosci.2025.106556.

Multi-objective Bayesian Optimization for the Retrieval of Aggregated Aerosol Structures from Microscopic Images

Abhishek Singh[1], Smruti Ranjan Jena[2], Abhishek Gupta[1], Thaseem Thajudeen[1*]

[1]School of Mechanical Sciences, Indian Institute of Technology Goa, Goa 403401, India.

[2]Department of Aerospace Engineering, Indian Institute of Technology Kanpur, Uttar Pradesh 208016, India.

* Corresponding author. Tel: 91-832-2490-859

E-*mail address*: thaseem@iitgoa.ac.in

Office of the Dean (R&D)
Administrative Block,
Indian Institute of Technology Goa,
Goa College of Engineering Campus,
Farmagudi, Goa. 403401. India.

Highlights

- Multi-objective optimization to predict 3-dimensional morphology from microscopic images
- Bayesian Optimization for faster convergence
- Retrieval time reduced by up to one-tenth of existing methods.
- Accurately predicts 3-dimensional aggregate structures, achieving less than 10% deviation in mobility diameter.
- Retrieval of aggregate structures with monomers exhibiting both polydispersity and overlap.

Abstract

Aerosol particles are increasingly recognized for their significant impacts on human health and climate. Often found in aggregated form, the morphology of these particles plays a crucial role in influencing their physicochemical properties. Owing to the sub-micron size, electron microscopy is the most commonly used technique to visualize the aggregated particles. In a prior study, we proposed a combination of forward modelling coupled with optimization techniques for the prediction of 3-dimensional structures from microscopic images. Here, we extend the methodology to a multi-objective optimization approach for the specific cases where aggregated particles are classified and sampled based on specific properties such as mobility diameter, aerodynamic diameter etc. The comparison of 2-dimensional features of the microscopic image with the projections of computationally generated aggregates forms the first objective function, while the comparison of the measured 3-dimensional property, mobility diameter, is used as the second objective function. The estimation of the mobility diameter often requires the calculation of the hydrodynamic radius (R_h) and the orientationally averaged projected area (PA), which can be computationally expensive for larger aggregates and for frequent calculations. Bayesian optimization is used for the retrieval process, as it can provide much faster convergence with significantly fewer function evaluations compared to metaheuristic algorithms. The multi-objective Bayesian optimization-based retrieval algorithm has been validated using synthetically generated and experimentally collected microscopic images. The process is found to be about 5 to 10 times faster than previously reported methods. The algorithm is further extended to retrieve aggregates with polydisperse and overlapping monomers. The retrieval process demonstrated strong accuracy, with fractal parameters showing around 10-15% error compared to the original values. This includes a mobility diameter difference of less than 10%, indicating high similarity between retrieved and input structures. Furthermore, tests are conducted on welding fume particles of varying mobility diameters, where retrieved structures consistently exhibited mobility diameters within a 10% difference from original values.

Keywords: Multi-Objective optimization, Bayesian optimization, Aggregated aerosol structures, Morphological analysis

1. Introduction

Aerosol particles, originating from various sources such as industrial combustion processes and vehicular emissions, are a major concern due to their significant impact on local and global weather patterns (Cao, 2017; Lighty et al., 2000; Shiraiwa et al., 2017). These particles typically exhibit a wide range of sizes, compositions, and morphologies, which play a crucial role in shaping environmental factors and affecting human health.(David et al., 2019; Gaeckle et al., 2020). Aerosol particles, specifically those generated from combustion sources, frequently exist as clusters rather than individual particles due to continuous collisions and subsequent coagulation (M. Eggersdorfer & Pratsinis, 2014; Filippov et al., 2000; Meakin et al., 1984). The clusters, also referred to as aggregates, often display fractal-like structures, exhibiting self-similarity in their arrangement. However, unlike ideal fractals, these aggregates show self-similarity only over a limited range of length scales and are referred to as quasi-fractal aggregates (Filippov et al., 2000; Forrest & Witten, 1979). Quasi-fractal structures are found to follow the scaling law given by:

$$N = k_f \left(\frac{R_g}{r_{p_{geo}}}\right)^{D_f} \quad (1)$$

where N is the number of particles or monomers with length scales comprising the radius of gyration (R_g). D_f is the fractal dimension ranging from 1 (for linear aggregates) to 3 (fully compact structure), and k_f is the fractal prefactor, spanning from lower values around 0.3 to higher values exceeding 7 (China et al., 2015; Sipkens et al., 2023; Sorensen & Roberts, 1996). r_{pgeo} is the geometric mean radius of monomers, which equals to monomer radius 'a' for the case of monodisperse particles. Aggregates encountered in practice often consist of nearly spherical, polydisperse primary particles, commonly described by lognormal distributions. These details in the scaling law to include polydispersed size distribution using the $r_{p,geo}$ and the geometric standard deviation $\sigma_{p,geo}$ as:

$$\log\left(r_{p_{geo}}\right) = \frac{\sum_{i=1}^{N} \log(r_i)}{N} \quad (2)$$

$$\log\left(\sigma_{p_{geo}}\right) = \sqrt{\frac{\sum_{i=1}^{N} \left(\log(r_i) - \log(r_{p,geo})\right)^2}{N}} \quad (3)$$

where r_i is the radius of the i^{th} particle, and N is the total number of monomers. For monodisperse primary particles, the RHS term in Eq. (3) becomes zero, yielding $\sigma_{p,geo}$ =1. This allows the equation to be applicable to both monodisperse and polydisperse size distributions of monomers in an aggregate.

Morphology of the aggregates can significantly influence their properties, including surface reactivity, optical properties, and transport and deposition (China et al., 2015; Yon et al., 2021). Morphological parameters (e.g., effective density, monomer number, fractal dimension) of size-resolved soot particles from various combustion sources were investigated in some recent studies, and the results revealed size-dependent variations, highlighting the differences in soot particle properties across sources (Li et al., 2024; Pang et al., 2022, 2023). Given their nanoscale size, electron microscopy-based image analysis is among the most widely used techniques for studying and visualizing these structures (Park et al., 2003; Thajudeen, Jeon, et al., 2015).

In a typical image analysis method, 2-dimensional features from microscopic images are used to retrieve 3-dimensional properties of aggregates, including primary particle size distributions, number of monomers (N), and fractal dimension (D_f) (Bescond et al., 2014; Chakrabarty et al., 2011a; Park et al., 2004). Kruis et al. (Einar Kruis et al., 1994) developed an automated particle recognition technique to estimate particle size distribution and Fractal dimensions of the aggregates from Transmission Electron Microscope (TEM) images. An image characterization process was proposed by Brasil (Brasil et al., 1999) for extracting 2-dimensional features (such as projected area, perimeter, and maximum length) of microscopic images to estimate 3-dimensional properties (N, k_f, D_f). Several studies have attempted to develop relations between 3-dimensional properties, like monomer numbers and fractal dimensions, and 2-dimensional projection properties, such as 2-dimensional radius of gyration and maximum length (Chakrabarty et al., 2011b, 2011a). Most of these studies have relied on developing regression-based equations to represent the relationship. Hogan and coworkers (Thajudeen, Jeon, et al., 2015) introduced a technique to predict the 3-dimensional structures of aggregates from TEM images by comparing them with a database of test images created from computationally generated aggregates. The method was later extended to various applications, including the study on the growth and absorption rates of iron oxide nanoparticles (Jeon et al., 2016) and morphological analysis of emissions from electrosurgical pencils (Qiao et al., 2020). However, the efficacy of this method depends on the comprehensiveness of the database.

Considering these limitations, a retrieval process combining optimization methods and forward modelling techniques was developed (Singh & Thajudeen, 2023). In this method, the features obtained from microscopic images were compared to the projections of candidate aggregates generated by FracVAL, an open-source forward modelling method (Morán et al., 2019). Optimization techniques were used to minimize the difference between the 2-

dimensional features of microscopic images and candidate projections, using an objective function based on six image features: projected area, perimeter, maximum length, width, 2-dimensional radius of gyration, and 2-dimensional fractal dimension. This method was then improved using Machine learning (ML) techniques (Singh et al., 2024) to narrow down the search space for optimization, thereby accelerating the retrieval process. The proposed method was validated by comparing the retrieved structure against available information on the sampled aggregate using 3-dimensional aerosol properties, such as mobility and aerodynamic diameters.

In certain cases, aggregate samples are collected based on predefined aerosol properties, as is the case when a Differential Mobility Analyzer (DMA) is used to classify aerosol particles based on their mobility diameter (d_m) (Lapuerta et al., 2003; Park et al., 2003; Trivanovic et al., 2019). Prior retrieval methods typically compare these 3-dimensional properties post-retrieval. However, incorporating these known properties directly into the retrieval process enhances its efficiency, as each iteration utilizes this information to guide the optimization. This approach ensures that the retrieved structure aligns with both the 2-dimensional projection features and the 3-dimensional properties of aerosol particles measured during sampling. Building on the previous methods, we propose a novel multi-objective optimization-based retrieval framework by incorporating known 3-dimensional properties as an additional objective function alongside 2-dimensional image features.

It could also be possible that an aggregate structure may be inaccurately designated as optimal if its projection closely resembles the input image due to a favourable viewing angle despite differences in its true 3-dimensional form. For example, in an ideal case, various candidate aggregate structures may yield identical projection properties, resulting in equivalent objective function values. It is then challenging to determine the best structure. By incorporating an additional objective function, such as a 3-dimensional property, the retrieval process can identify the structure with the closest 3-dimensional property compared to the aggregate corresponding to the microscopic image. This approach ensures that the structure with the closest match in 3-dimensional properties shall be selected as the optimal structure. Additionally, multiple aggregates collected at a fixed mobility diameter may exhibit varying projection features due to differences in structural shape. However, if these aggregates do not match the input image projections, they will be discarded. Thus, a multi-objective optimization (MOO) approach is particularly advantageous for retrieval, as it ensures that both 2-dimensional and 3-dimensional properties are considered in tandem.

An important point to consider is the need for an MOO approach rather than a single combined objective function for 2-dimensional features and 3-dimensional properties. The rationale for this choice lies in the varying importance of properties such as mobility diameter and projected features. For a single objective optimization, six 2-dimensional features and one 3-dimensional property must be combined into a single objective with a total of seven parameters. Each parameter has varying importance, requiring distinct weights to be added to the objective function. Determining these weights is not trivial, as improper assignment could introduce bias and could fail to accurately capture the trade-offs between the properties, potentially leading to suboptimal results (Cho et al., 2017; Coello, 2000). Moreover, separating each of the six 2-dimensional properties into individual objective functions would further complicate the process, making it computationally slow. Increasing the number of objectives in multi-objective problems also impacts the difficulty of convergence, an increase in the computational effort of many operations, or the memory requirements for storing non-dominated solutions (Allmendinger et al., 2022; Curry & Dagli, 2014; Schütze et al., 2011). Therefore, we propose a MOO approach with two objective functions, where 2-dimensional properties are aggregated into a single objective function, while the 3-dimensional property is used as a separate objective function.

