Machine Learning and Physics-Driven Modelling and Simulation of Multiphase Systems

Nausheen Basha[a], Rossella Arcucci[b], Panagiota Angeli[c], Charitos Anastasiou[c], Thomas Abadie[d], César Quilodrán Casas[e], Jianhua Chen[af], Sibo Cheng[e], Loïc Chagot[c], Federico Galvanin[c] ,Claire E. Heaney[bg], Fria Hossein[c], Jinwei Hu[b], Nina Kovalchuk[d], Maria Kalli[c], Lyes Kahouadji[a], Morgan Kerhouant[a], Alessio Lavino[a], Fuyue Liang[a], Konstantia Nathanael[d], Luca Magri[hi], Paola Lettieri[c], Massimiliano Materazzi[c], Matteo Erigo[c], Paula Pico[a], Christopher C. Pain[beg], Mosayeb Shams[a], Mark Simmons[d], Tullio Traverso[hi], Juan Pablo Valdes[a], Zef Wolffs[jk], Kewei Zhu[l], Yilin Zhuang[a], and Omar K Matar[a]

[a] Department of Chemical Engineering, Imperial College London, UK

[b] Department of Earth Science & Engineering, Imperial College London, UK

[c] Department of Chemical Engineering, University College London, UK

[d] School of Chemical Engineering, University of Birmingham, UK

[e] Data Science Institute, Department of Computing, Imperial College London, UK

[f] State Key Laboratory of Multiphase Complex Systems, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China

[g]Centre for AI-Physics Modelling, Imperial-X, White City Campus, Imperial College London, UK

[h] The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

[i] Department of Aeronautics, Imperial College London, UK

[j] Institute of Physics, University of Amsterdam, Science Park 904, Amsterdam

[k] Nikhef, Science Park 105, Amsterdam

[l] Department of Computer Science, University of York, UK

We highlight the work of a multi-university collaborative programme, PREMIERE (PREdictive Modelling with QuantIfication of UncERtainty for MultiphasE Systems), which is at the intersection of multi-physics and machine learning, aiming to enhance predictive capabilities in complex multiphase flow systems across diverse length and time scales. Our contributions encompass a variety of approaches, including the Design of Experiments for nanoparticle synthesis optimisation, Generalised Latent Assimilation models for drop coalescence prediction, Bayesian regularised artificial neural networks, eXtreme Gradient Boosting for microdroplet formation prediction, and a sub-sampling based adversarial neural network for predicting slug flow behaviour in two-phase pipe flows. Additionally, we introduce a generalised latent assimilation technique, Long Short-Term Memory networks for sequence forecasting mixing performance in stirred and static mixers, active

learning via Bayesian optimisation to recover coalescence model parameters for high current density electrolysers, Gaussian process regression for drop size distribution predictions for sprays, and acoustic emission signal inversion using gradient boosting machines to characterise particle size distribution in fluidised beds. We also offer perspectives on the development of a shape optimisation framework that leverages the use of a multi-fidelity multiphase emulator. The results presented have applications in chemical synthesis, microfluidics, product manufacturing, and green hydrogen generation.

Keywords: Machine Learning, Numerical Simulations, Multiphase, Multi-fidelity, Microfluidics, Hybrid Methods

1 Introduction

Multiphase flow phenomena abound in nature and are also of central relevance to numerous environmental, healthcare, industrial, and daily-life applications. These applications and settings include chemical reactors, separators and condensers, mixers, boilers and heat exchangers, and electrolysers for green hydrogen production. Their transient, three-dimensional dynamics invariably involve mechanisms that operate over a large range of length and time scales: from the nanometers to kilometers, and from sub-micro-seconds to hours and days. These dynamics, in turn, involve the motion and interaction of dispersed entities such as bubbles, drops and solid particles, the creation of three-dimensional, large-amplitude waves, changes of interfacial topology following break or coalescence events, and interactions with boundaries that have physical and physico-chemical heterogeneities. Multiphase flows may also be influenced by heat transfer and phase change, the presence of body forces, e.g. gravitational and/or electromagnetic, surface forces, e.g. due to surfactants, and, importantly, turbulence.

Given the practical importance of multiphase flows, it is imperative to have a framework for making efficient and reliable predictions of their dynamics, which have a direct bearing on the design of devices and industrial units, their scale-up as well as optimisation and control. Due to the complexity of these flows, it is often not feasible to rely solely on detailed physics-driven methods which invariably involve the use of computational fluid dynamics (CFD). These include direct numerical simulations (Deen et al. 2014; Seyed-Ahmadi and Wachs, 2020) which resolve all relevant scales without applying to closures, large eddy simulations (Fox, 2012) that involve the solution of filtered

versions of the equations of motion wherein the filtered terms require closures, or Reynolds-averaged Navier-Stokes type methods for which closures for the Reynolds stresses are also required. Approximate methods that rely on ensemble-averaging are also used which Eulerian-Eulerian approaches (Gidaspow, 1994) that include two- and multi-fluid models, and Eulerian-Lagrangian approaches (Esteghamatian et al., 2017; Lei et al., 2021) which include a combination of CFD and discrete particle or discrete element methods all of which also require closures. The closures needed for the above-mentioned methods are often empirical in nature which require calibration using large datasets obtained from physical or computationally-expensive DNS-based numerical experiments, and their reliability constitutes a challenge.

Recent years have witnessed an explosive uptake in the development and implementation of algorithms based on machine learning, which is a branch of artificial intelligence, with applications in chemistry, chemical engineering, materials science, physics, and medical imaging and visualisation (Venkatasubramanian, 2019). Fluid mechanics as a discipline has benefited from the rise of machine learning albeit for single-phase flows primarily (Duraisamy, Iaccarino, and Xia, 2019; Brunton, Noak and Koumoutsakos, 2020; Wang and Wang, 2021). In contrast, there is a relative dearth of studies featuring the application of machine learning for multiphase flows (Zhu et al., 2022). In order to address this imbalance, the PREMIERE (PREdictive Modelling with QuantIfication of UncERtainty for MultiphasE Systems) programme sought to develop a framework to blend data-centric and physics-driven methodologies featuring CFD, Design of Experiments, various types of neural networks, data-assimilation, ensemble methods, and uncertainty quantification. One of the aims of this programme is to engender a paradigm-shift in multiphase flow research worldwide towards the use of these hybrid approaches. In this paper, we provide highlights of the research produced as part of PREMIERE for this Special Issue on Machine Learning for Multiphase Flow.

Among the PREMIERE highlights, we showcase here the use of Design of Experiments (Nathanael, Galvanin, et al., 2023) in tandem with a CFD-population balance model (Pico et al., 2023) to optimise the synthesis protocols of silver nanoparticles, which are central to electronics, catalysis, pharmaceuticals, foods, and a range of products critical to the transition to green chemistry and environmentally friendly processes (Nathanael et al., 2022). This approach, which accounts for the simultaneous effect of various factors and their interactions to achieve global optimality, replaces the

one commonly used which relies on trial-and-error and resource-consuming experiments that proceed by changing one factor at a time and lead to a local optimum in parameter space.

Drop coalescence is of great importance for many industrial applications. In processes of emulsification, emulsion transportation and storage, coalescence is to be avoided, whereas it is to be facilitated in separation processes. In microfluidics, drop coalescence is one of the basic operations used, for example, to trigger chemical reactions by bringing together drops with reagents, to quench reaction or to add some chemicals of interest to cell environment during the cell screening processes. Usually drop coalescence in microfluidics is carried out under confinement conditions and confinement can be different in different directions. This brings additional variables affecting coalescence probability and therefore complicates reliable prediction of coalescence. Here, we demonstrate the use of Generalised Latent Assimilation (GLA) models for drop coalescence predictions (Zhuang et al., 2022; Cheng et al., 2024). We also show how machine learning methods involving Bayesian-regularised artificial neural networks and eXtreme Gradient Boosting can be used in combination with experiments for droplet formation prediction in microfluidics devices in the presence and absence of surfactants (Chagot et al., 2022).

We have provided examples of PREMIERE work on multiphase flows with significant inertia. One such example is a novel use of sub-sampling and adversarial neural networks to predict slug flow characteristics for two-phase flows in long horizontal pipes (Heaney, Liu et al., 2022; Heaney, Wolffs, et al., 2022). Another example involves the development of inexpensive machine learning surrogate models that can leverage data gathered from high-fidelity simulations of mixing characteristics in stirred and static mixers (Liang et al., 2023; Valdes et al., 2023b) to provide efficient performance predictions using Long Short-Term Memory (LSTM) network architectures. This methodology extrapolate future behaviour based on early-stage mixing performance with applications to mixing operations in a wide range of industries, ranging from FMCG to the energy and chemical sectors (Paul et al., 2004; Valdés et al., 2022).

Yet another example includes the use of active learning through Bayesian optimisation to determine the value of parameters governing coalescence in PBMs coupled to a multi-fluid Eulerian-Eulerian model that accounts for electrochemistry, heat transfer, and turbulence to simulate green hydrogen production in alkaline water electrolysers; these are commonly employed for industrial-scale

production of low-carbon hydrogen. This approach obviated the need to create new coalescence kernels for the PBM for use with electrochemistry and allowed the adaption of existing ones which were designed for air-water systems without electrochemical effects (Orvalho et al., 2021). We have also used Gaussian Process regression to predict the droplet size distributions of sprays as a function of global parameters such as the Reynolds and Weber numbers as well as the nozzle geometry (Traverso et al. 2023). The regression model was trained on volume of fluid simulations of the complex atomisation interfacial dynamics which leveraged adaptive mesh refinement. Lastly, we present results for gas-solid fluidised beds which constructed a machine learning regression model based on random forests and gradient boosting which predicts the particle size distribution using as input the frequency and kinetic energy from acoustic emission signals (Hossein et al., 2021; Hossein et al., 2022).

Finally, we offer perspectives for future work, which will involve the development of a multi-fidelity shape optimisation framework (Savage et al., 2023). This, in turn, will feature the training of deep neural networks for the optimisation of design parameters via interactions with a multi-fidelity, multiphase flow emulator; the latter will comprise a hierarchy of data- and physics-driven predictive models of varying degrees of fidelity. The interactions between the optimiser and the emulator will depend on the need and will range from cheap to expensive, low- to high-fidelity learning loops; this will enable rapid learning and decision-making in relatively simple situations, and more reliable decisions with increasing information flow from the higher-fidelity models. We hope that these perspectives, and the highlights provided earlier, will serve as departure points for exciting avenues of research in collaboration with the multiphase flow community.

2 Machine Learning for microfluidics

2.1 Computer vision for droplet image analysis

Expanding the scope of data-driven models in microfluidics, a collaborative study conducted by UCL and Imperial College London (ICL) explored their utilisation of computer vision techniques for image analysis of droplets in microfluidic channels (Gelado et al., 2023). This research addressed challenges posed by low-resolution images, enhancing spatial resolution to enable precise droplet detection and measurement. Notably, the Segment Anything Model (SAM) surpasses the widely

utilised Circular Hough Transform in imaging, demonstrating superior droplet detection and reduced uncertainty in droplet diameter measurements.

The robust performance of SAM is particularly evident in scenarios with low image contrast between inner and outer fluids. Additionally, for low-quality images, the study illustrates the effectiveness of training super-resolution methods, such as MSRN-BAM (multi-scale residual network-bottleneck attention model), on a dataset containing droplets to enhance detection capabilities. Finally, the DnCNN (denoising convolutional neural network) model (Zhang et al., 2017) proves its efficiency in removing Gaussian noise commonly observed in optical imaging. These results highlight the potential of deploying and combining data-driven methodologies for images analysis in microfluidic channels (Figure 1), offering valuable insights for advancing research and applications for droplet generation.

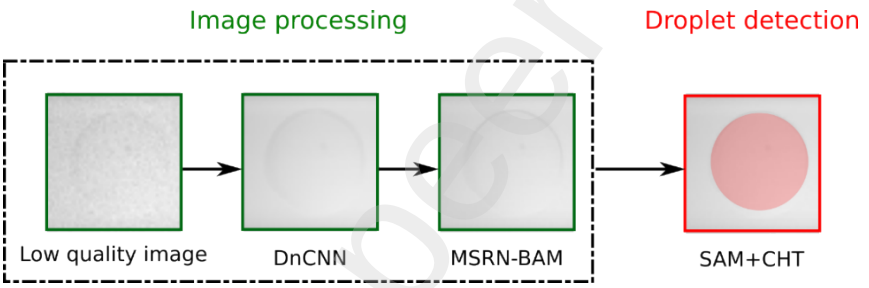


Figure 1. An example involving the combination of denoising convolutional neural networks (DnCNN) and multi-scale residual network-bottleneck attention model (MSRN-BAM) to extract microdroplet diameter information from a low-quality image (adapted from Gelado et al., 2023).

In a related case study, high-speed and high-resolution micro-PIV experiments were developed at UCL (Chagot et al., 2022) to measure velocity profiles within aqueous microdroplets in an organic medium in flow-focusing microfluidics devices laden with surfactants above and below the critical micelle concentration (CMC); interface-tracking CFD code was used at ICL to simulate, for the first time, the details of the droplet formation, flow field, and surfactant transport in the UCL experiments trans-CMC (Kalli et al., 2023). Chagot et al. (2022) obtained a large-scale dataset of droplet size from high-speed imaging experiments and used it to develop data-driven methods for the prediction of the drop sizes as a function of surfactant type and concentration as well as flow rates. They conducted the experiments on a flow-focusing microchannel where aqueous surfactant-laden droplets were generated in silicone oil. They utilised anionic (sodiumdodecylsulphate, SDS), cationic (cetyltrimethylammonium bromide, CTAB), and non-ionic (TX100) surfactants at concentrations

above and below the CMC and used these data to develop two data-driven models to predict the final droplet size as a function of flow rates, surfactant type, and concentration.

