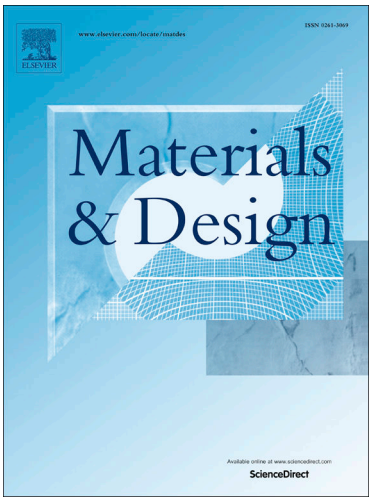
Journal Pre-proofs




Machine Learning Predicts Electrospray Particle Size

Fanjin Wang, Moe Elbadawi, Scheilly Liu Tsilova, Simon Gaisford, Abdul W. Basit, Maryam Parhizkar

PII: S0264-1275(22)00357-4
DOI: https://doi.org/10.1016/j.matdes.2022.110735
Reference: JMADE 110735

To appear in: *Materials & Design*

Received Date: 17 January 2022
Revised Date: 6 May 2022
Accepted Date: 7 May 2022

Please cite this article as: Wang, F., Elbadawi, M., Liu Tsilova, S., Gaisford, S., Basit, A.W., Parhizkar, M., Machine Learning Predicts Electrospray Particle Size, *Materials & Design* (2022), doi: https://doi.org/10.1016/j.matdes.2022.110735

# Machine Learning Predicts Electrospray Particle Size

*Fanjin Wang[1], Moe Elbadawi[1], Scheilly Liu Tsilova[1], Simon Gaisford[1], Abdul W. Basit[1*], Maryam Parhizkar[1*]*

[1]University College London, 29-39 Brunswick Square, London WC1N 1AX, UK.

* Corresponding author: Parhizkar, Maryam (m.parhizkar@ucl.ac.uk); Basit, Abdul. (a.basit@ucl.ac.uk).

**Abstract**

Electrospraying (ES) is a state-of-the-art processing technique with the promise of achieving key nanotechnology and contemporary manufacturing needs. As a versatile technique, ES can produce particles with different sizes, morphologies, and porosities by tuning a list of experiment parameters. However, this level of precision demands an exhaustive trial-and-error approach, at high costs and heavily relies on processing expertise. The present study demonstrates how machine learning (ML) can expedite the optimization process by accurately predicting particle diameter, for both nano- and micron-sized particles. This was achieved by constructing an informative electrospraying database containing 445 records from the literature, followed by the development of predictive ML models. Feature engineering techniques were explored, where ultimately it was found that solvent physiochemical properties as the molecular representation and data with imputation provided models the highest performance. The top two models were XGBoost and Random Forest (RF), which yielded root-mean-squared errors (RMSE) of 3.91 μm and 6.19 μm evaluated by 5-fold cross-validation (CV), respectively. These models were experimentally validated in-house with different combinations of experiment parameters, where RMSE between the predicted and actual particle size was found to be 1.30 μm for the XGBoost model and 1.62 μm for the RF model. Therefore, it was concluded that data generated by the ES literature, in addition to being both cost- and material-free, can yield high-performing ML models for predicting particle size. The ML models were also consulted to determine the key processing parameters that govern particle size, where it was concluded that the models learnt similar attributes identified by scaling laws.

**Keywords:** continuous manufacturing; nanomedicines & nanomaterials; digital fabrication technologies; *in silico* modelling; artificial intelligence.

# 1. Introduction

Electrospray (ES) is a powder fabrication technique expected to be a key driver in the nanotechnology sector for its ability to seamlessly produce nano-sized particles [1–5]. Whilst other technologies provide limited control over particle size and morphology, ES has been successfully demonstrated to produce various morphologies, including nanorods, nanofibers, and nanoribbons[6–8]. In addition, the fast-drying nature and one-step procedure of ES presents an opportunity for the technology to be integrated into continuous manufacturing applications [9–15]. Furthermore, ES allows for mild processing conditions where no high heat treatment nor high pressure are required throughout the process, which is especially favoured in the drug delivery and tissue engineering research for the production of particles loaded with sensitive materials [16–18]. Collectively, these traits of ES provide it with the potential to address unmet needs in a variety of fields, and thus a fabrication method of topical significance [19–23].