The cost of the retrieval process is yet another challenge that stems largely from the inclusion of the mobility diameter (d_m) as the second objective function. Two geometric properties, hydrodynamic radius (R_h) and orientationally averaged projected area (PA), are essential for calculating d_m across different flow regimes, from continuum to free molecular (Gopalakrishnan et al., 2015; C. Zhang et al., 2012). However, calculating R_h and PA is a computationally expensive process, especially when frequent calculations are required, which can significantly increase the simulation time. The process was found to be particularly prohibitive when utilizing population-based metaheuristic optimization techniques (Singh et al., 2024; Singh & Thajudeen, 2023). These methods required evaluating objective function values for the entire population in each iteration, placing increased demands on the computational resources.

Bayesian Optimization (BO) offers a principled alternative to address such computational inefficiencies (Jones et al., 1998; X. Wang et al., 2023). It can be adapted to deal with multiple, expensive black-box functions of potentially incommensurable objectives under constraints on the total number of evaluations due to time limitations (Garnett et al., 2010; Mahendran et al., 2012; Marchant & Ramos, 2012; Shahriari et al., 2016). In this study, the calculation of the objective function based on d_m is computationally expensive. BO offers advantages over the

previous methods due to its surrogate modelling technique. Unlike the metaheuristic methods, BO does not depend on the randomized sampling and evaluation of populations of candidate solutions in every iteration. Instead, it uses the predictive distribution of a probabilistic surrogate model (Rasmussen, 2003) trained on an evaluated dataset of solutions to determine the most promising solution to evaluate next. This approach leverages the uncertainty of the surrogate model to help balance exploration and exploitation of the search space, thereby reducing the number of expensive evaluations needed in the retrieval process.

Previously developed retrieval processes have primarily been applied to aggregates with monomers characterized by point-contact and monodispersity. Studies have shown that an aggregate can contain a range of monomer sizes with a degree of overlap (Dastanpour & Rogak, 2016; M. Eggersdorfer et al., 2012; Jourdain et al., 2023). These variations can significantly affect properties such as size, surface area, and total mass. Consequently, integrating these aspects into the retrieval process is essential to enhance its accuracy and broader applicability, which is also the focus of this study. The following section explains the methodology of the MOO-based retrieval process, outlining the steps involved in the integration of the BO technique for optimization.

2. Methodology

2.1. Multi-objective Optimization Preliminaries

A general MOO may be formulated with the goal of minimizing a vector-valued objective function F(x), i.e., $F(x) = [f_1(x),\ldots,f_m(x)]$, where x is a candidate solution belonging to a search space X (Ehrgott, 2005). Each component of F(x) symbolizes a mathematical description of a performance criterion that may be in conflict with the others. In this study, $x = [N, k_f, D_f]$ forms the fractal parameter vector used to generate candidate aggregate structures whose 2-dimensional and 3-dimensional properties give $F(x) = [f_1(x), f_2(x)]$; $f_1(x)$ is based on 2-dimensional image features whereas $f_2(x)$ is related to 3-dimensional properties of an aggregate.

Unlike traditional single-objective optimization methods, such multi-objective problems typically yield a set of optimal solutions rather than a single unique solution. Some solutions may perform better with respect to $f_1(x)$, while others may excel in $f_2(x)$. Thus, a collection of trade-off solutions is obtained, known as the Pareto set. Key concepts that formalize the distinctive facets of MOO are defined below (Coello, 2000).

Pareto Dominance: Solution vector x_a is said to dominate x_b if $f_i(x_a) \leq f_i(x_b)$, $\forall i$, and $\exists j$ such that $f_j(x_a) < f_j(x_b)$.

Pareto-Optimal Solution: A solution is considered Pareto-optimal if it is not dominated by any other solution within the search space. In other words, no alternative solution exists that is equally effective across all objectives and strictly superior in at least one objective. Such a solution is also referred to as a non-dominated solution.

Pareto Set: The set of all Pareto-optimal solutions is called the Pareto set.

Pareto Front: The image of the Pareto set in the objective space forms the Pareto front, showing trade-offs between the objectives.

2.2. Multi-Objective Optimization Framework for Aerosol Structure Retrieval

A bi-objective formulation of the aerosol structure retrieval problem is considered here. The first objective function is a combination of 2-dimensional properties of a candidate structure, whereas the second objective is derived from its 3-dimensional property. In what follows, details of the two objectives and their integration into a unified optimization framework are presented.

2.2.1. Objective function based on 2-dimensional features

Projected area (A_{proj}), perimeter (P), maximum end-to-end length of projection (L_{max}), maximum width perpendicular to maximum length (W_{max}), 2-dimensional radius of gyration

($R_{g,2D}$), and 2-dimensional fractal dimension ($D_{f,2D}$) are the 2-dimensional features used in the study (Singh et al., 2024; Singh & Thajudeen, 2023). These properties are calculated by image analysis using the box-counting method, a widely used technique (Panigrahy et al., 2020; So et al., 2017; R. Wang et al., 2022). The complete procedure is detailed in our previous study for reference (Singh & Thajudeen, 2023). The objective function (f_1) based on these properties is evaluated as,

$$f_1(x) = \left[\frac{A_{proj,T} - A_{proj}}{A_{proj,T}}\right]^2 + \left[\frac{L_{max,T} - L_{max}}{L_{max,T}}\right]^2 + \left[\frac{P_T - P}{P_T}\right]^2 + \left[\frac{R_{g2D,T} - R_{g2D}}{R_{g2D,T}}\right]^2 + \left[\frac{W_{max,T} - W_{max}}{W_{max,T}}\right]^2 + \left[\frac{D_{f2D,T} - D_{f2D}}{D_{f2D,T}}\right]^2 \quad (4)$$

where all notations are the same as defined earlier, with the subscript 'T' representing true or input image properties, while properties without the 'T' subscript refer to the projection properties of the candidate aggregate projection.

2.2.2. Objective function based on 3-dimensional properties

In this study, mobility diameter (d_m) is chosen as the 3-dimensional property, forming the basis of the second objective function. It can be easily extended to other properties, including aerodynamic diameter. Aggregate samples are often collected using a DMA based on their electrical mobility diameter for subsequent image analyses. Similar to the first objective function, the second one is also the squared relative error between the input $d_{m,T}$ and the d_m obtained from the candidate aggregate generated during optimization,

$$f_2(x) = \left[\frac{d_{m,T} - d_m}{d_{m,T}}\right]^2. \quad (5)$$

The relationship between geometrical parameters, hydrodynamic radius (R_h), orientationally averaged projected area (PA), and the mobility of aerosol particles is very well established (Thajudeen, Jeon, et al., 2015). We assume that all particles are singly charged. For non-spherical particles in different flow regimes, the expression for mobility is,

$$B = \frac{1+\left(\frac{\lambda\pi R_h}{PA}\right)\left(1.257+0.4*e^{-\frac{1.1PA}{\lambda\pi R_h}}\right)}{6\pi\mu R_h} \quad (6)$$

where μ is dynamic viscosity, and λ is the mean free path of the background medium, in this case, air. Mobility of a non-spherical particle can also be given in terms of the d_m as,

$$B = \frac{1+\left(\frac{2\lambda}{d_m}\right)\left(1.257+0.4*e^{-\frac{0.55d_m}{\lambda}}\right)}{3\pi\mu d_m}. \quad (7)$$

d_m is calculated by solving Eqs. (4) and (5) as implicit functions. To compute the R_h, the Smoluchowski radius (R_S) estimation process is used, following the methodology proposed in prior studies (Thajudeen et al., 2012; C. Zhang et al., 2012). The orientationally averaged projected area (PA) is determined through the Monte Carlo simulation technique. These processes become computationally expensive when frequent calculations are required for aggregates with a large number of monomers. Thus, the time required to calculate d_m can range from minutes to hours as the size of the aggregate increases. In studies where aggregates are generated multiple times, such as Langevin simulations (Thajudeen, Deshmukh, et al., 2015) or iterative optimization processes, the calculation of properties like d_m, R_h, and PA is required at each iteration. This frequent computation significantly increases the overall computational cost. BO is known to be an effective algorithm for such kinds of optimization problems as its use of computationally cheap surrogates of expensive objective function(s) helps accelerate the search process.

2.3. Multi-Objective Bayesian Optimization

In contrast to metaheuristic algorithms-based optimization, in any iteration of BO, the available dataset of evaluated solutions is used to train probabilistic surrogate models (Rasmussen, 2003) that serve as substitutes to the true (costly) objective functions. An *acquisition function* (ascribing an approximate figure of merit to any candidate solution) is defined based on the predictive probability distribution of the surrogate, which, when optimized, yields a single solution for evaluation in the next BO iteration. The evaluated solution is then appended to the dataset on which the surrogate model is retrained. These iterations continue until convergence or until some predefined computational budget is exhausted. The accuracy of the surrogate model improves as the model refines with each iteration, provably guiding the search to converge to an optimal solution (Srinivas et al., 2010). Compared to metaheuristic methods, BO entails significantly fewer expensive function evaluation calls, and hence can be much faster in reaching the optimal solution.

Algorithm 1 outlines the steps for extending basic BO to the multi-objective setting, following a framework similar to the Pareto extension of the efficient global optimization algorithm (Knowles, 2006). A detailed discussion of each step is provided in the subsequent sections. The algorithm begins with a Latin Hypercube Sampling (LHS) of the search space, generating ‘m’ initial candidate solutions [N, k_f, D_f] that ensure a representative coverage of the parameter search space. These solutions are used to generate aggregate structures to get projections to estimate 2D features, which are compared with the input values along with the

3D properties of the aggregate structures. The number of initial candidate solutions is fixed as $(11|x| - 1)$, where $|x|$ represents the number of variables; in our case, $|x| = 3$ for $[N, k_f, D_f]$. This formula is commonly used in previous BO studies, initially given in the works of Jones et. al. (1998) and Knowles (2006). LHS is a stratified Monte Carlo sampling method that divides each dimension into intervals and samples points from each interval, thereby resulting in a well-distributed dataset of initial solution samples (Iman, 2008).

Algorithm 1: Multi-Objective Bayesian Optimization for Structure Retrieval

Inputs: Microscopic image features (2-dimensional properties), Aerosol properties of sampled aggregate (3-dimensional properties)

Output: Retrieved aggregate structure

1. Initialization: Apply LHS to generate an initial set of $m = 11|x| - 1$ candidate solutions (x) with different N, k_f, and D_f values and generate corresponding aerosol structures using FracVAL. Evaluate objective functions for each candidate structure: $f_1(x)$ based on projected features and $f_2(x)$ based on mobility diameter, to form a dataset $D = \{(x_1, F(x_1)), (x_2, F(x_2)), \ldots, (x_m, F(x_m))\}$

While the evaluation budget is not met, do

2. Objective aggregation: Aggregate the two objective functions $f_1(x)$ and $f_2(x)$ into a single scalar value using a randomly sampled weight vector by Tchebycheff's method (see Section 2.3.1.)