Using a Bayesian regularised artificial neural network and eXtreme Gradient Boosting (XGBoost), these models were initially based on four inputs: the flow rates of the two phases, equilibrium interfacial tension, and the normalised surfactant concentration. To overcome experimental difficulties in acquiring accurate interfacial tension values, both models were also trained by removing this parameter from the training dataset. Finally, over 10^4 synthetic data points were generated (based on the initial dataset) with a variational autoencoder (see Figure 2). The high-fidelity of the extended synthetic dataset highlights that this method can be a quick and low-cost alternative to study microdroplet formation where experimental data may not be readily available.

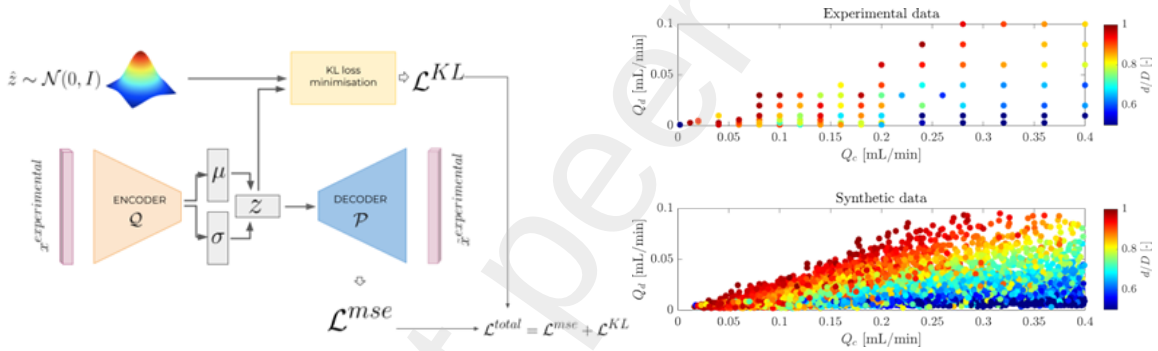


Figure 2. Synthetic experimental data generation using variational autoencoders (adapted from Chagot et al. 2022).

2.2 Synthesis of silver nanoparticles guided by design of experiments

Motivated by the need to develop methods for optimal routes of silver nanoparticle (AgNP) design, as discussed in the Introduction, a *D*-optimal Design of Experiments (DoE) was used to predict the size of AgNPs synthesised in a microfluidics flow reactor, shown in Figure 3a; here, the aim is to determine the conditions enabling the synthesis of AgNPs of minimal size (Nathanael, Galvanin, et al., 2023). Nanoparticles were synthesised using silver nitrate ($AgNO_3$) as the silver precursor, tannic acid as the reducing agent, and trisodium citrate, which serves both as a reducing as well as a stabilizing agent. To study the effect of mixing on particles size, the performance of a straight output microfluidics channel was compared with channels of helical shape. The studied factors included pH, concentration of trisodium citrate (*TC*), collection and storage temperature (*ST*), flow rate (*FR*), and curvature of the helical channel (*CR*).

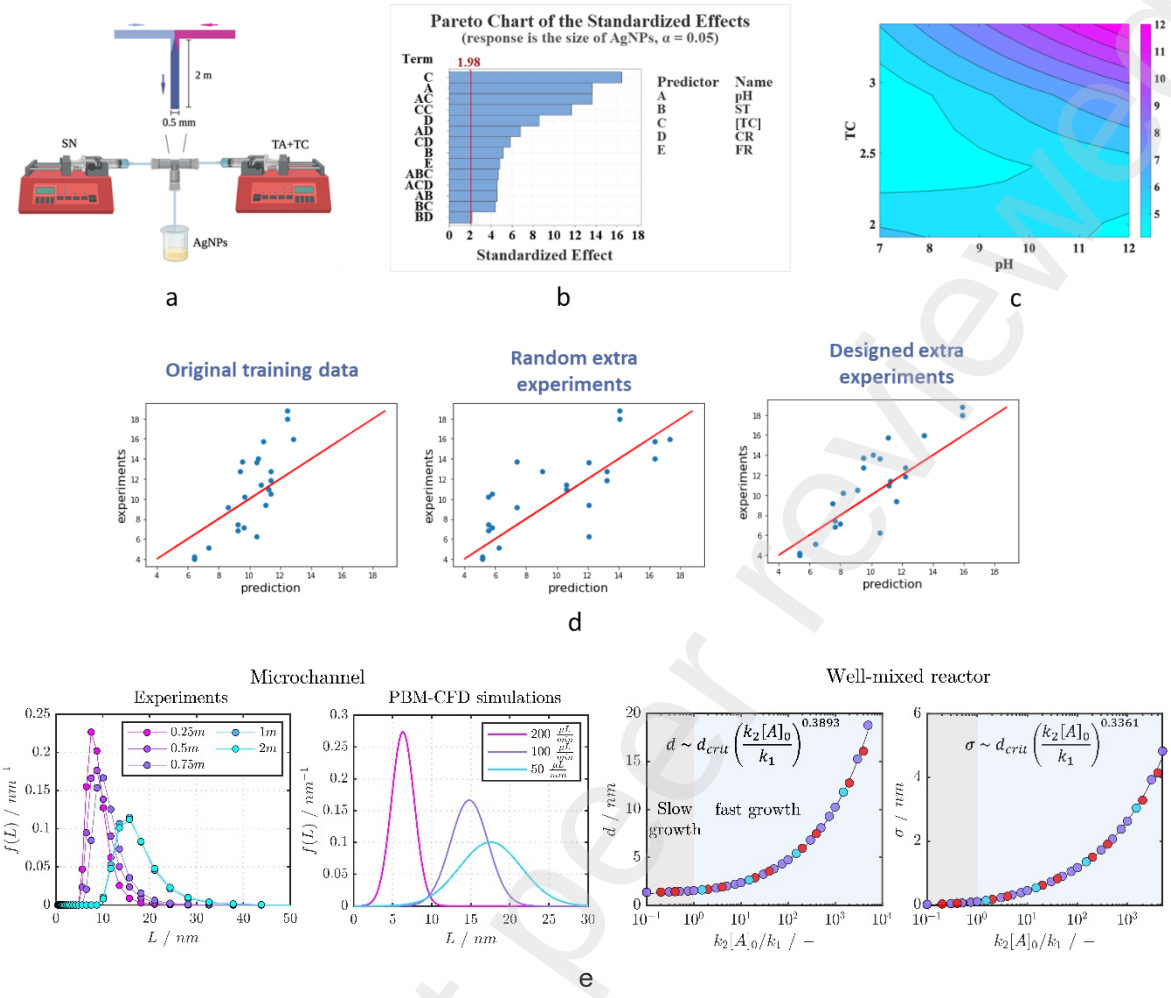


Figure 3. Using a microfluidics setup for AgNPs synthesis: a) the microfluidics device; b) Pareto chart demonstrating statistical significance of factors and their interactions, the red line showing the limiting value for statistical significance; c) response surface demonstrating the combined effect of pH and trisodium citrate (TC) on AgNPs size, the TC values are given in mM, and the colour bar is for the particle size in nm; d) comparison of the performance of the model trained on the initial set of data with those using randomly selected additional experiments, and guided DoE based on decision tree; e) comparison of the PBM-CFD simulation results to experimental data from the microfluidics channel and a well-mixed reactor. The figures shown have been adapted from (Nathanael, Galvanin, et al., 2023)(a – c), (Nathanael, Cheng, et al., 2023)(d), and Pico et al. (2023).

The optimal polynomial model was chosen using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). A Pareto chart presented in Figure 3b shows that the most important factors for the size of synthesised AgNPs are TC and pH as well as their interaction. The curvature of the helical device defining mixing in the system is the next most important parameter. The combined effect of the various factors was visualised using a response surface methodology, as shown in Figure 3c, which demonstrates the complex interplay between [TC] and pH in their contribution to the final particle size. The DoE approach provided a good prediction of the particle

size based on both the training and testing datasets. The results were validated experimentally using dynamic light scattering and transmission electron microscopy (Nathanael, Galvanin, et al., 2023).

In addition to the DoE-based experimental approach, we have also developed a numerical methodology to identify the parameters reflecting the process chemistry and the kinetic models for AgNP synthesis using computational fluid dynamics (CFD) and population balance modelling (PBM) (Pico et al., 2023). The two-step Finke-Watzky (F-W) mechanism was chosen to describe kinetics of nucleation followed by nanoparticle growth and has been shown to be the most reliable mechanism to describe the synthesis of metallic nanoparticles. This mechanism assumes that the nucleation process is relatively slow, whereas particle growth is a fast auto-catalytic process. Reduction of metal ions to metal atoms is not considered in this F-W model but is implicitly included in the nucleation process. The results of the PBM-CFD simulations of AgNP synthesis in microfluidic device presented in Figure 3a have highlighted the delicate interplay between mass transfer, precursor reduction, nucleation and growth of nanoparticles and a good qualitative agreement with experiments; this is illustrated in Figure 3e. Furthermore, these results were also in very good agreement with experimental studies for a well-mixed reactor where instantaneous mass transfer was observed. Wherein it was shown that the AgNP size is governed by the ratio of nucleation to growth time scales; here, the former time scale is the reciprocal of nucleation constant, k1, whereas the latter comprised the initial precursor concentration (A_o) as well as the growth rate constants. Thus, the combined DoE-PBM-CFD-based approaches has successfully teased out the most important factors governing AgNP synthesis.

Further work (Nathanael, Cheng, et al., 2023) has exploited the model-guided choice of parameter values which resulted in a vastly reduced number of additional experiments needed to improve the prediction accuracy. This study utilised a decision tree-guided DoE for predicting AgNP size using dimensionless parameters such as the Reynolds and Dean numbers, k_1 and k_2 (with fixed (A_o)), and collection/storage temperature. Importantly, the regions in parameter space targeted by this methodology in which to carry out the experiments corresponded to those associated with the highest mean squared error. As shown in Figure 3d the decision tree-guided DoE improves considerably the model prediction accuracy and that trained on randomly selected additional experimental data. It is also possible to generalise this approach for different chemistries as well as

different device design and related hydrodynamics. Another promising avenue is to use inverse modelling enabling to determine the optimal parameters to synthesis NPs of desired size.

2.3 Deep Learning models for prediction of drop coalescence under confinement

Here, we present machine learning-based approaches to predict drop coalescence in a microfluidics device shown in Figure 4a. Water drops in oil were formed in two symmetrical cross-junctions and driven to approach one another in a coalescence chamber via a compression-expansion flow field (see Figure 4b). The evolution of each doublet formed was recorded by a high-speed camera starting from the moment when the first drop enters the chamber until the drops either coalesce or detach. The dataset created includes 1531 and 1201 doublets in silicone oil of viscosity 4.6 mPa s and mineral oil of viscosity 31 mPa s, respectively, with a range of capillary numbers of 0.001 - 0.004.

The raw experimental data presented as video-recordings were used for reduced order (surrogate) modelling (ROM) with data assimilation in the latent space (LA), as shown in Figure 4c (Zhuang et al., 2022). ROM enables a considerable reduction in computational cost for high-dimensional raw images by compressing them into low-dimensional spaces, where recurrent neural networks (RNNs) are employed to build a surrogate model in the reduced-order space. LA with a developed ensemble-based correction determination algorithm to implement the optimum frequency of performing LA, improves the prediction accuracy to maximise agreement with experimental data. Further model improvement is achieved by using CFD data (see Figure 4d). To fully address the challenge of non-explicit transformation function in data assimilation, we have developed Generalised Latent Assimilation (GLA) designed to manage various data sources through a shared latent space. Notably, this novel methodology decreases computational demands by using a multi-domain encoder-decoder neural network to replicate complex mapping functions. Furthermore, a CNN-based approach with voronoi-tessellation is also introduced to handle time-varying sensor placement in the CFD simulations (Cheng et al., 2024)

Another ML approach uses tabulated data from pre-processed video-recordings. In this case, numerical parameters representing confinement of drops in the field of view and in the perpendicular direction, drop size difference, total flow rate (characterised by a capillary number) and time-delay between drops were used as model factors. Two tree-based models, a random forest (RF), and a XGBoost model, were applied to the dataset based on silicone oil (K. Zhu et al., 2023). This dataset

was imbalanced with the majority of the data representing coalescence events. To improve the model prediction, a novel generative model, the double-space conditional variational autoencoder, was developed and used to generate synthetic tabular data. The dataset based on mineral oil was well balanced, and three models were used to predict drop coalescence for this dataset; the RF, XGBoost, and Deep Neural Networks with a Multilayer Perceptron (MLP) all led to good results.

To better understand the importance of the various features and clarify the model's decision-making, explainable AI models were further applied to the dataset in (Hu et al., 2024). SHapley Additive exPlanations (SHAP) was used to understand contribution of each feature into model output. The TreeExplainer was employed for tree-based models, while the KernelExplainer was used for the MLP model. The SHAP values were calculated for each datum point, as shown in Figure 4e. It is seen that the most important feature is parameter H_{eff}, which characterises the confinement in the cross-stream direction: the coalescence probability decreases with H_{eff}. The second important feature in the RF model is confinement in the field of view, whereas for the MLP model this is a third important feature, while the flow rate is the second important. It can be concluded therefore that these features have roughly similar importance for coalescence. The least important feature for both models is the time interval between the drops.