The versatility of ES stems from a range of tuneable parameters that provides users with the flexibility to achieve varying particle properties, making it a favourable fabrication technique. However, this versatility becomes a double-edged sword in the designing phase. This is because the process of finding out a suitable combination of interrelated parameters to achieve the desired properties is highly complicated. On the one hand, the number of possible combinations of parameters grows exponentially with the number of variables associated with the ES processing. Furthermore, the extra time and efforts to characterize the products, which are normally in the nano and microscale, make it a time- and resource-consuming process to optimize ES products. Therefore, there is a pressing need for tools that enable real-time prediction of the ideal parameters to expedite processing of the desired products.

A standard ES setup design contains a high voltage power source, a syringe pump, a metal collector, tubing and a syringe with a metal nozzle [24]. During the ES process, a viscous solution is pumped to the metal nozzle and charged by the electric field. The presence of high voltage will lead to the formation of liquid jets and these jets further break down into tiny liquid droplets. Then, the solvent in these droplets will evaporate *in situ*, which subsequently forms solid particles comprising solutes [25,26]. The processing parameters include the applied voltage, the flow rate of the liquid, the collection distance between the nozzle and the plate, and the diameter of the metal nozzle [27]. Other than processing parameters, the properties of the ES solution can also affect the product. For example, the type of solvent and the concentration of the solute determines the evaporation rate, viscosity, and other solution properties. These solution parameters will impact the behaviour of liquid droplets and eventually change the properties of particles. Furthermore, environmental parameters like temperature and humidity during spraying can shift the evaporation speed of the liquid and likewise affect the formation of particles.

Currently, the parameter selection process relies heavily on prior experiences and is conducted through trial-and-error, which is costly, time consuming and resource intensive. The problem is further compounded when expensive materials, such as poly(lactic-co-glycolic acid) (PLGA) are being processed. Thus, methods that can reduce cost, material expenditure and time will accelerate ES product developments. Alternative to this empirical approach is to leverage *in silico* tools to diminish the number of experiments needed to achieve the desired products [28–31]. An emerging *in silico* tool is machine learning (ML), a subfield of artificial intelligence (AI), which uses historical data to predict future outcome. As a computational tool, ML offers several advantages that other computational methods lack, such as handling high-dimensional data, large datasets, and computing a range of data formats, including numeric, texts and images [32–34].

Moreover, in comparison to numerical modelling techniques, ML is computationally fast, which obviates the need for high-end computers. Collectively, these attributes explain why ML has attracted considerable attention across many sectors.

While ML has gained fame in many areas including autopilot cars, face recognition, and online translation, as well as outperforming clinicians in diagnostic tests, its application in fabrication technologies remains thoroughly underexplored [35–37]. When it comes to fabrication processes, ML was exploited to predict the product properties of three-dimensional (3D) printing and electrospinning [38–40]. These studies have reported that ML has the potential to accelerate developments. Nevertheless, fewer than a handful of applications using ML to assist ES particle fabrication can be found in previous literature [41,42]. A challenge with ML is the need for sufficient data for ML algorithms to learn the patterns associated between the input and output variables. In material science, data collection can be costly, time consuming and resource intensive. Even for people who are willing to build ML models from scratch to expedite their research, the cost is undesirable and prohibitive, let alone other users in the ES community who only need a tool to guide experiment design. Despite previous studies provided some ready-to-use ML models for ES, these models are only applicable to the specific formulation from which the model was trained. Thus, there is an unmet need for a model that is able to consider various formulations used by different researchers for different purposes. This again exacerbate the difficulty of data generation. Fortunately, the scientific literature contains copious data that can be extracted by researchers to begin developing ML models. Recently it was revealed that a large number of formulations, over 900, can be extracted from the literature, which was subsequently used to develop ML models [43]. The authors of the study reported that the literature presented with a greater number of formulations than what was available within their in-house dataset. Moreover, the data was generated by several research groups, which reduces experimental bias and consequently results in ML models with better generalisability. Overall, the researchers concluded that the literature was a viable approach to rapidly gain a large dataset for developing effective ML models.