3. Surrogate model building: Train a Gaussian Process (GP) regression model using dataset D to predict the aggregated objective value from fractal parameters (see Section 2.3.2.)

4. Determining next solution to evaluate: Search for values of N^*, k_f^*, and D_f^* that optimize the acquisition function derived from GP predictions (see Section 2.3.3)

5. Dataset update: Use $x^* = [N^*, k_f^*, D_f^*]$ to generate a new candidate structure with FracVAL and calculate f_1 and f_2. Update $D \leftarrow D \cup \{(x^*, F(x^*))\}$ and $m \leftarrow m + 1$

End While

6. Pareto Solution Selection: Identify the subset of non-dominated solutions in D and select the "knee" point as the final retrieved aggregate structure (see Section 2.3.4.)

2.3.1. Objective Aggregation Method

The fundamental idea behind Algorithm 1 is to decompose the multi-objective optimization problem into a set of single-objective subproblems, with a random subproblem considered in every iteration. Given a candidate solution x, represented by the parameters (N, k_f, and D_f), the original multi-objective problem involves two objective functions, $f_1(x)$ and f_2

(x), based on 2-dimensional and 3-dimensional properties of the aerosol structure generated by FracVAL, respectively. These objectives may however be of incommensurable scale. Therefore, to enable meaningful aggregation of the objectives, each function $f_i(x)$ is first normalized to the range [0, 1] using the values observed in the dataset D. Following normalization, the Tchebycheff scalarization method is applied in Step 2 of Algorithm 1 to aggregate the objectives into a single weighted objective function (Knowles, 2006; Min et al., 2019; Q. Zhang et al., 2010). In each iteration of the proposed multi-objective BO algorithm, a weight vector, i.e., $\mathbf{w} = [w_1, w_2]$, is randomly and independently sampled, ensuring $\sum w_j = 1$ and $w_j \geq 0 \; \forall j$. These weights are incorporated into forming the following optimization subproblem with the Tchebycheff scalarized objective function,

$$\text{minimize: } y_{\mathbf{w}}(x) = \max_j \{w_j f_j(x)\}; \quad j = [1,2]. \tag{8}$$

In simple terms, the Tchebycheff scalarization guides the search to move downwards (in the case of minimization problems) along the directional vector $[w_2, w_1]$ in objective space. An important property of the resulting optimization dynamics is that the Tchebycheff method is compatible with both convex and non-convex Pareto fronts. The applicability of the more straightforward linear weighted sum of objectives, on the other hand, is limited to convex Pareto fronts (Min et al., 2021; Q. Zhang et al., 2010).

2.3.2. Surrogate Modelling with Gaussian Process

Notice that the Tchebycheff scalarized objective function $y_{\mathbf{w}}(x)$ has no simple mathematical form as $f_1(x)$ and $f_2(x)$ are obtained from FracVAL simulations. These simulations may moreover be expensive, adding to the challenge of solving the subproblem in Eq. (8). To address this, BO involves training a probabilistic surrogate model to act as a computationally cheap substitute to the true costly function $y_{\mathbf{w}}(x)$. The Gaussian process (GP) regression model is a popular choice in this regard, due to its principled modelling of predictive uncertainties (Min et al., 2019; Rasmussen & Williams, 2006; S. Zhang et al., 2022). Uncertainty estimation is especially important in simulation-based global optimization environments where an effective balance of exploration (in regions of large predictive uncertainty due to data sparsity) and exploitation (in regions where the optimum is predicted to exist) of the search space is sought. With more data being gathered with each iteration in Algorithm 1, the GP improves its predictions across the search space, reducing uncertainty in evaluated regions. By leveraging the surrogate model in this manner, BO substantially lowers the need for costly function evaluation calls to $f_1(x)$ and $f_2(x)$.

A GP is a stochastic approach to regression that extends the concept of multivariate Gaussians to infinite dimensions (Rasmussen, 2003). It places a Gaussian distribution prior over possible functions $y_{\mathbf{w}}$, with the distribution completely specified by its mean (usually set to zero) and covariance function $k(x, x')$ (whose hyperparameters are tuned to maximize the log marginal likelihood of the observed aggregated data). Once the original dataset, *D* is transformed into the aggregated form $D_{\mathbf{w}} = \{(x_i, y_{\mathbf{w}}(x_i))\}_{i=1}^{m}$, the Gaussian prior can be updated to give the posterior predictive distribution over function values at any new query point x_q. The predictive mean $\mu(x_q)$ and variance $\sigma^2(x_q)$ for the function at a new query point x_q are then given by,

$$\mu(x_q) = k_q(K + \sigma_n^2 . I)^{-1} Y \quad (9)$$

$$\sigma^2(x_q) = k(x_q, x_q) - k_q(K + \sigma_n^2 . I)^{-1} k_q^T \quad (10)$$

where $Y = [y_{\mathbf{w}}(x_1), y_{\mathbf{w}}(x_2), \ldots y_{\mathbf{w}}(x_m)]^T$, $k_q = \left[k(x_1, x_q), k(x_2, x_q), \ldots, k(x_m, x_q)\right]$ is the covariance between the test point x_q and the training points (x_1, x_2, …, x_m), K is an m×m matrix whose entries are the covariance between all training points, σ_n is a noise parameter, and I is the identity matrix. In this study, the squared exponential function is chosen as the covariance function $k(x, x')$ and the GPyTorch library (Gardner et al., 2018) is used for building the GP model.

2.3.3. Acquisition function

Given the trained probabilistic surrogate model, one way to approach the optimization subproblem in Eq. (8) is to solve for x* that minimizes $\mu(x_q)$ as given by Eq. (9). However, this ignores sparsely evaluated / underexplored regions of the search space where the predictive uncertainty, $\sigma^2(x_q)$, of the GP may be high, and therefore has the tendency of getting trapped in spurious local minima. In order to address this shortcoming, an *acquisition function* that combines both the predictive mean and the variance of the GP is typically defined in BO frameworks. The acquisition function serves as a figure of merit, which, when optimized, yields an uncertainty-informed solution x* that offers a trade-off between exploration and exploitation of the search space. While various acquisition functions have been proposed over the years (Shahriari et al., 2016; X. Wang et al., 2023), many of which may be applicable to the structure retrieval problem, in this work, we employ the relatively simple lower confidence bound (LCB) acquisition function. Specifically, instead of naively minimizing $\mu(x_q)$, with LCB, the optimization in Eq. (8) is approached by minimizing:

$$\mathrm{LCB}(x_q) = \mu(x_q) - \kappa \cdot \sigma(x_q) \quad (11)$$

where κ is a user-defined parameter that controls the trade-off between exploration and exploitation. The search for $x^* = [N^*, k_f^*, D_f^*]$ minimizing the LCB can ultimately be carried out using any preferred global optimization heuristic. In our implementation, the state-of-the-art exponential natural evolution strategies (xNES) (Wierstra et al., 2008) is used for its known efficiency in low-dimensional search spaces.

One advantage of the LCB over many other acquisition functions is that κ can be freely tuned to instil BO with desired search dynamics. Higher values of κ emphasize exploration, while lower values focus more on exploitation. In this study, κ was chosen between 0.3 and 0.7, with any value within this range yielding effective results for the test cases considered.

2.3.4. Pareto Solution Selection

By the time the available function evaluation budget is exhausted, the iterative multi-objective BO algorithm yields a set of approximately Pareto optimal solutions. Figure 1 provides an illustration of a two-dimensional objective space containing all solutions generated during a single run of the BO algorithm, with the subset of non-dominated solutions (closest to being Pareto optimal) marked in red.

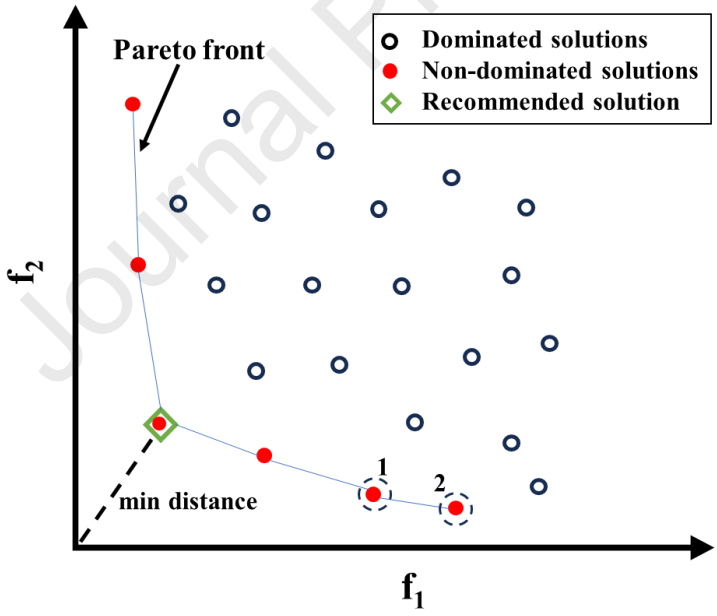


Figure 1. Representation of a Pareto front in multi-objective optimization: dominated solutions (blue circles), non-dominated solutions (red dots), and the recommended best trade-off point (green diamond)

In Figure 1, six non-dominated solutions are identified for further analysis. To illustrate this, consider solutions 1 and 2 from the figure. Both solutions are non-dominated because neither is superior in both objectives; solution 1 offers better performance for f_1, while solution 2 is slightly better with respect to f_2. The six non-dominated solutions, including solutions 1 and 2, collectively constitute an approximation of the Pareto set. Each point in the set offers a

unique balance between the two objectives, thereby requiring an additional stage of decision-making to determine a single preferred solution to the problem at hand.

In this study, the Pareto front is found to often be irregular and non-convex, indicating a complex trade-off landscape among the objectives. Since it is difficult to provide explicit preference information in such cases, a common heuristic strategy to determine a single preferred solution from the Pareto set is to select a point in the middle of the Pareto front, where the surface bulges out the most. This point is marked in Figure 1 with a green diamond. The rationale behind the heuristic strategy is that such points are generally furthest from the extremes of the Pareto front, and thus represent a well-balanced compromise among the various objectives. In some disciplines, this bulge is informally referred to as the "knee" of the Pareto front (Heidari et al., 2022). To locate the knee, both objectives of all non-dominated solutions were normalized to the range [0, 1], following which the point with the minimum Euclidean distance to the origin was selected as the preferred solution. Thus, the aerosol aggregate corresponding to this point was identified as the final retrieved structure.