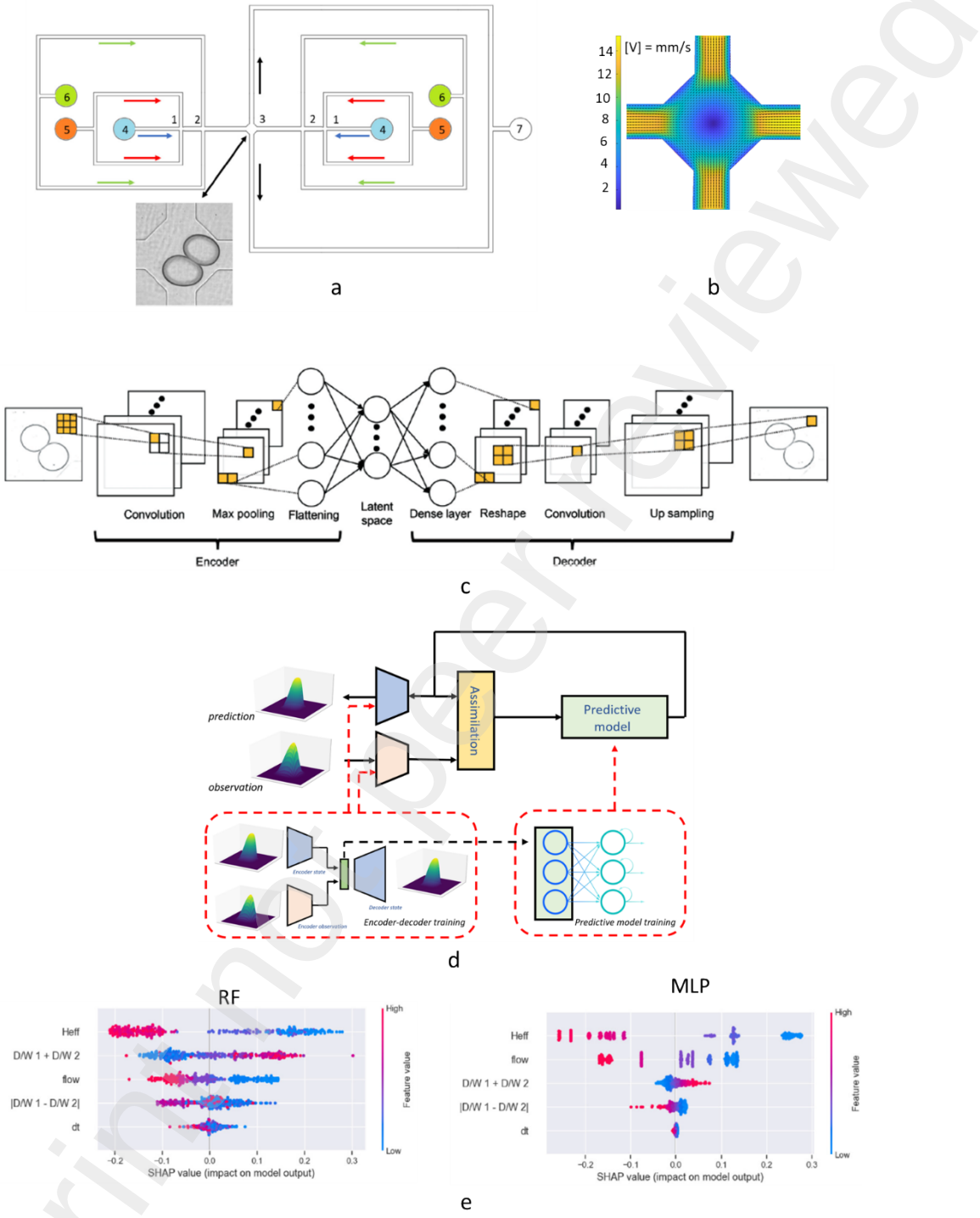


Figure 4. a) microfluidic device used to collect experimental data on drop coalescence: 1 – cross-junctions for drop formation, 2 – additional oil inputs through side channels, 3 –coalescence chamber, 4 – water inlets, 5 – main oil inlets, 6 – additional oil inlets, 7 – outlet; b) flow field in the middle plane of coalescence chamber measured by ghost particle velocimetry; c) ensemble latent assimilation with deep learning surrogate model for drop interaction in microfluidic device; d) Flow chart of the new method GLA; e) SHAP plots for RF and MLP models. The figures are adapted from (Hu et al., 2024) (a, e), (K. Zhu et al., 2023) (b), (Zhuang et al., 2022) (c).

Another approach which highlights the impact of specific features used in (Hu et al., 2024) was the Feature Ablation Test, where one of the features was dropped at a time and model predictions without this feature were compared with those from models operating on the full features set. To understand in more details the model decision making for a specific set of factors, Local Interpretable Model-agnostic Explanations, LIME, was used in (Hu et al., 2024). LIME work by sampling new data points in the vicinity of a chosen point in the parametric space represented by experimental data and creating a new local surrogate model for this local dataset. Usually, the surrogate model is very simple, often linear, and therefore easily interpretable in contrast to a much more sophisticated, general model. It provides a reasonable approximation of the behaviour of the global model in the vicinity of this particular point in the parametric space. Importantly, the LIME algorithm generates probability of coalescence for any chosen factor combination, which is of great practical interest. So far, tabulated data for two datasets, based on silicone oil and mineral oil, were used separately due to considerable imbalance of the former one. The future work aims at collecting more data to balance the silicone oil-based dataset and develop models describing all data. This will enable us to consider the effect of factors such as viscosity and interfacial tension, and to generalise the analysis to include dimensionless parameters such as the capillary number.

3 Machine Learning for inertial flows

3.1 Extending the generalisation capabilities of surrogate models for two-phase pipe flows

Until recently, machine-learning surrogates have been tied to the mesh on which the training data or snapshots are held. This means that if one has a surrogate model trained on one geometry, predictions can only be made for that particular geometry. A new method proposed the use of sub-sampling and adversarial networks to train using one geometry and predict for a different geometry (Heaney, Liu, et al., 2022). This method involved generating snapshots by sub-sampling the solutions on small regions within the domain rather than by using the CFD solution over the entire mesh. Therefore, neural networks can be taught how to predict the flows within a small region of the domain, given the flows from neighbouring regions. This way of training the neural networks leads to independence from the original geometry.

For the inference stage, a different domain (larger or with a different configuration or both) can be decomposed into regions of the same size as that used to generate the training data. The network can be used to make predictions for each small region, which are linked through an iterative process ending when the solution across the entire domain does not change significantly; see Figure 5 for a schematic of the steps required in the sub-sampling approach in addition to those of a typical machine-learning based surrogate model. Originally developed for urban air flows (Heaney, Liu, et al., 2022), this method has also been applied to multiphase pipe flow (Heaney, Wolffs, et al., 2022). The latter demonstrated that predictions of volume fraction and velocity could be made for a 98m pipe given training data from a 10m pipe simulation. Here, we use the same approach as in (Heaney, Liu, et al., 2022) to generate predictions of a 1000m pipe based on training data from a 10m pipe.

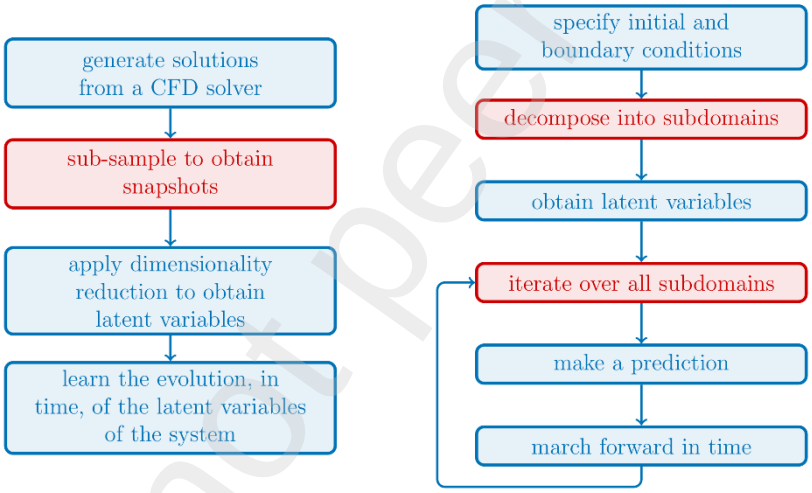


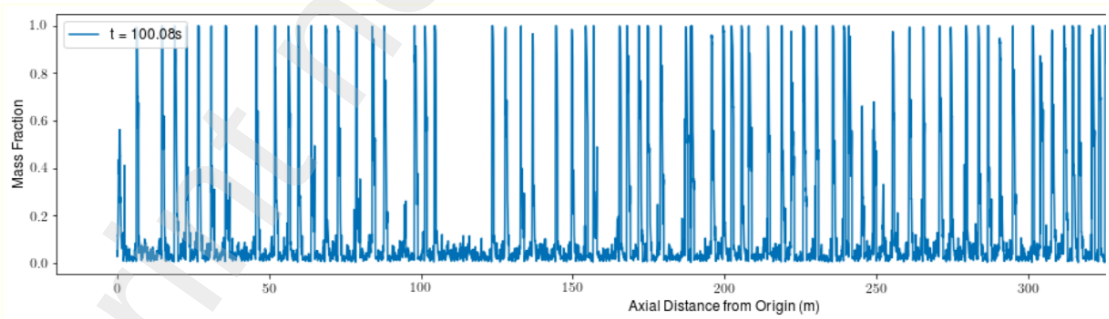
Figure 5. Typical steps of a machine learning-based surrogate model (blue) and additional steps for the sub-sampling strategy (red). Left: offline or training stage; right: online or predicting stage.

We use an autoencoder for dimensionality reduction to obtain the latent variables; this is followed by three stages for the generation of a surrogate model which involve departures from a conventional machine learning approach. The first one involves a sub-sampling approach and collecting training data from CFD solutions sampled over a small region and a random time level. The size of the small region is fixed and the location is obtained from splitting the pipe up into equally-sized sub-domains. In this work, the training data was augmented by random sampling. The usual min-max scaling (normalisation), in which the input and output variables are re-scaled linearly, was applied to the data before training an autoencoder. In the second stage, in order to generate training data for the network used to predict in time, the same sized sub-domain was used for sampling as in the first stage,

however, the network also needs information from the sub-domains that are upwind and downwind of the current sub-domain. After normalisation, an adversarial network was trained to take latent variables from the current sub-domain at time level *n-1* and latent variables from the neighbouring two sub-domains at time level *n*, in order to make a prediction for the latent variables in the current sub-domain at time level *n*. The training data can be sampled from random locations and a randomly chosen time level paired with the successive time level.

A third change to the typical machine learning surrogate modelling approach is made when using the trained networks for inference. A pipe of arbitrary length can be modelled by splitting the domain into sub-domains of the same size as those used in the sampling process. Initial and boundary conditions must be supplied or generated, and then, the adversarial network generates a prediction for one sub-domain in the pipe, given estimates of the solutions in the neighbouring sub-domains. Similarly, a solution is generated for the next sub-domain, using the most up-to-date solutions available for the neighbouring sub-domains. If a sweep is said to be completed when all the sub-domains have a prediction, multiple sweeps are performed (to allow information to propagate from one end of the domain to the other) in order to improve the solution. Once the global solution does not change (to within a specified tolerance), the sweeping process stops.

a



b

Figure 6. a) Two slugs in a section of the 10m pipe indicated through the volume fraction field (red represents water and blue represents air); b) spatial development of the liquid mass fraction inside a 1000m pipe at 100 seconds obtained using the surrogate model.

Here, we demonstrate our method by modelling the flow in a pipe of length 1000m and radius 0.039m. We split this into sub-domains of length 1m and used initial conditions for each sub-domain based on those of the 10m pipe. For the inlet condition, three methods were experimented with for

the surrogate model of the 98m pipe reported in Heaney, Liu et al. (2022). In generating results for a 1000m pipe, one of the three methods was used, which takes a series of solutions from the CFD results that led to the formation of a slug; this series is repeated as needed. To generate the CFD solutions for the machine-learning models, a finite-element control-volume code (Salinas et al., 2017), (Obeysekara et al., 2021) was used, which solves an advection equation for the volume fraction in order to track the interface.

The pipe was initially filled with water and once the simulation starts, air is injected in the top half of the pipe. The air speed is 4.162m/s and the water speed is 2.081m/s in the axial direction. Solutions at 800 time levels were generated with a fixed time interval of 0.01s. Figure 6a depicts a plot of the volume fraction field showing two slugs advecting down a section of the pipe. To create the surrogate model, we choose a sub-domain size of 1m in the axial direction and use the entire domain in the other two directions due to the aspect ratio of the problem. The velocity and volume fraction solutions are then interpolated onto grids of resolution 60×20×20 at 800 time levels and from various locations along the pipe to produce snapshots or training data. A convolutional autoencoder is trained which has just 10 latent variables, representing a compression ratio of 9600 (4 solution fields on a grid of 60×20×20 maps to 10 variables).

The latent variables from the 800 time levels are then used to train a network which will model the time-dependent behaviour of the system. We chose an adversarial network (Makhzani et al., 2015) for this task, as standard time prediction networks have been shown to struggle with predicting for a long period in time, as the error in their prediction accumulates over time and ultimately leads to unphysical behaviour (either tending to a false steady state or diverging) (Maulik et al., 2020), (Arcucci et al., 2021). The adversarial network has as an input, the latent variables at the current time level in neighbouring sub-domains and latent variables at the previous time level in the sub-domain of interest. Its output is the latent variables in the sub-domain of interest at the current time level. With initial and boundary conditions, we map from these to the latent variables and then begin the process of sweeping through the sub-domains generating a global prediction for the velocities and volume fraction. In Figure 6b, we plot the volume fraction against axial position, and the slugs can be seen clearly. This machine-learning surrogate stably predicts slugs over a pipe of length 1000m after having been trained from CFD solutions of multiphase flow in a 10m pipe.