To that end, the present work investigated exploiting the literature to build ML models to predict ES particle size, as the technology remains a nascent powder fabrication technology. The largest ES product database to date was compiled, with 445 data records by extracting information from 45 selected previous publications on ES. The database covered 5 commonly used polymers in 13 different solvent systems and was enriched with detailed documentation of experiment parameters. From the database, records with PLGA as the polymer was selected to benchmark the performance of several ML models with different feature engineering techniques in the task of predicting the product's particle diameter. The study concluded by applying the best trained ML models to predict PLGA particle sizes produced via an in-house ES setup. For the first time, it was demonstrated that ML successfully learnt the correct ES processing features as traditional mechanistic models, and in the process yielded highly-accurate predictions when paired with literature-acquired data. The study highlights the promise of ML in the design of automated ES technologies, thereby facilitating the manufacturing process.

# 2. METHODS

## 2.1. Data Acquisition

All articles used for data extraction were retrieved from the Web of Science Core Collection by querying the search engine with the following term '(ALL=((ELECTROSPRAY* OR (ELECTROHYDRODYNAMIC ATOMIZATION)) AND (MICROSPHERE OR PARTICLE)))'. The results were refined by limiting the year of publication to after 2000 and excluding publications in the analytical chemistry field. In addition, only articles reporting experiments with PLGA, polycaprolactone (PCL), poly(lactic acid) (PLA), chitosan (CS), and polyvinylpyrrolidone (PVP) were included. Manual data extraction was conducted to construct the ES database. In short, the experiment parameters along with the yielded particle diameters were collected in a spreadsheet. Experiment parameters were divided into three parts: 1) processing parameters including the applied voltage, flow rate, collection distance, and needle diameter, 2) solution parameters such as the type of the polymer and solvent, the concentration of the polymer and the properties of the solution, and 3) environmental parameters including temperature and humidity during the experiment. For mixed solvent systems, only the primary solvent with the highest volume ratio were recorded. All units were converted to the same in the database. For incomplete reports of experiment parameters in the articles, the missing value was left blank for further processing. Two examples of data records in the database were shown in **Table S1**. Data visualization was carried out on the whole database using statistical data visualization library Seaborn (v0.11.2).

## 2.2. Feature Engineering

PLGA was selected as the model polymer to carry out feature engineering and ML model building due to the abundance of data when compared with other polymers. Since some parameters had insufficient records, only six parameters were chosen to be the input of the ML models: the concentration of the polymer, the type of the solvent, the flow rate, the voltage applied, the diameter of the needle, and the collection distance. Noteworthy, techniques in cheminformatics were applied to represent the type of the solvent in a computer-understandable way, details of which will be further explained in the following section. And a flow chart of procedures in feature engineering is shown in **Figure 1**.