The discussions heretofore outlined the multi-objective optimization approach via probabilistic surrogate model-based BO, incorporating both 2-dimensional image features and 3-dimensional properties of candidate aggregates generated by FracVAL. The algorithm can be further tailored to utilize any other 3-dimensional property-based objective function, such as the aerodynamic diameter. The detailed results and corresponding analysis are discussed in the following section.

3. Results and discussion

The proposed method was tested using both synthetic images generated from computationally simulated aggregates and microscopic images of mobility-classified aggregates. The synthetic test aggregates were created with varying fractal parameters, accommodating up to 500 monomers, with D_f ranging from 1.3 to 2.4 and k_f from 0.9 to 1.6. These synthetic images were derived from aggregates featuring monomers of a fixed radius of 15 nm. Test cases were randomly selected from the defined parameter ranges, with the assumption that all aggregates were constructed with monodispersed monomers with point contact. The testing was subsequently expanded to include synthetic aggregates with monomers having polydispersity and overlapping. Finally, the methodology was applied to microscopic images collected from welding fumes, further validating its applicability in practical scenarios.

3.1. Bayesian optimization-based retrieval with single objective function

The proposed method was first tested with the single-objective optimization (SOO)-based retrieval process and compared with previous retrieval methods to evaluate its effectiveness. The effectiveness of the Bayesian optimization-based retrieval process was evaluated using the same test cases. The retrieved fractal parameters and mobility diameters of the resulting structures are presented in Table 1.

Table 1. Retrieved fractal parameters and mobility diameters for synthetic aggregates using Bayesian optimization-based retrieval.

Case no.	Fractal Parameter(s)						d_m(I)	d_m(O)	Function evaluations
	N(I)	D_f(I)	k_f(I)	N(O)	D_f(O)	k_f(O)			
1	50	1.8	1.3	50	1.80	1.35	183.6	185.1	55
2	50	2.0	1.1	55	1.90	1.46	177.5	191.6	65
3	100	1.8	1.3	106	1.96	1.05	259.6	263.3	48

The results are compared based on the number of function evaluations, as illustrated in Figure 2. The comparison was made between the initial retrieval method, based on Particle Swarm Optimization (PSO) and forward modelling method (Singh & Thajudeen, 2023), and ML-based retrieval methods (Singh et al., 2024) developed in earlier studies. The number of function evaluations was recorded when the objective function values reached an order of 10^{-3}. For both prior retrieval methods, the number of function evaluations was determined by multiplying the required iterations by the population size, which was consistently set at 20 for each case.

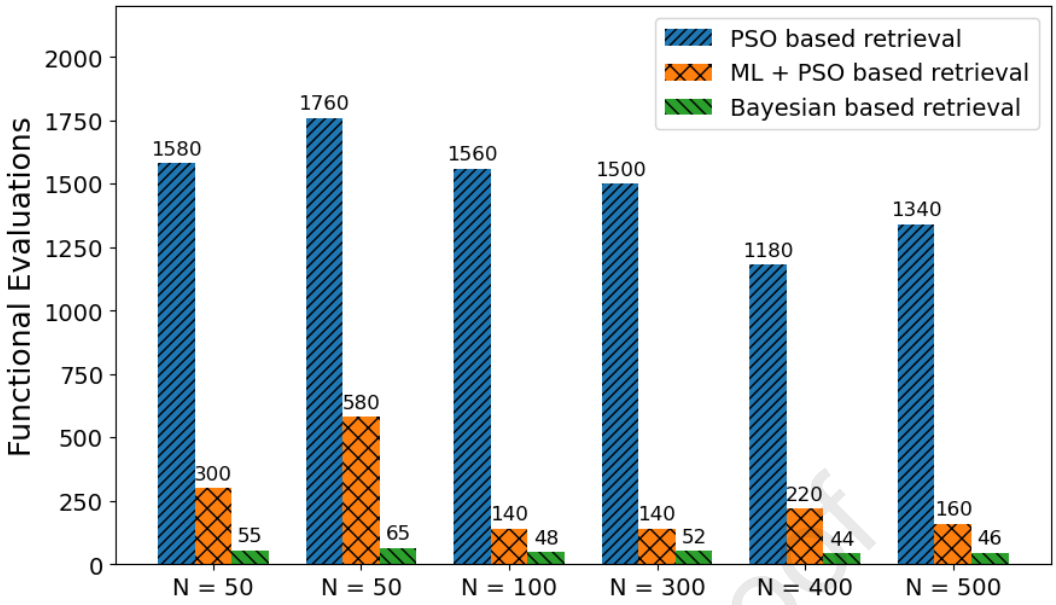


Figure 2. Comparison of the number of function evaluations required by the Bayesian optimization-based retrieval process and the PSO-based retrieval process

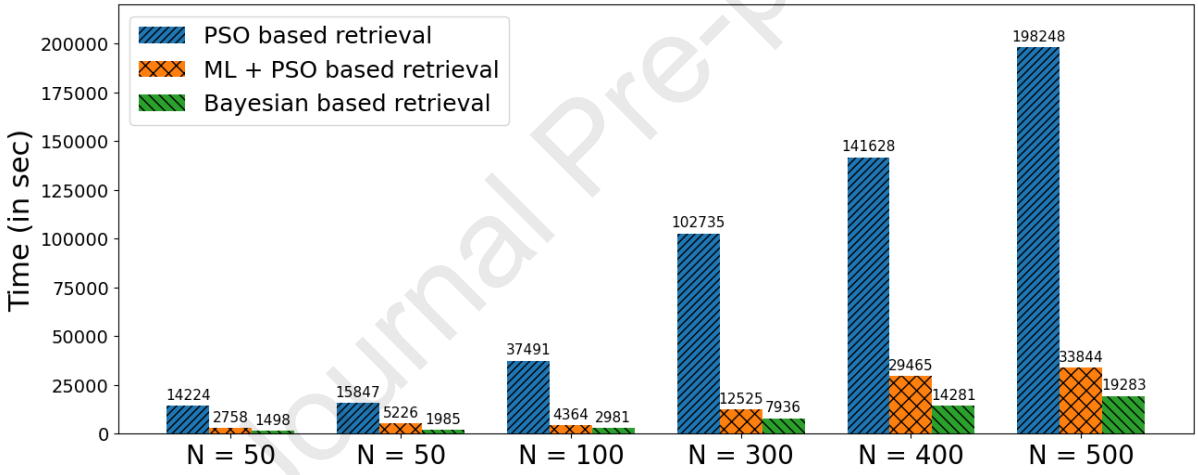


Figure 3. Comparison of total retrieval time between different retrieval methods

The results indicate that the Bayesian optimization-based retrieval method requires the least number of function evaluations across all tested aggregates across a wide range of fractal parameters. This efficiency is further illustrated in Figure 3, where the computational time for retrieval is compared. The total retrieval time encompasses the entire process, from aggregate generation and projection property calculation to optimization. As the size of the aggregate increases, so does the generation time, leading to a proportional increase in total retrieval time. For instance, obtaining the optimal structure of an aggregate with 500 monomers using traditional methods would take approximately 50 to 60 hours. In contrast, employing the Bayesian method can reduce this time to just 5 to 6 hours, representing a significant improvement over previous methods. This demonstrates the superiority of the Bayesian optimization-based retrieval method for determining aggregate structures for further analysis.

3.2. Multi-objective Bayesian optimization for aggregate retrieval

Given the demonstrated effectiveness of BO in the single-objective method, it was extended to MOO. Two objective functions based on projected properties and mobility diameter were used for evaluation, resulting in the Pareto set. The Pareto set was developed with non-dominated (Pareto optimal) solutions in the search space. A solution is nondominated if no solution improves one objective without causing a deterioration in at least one other objective. The final candidate was chosen by identifying the solution on the Pareto front with the minimum distance from the origin after normalization. The process was initially tested on synthetic aggregates.

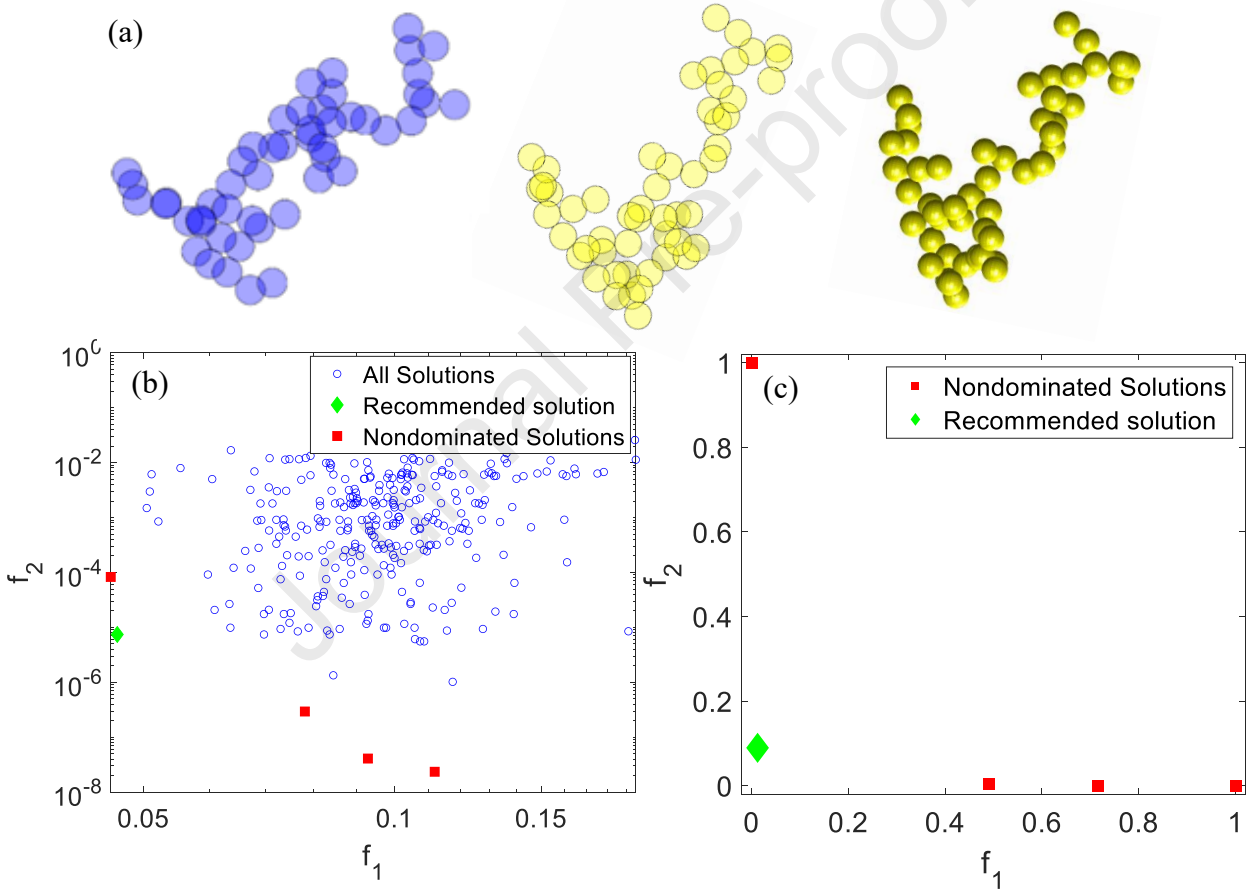


Figure 4. Retrieval based on Bayesian multi-objective optimization: (a) Input image (blue), most similar projection and retrieved structure (yellow), (b) solutions plot (blue circles) and non-dominated solutions (red squares), (c) normalized plot with recommended solution (green diamond) and non-dominated solutions (red squares).