3.2 Generalised Latent Assimilation for multiphase pipe flow

Our work presents a novel system that combines reduced-order surrogate models with a new data assimilation technique for real-time observation incorporation in different physical spaces (Cheng et al., 2023). The focus is on the development of Generalised Latent Assimilation (GLA) and its application in multiphase flow pipes, specifically in a high-dimensional CFD setting for two-phase liquid flow (Chen et al., 2023). GLA is designed to work efficiently with reduced-order modelling and low-dimensional surrogate models generated using machine learning algorithms. This approach benefits from the efficiency of reduced-order modelling and the accuracy of data assimilation (Cheng, Quilodran-Casas, et al., 2023). This work outlines the theoretical framework and provides a detailed analysis of the innovative loss function in the new data assimilation algorithm. The GLA approach is tested and validated numerically, demonstrating significant improvements in reconstruction and prediction accuracy of deep learning-based surrogate models. The development and application of GLA highlights the importance of real-time data incorporation to enable efficient and reliable predictions in complex, high-dimensional dynamical systems such as multiphase flow in pipes.

The study examines the flow separation characteristics of a silicone oil and water mixture in a pipe 4m long and 26mm in diameter (Chen et al., 2023). Eulerian–Eulerian simulations were conducted using OpenFOAM, an open-source CFD platform, to solve equations of mass and momentum conservation for oil and water phases taking into account their respective properties. Population balance models were employed to model droplet size and coalescence behaviour. Numerical solutions were obtained on a structured mesh with 180,000 nodes, and the flow time was set to 10 seconds with a time-step of 0.005 seconds to ensure convergence.

Regarding the machine learning phase, 1000 snapshots of the single-trajectory simulation data were divided into training (80%) and testing (20%) datasets (Cheng, Quilodran-Casas, et al., 2023). The data, originally from cylindrical meshes, were first flattened to 1D vectors before being auto-encoded. Our study focused on building machine learning surrogate models to predict the evolution of the oil volume fraction within the test section. Techniques such as autoencoders and Long Short-Term Memory (LSTM) networks were employed for training on the simulation snapshots. LSTM network-based models have been shown to provide a three orders of magnitude increase in computational

efficiency as compared to traditional CFD simulations (Cheng, Quilodran-Casas, et al., 2023). Our work introduces a novel Generalised Latent Assimilation algorithm that performs data assimilation (DA) with heterogeneous latent spaces for state and observation data. This approach allows for a more flexible and efficient assimilation process compared to the current state-of-the-art (Cheng, Quilodran-Casas, et al., 2023), as shown in Figure 7.

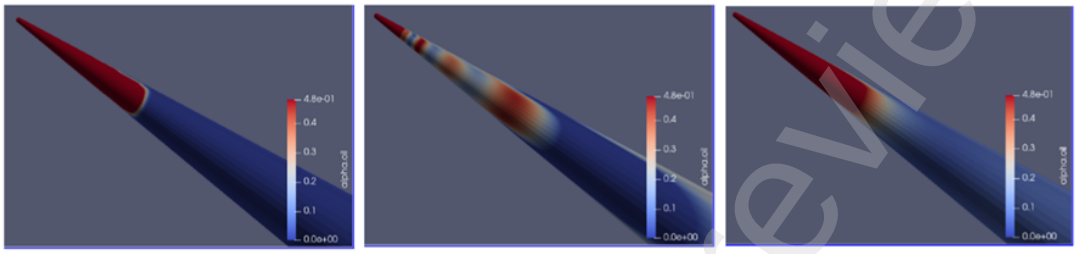


Figure 7. Comparison of CFD simulation against deep learning prediction and GLA correction with real-time observations (Cheng, Chen, et al., 2023)

3.3 Recurrent Neural Networks for mixing performance prediction

The numerical framework we developed to predict the performance of stirred and static mixers in forming dispersions, with and without surfactants, consists of three main stages, as depicted in **Figure** 8: 1) data generation, 2) data re-conditioning, and 3) model deployment. At its core, we built a multi-step, multivariate LSTM RNN which predicts the temporal evolution of key performance metrics, whilst remaining agnostic to the mixing process itself. The datasets used consisted of 43 cases which encompass different surfactant profiles, rotational speeds, inlet configurations and mixer arrangements for both stirred and static mixers. These cases, which have been featured in previous publications [(Valdes et al., 2023a) (Valdes et al., 2023b)(Liang et al., 2022)(Liang et al., 2023)], were generated via an in-house high-fidelity CFD code which features a robust interface-tracking algorithm (Shin et al., 2018).

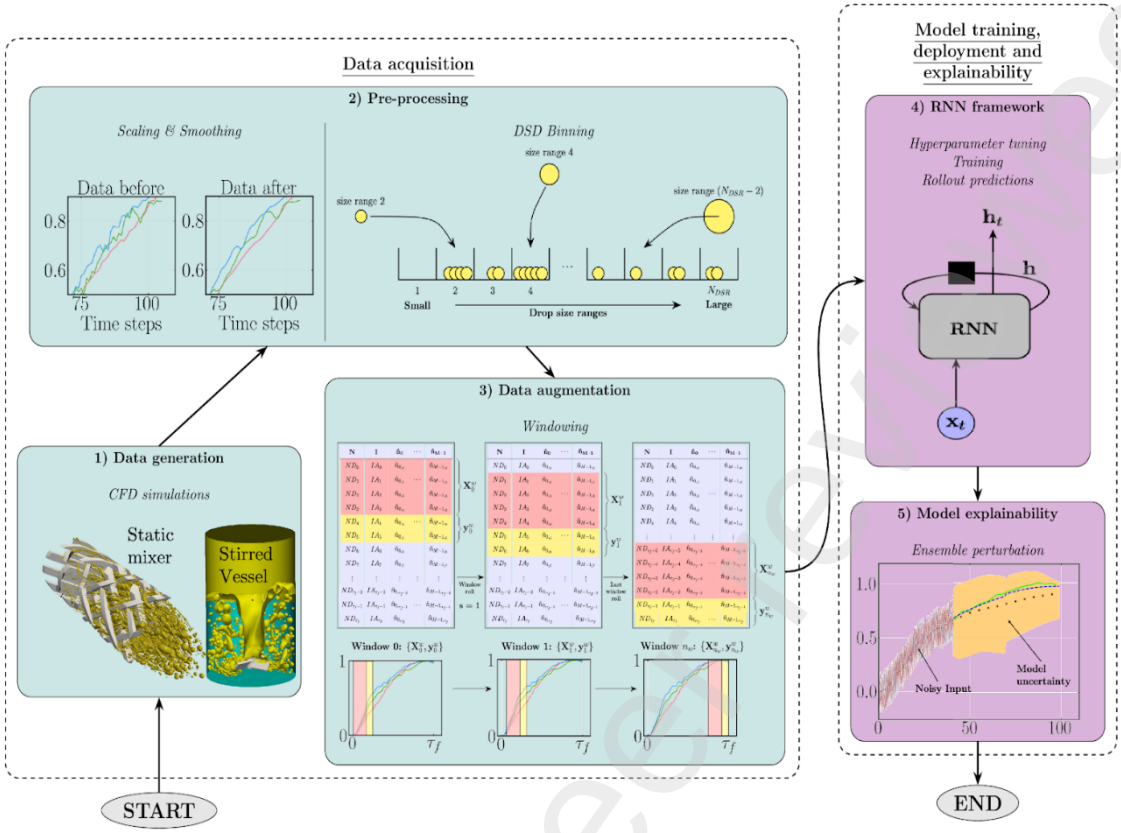


Figure 8. Flowchart detailing the workflow adopted in this study, consisting of the deployment of a time-series RNN model for multi-step performance predictions based on data from high-fidelity CFD simulations. Teal-coloured boxes refer to the process of data acquisition, while model deployment and explainability are shown in violet-coloured boxes.

The set of features considered in this study consists of two scalar quantities, namely interfacial area growth (*IA*) and number of drops (*ND*), and a list of scalars representing the droplet size distribution (DSD), whose size changes over time. This characteristic poses a challenge when dealing with standard LSTM networks as they are designed to have a fixed topology *a priori* (i.e., invariant parameter size) [(Sherstinsky, 2020)]. Consequently, a DSD binning procedure was implemented to transform drop volume lists into discrete drop counts across different size ranges, normalised through a probability density estimation function, as represented in Step 2 in **Figure** 8. This procedure introduces *M* (i.e., number of bins) as additional scalar features, from which the DSD can be later reconstructed.

In addition, a time domain sequence-to-sequence rolling window augmentation technique was implemented herein to compensate for the small dataset available for training without altering the data semantics. As shown in Step 3 in **Figure** 8, this procedure consists of splitting the original

dataset into $n_w + 1$ input and target sequences or 'windows' (coloured red and yellow, respectively in **Figure** a) with dimensions n_i, n_k, respectively. For each case, a window is initialised at $t = 0$ and rolled forward by one time-step n_w times. After each roll, the new input/target sequence pairs generated are stacked separately into new tensors $W_x \in \mathbb{R}^{n_w \times n_i \times (M+2)}$, $W_y \in \mathbb{R}^{n_w \times n_k \times (M+2)}$, which are fed during training to the LSTM network. The values for n_i, n_k, n_w impact training and predictions by influencing error propagation but these parameters were not tuned due to the impracticality of re-processing the entire dataset at every iteration.

The RNN was built using PyTorch where two different network architectures were evaluated: Fully connected (FC) and Encoder-Decoder (ED). The former refers to a single LSTM layer followed by a fully-connected linear layer which reads the hidden state from the LSTM layer and yields the predicted sequence. The latter provides a more robust framework specifically designed for multi-step forecasting (Cho et al., 2014) which consists of two LSTM sub-networks: an *Encoding* layer which derives a compressed representation of the input sequence, followed by a *Decoding* layer which interprets the encoded representation and exploits it to generate the predicted target sequence. This latter architecture provides additional flexibility when it comes to training methodologies, as it can be taught to work *recursively* on its own output, *forced* to use true data or a combination of both. Both architectures were fitted with a custom loss penalty term of the form $w_p * 1/N \sum_{i=1}^{N} (ReLu(-y))$ to avoid negative predictions (i.e., non-physical estimations), where w_p is a tuneable weight and y denotes a predicted sequence. A comprehensive hyperparameter tuning procedure was executed, sweeping over 2000 possible combinations and testing different values for batch size, loss penalty weight, hidden size, learning rate, regularisation coefficients (L_1,L_2) and training schemes for the ED architecture.

The last step displayed in **Figure** 8 is to estimate the model prediction uncertainty using an ensemble-based method from numerical weather prediction (Y. Zhu, 2005; Parker, 2013). This method involved generating an ensemble of perturbed inputs by introducing noise (ε) to feature values at each time-step, which are subsequently fed into the trained models, producing corresponding predicted sequences whence we can compute a prediction interval, which provides an overview of the range wherein the model prediction is likely to occur given the input noise presented. We evaluated the model performance on the training and validation datasets by

computing the root mean squared error (RMSE) and the coefficient of determination (R^2), which provide measures of the model's robustness against overfitting and reliability of predictions on unseen data. Following training, an exploration of model performance was undertaken on the testing datasets for the LSTM-FC for both mixers in terms of *ND, IA,* and DSD.

As displayed in **Figure** 9, the trained LSTM-FCs for both mixers capture the correct hierarchy of the simulated cases for feature *ND* and *IA*, especially at the early times. For instance, in the case of the stirred mixer, the targeted order for *ND* (see **Figure** b), *CI: 8 Hz* >$Bi = 0.002$ > $Bi = 1$, is recovered before time-step 200. However, the trained model fails to predict this ranking in the subsequent time range that the predicted ND for $Bi = 0.002$ is higher than that for *CI: 8 Hz*. Likewise, for feature *IA*, the trained model loses track of the increasing trend for *CI: 8 Hz* since time-step 300. This can be related to the model's inability to extract the underlying physical knowledge from the case of *CI: 8 Hz*, failing to be aware of the higher agitating speed in this case and thus the extended duration of drop generation. Similar observation occurs in the case of static mixer (see **Figure** 9a,c) wherein the trained LSTM-FC correctly reproduces the hierarchy at early times, though the model starts to perform poorly in the time frame further down.