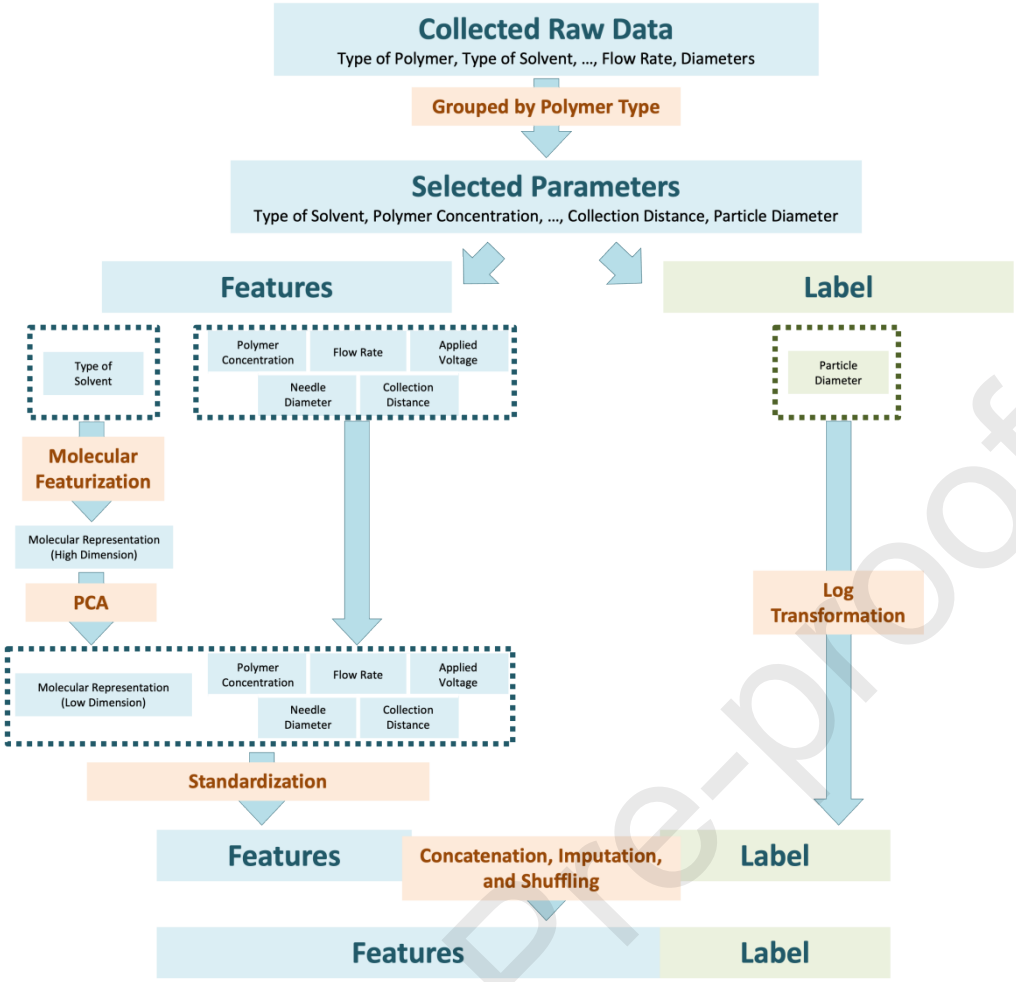



**Figure 1**. A flow chart of procedures in feature engineering. The raw data collected from previous publications were categorized by the type of polymer and only the records using PLGA was selected. Only six selected parameters (the "features") and the product mean diameter (the "label") were used for ML. The type of solvent was further represented by molecular featurization techniques before standardization was carried out. Finally, features and labels were concatenated, imputed, and shuffled.

## 2.3. Solvent Featurization

All other input parameters ("features") except the type of solvents can be expressed by numerical values, which can be easily processed by ML algorithms. However, the type of solvents, normally represented by their chemical names, are texts that cannot be understood by computers unless ML algorithms for texts are used [37]. In order to represent different molecules, in cheminformatics, molecular featurization techniques were developed [44,45]. These techniques are widely applied in various ML applications in drug development and quantitative structure-property relationship (QSPR) modelling [46,47]. Briefly, molecular featurization techniques use a series of numbers that describes the property of a molecule (e.g., molecular weight, number of acid groups, and number of electron donors) as the representation of the molecule. In this study, four molecular featurizations were performed through the DeepChem (v2.5.0) Python library: *Mol2Vec*, *Mordred*, *RDKit*, and *ECFP*. A detailed introduction of these featurization techniques can be found in their documentations, respectively [48–51]. In addition, two extra featurization methods were implemented manually: *EHDProperties* and *one-hot* featurization. The *EHDProperties* featurization used 8 important solvent properties collected from PubChem

including boiling point, density, dipole moment, dielectric constant, viscosity, surface tension, relative evaporation rate (where Butyl acetate=1), and the Hansen solubility distance calculated with respect to PLGA [52,53]. For *one-hot* featurization, a vector consisting of 13 entries (corresponding to 13 solvents) was used to represent a solvent where the presence of the solvent was indicated by 1 of that specific entry and the rest of other entries were filled by 0. To address the problem of high dimensionality, principal component analysis (PCA) was performed on the features generated by molecular featurization techniques. The dimension of different featurizations were all compressed to 8 for comparison.