Figure 4 shows an input image of a synthetic aggregate with monodispersed and point-contacted monomers. The input fractal parameters were N = 50, k_f = 1.3, and D_f = 1.8. The first step in the retrieval process involved the calculation of relevant 2-dimensional features. The mobility diameter of the input aggregates was d_m = 183.6 nm. The plot in Figure 4(b) represents all the solutions (in blue circles), where $f_{obj,1}$ corresponds to the values of the objective function

of the 2-dimensional properties, and $f_{obj,2}$ corresponds to the objective function based on mobility diameter. The point identified as the green diamond on the plot is considered as the preferred solution, and the red squares represent non-dominated solutions. The fractal parameters associated with this point were N = 50, k_f = 1.26, and D_f = 1.8. The most similar aggregate and retrieved structure is presented in yellow. The predicted d_m of the retrieved aggregate was found to be 183.1 nm. The final retrieved structure is chosen from the set of non-dominated solutions, as presented in Table 2. The table demonstrates the progression of the results towards the preferred solution. This case was executed for approximately 200 iterations, with convergence towards these solutions starting after 50 iterations.

Table 2. Selection of the final retrieved structure from non-dominated solutions

No.	k_f	D_f	N	f_1	d_m	f_2
1	1.35	1.74	52	0.077266	183.1	0.0000074
2	1.31	1.73	53	0.045675	181.8	0.0000961
3	**1.26**	**1.80**	**50**	**0.046504**	**183.1**	**0.0000074**
4	1.26	1.79	49	0.078079	183.4	0.0000012
5	1.26	1.80	49	0.092937	183.5	0.0000003

3.3. Comparison between single and multi-objective optimization-based retrieval

The method was initially applied to different aggregate structures that were synthetically generated, and the results are presented in Table 3a, and predictions of d_m are in Table 3b. Retrieval was performed using both single and multi-objective optimization (MOO) approaches, with most retrieved parameters falling within a 10% error margin, demonstrating the effectiveness of the MOO process in improving accuracy. In each case, the best solution has been highlighted in bold for emphasis. The multi-objective approach showed an advantage by selecting aggregates with 3-dimensional properties more closely resembling the input aggregate, as presented in Table 4. This observation is supported by Figure 4(c), where the optimal solution aligned well with both objectives.

While MOO outperforms SOO in most cases, SOO still performs well in a few instances. This is expected, as SOO-based retrieval is a well-developed approach and remains effective when no prior 3D information is available. However, when samples are collected based on a specific 3D property, MOO becomes the preferred method, as it integrates both 2D and 3D information for improved accuracy. Table 3b further supports this, showing that MOO-based retrieval consistently produces aggregates with better d_m values compared to SOO. Therefore, MOO is recommended when 3D information is available, while SOO remains useful when only microscopic images are provided without prior 3D data.

Table 3a. Retrieval of test aggregates with various fractal parameters (input) and retrieved parameters using both single and multi-objective optimization-based retrieval processes (best results are in boldface)

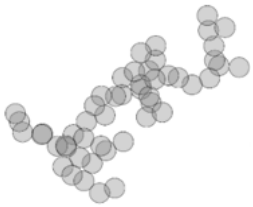
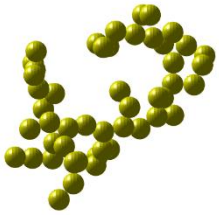
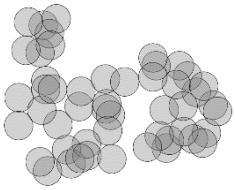
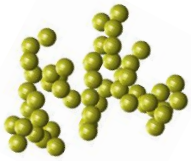
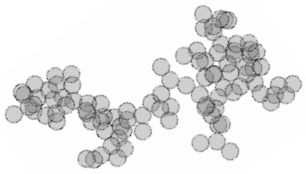
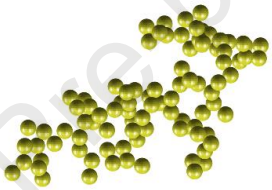
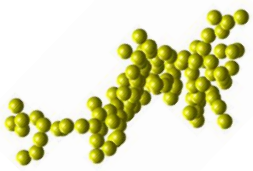
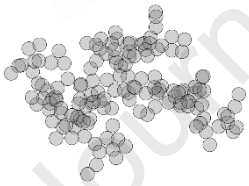
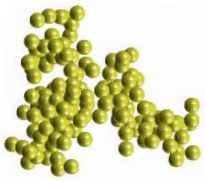
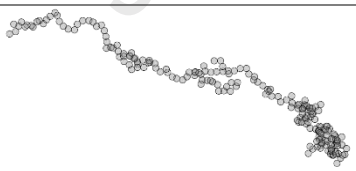
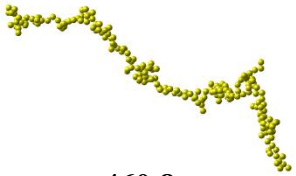
Case no.	Fractal Parameter(s)			Single objective optimization			Multi-objective optimization		
	N(I)	D_f(I)	k_f(I)	N(O)	D_f(O)	k_f(O)	N(O)	D_f(O)	k_f(O)
1	50	1.8	1.3	50	1.80	1.35	**50**	**1.80**	**1.26**
2	50	2.0	1.1	55	1.90	1.46	**52**	**2.05**	**0.91**
3	100	1.3	1.4	103	1.41	1.5	**100**	**1.37**	**1.39**
4	100	1.8	1.3	106	1.96	1.05	**98**	**1.86**	**1.11**
5	150	2.2	0.9	138	2.15	1.28	**149**	**2.21**	**1.22**
6	200	1.3	1.4	188	1.52	1.52	**210**	**1.31**	**1.29**
7	250	1.8	1.3	**250**	1.85	**1.32**	239	**1.79**	1.27
8	250	2.3	1.0	229	**2.28**	1.11	**255**	2.27	**1.04**
9	300	1.3	1.4	287	1.2	1.53	**313**	**1.34**	**1.41**
10	300	1.8	1.3	296	1.91	0.93	**297**	**1.85**	**1.32**

Table 3b. Comparison between input and retrieved mobility dimeters using both single and multi-objective optimization-based retrieval processes (best results are in boldface)

Case no.	N(I)	D_f(I)	k_f(I)	d_m(I)	d_m(O,soo)	d_m(O,moo)
1	50	1.8	1.3	183.61	185.11	**183.10**
2	50	2.0	1.1	177.55	191.58	**182.87**
3	100	1.3	1.4	312.15	287.47	**297.96**
4	100	1.8	1.3	259.61	**263.26**	255.29
5	150	2.2	0.9	291.33	255.95	**278.27**
6	200	1.3	1.4	473.85	400.38	**460.75**
7	250	1.8	1.3	414.51	381.66	**420.23**
8	250	2.3	1.0	342.57	335.71	**337.75**
9	300	1.3	1.4	628.73	652.64	**624.85**
10	300	1.8	1.3	458.68	445.36	**456.28**

The visual representation in Table 4 highlights the comparison between the retrieved structures obtained through single and multi-objective optimization-based retrieval processes. The results demonstrate that the multi-objective approach yields more structurally similar aggregates. Additionally, aggregates retrieved using the multi-objective method show better similarity in both 2-dimensional and 3-dimensional properties and the fractal parameters, further emphasizing its ability to align with the input data compared to the single-objective approach.

Table 4. Comparison of retrieved structures using single and multi-objective optimization-based retrieval processes

Case no.	Input image and parameters (N, D_f, k_f, d_m)	Single-objective BO optimization & $d_{m,out}$	Multi-objective BO optimization& $d_{m,out}$
Case 1	 50, 1.8, 1.3, 183.6	 185.1	 **183.1**
Case 2	 50, 2.0, 1.1, 177.6	 191.6	 **182.9**
Case 3	 100, 1.8, 1.3, 259.6	 **263.3**	 255.3
Case 4	 150, 2.2, 0.9, 291.3	 256.0	 **278.3**
Case 5	 200, 1.3, 1.4, 473.9	 400.4	 **460.8**

Additional validation has been done, including the calculation of relative errors between input and retrieved parameters (N, k_f, D_f) as well as actual and predicted d_m values for SOO and MOO-based retrieval process, presented in Table 5; the test cases are based on Table 3.

Table 5. Relative errors between input and retrieved parameters (N, k_f, D_f) and actual and predicted d_m values for SOO and MOO-based retrieval processes.

Case no.	% Relative Error							
	N(SOO)	D_f(SOO)	k_f(SOO)	N(MOO)	D_f(MOO)	k_f(MOO)	d_m(SOO)	d_m(MOO)
1	0.0	0.0	3.8	**0.0**	**0.0**	**3.1**	0.81	**0.28**
2	10.0	5.0	32.7	**4.0**	**2.5**	**17.3**	7.90	**3.00**
3	3.0	8.5	7.1	**0.0**	**5.4**	**0.7**	7.91	**4.55**
4	6.0	8.9	19.2	**2.0**	**3.3**	**14.6**	**1.41**	1.66
5	8.0	2.3	42.2	**0.7**	**0.5**	**35.6**	12.15	**4.48**
6	6.0	16.9	8.6	**5.0**	**0.8**	**7.9**	15.51	**2.76**
7	**0.0**	2.8	**1.5**	4.4	**0.6**	2.3	7.93	**1.38**
8	8.4	**0.9**	11.0	**2.0**	1.3	**4.0**	2.00	**1.41**
9	4.3	7.7	9.3	**4.3**	**3.1**	**0.7**	3.80	**0.62**
10	1.3	6.1	28.5	**1.0**	**2.8**	**1.5**	2.90	**0.52**

3.4. Retrieval of aggregates with polydispersity and overlapping

Aggregated aerosol particles are composed of monomers that vary in size, a characteristic known as polydispersity. These particles exhibit a broad size distribution, ranging from a few nanometers to several hundred nanometers in diameter (Dastanpour & Rogak, 2016; M. L. Eggersdorfer & Pratsinis, 2012). Subsequently, these particles aggregate to form larger structures composed of primary particles that also display polydispersity. Therefore, incorporating this aspect into the retrieval process is essential to enhance the versatility and accuracy of the analysis, since real aggregates exhibit some level of polydispersity in the monomer size. The polydispersity can be determined with a geometric standard deviation $\sigma_{p,geo}$ and the geometric mean of the monomers size $r_{p,geo}$ based on Eqs. (2), and (3) Monomers of different sizes are generated and attached to the growing cluster by following the scaling law. This feature is available in FracVAL and was exploited to generate candidate aggregates with polydispersity.