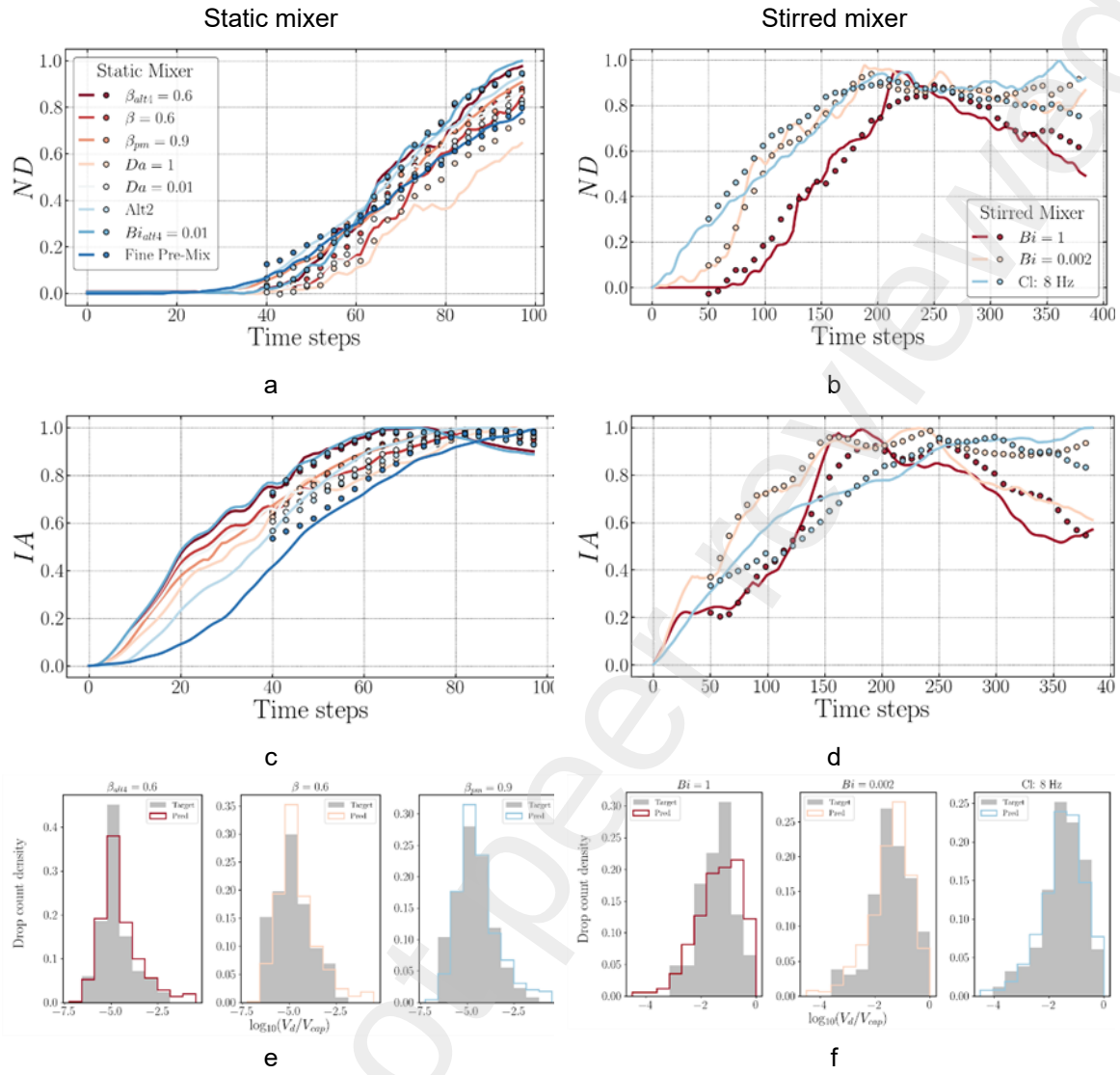


Figure 9. Plots presenting the model prediction generated via LSTM-FC for static mixer ((a),(c), and (e)) and stirred mixer ((b),(d), and (f)). Histograms (e)-(f) exemplify the predicted drop size distribution, where three test cases are shown for stirred mixer at t=280 while the results relevant to static mixer are demonstrated using test cases, β_alt=0.6, β=0.6 , and β_pm=0.9 at t=74.

Similarly, we explored the model performance relevant to the DSD, tracking the temporal deviation between targeted and predicted values for all the bins, which were introduced through data pre-processing previously. Snapshots of the DSD predictions are presented in Figure 9e and Figure 9f at a single time-step in the form of histograms (drop size density with regard to the drop size in terms of a log-scale drop volume, $\log_{10}(V_d/V_{cap})$, with V_d representing the drop volume and V_{cap} the volume of a spherical drop whose diameter corresponds to the capillary length scale, $\lambda_c = \sqrt{\frac{\alpha_{cl}}{(\rho_a - \rho_o)g}}$). These plots provide a direct insight into the degree of agreement between the targeted and predicted

distributions. In addition, to quantify the resemblance of the two distributions, the Wasserstein distance is computed; lower values of this metric indicate higher similarity.

Though it is not presented herein, it is found that LSTM-ED performs similarly to LSTM-FC in that the prediction accuracy is initially high and deteriorates subsequently. However, LSTM-ED appears to outperform LSTM-FC, to some level. For example, the abovementioned increasing trend in the case of *Cl: 8 Hz* for the stirred mixer is well captured via LSTM-ED, as well as the descending one presented for $Bi = 0.002$ and $Bi = 1$. Recall from above that LSTM-FC is trained to map an entire output sequence with corresponding input, whereas LSTM-ED learns to progressively generate the target sequence using its output from the previous time-step. This could explain the superiority of the latter, especially when dealing with the sequence generation since LSTM-FC has not been trained to utilise its previous prediction for the next target. Nevertheless, it is noticed that the trained LSTM-ED is unable to provide comparable predictions to LSTM-FC on the feature *ND* for the stirred mixer. Similarly, obvious discrepancies are presented in some cases of the static mixer. These observations suggest that the trained LSTM-ED could be further improved to achieve a better understanding of the mechanisms embedded in the current data. From above, though simply mapping between input and target sequence makes it relatively easier to train an LSTM-FC, its performance would be globally inferior to that of a well-trained LSTM-ED, since the feature evolution underlying the sequence is preserved in the latter.

Finally, we presented the corresponding results of the trained model's uncertainty obtained via the ensemble-based approach, which is illustrated in Figure 10 using features ND and IA from one testing case, i.e., β=0.6 for a static mixer and *Bi* = 0.002 for a stirred mixer. As displayed, Figure 10 compares the model target and prediction from an unperturbed input sequence, and the prediction interval region from an ensemble of perturbed inputs. For both mixers, the spread of the prediction ensemble via LSTM-FC is narrower compared to that via LSTM-ED signifying a higher sensitivity of the LSTM-ED to the added noise. This can be linked to the different predicting processes of the two models illustrated above that the latter is trained to generate target sequence step-by-step; therefore, fluctuations in feature values exert a profound effect on the model prediction for the next time-step.

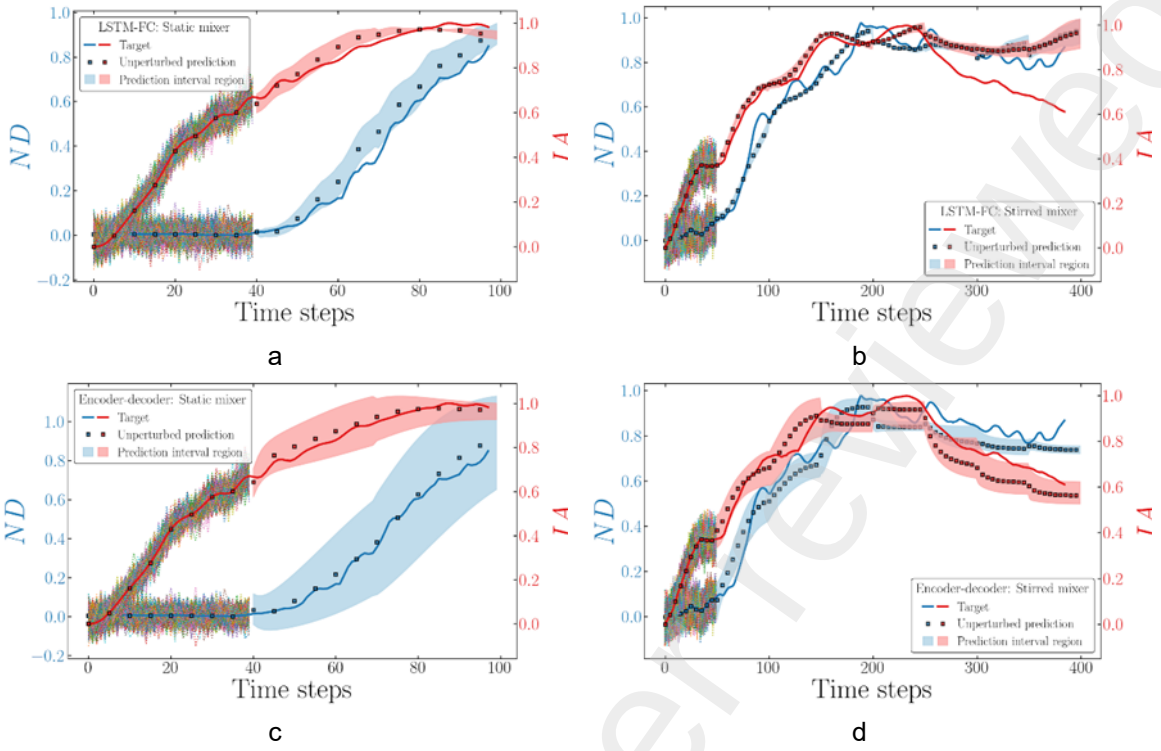


Figure 10. Plots comparing the targeted sequences (lines), predicted sequences (dots) and the ensemble prediction interval region (bands) for LSTM-FC (a),(b)) and LSTM-ED ((c),(d)). The cloud of dots at the lower left corner refer to the ensemble of perturbed input sequences.

In this section, we have provided a baseline procedure of data pre-processing, RNN-LSTM deployment, and uncertainty quantification in the field of complex multi-phase mixing, showcasing the application of the ML-based tool to predict the evolution of dispersion performance metrics of interest for two mixing systems, generated via a static and a stirred mixer.

3.4 Active learning of PBM coalescence parameters for electrolytic flows

Alkaline water electrolysers are commonly employed for industrial-scale production of low-carbon hydrogen. Operating at high current densities leads to enhanced hydrogen production, but reduced cell efficiency, partly due to ohmic losses from bubble coverage of the electrode and reduced electrolyte conductivity. To gain further insights into these phenomena, we carried out three-dimensional transient multi-physics simulations with the OpenFOAM libraries. We simulate the bubbly flow in a forced convection electrolyser with a multi-fluid Eulerian model and consider electrochemistry, heat transfer, turbulence, and population balance modelling (PBM). A major drawback of common coalescence models is that they were originally designed for air-water systems, and thus cannot be applied directly to the present flow as coalescence is inhibited in

electrolytic solutions (Orvalho et al., 2021). Rather than creating new kernels for the current application, it is possible to adapt the existing models by adjusting a constant that regulates the coalescence efficiency. Consequently, since the PBM parameters cannot be derived from first principles, we use here a Bayesian optimisation (BO) approach to avoid the need for extensive parametric investigations, allowing us to search the parameter space and optimise our model in a significantly lower number of iterations and at lower computational costs. Our OpenFOAM solver is coupled to the *scikit-optimize* library and runs autonomously to determine optimal input parameters by minimising an objective function based on the error between predicted gas volume fraction distributions, and experimental results (Riegel et al., 1998). A Gaussian process is employed here as a surrogate of the objective function.

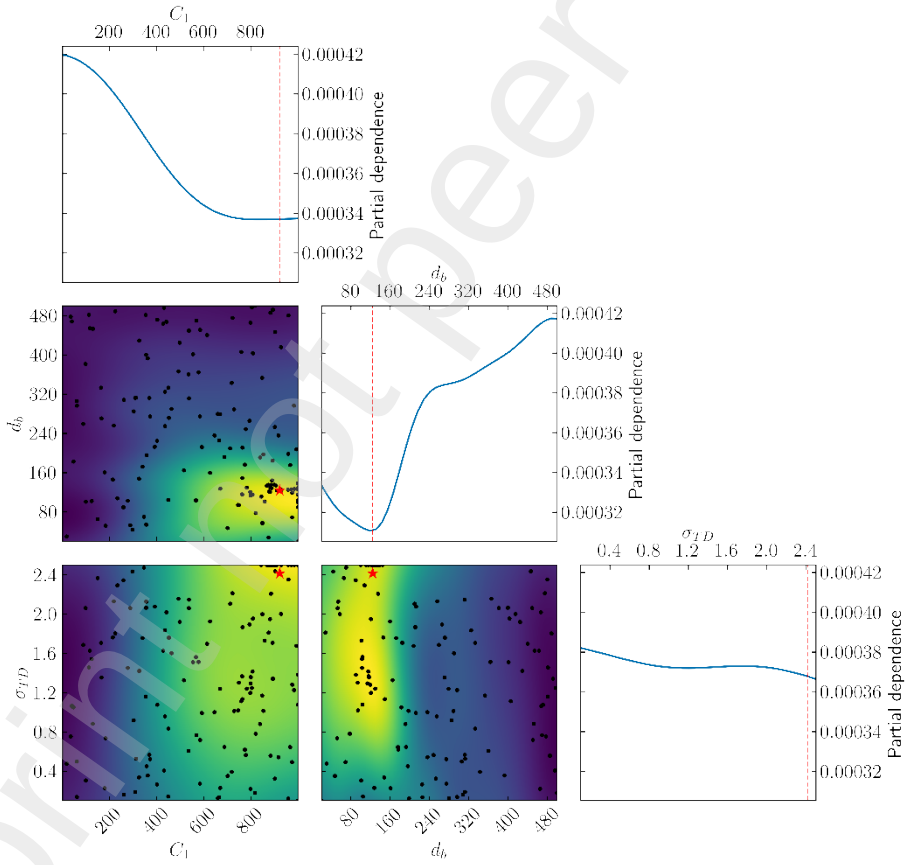


Figure 11. Partial dependence and objective function plots at imposed current density of 500 A m^{-2} and Reynolds number of 6885, using the Luo coalescence model. Here, C_1 is a constant controlling coalescence efficiency, d_b is the initial bubble diameter at the electrode, and σ_{TD} is the turbulent dispersion constant.

We carry out Bayesian optimisation of the C_1 constant which regulates the coalescence efficiency, and the initial bubble diameter at the electrode, d_b. Due to interactions between the bubble diameter

and the turbulent dispersion constant, denoted σ_{TD}, the latter is also included in the optimisation procedure. The bubble size distribution is discretised into groups of 10 μm in width, and boundary conditions are set for each size group fraction. The size group fraction of the bin closest to the diameter proposed by the acquisition function is set to one at the electrode and Neuman everywhere else, and Neuman conditions are applied everywhere to all other size groups. For each optimisation case, 50 initial simulations were performed prior to beginning the optimisation, and the BO was run for 100 iterations.

Figure 11 displays the BO results at a current density of 500 $\mathrm{A\,m^{-2}}$ and Reynolds number of 6885, using the coalescence model of (Luo, 1993). The partial dependence plots showcases the influence of a single parameter when all other parameters are kept constant, while the objective function is plotted in two dimensions and is averaged over the third dimension. The black points represent each sampling point selected by optimising the acquisition function, while the red star is the optimal point. Finally, the zones coloured yellow represent low objective function values, and the opposite holds for the areas coloured dark blue. It is important to note that the objective function and partial dependence plots are obtained from the surrogate model predictions, as opposed to the forward simulator, since the latter is too expensive to evaluate a large number of times.