## 2.4. Data Standardization, Transformation, and Imputation

Standardization was performed on features to remap the distribution, which could help accelerate training [54]. This was achieved by:

$$x_{std} = \frac{x - mean}{standard\ deviation} \qquad \text{Equation 1}$$

where $x$ is the value before standardization and $x_{std}$ is the value after standardization. In addition, the particle's diameter, denoted as the prediction "label" of an experiment record, was transformed into logarithmic form. Then, all features including polymer concentration, flow rate, applied voltage, needle diameter, collection distance, and the type of solvent (represented by molecular featurization techniques after PCA dimensionality reduction) were concatenated into a long vector denoted as "features" for an experiment record. Furthermore, data extracted from previous publications contained missing values due to incomplete reporting of parameters. The intentionally-left-blank values in the spreadsheet were filled by "NaN" as the identification of missing data. Four different strategies were implemented to handle the missing values. Two of them were actively filling the blank values through different data imputation techniques provided by the Python library Sci-kit learn (v0.24.2) [55]. By using the k-nearest neighbour (*kNN*) algorithm, the *kNN* method imputed the missing values with the average of *k*-th most "similar" experiment. The *Mean* method, as suggested by its name, used the mean value of that variable in the database to fill the blank. The rest two strategies didn't deal with the missing values as controls. The *None* method directly deleted all records that contained missing values for comparison of performance. The *leave_empty* method left the records with missing values as "NaN" in the database without any processing. Although most of ML methods cannot hanle data with missing values, certain algorithms are capable of utilizing these incompelete information to build the model. Thus, the *leave_empty* method was used as a control for these models in this study. Finally, the sequence of experiment records consisting of *(features, label)* pairs was shuffled randomly for further model building.

## 2.5. Model Building

The whole PLGA dataset (n=248) was treated as the full dataset for ML. To compare the performances of different ML models on the PLGA dataset, seven ML algorithms with different specializations were chosen for benchmarking. The models included Support Vector Regression (SVR), Kernel Ridge Regression (KRR), kNN, Multilayer Perceptron (MLP), RF, XGBoost, and Light Gradient Boosting Machine (LGBM). The XGBoost model used py-xgboost library (v1.3.3) [56]. And LGBM used lightgbm Python library (v3.1.1) [57]. The other six models were implemented through Sci-kit learn library (v0.24.2). Performances of ML models were evaluated through 5-fold CV, which allowed better representation of model performances in a small dataset [58,59].

RMSE, mean absolute percentage error (MAPE), and coefficient of determination ($R^2$)were set as the metrics. They were calculated by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \quad \text{Equation 2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|} \quad \text{Equation 3}$$

$$R^2 = \sqrt{\frac{\sum_{i=1}^{N}\left(\overline{y_i} - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(\overline{y_i} - y_i\right)^2}} \quad \text{Equation 4}$$

where $y_i$ is the ground truth value obtained from articles, $\hat{y}_i$ is the predicted value provided by the ML model, and *N* is the total number of samples. Hyperparameter tuning was also carried out for all models through randomized searching 100 times in the parameter space. The potential parameters are listed in **Table S2**.

## 2.6. Model interpretation

For interpretable models like RF and XGBoost, feature importance assigned by the model can be plotted to understand the relationship between the particle diameter and experiment parameters. In our study, PCA dimensionality in the feature engineering step was not carried out for better interpretability for *EHDProperties* and *one-hot* featurization in the model interpretation step. For other molecular featurization techniques like *Mol2Vec*, *Mordred*, *RDKit*, and Extended-connectivity fingerprints *(ECFP)*, the representation of molecules was usually in very high dimensions that were difficult to interpret. Thus, PCA was applied to the representation to reduce them to only 1 dimension during the training for model interpretation purposes.