Additionally, the monomers can overlap with each other due to sintering (M. Eggersdorfer et al., 2012). The aggregates formed from combustion-generated nanoparticles often exhibit overlapping between the monomers. The morphology can vary significantly due to necking or overlap between monomers. Necking or overlap refers to non-point contacts between adjoining monomers due to sintering or variations in combustion temperature, which alter the overall morphology. This overlapping phenomenon requires careful consideration, as excessive overlapping can result in the merging of adjacent monomers. The overlapping can be quantified with the overlapping coefficient (V) as:

$$V = 1 - \left|\frac{AB}{2r_p}\right| \quad (12)$$

where AB is the centre distance between the two adjoined monomers and the r_p is the radius of each monomer. Prior studies used this factor to limit the overlapping between the monomers (Morán et al., 2018; Oh & Sorensen, 1997). It was integrated into the FracVAL code to introduce controlled overlapping between monomers.

Initial tests were conducted using synthetic aggregates generated with FracVAL, which was adjusted to produce aggregates with varying degrees of polydispersity and overlapping. The polydispersity index ranged from 1 to 2, and the overlapping index ranged from 0 to 0.5. The results are presented in Table 6, which includes different polydispersity indices and overlapping coefficients. This analysis provides insight into prediction accuracy by evaluating the retrieved values of N, k_f, D_f, and d_m for different test cases. It shows that the MOO-based retrieval processes perform well with structures having polydispersed monomers with a certain degree of overlap between them.

Table 6. Retrieved values of N, k_f, D_f, and d_m for aggregates with varying polydispersity indices and overlapping coefficients

Case no.	N(I)	D_f(I)	k_f(I)	$\sigma_{p,geo}$	ov	N(O)	D_f(O)	k_f(O)	d_m(I)	d_m(O)
1	50	1.8	1.3	**1.2**	**0.2**	50	1.91	1.15	180.5	184.6
2	50	2.0	1.1	**1.2**	**0.2**	49	1.94	1.14	174.1	179.1
3	100	1.6	1.6	**1.2**	**0.2**	99	1.65	1.37	263.6	254.1
4	100	1.8	1.3	**1.1**	**0.2**	95	1.77	1.19	247.8	251.1
5	100	1.8	1.3	**1.2**	**0.1**	99	1.76	1.06	258.1	247.4
6	100	1.8	1.3	**1.2**	**0.3**	99	1.84	1.20	242.4	254.2

Two test cases were used to demonstrate this retrieval process: In Case 2, the inputs were N = 50, D_f = 2.0, k_f = 1.1, σ_{geo} = 1.2, with an overlapping coefficient of 0.2 and an input mobility diameter for the second objective function of 174.2 nm. In Case 6, the inputs were N = 100, D_f = 1.8, k_f = 1.3, σ_{geo} = 1.2, with an overlapping coefficient of 0.3 and an input mobility diameter for the second objective function of 242.4 nm. The input images are presented in blue in Figure 5, while the retrieved projections and aggregates are shown in yellow. The parameters of the retrieved structures are as follows: for case (a), N = 49, D_f = 1.94, k_f = 1.14 and the $d_{m,out}$ = 179.1 nm; for case (b), N = 99, D_f = 1.84, k_f = 1.20 and $d_{m,out}$ = 254.2 nm. These predictions demonstrate good agreement with the input values.

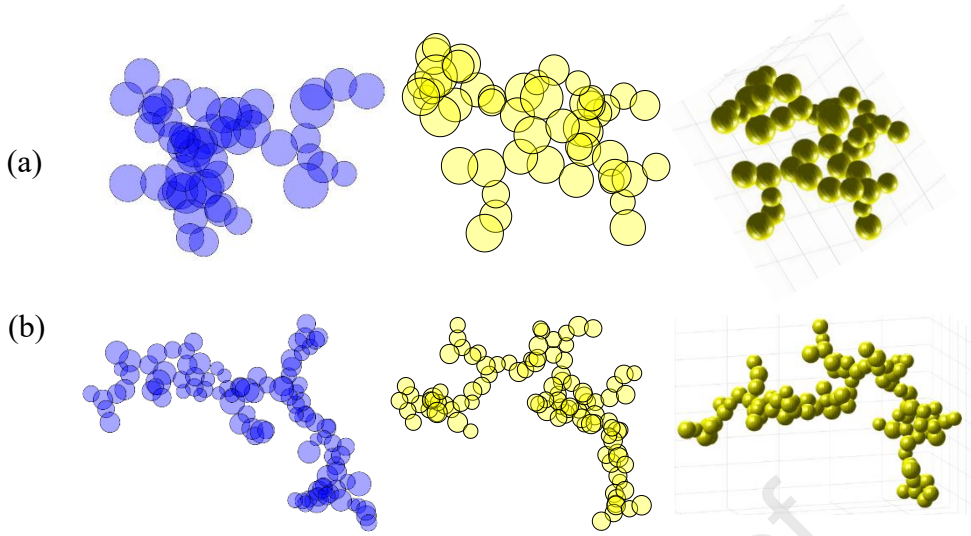


Figure 5. Retrieval of aggregates with polydispersity and overlapping of input (blue) and retrieved (yellow) aggregate structures.

3.5. Retrieval from microscopic images

The proposed retrieval method was further evaluated using microscopic images of aggregates obtained from welding processes through scanning electron microscopy (SEM). The test images are collected from fume particles generated by a wire additive arc welding process. Aggregate samples are collected using a DMA with a fixed voltage, corresponding to a fixed electrical mobility value. Silica wafer is positioned at the outlet of the DMA to collect the particles, which is utilized to collect images using SEM.

The obtained images are processed using ImageJ software, and a Fortran code to extract the required features. The 2-dimensional features, along with the measured mobility diameter, are integrated into the retrieval process as an input. The retrieval process follows the proposed methodology to obtain the aggregate structures. The test images used for retrieval were collected from aggregates exhibiting polydispersity and monomer overlap, which are key factors for evaluation. ImageJ software is used to assess the monomer size distribution and calculate the polydispersity and overlapping index. Figure 6 presents two test images used in the retrieval process. The retrieval was configured for 5% monomer overlap, with a polydispersity index of 1.5, estimated through ImageJ by measuring the sizes of individual monomers.

Figure 6 shows the best-matched projection highlighted in blue and the corresponding 3-dimensional aggregate presented in yellow. The predicted fractal parameters for the image (a) with measured d_m of 500 nm are N = 45, D_f = 1.69, and k_f = 1.12. The mobility diameter of the retrieved structure is found to be 537.4 nm. Image (b) is collected with a d_m of 270 nm and the

predicted parameters of N = 42, D_f = 1.65, and k_f = 1.31. The corresponding predicted mobility diameter for the retrieved structure is 303.9 nm. These results highlight the effectiveness of the proposed method for real-time analysis of aggregates.

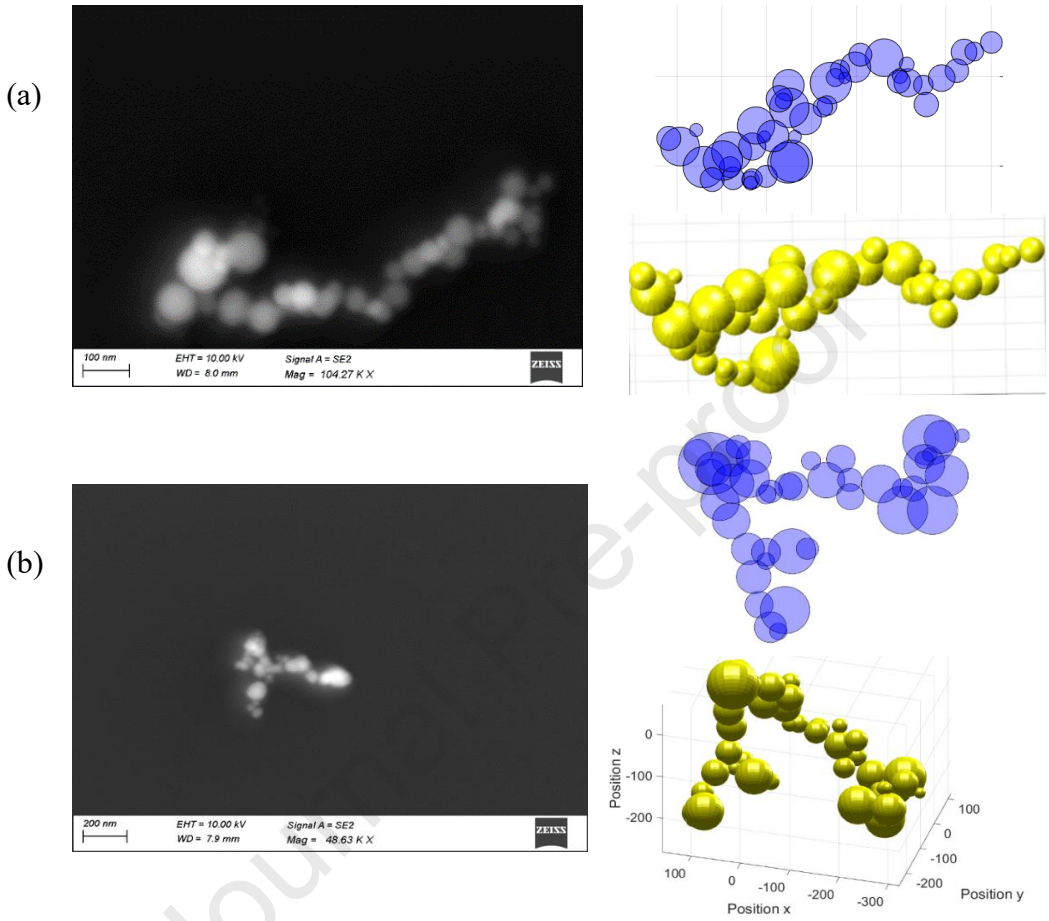


Figure 6. Retrieved structures of two mobility-classified aggregate images, a) d_m = 500nm and b) 270nm, with the best-matched projection (in blue) and corresponding 3-dimensional aggregate (in yellow).