The model's objective function reveals an optimal region with small bubble diameters and large C_1 values, which are associated with typical d_b values from the literature, and minimal coalescence. Examining the correlation between σ_{TD} and d_b, the optimal region is characterised by high values for both parameters, indicating low turbulent dispersion and minimal coalescence. The partial dependence plots show that larger values of C_1 and σ_{TD} values lead to a decreased loss, while an optimal region for d_b is highlighted, with 130 μm emerging as the best bubble size. The clustering of points in these optimal areas demonstrates the successful convergence of the Bayesian optimisation approach, and highlights the ability to recover experimentally uncertain parameters.

3.5 Gaussian Process Regression for prediction of droplet size distributions in sprays

Liquid atomisation is crucial for various applications, from agriculture to fuel injection and pharmaceutical inhalers. The efficient performance for these applications hinges on the understanding of the droplet size distribution (DSD) resulting from liquid jet atomisation under varied operating conditions. However, exploring the design space for optimal performance is challenging due to the complex interfacial physics governing DSD. The high-fidelity numerical simulations and experiments needed to comprehensively explore this design space are not only expensive but also often impractical for systematic and extensive investigations.

In response to these challenges, in this study, we apply a surrogate model to map the working conditions of a flat-fan liquid spray, parametrised by the spray angle *(α)*, Reynolds number *(Re)*, and Weber number *(We)*, to the resulting DSD. Data to train Gaussian process regression (GPR) machines is generated via high-fidelity numerical simulations utilising the Basilisk flow solver, incorporating an octree adaptive mesh and volume-of-fluid interface tracking. Post simulation filtering is conducted based on the equivalent diameter and sphericity to choose droplets needed for the further statistical analysis (Traverso et al., 2023).

This analysis leverages GPR to scrutinise and estimate the drop size distribution (DSD) of a liquid spray obtained from numerical simulations. In this approach, the drop population is treated as an independent and identically distributed sample of events. The unknown distribution from which the sample is drawn is denoted as *p(d)*, representing the probability density function (PDF) of drop diameters. GPR is employed to infer both the PDF and its uncertainty from the obtained sample. Figure 12a shows liquid spray for *α=65°*, *Re=20* and *We=9*. Figure 12b shows the adaptive mesh refinement capable of capturing droplets and Figure 12c shows the DSD predicted by the GPR by two different methods (Traverso et al., 2023).

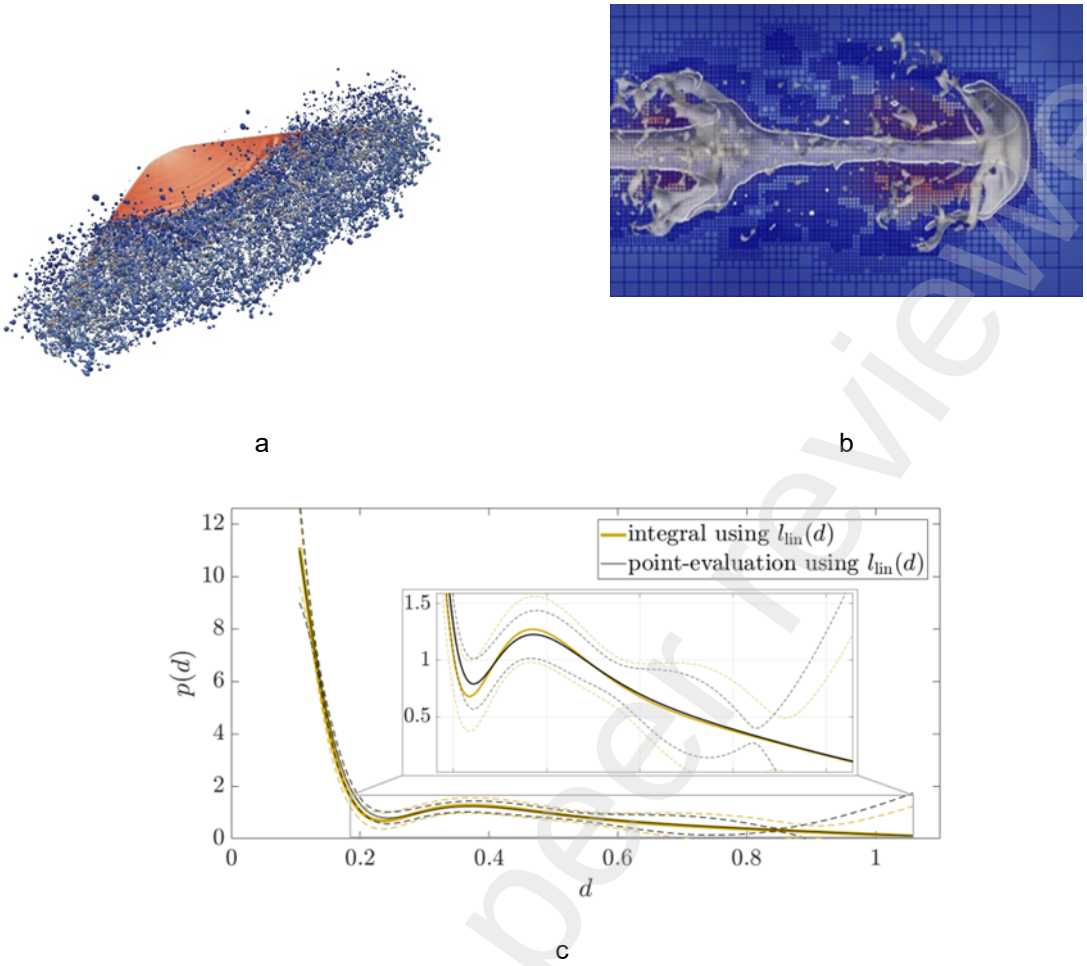


Figure 12. a) Representation of atomisation, b) Adaptive mesh refinement capturing liquid jet disintegration, and c) DSD predicted by the GPR at the condition α=65°, Re= 20 and We=9.

For predicting the DSD with respect to various operational parameters involves a two-step process. In the first step, multiple GPR models are trained to predict how each feature of the DSD changes with the input parameters. These parameters, within the range of *α= [10-65]*, *Re= [20-59]* and *We= [9-45]* are explored using 38 numerical simulations An automatic relevance determination (ARD) squared exponential kernel effectively captures the changes in DSD features based on the input parameters, with higher sensitivity to changes in We, followed by α and then Re . In the second step, the design space exploration produces DSDs at unseen conditions within the input space, such as *We>45* , successfully capturing trends like the effect of on the DSD tail and the shift in the peak diameter with changing We.

This study presents a comprehensive framework for exploring the intricate design space of liquid atomisation using a surrogate model based on Gaussian process regression. The proposed

methodology, integrating high-fidelity numerical simulations, meticulous post-simulation processing, and advanced statistical analysis, is adaptable to other complex processes beyond liquid atomisation, holding potential for widespread applications.

4 Machine Learning for particle-laden flows

A range of fields have been investigated that will benefit from surrogate models, including acoustic techniques, advanced signal processing, and machine learning-CFD hybrid approaches. Our researchers from UCL and ICL have developed an acoustic emission (AE) technique in combination with a machine learning (ML) algorithm to characterise the particle size distribution in a gas- solid fluidised bed. The AE signal is generated in solid-gas fluidised beds due to the collision and friction between fluidised particles as well as between particles and the bed inner wall, as shown in Figure 13. The generated AE signal depends on the properties of the fluid, particles, and the fluidised bed material. We have also developed a theoretical approach to explain the generation of AE signals in gas-solid flows and proposed an inversion algorithm to obtain the particle size distribution from the AE energy spectrum. This methodology was applied to AE measurements obtained in a pseudo-2D fluidised bed, characterising the particle size distribution for different particle sample sizes. The fluidised-bed walls are 10mm thick and made of polymethyl methacrylate (PMMA). The fluidised bed dimensions are 1000mm (height), 100mm (width), and 10mm (depth) and the wall is made of polymethyl methacrylate (PMMA) and has a thickness of 10mm. The particles used in this work are spherical glass beads (ballotini) with a density of 2500 kg m^{-3}, and the fluidising gas is air. Air injection at the bottom of the fluidised bed is controlled by a rotameter, while the distributor is a porous plate with a hole size of 25 μm. Before entering the fluidised bed, the flow of air is homogenised in a windbox, a plenum filled with ceramic beads approximately in diameter. All experiments were conducted at room temperature and pressure. The density and viscosity of air are 1.2 kg m^{-3} and 0.018 mPa s^{-1}, respectively.

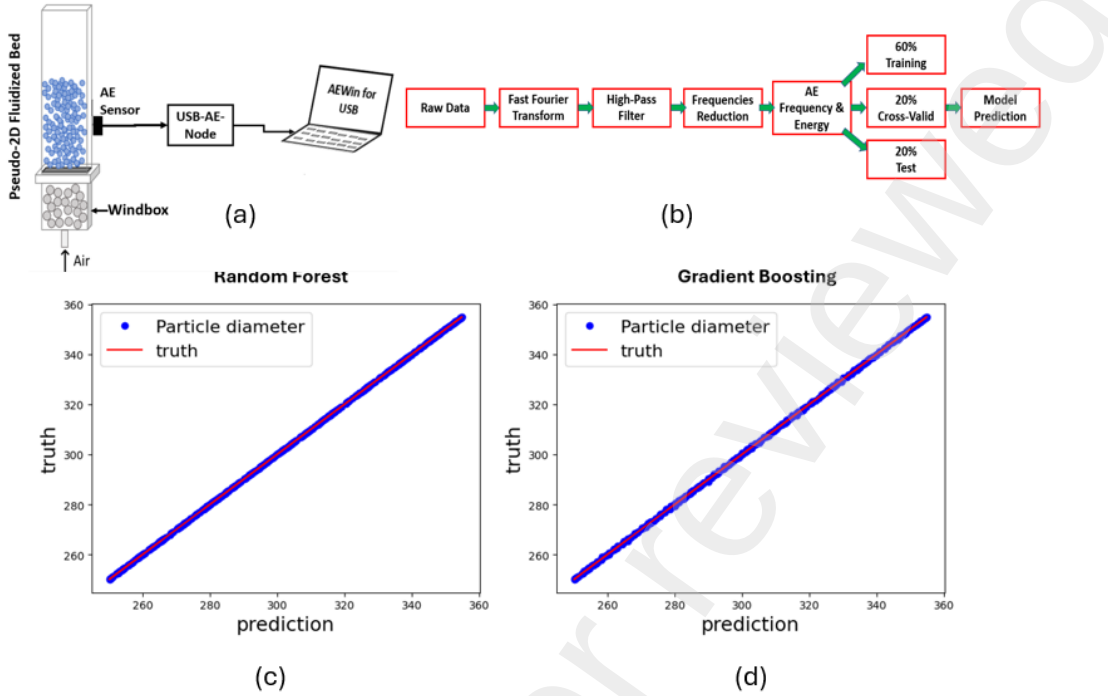


Figure 13. a) Experimental setup pseudo-2D flat fluidised bed and acoustic emission sensing diagram, b) Data preparation and ML training. Prediction vs truth on the test dataset [c) Random Forest and d) Gradient Boosting]

Additionally, for the first time, we investigated the application of machine learning techniques to invert the AE signal in the gas-solid fluidised-beds. Two ensemble machine learning algorithms, Random Forest (RF) and Gradient Boosting Regressor (GBR), were applied to the AE signal to predict the particle size distribution.

We built a machine learning regression model which takes frequency and kinetic energy from the acoustic signal as inputs and predicts the particle size distribution. The training, validation, and test datasets, each consisting of 2497 data points from each sample, were used in the machine learning process. The data post-processing and machine learning steps are illustrated in Figure 13b. We evaluated the performance of RF and GBR on the test dataset by comparing the model output against the particle diameter measured from the AE technique, as shown in Figure 13c and d. To further quantify the prediction error, we estimated two regression metrics, the RMSE and R^2 which measure the relative error of the machine learning prediction and evaluate how well the test data fits the regression model. For RF, we found that RMSE = 0.017% and $R^2 = 0.9999$, and for GBR, RMSE = 0.217% and $R^2 = 0.9999$. It is evident that both approaches, are lower than 0.5% in terms of RMSE and with $R^2 >$ 0.999, indicating a robust performance of the regression. The machine learning

methods implemented in this study not only provide an accurate prediction of the particle size distribution but also offer more insight into understanding the relationship between particle size, kinetic energy, particle velocity, and AE frequency.

5 Future perspectives

The preceding sections covered a range of multiphase flows, spanning droplets, liquid-liquid, gas-liquid, and solid-gas systems. Within these systems, the impact of geometry was highlighted. Specifically, the T-mixer and helical coil reactors used in nanoparticle synthesis showed distinct mixing behaviour and yielded varying particle sizes (Section 2.2). In the context of microfluidic droplet coalescence, the initiation of reactions was found to be critically dependent on the channel height, which determines drop interface contact (Section 2.3). Moreover, the formation of droplets was shown to be influenced by nozzle geometries (Section 2.1), as well as the geometries of flow-focusing channels (Section 2.3). Broadly, this shows that the interfacial tension effects, fluid inertia, and hydrodynamics are sensitive to the specific geometry of the domain, which affects the process efficiency, selectivity, and stability.

However, manually identifying optimal designs becomes impractical when dealing with large sets of design parameters. Therefore, the adoption of data-driven methods becomes imperative to navigate the complexity of these expansive design spaces. The representation of geometries, achieved through shape parametrisation methods like standard design parameters (Mansour et al., 2020), free-form geometry deformation (Kurenkov et al., 2018), or curve models (ex. nonuniform rational B-splines (Najafi et al., 2017)), demands parametrisations that can range from less than 10 to millions, depending on the desired level of flexibility. However, this wide range of parametrisations can significantly escalate computational expenses. Given the prohibitively high computational cost associated with a single multiphase flow problem, even with a parameter size as modest as 50, the optimisation cost becomes impractical. Consequently, there is a clear need for the implementation of multi-fidelity frameworks in the optimisation of designs for multiphase flow problems.