## 2.7. Experiments

### 2.7.1. Materials

PLGA (50:50, PURASORB PDLG 5002) was provided by Corbion (Amsterdam, Netherlands). Solvents, including acetone, dimethylacetamide (DMA), dichloromethane (DCM), were purchased from Sigma-Aldrich (Poole, UK).

### 2.7.2. Preparation of the ES solution

The ES solutions were prepared by dissolving PLGA powders in acetone, DMA, DCM at 2, 4, and 8% (w/v), respectively. Magnetic stirring was applied for 1 h to dissolve the polymer.

### 2.7.3. Electrospraying

To better explore the validation space with least number of experiments, orthogonal experiment design was used. Four factors with each at three levels were filled in an $L_9$ table as shown in **Table**

**1 (Experiment No. 1-9)**. Briefly, the polymer concentration factor was chosen from 2, 4 and 8% (w/v). The solvent factor was chosen from acetone, DMA, and DCM. The flow rate was running at 2, 4, or 8 μL/min. The applied voltage was set at 10, 12.5, or 15 kV. Other experiment parameters were fixed: the collection distance was maintained at 195 mm and the gauge used was 22G with an outer diameter of 0.7 mm. The flow rate was controlled by a syringe pump (World Precision Instruments). A voltage generator (Genvolt) provided electric potential between the metal needle and the collection plate. In addition, experiment No**.** 10-13 were carried out to evaluate the generalisability of the model. The experiments were conducted at ambient temperature (22-23 °C) and relative humidity around 50%. Each experiments were carried out two times (n=3). And the particles were collected on glass slides for further characterizations.

**Table 1.** Orthogonal design of validation 9 experiments and an extra of 4 experiments to evaluate generalisability of ML models

| No. | PLGA Concentration (% (w/v)) | Flow rate (μL/min) | Applied Voltage (kV) | Solvent |
|---|---|---|---|---|
| 1 | 2 | 2 | 10 | Acetone |
| 2 | 4 | 4 | 15 | Acetone |
| 3 | 8 | 8 | 12.5 | Acetone |
| 4 | 2 | 8 | 15 | DCM |
| 5 | 4 | 2 | 12.5 | DCM |
| 6 | 8 | 4 | 10 | DCM |
| 7 | 2 | 4 | 12.5 | DMA |
| 8 | 4 | 8 | 10 | DMA |
| 9 | 8 | 2 | 15 | DMA |
| 10 | 2 | 2 | 7 | Acetone |
| 11 | 2 | 8 | 8 | Acetone |
| 12 | 4 | 4 | 7 | Acetone |
| 13 | 8 | 8 | 7 | Acetone |

## 2.7.4. Particle Characterization

The diameters of the particles produced by ES were characterized by a benchtop scanning electron microscopy (SEM; Phenom Pro, Phenomworld). Images obtained were further analyzed by *ImageJ* (National Institute of Health, USA) software. A collection of SEM images can be found in the Supplementary Materials (**Figure S1**).

# 3. Results and Discussion

## 3.1. Data Acquisition

After a comprehensive examination of the publication literature retrieved from the Web of Science (WOS) Core Collection, 45 publications that satisfied requirements were chosen for manual data extraction. From these articles, 442 experiment records were successfully extracted as the ES database. The database contained 248 records with PLGA, 114 records with PCL, 16 records of PLA, 3 records of CS, 12 records of PVP, and 49 failed experiment records that produced no particles. To the best of authors' knowledge, the database of ES experiments herein is the largest to date in the field (**Table S3**). Furthermore, the study collated a broad collection of various solvent information retrieved from previous publications, which enabled us to construct more generalized models that are applicable to different solvent systems, and also to further analyze the effect of solvents via model interpretation.