4. Conclusion

In this study, a multi-objective Bayesian optimization-based retrieval method is developed to retrieve 3-dimensional structures from microscopic images of aggregated aerosol particles. The objective was to identify the optimal aggregate structures that exhibit the maximum similarity to the input microscopic image based on projection or image features and 3-dimensional property, mobility diameter. Initial testing was conducted using a Bayesian optimization-based retrieval process and compared with previously developed metaheuristic optimization-based retrieval. BO was found to be better than prior methods. It required fewer function evaluations and reduced retrieval time. It showed significant improvement in computational efficiency, reducing the time required for aggregate retrieval from approximately 50-60 hours to 5-6 hours for aggregates with 500 monomers.

The following tests involved comparing single-objective and multi-objective Bayesian methods focused on assessing their effectiveness in predicting fractal parameters and mobility diameters. The results suggested that the multi-objective approach, which considers two objective functions, consistently outperforms the single-objective method, particularly in terms of predicting mobility diameter (d_m). Further testing involved synthetic aggregates with polydispersity and overlapping, confirming the adaptability of the proposed method. The method effectively handled aggregates with varying polydispersity indices (up to $\sigma_{P,geo} = 2$) and overlapping coefficients (ov = 0.5), demonstrating accurate predictions of fractal parameters and mobility diameters with errors under the 10% range. Additionally, the methodology was validated with real-time aggregates collected from welding conditions. The Bayesian optimization-based retrieval method successfully predicted the parameters of aggregates with a known mobility diameter, again with an error of less than 10%.

In conclusion, MOO, with Bayesian techniques, represents a fast and efficient method for retrieving aggregate structures. This method has proven successful across various cases, from synthetic test images to real microscopic images of aggregates. The study can also be extended to aggregates sampled based on other aerosol properties like aerodynamic diameter as the second objective function. The method demonstrates its versatility and potential for application in multiple areas requiring morphological analysis.

Acknowledgements

This work was supported in part by the Ramanujan Fellowship from the Science and Engineering Research Board, Government of India (Grant No. RJF/2022/000115).

The authors gratefully acknowledge the support of Mr. Rubal and Mr. Adarsh Prakash, School of Mechanical Sciences, IIT Goa, in providing the SEM images of particles generated from welding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

CRediT authorship contribution statement

Abhishek Singh: Writing –original draft, simulations, Algorithm development, Methodology, Data analysis. **Smruti Ranjan Jena:** Algorithm development, simulations. **Abhishek Gupta:** Conceptualization, Supervision, Review & Editing, Funding acquisition. **Thaseem Thajudeen:** Conceptualization, Supervision, Review & Editing, Funding acquisition.

References

Allmendinger, R., Jaszkiewicz, A., Liefooghe, A., & Tammer, C. (2022). What if we increase the number of objectives? Theoretical and empirical implications for many-objective combinatorial optimization. *Computers and Operations Research*, *145*(October 2020), 105857. https://doi.org/10.1016/j.cor.2022.105857

Bescond, A., Yon, J., Ouf, F. X., Ferry, D., Delhaye, D., Gaffié, D., & Coppalle, A. (2014). Automated Determination of Aggregate Primary Particle Size Distribution by TEM Image Analysis : Application to Soot. *Aerosol Science and Technology*, *6826*. https://doi.org/10.1080/02786826.2014.932896

Brasil, A. M., Farias, T. L., & Carvalho, M. G. (1999). A recipe for image characterization of fractal-like aggregates. *Journal of Aerosol Science*, *30*(10), 1379–1389. https://doi.org/10.1016/S0021-8502(99)00026-9

Cao, J. (2017). The Importance of Aerosols in the Earth System: Science and Engineering Perspectives. *Aerosol Science and Engineering*, *1*(1), 1–6. https://doi.org/10.1007/s41810-017-0005-1

Chakrabarty, R. K., Garro, M. A., Garro, B. A., Chancellor, S., Moosmuller, H., & Herald, C. M. (2011a). Simulation of aggregates with point-contacting monomers in the cluster-dilute regime. Part 1: Determining the most reliable technique for obtaining three-dimensional fractal dimension from two-dimensional images. *Aerosol Science and Technology*, *45*(1), 75–80. https://doi.org/10.1080/02786826.2010.520363

Chakrabarty, R. K., Garro, M. A., Garro, B. A., Chancellor, S., Moosmuller, H., & Herald, C. M. (2011b). Simulation of aggregates with point-contacting monomers in the cluster-dilute regime. Part 2: Comparison of two-and three-dimensional structural properties as a function of fractal dimension. *Aerosol Science and Technology*, *45*(8), 903–908. https://doi.org/10.1080/02786826.2011.568022

China, S., Scarnato, B., Owen, R. C., Zhang, B., Ampadu, M. T., Kumar, S., Dzepina, K., Dziobak, M. P., Fialho, P., Perlinger, J. A., Hueber, J., Helmig, D., Mazzoleni, L. R., & Mazzoleni, C. (2015). Morphology and mixing state of aged soot particles at a remote marine free troposphere site: Implications for optical properties. *Geophysical Research Letters*, 1243–1250. https://doi.org/10.1002/2014GL062404.Received

Cho, J. H., Wang, Y., Chen, I. R., Chan, K. S., & Swami, A. (2017). A Survey on Modeling

and Optimizing Multi-Objective Systems. *IEEE Communications Surveys and Tutorials*, *19*(3), 1867–1901. https://doi.org/10.1109/COMST.2017.2698366

Coello, C. A. C. (2000). An Updated Survey of GA-Based Multi-objective Optimization Techniques. *ACM Computing Surveys*, *32*(2), 109–143. https://doi.org/10.1145/358923.358929

Coello Coello, C. A. (2000). Constraint-handling using an evolutionary multi-objective optimization technique. *Civil Engineering and Environmental Systems*, *17*(4), 319–346. https://doi.org/10.1080/02630250008970288

Curry, D. M., & Dagli, C. H. (2014). Computational complexity measures for many-objective optimization problems. *Procedia Computer Science*, *36*(C), 185–191. https://doi.org/10.1016/j.procs.2014.09.077

Dastanpour, R., & Rogak, S. N. (2016). The effect of primary particle polydispersity on the morphology and mobility diameter of the fractal agglomerates in different flow regimes. *Journal of Aerosol Science*, *94*, 22–32. https://doi.org/10.1016/j.jaerosci.2015.12.005

David, L. M., Ravishankara, A. R., Kodros, J. K., Pierce, J. R., Venkataraman, C., & Sadavarte, P. (2019). Premature Mortality Due to PM2.5 Over India: Effect of Atmospheric Transport and Anthropogenic Emissions. *GeoHealth*, *3*(1), 2–10. https://doi.org/10.1029/2018GH000169

Eggersdorfer, M., Kadau, D., Herrmann, H. J., & Pratsinis, S. E. (2012). Aggregate morphology evolution by sintering: Number and diameter of primary particles. *Journal of Aerosol Science*, *46*, 7–19. https://doi.org/10.1016/j.jaerosci.2011.11.005

Eggersdorfer, M. L., & Pratsinis, S. E. (2012). The structure of agglomerates consisting of polydisperse particles. *Aerosol Science and Technology*, *46*(3), 347–353. https://doi.org/10.1080/02786826.2011.631956

Eggersdorfer, M., & Pratsinis, S. E. (2014). Agglomerates and aggregates of nanoparticles made in the gas phase. *Advanced Powder Technology*, *25*(1), 71–90. https://doi.org/10.1016/j.apt.2013.10.010

Ehrgott, M. (2005). *Multicriteria Optimization* (Second edi, Vol. 491). Springer Science & Business Media. https://doi.org/10.1007/978-0-387-95865-1_8

Einar Kruis, F., van Denderen, J., Buurman, H., & Scarlett, B. (1994). Characterization of

Agglomerated and Aggregated Aerosol Particles Using Image Analysis. *Particle & Particle Systems Characterization*, *11*(6), 426–435. https://doi.org/10.1002/ppsc.19940110605

Filippov, A. V., Zurita, M., & Rosner, D. E. (2000). Fractal-like aggregates: Relation between morphology and physical properties. *Journal of Colloid and Interface Science*, *229*(1), 261–273. https://doi.org/10.1006/jcis.2000.7027

Forrest, S. R., & Witten, T. A. J. (1979). Long-range correlations in smoke-particle aggregates. *J. Phys. A: Math. Gen.*, *109*(12).

Gaeckle, N. T., Lee, J., Park, Y., Kreykes, G., Evans, M. D., & Hogan, C. J. (2020). Aerosol generation from the respiratory tract with various modes of oxygen delivery. *American Journal of Respiratory and Critical Care Medicine*, *208*(8), 1115–1124. https://doi.org/10.1164/rccm.202006-2309OC

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., & Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. *Advances in Neural Information Processing Systems*, *2018-Decem*(NeurIPS), 7576–7586.

Garnett, R., Osborne, M. A., & Roberts, S. J. (2010). Bayesian optimization for sensor set selection. *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN '10*, 209–219. https://doi.org/10.1145/1791212.1791238

Gopalakrishnan, R., McMurry, P. H., & Hogan, C. J. (2015). The electrical mobilities and scalar friction factors of modest-to-high aspect ratio particles in the transition regime. *Journal of Aerosol Science*, *82*, 24–39. https://doi.org/10.1016/j.jaerosci.2015.01.001

Heidari, A., Couckuyt, I., Qing, J., Gonzalez, S. R., & Dhaene, T. (2022). Finding Knees in Bayesian Multi - objective Optimization. *International Conference on Parallel Problem Solving from Nature*, 104–117.

Iman, R. (2008). Latin Hypercube Sampling. *Encyclopedia of Quantitative Risk Analysis and Assessment*. http://onlinelibrary.wiley.com/doi/10.1002/9780470061596.risk0299/full

Jeon, S., Hurley, K. R., Bischof, J. C., Haynes, C. L., & Hogan, C. J. (2016). Quantifying intra- and extracellular aggregation of iron oxide nanoparticles and its influence on specific absorption rate. *Nanoscale*, *8*(35), 16053–16064. https://doi.org/10.1039/c6nr04042j

Jones, D. R., Schonlau Matthias, & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-box Functions. *Journal of Global Optimization*, *13*, 455–492.