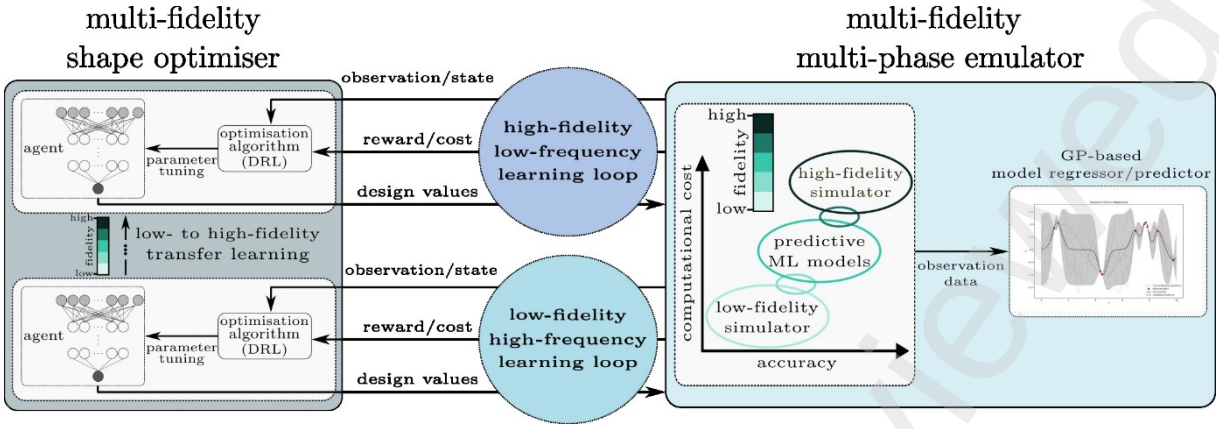


Figure 14. Schematic of a closed-loop multi-fidelity shape optimisation framework, where deep neural networks are trained to optimise design parameters through unsupervised input-output interaction with a multi-fidelity multi-phase emulator. The emulator comprises a unified hierarchy of physics-based and data-driven predictive models of different fidelities. The optimiser interacts with the emulator across a spectrum of fidelities, ranging from cost-efficient low-fidelity to resource-intensive high-fidelity learning loops, allowing learned knowledge to flow up through the varying levels of fidelity. Transfer learning in the context of multi-fidelity optimisation enables the optimiser to establish fast learning in simpler environments and progressively enhance its decision-making capabilities as it receives more detailed flow scenarios from high-fidelity multi-phase models.

Multi-fidelity modelling bridges the gap between rapid computation and high accuracy by combining different levels of fidelity (see Figure 14). This allows exploring large design spaces at lower overall cost. Multi-fidelity models have gained significant attention since the 2000s due to their capacity to balance between accuracy and cost. Recent literature reflects an increase in the adoption of multi-fidelity techniques in both optimisation and fluid mechanics (M. Giselle Fernández-Godino, 2023; Peherstorfer et al., 2018). Our prior research shows the effectiveness of multi-fidelity modelling in uncovering intricate flow phenomena (Savage et al., 2023; Savage, Basha, McDonough, Matar, & del Rio Chanona, 2023). Specifically, the revelation of Dean vortices, which enhance radial mixing and reduce axial mixing, in coiled tube reactors at a low Reynolds number (50) was a result of this approach (Basha et al., 2023). These vortices were not previously identified without the superimposition of oscillatory velocities at the coil inlet . The success of this discovery can be attributed to the exploration of shape using 72 shape-related parameters in both the cross-sectional and axial directions of the coil. The various fidelities in this study are discretisation in both radial and axial directions, leading to mesh node counts as low to high fidelities. The opportunity now lies in extending and adapting these techniques for multiphase flows (see Figure 14).

We are also currently exploring an approach that has the potential to seamlessly integrate multi-fidelity models, for example surrogate models with physics-based models. The idea stems from the realisation that many discretisations can be written as discrete convolutions, which, in turn means that discretised systems can be written as convolutional neural networks whose weights are determined by the discretisation method and not by training (Chen et. al, 2024). These systems can be solved by multigrid methods expressed in a U-Net architecture. The advantages of writing physics-based models in this way are (1) the models can be run any platform (CPUs, GPUs, AI processors) with barely any code modification; (2) coupled multi-physics models can easily be developed as neural networks can be linked together; (3) sub-grid-scale approaches can be easily developed which combine physics-based approaches (written as neural networks with analytically determined weights) with learned closure models (expressed as trained neural network); (4) the optimisation techniques used in training neural networks can now easily be applied to physics-based models, which are now fully differentiable.

Therefore, fidelities within the computational framework can be influenced by discretisation parameters such as mesh count, type, and size. Additionally, the choice of numerical solution methods, including convergence tolerances, Courant numbers, iterative approaches, and discretisation schemes spanning from low to higher order, adds an additional spectrum of accuracy-cost considerations. Moreover, the choice of fluid-fluid interface modelling methods introduces different fidelity levels. These methods range from low-fidelity interface capturing techniques, such as the volume of fluid (VOF) and level-set methods, to high-fidelity interface tracking approaches, including interface-fitted moving mesh and interface tracking methods. Turbulence models, such as Direct Numerical Simulation (DNS), Reynolds-Averaged Navier-Stokes (RANS), Large Eddy Simulation (LES), and algebraic stress models, also contribute to a spectrum of fidelities that cover resolved physics.

Data-driven predictive ML models such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), as demonstrated in previous sections for predicting microfluidic droplet data or droplet dynamics, can serve as low-fidelity surrogates. These ML approaches can rapidly predict multiphase data once trained, facilitating efficient navigation of the design space. However, it is important to note that they may lack the accuracy of physics-based models when extrapolating

beyond the training coverage or representing fine changes to geometric details. Thereby, fidelities individually, in various combinations, can form a spectrum that balances solution accuracy against cost within the multiphase flow framework.

We give a couple of examples to illustrate modelling level fidelities for gas-liquid problems, such as liquid jet atomisation and gas bubbles in a liquid medium. Liquid jet atomisation demands a certain droplet size distribution (DSDs) for which one of the influential parameters is the nozzle geometry. DSDs are a result of simulating jets fragmenting into myriad droplets as interfaces dynamically distort and demand excessive resolution stretching across vast length scales. Adaptive mesh refinement (AMR) alleviates this by localising grid resolution. Coarser base meshes with diffuse interface capturing establish low-fidelity, inexpensive simulations. Adding AMR refinement criteria triggered by velocity and volume fraction gradients then targets emerging ligaments and droplets for higher fidelity representation (Traverso et al. 2023). Second, gas bubbles in a liquid medium, where interface capturing models range from the lowest fidelity, while state-of-the-art reconstructed interface tracking offers the highest accuracy.

Recent advancements in multi-fidelity optimisation have harnessed machine learning techniques to bridge models, ensuring a balance between cost and accuracy. Techniques such as multi-fidelity Bayesian optimisation using deep Gaussian processes for fidelity mapping, exploring fidelities in continuous space, applying deep attention mechanisms to capture intricate dependencies across fidelities, and utilising Kriging models, or Bayesian inference methods have shown promise. Nevertheless, these methods heavily depend on the predominant paradigm of supervised learning or predictive ML, which either requires a substantial amount of labelled data or assumes that the underlying model of the dynamic problem can be estimated as smooth and differentiable functions sampled stochastically from multivariate Gaussian distributions.

The ideal approach involves creating a comprehensive autonomous optimisation framework that integrates CFD simulators, data-driven surrogate models, and analytical/statistical models into a unified predictive multiphase flow emulator within the optimisation workflow. Within this framework, the optimisation algorithm should effectively engage in a two-way interaction with the emulator to gain information and adjust optimisation parameters, while leveraging the strengths across various fidelity levels to efficiently explore complex design spaces. Deep Reinforcement Learning (DRL) is

a promising outer-loop intelligence paradigm for constructing such an optimisation framework (see Figure 14). Model-free DRL, operating within a closed-feedback loop, trains a stochastic deep neural network (agent) to make sequential optimal decisions solely through input-output interactions with the underlying dynamic system (environment) (Sutton & Barto, 1998). This approach remains effective irrespective of the complexity and stochasticity of the system being optimised. Consequently, DRL can adeptly address the curse of dimensionality, marked by exponential computational complexity growth as the number of optimisation parameters increases, and the curse of modelling, where explicit well-defined mathematical models for predicting the transition of underlying dynamics are lacking, a common issue in multi-phase flow systems (Garnier et al., 2021).

The evolving research on automated multi-fidelity models within the machine learning community shows great promise for further enhancing shape optimisation strategies. These advancements are expected to play a crucial role in advancing the future of discoveries through shape optimisation within the realm of multiphase flows. Three key aspects need immediate attention in this process. Firstly, it is crucial to establish an intelligent mechanism for propagating fidelity relationships throughout the optimisation models. Secondly, there is a need to strike a balance between the cost and information gained through optimisation models, ensuring the selection of an appropriate fidelity level for evaluation. Thirdly, addressing the data challenges associated with these models to harness their full potential for efficient and computationally tractable shape optimisation.

6 Conclusions

In this article for the Special Issue on Machine Learning for Multiphase Flows, we have provided some highlights from the PREMIERE (PREdictive Modelling with QuantIfication of UncERtainty for MultiphasE Systems) programme that showcased the use of machine learning techniques combined with physics-driven approaches based on the use of experiments and/or computational fluid dynamics (CFD) simulations. For multiphase flows with low inertia, we included examples featuring the use of Design of Experiments for nanoparticle synthesis optimisation, Generalised Latent Assimilation (GLA) models for drop coalescence predictions, and Bayesian regularised artificial neural networks, and eXtreme Gradient Boosting (XGBoost) for microdroplet formation prediction in in the presence of surfactants microfluidic devices.

For inertial multiphase flows, we highlighted the use of novel sub-sampling allied to adversarial neural network architectures to predict the characteristics of air-water slug flows in long pipelines. Furthermore, we introduced a generalised latent assimilation technique, Long-Short-Term Memory networks for predicting the performance of stirred and static mixers, and the use of active learning via Bayesian optimisation to determine parameters for coalescence kernels for PBM-CFD models for green hydrogen production in high current density electrolysers. We also showcased the deployment of Gaussian Process regression and CFD simulations with adaptive mesh refinement to predict drop size distributions for sprays, and the combined use of machine learning and acoustic emission experiments to predict the particle size distribution in gas-solid fluidised beds. Finally, perspectives were discussed on the development of a closed-loop shape optimisation framework that makes use of a multiphase flow emulator with a hierarchy of data- and physics-driven multi-fidelity models.

The work presented here shows that the combined use of machine learning and physics-driven methods offers enhanced modelling, characterisation, optimisation, and prediction for complex multiphase systems. Key benefits include significantly reduced computational costs, accelerated discovery of optimal operating conditions, faster convergence, and precise quantification of flow characteristics matching experimental validation data. The impact spans diverse applications such as chemical synthesis, microfluidics, electrolytic hydrogen generation, fast-moving consumer goods, and particle-laden systems across the energy and manufacturing industries. It is hoped that the PREMIERE work presented in this Special Issue on Machine Learning for Multiphase Flows will lead to exciting collaborations within the multiphase flow community.

Acknowledgements

This research is funded by the EP/T000414/1 PREdictive Modelling with QuantIfication of UncERtainty for MultiphasE Systems (PREMIERE), United Kingdom.

REFERENCES

Afshar Ghotli, R., Abbasi, M. R., Bagheri, A. H., Raman, A. A. A., Ibrahim, S., & Bostanci, H. (2019). Experimental and modeling evaluation of droplet size in immiscible liquid-liquid stirred vessel using various impeller designs. *Journal of the Taiwan Institute of Chemical Engineers*, *100*. https://doi.org/10.1016/j.jtice.2019.04.005

Quilodrán-Casas, C., Arcucci, R., Pain, C., & Guo, Y. (2021). *Adversarially trained LSTMs on reduced order models of urban air pollution simulations*. ArXiv. /abs/2101.01568

Basha, N., Savage, T., McDonough, J., del Rio Chanona, E. A., & Matar, O. K. (2023). Discovery of mixing characteristics for enhancing coiled reactor performance through a Bayesian optimisation-CFD approach. Chemical Engineering Journal, 473. https://doi.org/10.1016/j.cej.2023.145217

Brunton, S. L., Noak, B. R., Koumoutsakos, P (2020) Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech. 52, 477-508.*

Chagot, L., Quilodrán-Casas, C., Kalli, M., Kovalchuk, N. M., Simmons, M. J. H., Matar, O. K., Arcucci, R., & Angeli, P. (2022). Surfactant-laden droplet size prediction in a flow-focusing microchannel: a data-driven approach. *Lab on a Chip*, *22*(20), 3848–3859. https://doi.org/10.1039/d2lc00416j

Chen, B., Heaney, C. E., Gomes, J. M. L. A., Matar, O. K., Pain, C. C. (2024) Solving the Discretised Multiphase Flow Equations with Interface Capturing on Structured Grids Using Machine Learning Libraries, arXiv preprint 2401.06755. https://doi.org/10.48550/arXiv.2401.06755

Cheng, S., Liu, C., Guo, Y., & Arcucci, R. (2024). Efficient deep data assimilation with sparse observations and time-varying sensors. *Journal of Computational Physics*, *496*, *11258*. https://doi.org/10.1016/j.jcp.2023.112581

Chen, J., Anastasiou, C., Cheng, S., Basha, N. M., Kahouadji, L., Arcucci, R., Angeli, P., & Matar, O. K. (2023). Computational fluid dynamics simulations of phase separation in dispersed oil-water pipe flows. *Chemical Engineering Science*, *267*, 118310. https://doi.org/10.1016/j.ces.2022.118310

Cheng, S., Chen, J., Anastasiou, C., Angeli, P., Matar, O. K., Guo, Y. K., Pain, C. C., & Arcucci, R. (2023). Generalised Latent Assimilation in Heterogeneous Reduced Spaces with Machine Learning Surrogate Models. *Journal of Scientific Computing*, *94*(1). https://doi.org/10.1007/s10915-022-02059-4

Cheng, S., Quilodran-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., & Arcucci, R. (2023). Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review. In *IEEE/CAA Journal of Automatica Sinica* (Vol. 10, Issue 6). https://doi.org/10.1109/JAS.2023.123537

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/d14-1179

Deen, N. G., Peters, E. A. J. F., Padding, J. T., Kuipers, J. A. M. (2014) Review of direct numerical simulation of fluid−particle mass, momentum and heat transfer in dense gas−solid flows. *Chem. Eng. Sci., 116, 710−724.*

Duraisamy, K, Iaccarino, G., Xiao, H. (2019) Turbulence modelling in the age of data. *Annu. Rev. Fluid Mech. 51, 357-377.*

Esteghamatian, A., Bernard, M., Lance, M., Hammouti, A., Wachs, A. (2017) Micro/meso simulation of a fluidized bed in a homogeneous bubbling regime. *Int. J. Multiphase Flow, 92, 93−111*.