A bar plot of records using different solvents is presented in **Figure 2** (failed experiment records were excluded). Notably, 56 out of 393 successful records used mixed solvent systems. Here, only the primary solvent which owned the highest volume ratio was considered. It can be observed from the plot that chloroform is the most popular solvent in ES which had 97 records. Some specific solvents were used for polymers. For example, CS and PVP are electrosprayed in ethanol-based solvent systems (15 records). When examining the PLGA records, it appeared that halogenated solvents like DCM (48 records), chloroform (43 records), and trifluoroethanol (TFE, 35 records) were favoured. Also, acetone (40 records) acetonitrile (ACN, 35 records), and tetrahydrofuran (THF, 27 records) were used in PLGA electrospraying. These balanced records of solvents for PLGA data made it potentially possible for the ML algorithm to capture the effect of solvents. When it comes to PCL records, halogenated solvents were also preferred (chloroform: 38 records and DCM: 37 records), whereas other solvents had limited records (ACN: 2 records and THF: 2 records).

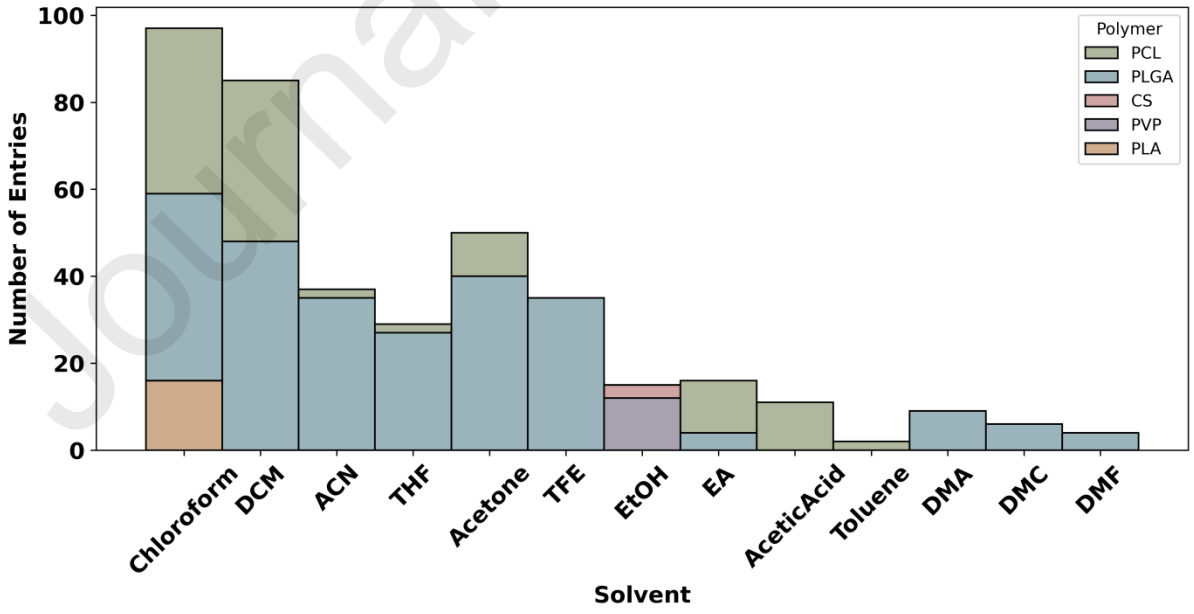



**Figure 2.** Solvents recorded in electrospraying database. (Solvents used in electrospraying in extracted data. (EtOH: ethanol, EA: ethyl acetate, and DMC: dimethyl carbonate)

Data distribution of experiment parameters in the database was visualized by histograms and kernel density estimation (KDE) plots (Figure 3). As depicted in the figure, some outliers existed in the database. For example, the majority of needle diameters chosen for ES fell in the range between 0.2-1.5 mm (outer diameter, O.D.) and most of the collection distances were between 75-200 mm. However, several records used large needle gauges above 2 mm and long collection distances around 400 mm. Other than experiment parameters, particle diameters of all the records were also plotted in (**Figure 3(f)**). The KDE plot of particle diameter had a skewed normal distribution spreading from $10^{-2}$ to $10^{2}$ μm with a peak around 10 μm.