Jourdain, C., Symonds, J. P. R., & Boies, A. M. (2023). Compact aerosol aggregate model (CA2M): A fast tool to estimate the aerosol properties of fractal-like aggregates. *Aerosol Science and Technology*, *57*(8), 797–809. https://doi.org/10.1080/02786826.2023.2206442

Knowles, J. (2006). ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multi-objective optimization problems. *IEEE Transactions on Evolutionary Computation*, *10*(1), 50–66. https://doi.org/10.1109/TEVC.2005.851274

Lapuerta, M., Armas, O., & Gómez, A. (2003). Diesel particle size distribution estimation from digital image analysis. *Aerosol Science and Technology*, *37*(4), 369–381. https://doi.org/10.1080/02786820300970

Li, W., Riemer, N., Xu, L., Wang, Y., Adachi, K., Shi, Z., Zhang, D., Zheng, Z., & Laskin, A. (2024). Microphysical properties of atmospheric soot and organic particles: measurements, modeling, and impacts. *Npj Climate and Atmospheric Science*, *7*(1), 1–14. https://doi.org/10.1038/s41612-024-00610-8

Lighty, J. A. S., Veranth, J. M., & Sarofim, A. F. (2000). Combustion aerosols: Factors governing their size and composition and implications to human health. *Journal of the Air and Waste Management Association*, *50*(9), 1565–1618. https://doi.org/10.1080/10473289.2000.10464197

Mahendran, N., Wang, Z., Hamze, F., & De Freitas, N. (2012). Adaptive MCMC with Bayesian optimization. *Journal of Machine Learning Research*, *22*, 751–760.

Marchant, R., & Ramos, F. (2012). Bayesian optimization for Intelligent Environmental Monitoring. *IEEE International Conference on Intelligent Robots and Systems*, 2242–2249. https://doi.org/10.1109/IROS.2012.6385653

Meakin, P., Majid, I., Havlin, S., & Eugene Stanley, H. (1984). Topological properties of diffusion limited aggregation and cluster-cluster aggregation. *Journal of Physics A: General Physics*, *17*(18). https://doi.org/10.1088/0305-4470/17/18/008

Min, A. T. W., Gupta, A., & Ong, Y. S. (2021). Generalizing Transfer Bayesian Optimization to Source-Target Heterogeneity. *IEEE Transactions on Automation Science and Engineering*, *18*(4), 1754–1765. https://doi.org/10.1109/TASE.2020.3017644

Min, A. T. W., Ong, Y. S., Gupta, A., & Goh, C. K. (2019). Multiproblem Surrogates: Transfer Evolutionary Multi-objective Optimization of Computationally Expensive Problems. *IEEE Transactions on Evolutionary Computation*, *23*(1), 15–28. https://doi.org/10.1109/TEVC.2017.2783441

Morán, J., Cuevas, J., Liu, F., Yon, J., & Fuentes, A. (2018). Influence of primary particle polydispersity and overlapping on soot morphological parameters derived from numerical TEM images. *Powder Technology*, *330*, 67–79. https://doi.org/10.1016/j.powtec.2018.02.008

Morán, J., Fuentes, A., Liu, F., & Yon, J. (2019). FracVAL: An improved tunable algorithm of cluster–cluster aggregation for generation of fractal structures formed by polydisperse primary particles. *Computer Physics Communications*, *239*, 225–237. https://doi.org/10.1016/j.cpc.2019.01.015

Oh, C., & Sorensen, C. M. (1997). The effect of overlap between monomers on the determination of fractal cluster morphology. *Journal of Colloid and Interface Science*, *193*(1), 17–25. https://doi.org/10.1006/jcis.1997.5046

Pang, Y., Chen, M., Wang, Y., Chen, X., Teng, X., Kong, S., Zheng, Z., & Li, W. (2023). Morphology and Fractal Dimension of Size-Resolved Soot Particles Emitted From Combustion Sources. *Journal of Geophysical Research: Atmospheres*, *128*(6). https://doi.org/10.1029/2022JD037711

Pang, Y., Wang, Y., Wang, Z., Zhang, Y., & Liu, L. (2022). *Quantifying the Fractal Dimension and Morphology of Individual Atmospheric Soot Aggregates Journal of Geophysical Research : Atmospheres*. 1–11. https://doi.org/10.1029/2021JD036055

Panigrahy, C., Seal, A., & Mahato, N. K. (2020). Image texture surface analysis using an improved differential box counting based fractal dimension. *Powder Technology*, *364*, 276–299. https://doi.org/10.1016/j.powtec.2020.01.053

Park, K., Cao, F., Kittelson, D. B., & McMurry, P. H. (2003). Relationship between particle mass and mobility for diesel exhaust particles. *Environmental Science and Technology*, *37*(3), 577–583. https://doi.org/10.1021/es025960v

Park, K., Kittelson, D. B., & McMurry, P. H. (2004). Structural properties of diesel exhaust particles measured by Transmission Electron Microscopy (TEM): Relationships to particle mass and mobility. *Aerosol Science and Technology*, *38*(9), 881–889.

https://doi.org/10.1080/027868290505189

Qiao, Y., Andrews, A. J., Christen, C. E., Olson, B. A., Khariwala, S., MacLachan, B., & Hogan, C. J. (2020). Morphological characterization of particles emitted from monopolar electrosurgical pencils. *Journal of Aerosol Science*, *142*(September 2019), 105512. https://doi.org/10.1016/j.jaerosci.2020.105512

Rasmussen, C. E. (2003). Gaussian Processes in machine learning. *Summer School on Machine Learning*, *3176*, 63–71. https://doi.org/10.1007/978-3-540-28650-9_4

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. In *MIT Press* (Vol. 7, Issue 5).

Schütze, O., Lara, A., & Coello, C. A. C. (2011). On the influence of the number of objectives on the hardness of a multi-objective optimization problem. *IEEE Transactions on Evolutionary Computation*, *15*(4), 444–455. https://doi.org/10.1109/TEVC.2010.2064321

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Shiraiwa, M., Ueda, K., Pozzer, A., Lammel, G., Kampf, C. J., Fushimi, A., Enami, S., Arangio, A. M., Fröhlich-Nowoisky, J., Fujitani, Y., Furuyama, A., Lakey, P. S. J., Lelieveld, J., Lucas, K., Morino, Y., Pöschl, U., Takahama, S., Takami, A., Tong, H., … Sato, K. (2017). Aerosol Health Effects from Molecular to Global Scales. *Environmental Science and Technology*, *51*(23), 13545–13567. https://doi.org/10.1021/acs.est.7b04417

Singh, A., Kailas, S. K., & Thajudeen, T. (2024). ML-assisted optimization method for the prediction of 3-dimensional fractal structures from microscopic images. *Journal of Aerosol Science*, *177*(December 2023), 106331. https://doi.org/10.1016/j.jaerosci.2023.106331

Singh, A., & Thajudeen, T. (2023). A Hybrid Particle Swarm Optimization-Tuning Algorithm for the Prediction of Nanoparticle Morphology from Microscopic Images. *Aerosol and Air Quality Research*, *23*(4), 220453. https://doi.org/10.4209/aaqr.220453

Sipkens, T. A., Boies, A., Corbin, J. C., Chakrabarty, R. K., Olfert, J., & Rogak, S. N. (2023). Overview of methods to characterize the mass, size, and morphology of soot. *Journal of Aerosol Science*, *173*(June), 106211. https://doi.org/10.1016/j.jaerosci.2023.106211

So, G. B., So, H. R., & Jin, G. G. (2017). Enhancement of the Box-Counting Algorithm for fractal dimension estimation. *Pattern Recognition Letters*, *98*, 53–58. https://doi.org/10.1016/j.patrec.2017.08.022

Sorensen, C., & Roberts, G. C. (1996). The prefactor of fractal aggregates. *Journal of Colloid and Interface Science*, *186*, 447–452.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 1015–1022. https://doi.org/10.1109/TIT.2011.2182033

Thajudeen, T., Deshmukh, S., & Hogan, C. J. (2015). Langevin Simulation of Aggregate Formation in the Transition Regime. *Aerosol Science and Technology*, *49*(2), 115–125. https://doi.org/10.1080/02786826.2015.1008971

Thajudeen, T., Gopalakrishnan, R., & Hogan, C. J. (2012). The collision rate of non-spherical particles and aggregates for all diffusive knudsen numbers. *Aerosol Science and Technology*, *46*(11), 1174–1186. https://doi.org/10.1080/02786826.2012.701353

Thajudeen, T., Jeon, S., & Hogan, C. J. (2015). The mobilities of flame synthesized aggregates/agglomerates in the transition regime. *Journal of Aerosol Science*, *80*, 45–57. https://doi.org/10.1016/j.jaerosci.2014.11.003

Trivanovic, U., Corbin, J. C., Baldelli, A., Peng, W., Yang, J., Kirchen, P., Miller, J. W., Lobo, P., Gagné, S., & Rogak, S. N. (2019). Size and morphology of soot produced by a dual-fuel marine engine. *Journal of Aerosol Science*, *138*(September 2019), 105448. https://doi.org/10.1016/j.jaerosci.2019.105448

Wang, R., Singh, A. K., Kolan, S. R., & Tsotsas, E. (2022). Fractal analysis of aggregates: Correlation between the 2D and 3D box-counting fractal dimension and power law fractal dimension. *Chaos, Solitons and Fractals*, *160*, 112246. https://doi.org/10.1016/j.chaos.2022.112246

Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent Advances in Bayesian Optimization. *ACM Computing Surveys*, *55*(13s). https://doi.org/10.1145/3582078

Wierstra, D., Schaul, T., Peters, J., & Schmidhuber, J. (2008). Natural Evolution Strategies. *2008 IEEE Congress on Evolutionary Computation, CEC 2008*, *15*, 3381–3387. https://doi.org/10.1109/CEC.2008.4631255

Yon, J., Morán, J., Lespinasse, F., Escudero, F., Godard, G., Mazur, M., Liu, F., & Fuentes, A. (2021). Horizontal Planar Angular Light Scattering (HPALS) characterization of soot produced in a laminar axisymmetric coflow ethylene diffusion flame. *Combustion and Flame*, *232*, 111539. https://doi.org/10.1016/j.combustflame.2021.111539

Zhang, C., Thajudeen, T., Larriba, C., Schwartzentruber, T. E., & Hogan, C. J. (2012). Determination of the scalar friction factor for non-spherical particles and aggregates across the entire knudsen number range by Direct Simulation Monte Carlo (DSMC). *Aerosol Science and Technology*, *46*(10), 1065–1078. https://doi.org/10.1080/02786826.2012.690543

Zhang, Q., Liu, W., Tsang, E., & Virginas, B. (2010). Expensive multi-objective optimization by MOEA/D with gaussian process model. *IEEE Transactions on Evolutionary Computation*, *14*(3), 456–474. https://doi.org/10.1109/TEVC.2009.2033671

Zhang, S., Yang, F., Yan, C., Zhou, D., & Zeng, X. (2022). An Efficient Batch-Constrained Bayesian Optimization Approach for Analog Circuit Synthesis via Multi-objective Acquisition Ensemble. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *41*(1), 1–14. https://doi.org/10.1109/TCAD.2021.3054811

Highlights

- Multi-objective optimization to predict 3-dimensional morphology from microscopic images
- Bayesian Optimization for faster convergence
- Second objective function includes apriori information of mobility diameter
- Retrieval time reduced by up to one-tenth of existing methods
- Accounts for polydispersity in particle size and overlapping

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.