Fox, R. O. (2012) Large-eddy-simulation tools for multiphase flows. Annu. Rev. Fluid Mech., *44, 47-76*.

Gidaspow, D. (1994) Multiphase flow and fluidization: continuum and kinetic theory descriptions. Academic Press.

Garnier, P., Viquerat, J., Rabault, J., Larcher, A., Kuhnle, A., & Hachem, E. (2021). A review on deep reinforcement learning for fluid mechanics. *Computers and Fluids*, *225*. https://doi.org/10.1016/j.compfluid.2021.104973

Gelado, S. H., Quilodrán-Casas, C., & Chagot, L. (2023). Enhancing Microdroplet Image Analysis with Deep Learning. *Micromachines*, *14(10)*. https://doi.org/10.3390/mi14101964

Heaney, C. E., Liu, X., Go, H., Wolffs, Z., Salinas, P., Navon, I. M., & Pain, C. C. (2022). Extending the Capabilities of Data-Driven Reduced-Order Models to Make Predictions for Unseen Scenarios: Applied to Flow Around Buildings. *Frontiers in Physics*, *10*. https://doi.org/10.3389/fphy.2022.910381

Heaney, C. E., Wolffs, Z., Tómasson, J. A., Kahouadji, L., Salinas, P., Nicolle, A., Navon, I. M., Matar, O. K., Srinil, N., & Pain, C. C. (2022). An AI-based non-intrusive reduced-order model for extended domains applied to multiphase flow in pipes. *Physics of Fluids*, *34*(5). https://doi.org/10.1063/5.0088070

Hossein, F., Materazzi, M., Lettieri, P., & Angeli, P. (2021). Application of acoustic techniques to fluid-particle systems – A review. Hossein, F., Materazzi, M., Lettieri, P., & Angeli, P. (2021). Application of acoustic techniques to fluid-particle systems – A review. *Chemical Engineering Research and Design*, *176*, 180-193. https://doi.org/10.1016/j.cherd.2021.09.031

Hossein, F., Materazzi, M., Errigo, M., Angeli, P., & Lettieri, P. (2022). Application of ultrasound techniques in Solid-Liquid fluidized bed. *Measurement*, *194*, 111017. https://doi.org/10.1016/j.measurement.2022.111017

Hu, J., Zhu, K., Cheng, S., Kovalchuk, N. M., Soulsby, A., Simmons, M. J., Matar, O. K., & Arcucci, R. (2024). Explainable AI models for predicting drop coalescence in microfluidics device. *Chemical Engineering Journal*, *481*, *14846*.

Kalli, M., Pico, P., Chagot, L., Kahouadji, L., Shin, S., Chergui, J., Juric, D., Matar, O. K., & Angeli, P. (2023). Effect of surfactants during drop formation in a microfluidic channel: a combined experimental and computational fluid dynamics approach. *Journal of Fluid Mechanics*, *961*. https://doi.org/10.1017/jfm.2023.213

Kurenkov, A., Ji, J., Garg, A., Mehta, V., Gwak, J., Choy, C., & Savarese, S. (2018). DeformNet: Free-form deformation network for 3D shape reconstruction from a single image. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-January*. https://doi.org/10.1109/WACV.2018.00099

Lebaz, N., & Sheibat-Othman, N. (2019). A population balance model for the prediction of breakage of emulsion droplets in SMX+ static mixers. *Chemical Engineering Journal*, *361*. https://doi.org/10.1016/j.cej.2018.12.090

Lei, H., Zhu, L. T., Luo, Z. H. (2021) Study of filtered interphase heat transfer using highly resolved CFD−DEM simulations. *AIChE J.*, *67(4), e17121 (32)*

Liang, F., Kahouadji, L., Valdes, J. P., Shin, S., Chergui, J., Juric, D., & Matar, O. K. (2022). Numerical study of oil-water emulsion formation in stirred vessels: effect of impeller speed. *Flow Measurement and Instrumentation*, *2*. https://doi.org/10.1017/flo.2022.27

Liang, F., Kahouadji, L., Valdes, J. P., Shin, S., Chergui, J., Juric, D., & Matar, O. K. (2023). Numerical simulation of surfactant-laden emulsion formation in an un-baffled stirred vessel. *Chemical Engineering Journal*, *472*. https://doi.org/10.1016/j.cej.2023.144807

Luo, H. (1993). Coalescence, breakup and liquid circulation in bubble column reactors. *Thesis*.

M. Giselle Fernández-Godino. (2023). Review of multi-fidelity models. *Advances in Computational*

Science and Engineering, *1(4)*, 351–400.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). *Adversarial Autoencoders*. ArXiv. /abs/1511.05644

Mansour, M., Zähringer, K., Nigam, K. D. P., Thévenin, D., & Janiga, G. (2020). Multi-objective optimization of liquid-liquid mixing in helical pipes using Genetic Algorithms coupled with Computational Fluid Dynamics. *Chemical Engineering Journal*. https://doi.org/10.1016/j.cej.2019.123570

Maulik, R., Lusch, B., & Balaprakash, P. (2020). Non-autoregressive time-series methods for stable parametric reduced-order models. *Physics of Fluids*, *32*(8). https://doi.org/10.1063/5.0019884

Najafi, A. R., Safdari, M., Tortorelli, D. A., & Geubelle, P. H. (2017). Shape optimization using a NURBS-based interface-enriched generalized FEM. *International Journal for Numerical Methods in Engineering*, *111*(10). https://doi.org/10.1002/nme.5482

Nathanael, K., Cheng, S., Kovalchuk, N. M., Arcucci, R., & Simmons, M. J. H. (2023). Optimization of microfluidic synthesis of silver nanoparticles: A generic approach using machine learning. *Chemical Engineering Research and Design*, *193*. https://doi.org/10.1016/j.cherd.2023.03.007

Nathanael, K., Galvanin, F., Kovalchuk, N. M., & Simmons, M. J. H. (2023). Development of a predictive response surface model for size of silver nanoparticles synthesized in a T-junction microfluidic device. *Chemical Engineering Science*, *279*. https://doi.org/10.1016/j.ces.2023.118907

Nathanael, K., Pico, P., Kovalchuk, N. M., Lavino, A. D., Simmons, M. J. H., & Matar, O. K. (2022). Computational modelling and microfluidics as emerging approaches to synthesis of silver nanoparticles – A review. In *Chemical Engineering Journal* (Vol. 436). https://doi.org/10.1016/j.cej.2022.135178

Obeysekara, A., Salinas, P., Heaney, C. E., Kahouadji, L., Via-Estrem, L., Xiang, J., Srinil, N., Nicolle, A., Matar, O. K., & Pain, C. C. (2021). Prediction of multiphase flows with sharp interfaces using anisotropic mesh optimisation. *Advances in Engineering Software*, *160*(March). https://doi.org/10.1016/j.advengsoft.2021.103044

Orvalho, S., Stanovsky, P., & Ruzicka, M. C. (2021). Bubble coalescence in electrolytes: Effect of bubble approach velocity. *Chemical Engineering Journal*, *406*. https://doi.org/10.1016/j.cej.2020.125926

Parker, W. S. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(3). https://doi.org/10.1002/wcc.220

Paul, E. L., Atiemo-Obeng, V. A., & Kresta, S. M. (2004). *Handbook of Industrial Mixing: Science and Practice - Wiley Online Library*. John Wiley & Sons.

Pico, P., Nathanael, K., Lavino, A. D., Kovalchuk, N. M., Simmons, M. J. H., & Matar, O. K. (2023). Silver nanoparticles synthesis in microfluidic and well-mixed reactors: A combined experimental and PBM-CFD study. *Chemical Engineering Journal*, *474*. https://doi.org/10.1016/j.cej.2023.145692

Riegel, H., Mitrovic, J., & Stephan, K. (1998). Role of mass transfer on hydrogen evolution in aqueous media. *Journal of Applied Electrochemistry*, *28*(1). https://doi.org/10.1023/A:1003285415420

Salinas, P., Pavlidis, D., Xie, Z., Jacquemyn, C., Melnikova, Y., Jackson, M. D., & Pain, C. C. (2017). Improving the robustness of the control volume finite element method with application to multiphase porous media flow. *International Journal for Numerical Methods in Fluids*, *85*(4). https://doi.org/10.1002/fld.4381

Savage, T., Basha, N., McDonough, J., Matar, O. K., & del Rio-Chanona, E. A. (2023). *Machine Learning-Assisted Discovery of Novel Reactor Designs via CFD-Coupled Multi-fidelity Bayesian Optimisation*. https://doi.org/10.48550/arXiv.2308.08841

Savage, T., Basha, N., McDonough, J., Matar, O. K., & del Rio Chanona, E. A. (2023). Multi-fidelity data-driven design and analysis of reactor and tube simulations. Computers and Chemical Engineering, 179. https://doi.org/10.1016/j.compchemeng.2023.108410

Seyed-Ahmadi, A., Wachs,A. (2020) Microstructure-informedproba-bility-drivenpoint-particlemodelfor hydrodynamic forces and torques in particle-laden flows. *J. Fluid Mech., 900, A12 (27).*

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*. https://doi.org/10.1016/j.physd.2019.132306

Shin, S., Chergui, J., Juric, D., Kahouadji, L., Matar, O. K., & Craster, R. V. (2018). A hybrid interface tracking – level set technique for multiphase flow with soluble surfactant. *Journal of Computational Physics*, *359*. https://doi.org/10.1016/j.jcp.2018.01.010

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, *9*(5). https://doi.org/10.1109/tnn.1998.712192

Traverso, T., Abadie, T., Matar, O. K., & Magri, L. (2023). Data-driven modeling for drop size distributions. *Phys. Rev. Fluids*, *8*(10), 104302. https://link.aps.org/doi/10.1103/PhysRevFluids.8.104302

Valdes, J. P., Kahouadji, L., Liang, F., Shin, S., Chergui, J., Juric, D., & Matar, O. K. (2023a). Direct numerical simulations of liquid–liquid dispersions in a SMX mixer under different inlet conditions. *Chemical Engineering Journal*, *462*. https://doi.org/10.1016/j.cej.2023.142248

Valdes, J. P., Kahouadji, L., Liang, F., Shin, S., Chergui, J., Juric, D., & Matar, O. K. (2023b). On the dispersion dynamics of liquid–liquid surfactant-laden flows in a SMX static mixer. *Chemical Engineering Journal*, *475*. https://doi.org/10.1016/j.cej.2023.146058

Valdés, J. P., Kahouadji, L., & Matar, O. K. (2022). Current advances in liquid–liquid mixing in static mixers: A review. In *Chemical Engineering Research and Design* (Vol. 177). https://doi.org/10.1016/j.cherd.2021.11.016

Venkatasubramanian, V (2019) The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE J. 65 (2), 466-478.*

Vikhansky, A. (2020). CFD modelling of turbulent liquid–liquid dispersion in a static mixer. *Chemical Engineering and Processing - Process Intensification*, *149*. https://doi.org/10.1016/j.cep.2020.107840

Wang, B and Wang, J. (2021) Application of artificial intelligence in computational fluid dynamics. *Ind. Eng. Chem. Res., 60 (7), 2772-2790.*

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7). https://doi.org/10.1109/TIP.2017.2662206

Zhu, K., Cheng, S., Kovalchuk, N., Simmons, M., Guo, Y. K., Matar, O. K., & Arcucci, R. (2023). Analyzing drop coalescence in microfluidic devices with a deep learning generative model. *Physical Chemistry Chemical Physics*, *25*(23). https://doi.org/10.1039/d2cp05975d

Zhu, L-T, Chen, X-Z, Ouyang, B., Yan, W-C., Lei, H., Chen, Z., Luo, Z-H (2022) Review of machine learning for hydrodynamics, transport, and reactions in multiphase flows and reactors. *Ind. Eng. Chem. Res. 61, 9901-9949.*

Zhu, Y. (2005). Ensemble forecast: A new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, *22*(6). https://doi.org/10.1007/BF02918678

Zhuang, Y., Cheng, S., Kovalchuk, N., Simmons, M., Matar, O. K., Guo, Y. K., & Arcucci, R. (2022). Ensemble latent assimilation with deep learning surrogate model: application to drop interaction in a microfluidics device. *Lab Chip*, 3187–3202. https://doi.org/10.1039/d2lc00303a