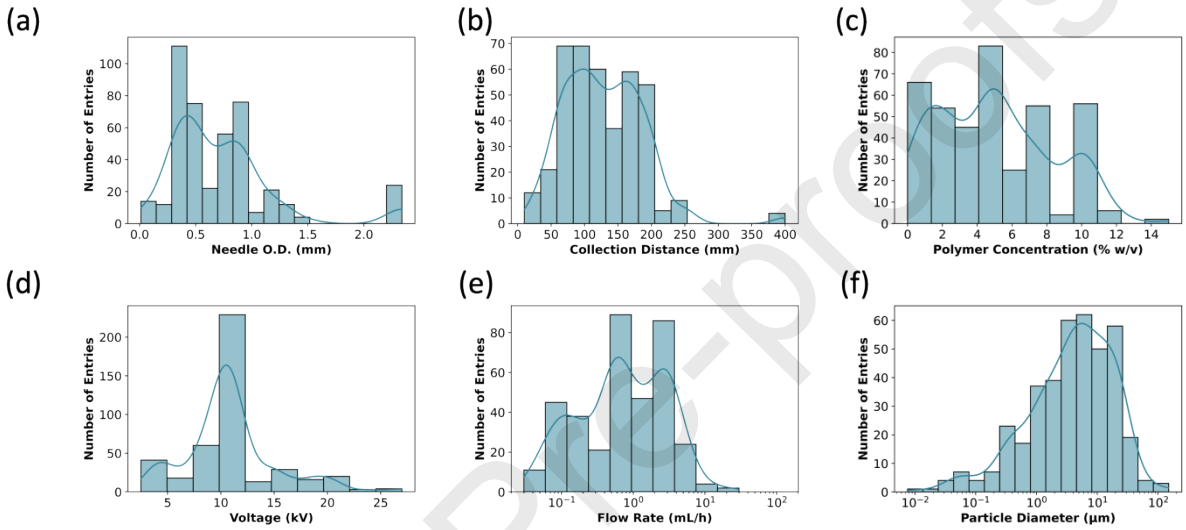



**Figure 3**. Data distributions of (a) needle O.D., (b) collection distance, (c) polymer concentration, (d) applied voltage, (e) flow rate, and (f) particle diameter.

An interesting observation from **Figure 3** is that the voltage used for ES, as indicated by the KDE plot, formed a sharp peak around 10-12 kV. This specific voltage seems to be a "safe" choice that suits most ES applications. Nevertheless, considering the electrical properties of different solvents, a "one-size-fits-all" choice of voltage is only workable, but not optimized [60]. As suggested by Borra *et al.*, the optimized voltage range for spraying is indeed between 11-20 kV but varies largely depending on the solvent conductivity [60]. Therefore, for acetone which has a conductivity around 0.48 μS/cm (with PLGA 2%) and ACN which has a more than 20 times higher conductivity at 18.1 μS/cm (with PCL 3%), an ES voltage between 10-12 kV should not be an optimized value for both solutions. The relationship between the type of solvents and ES voltage was also studied and similar conclusions could be found in another study by Zhang *et al.* [61]. These results suggested that finding the optimized voltage for ES empirically through experiments is not practical in this wide range, highlighting the need for modelling techniques to assist the experiment process to find optimized experiment conditions.

A more detailed observation of particle diameter was conducted by plotting its distribution over different solvents and polymers, as shown in **Figure 4**. Strip and box plots intuitively portray the distribution and the statistical summary of the data. Here, each scatter point in the plot represents a record in the database. The plot portrayed a clear evidence of data deficiency for several less-explored solvents like toluene and DMF. In addition, the particle diameter distribution varied

between different polymers. Since the particle diameter was used as the prediction target in further ML tasks, data distribution with good quantity and quality were both desired. CS, PVP, and PLA all had limited number of records. Thus, they were not preferred for ML developments. The PCL data concentrated between its first quartile (6.01 μm) and third quartile (16.5 μm), whereas PLGA records spread within a larger range between first quartile (1.29 μm) and third quartile (10.0 μm) after logarithm transformation. Therefore, considering the quantity and quality of the data, PLGA was chosen as the model polymer for further ML processing.

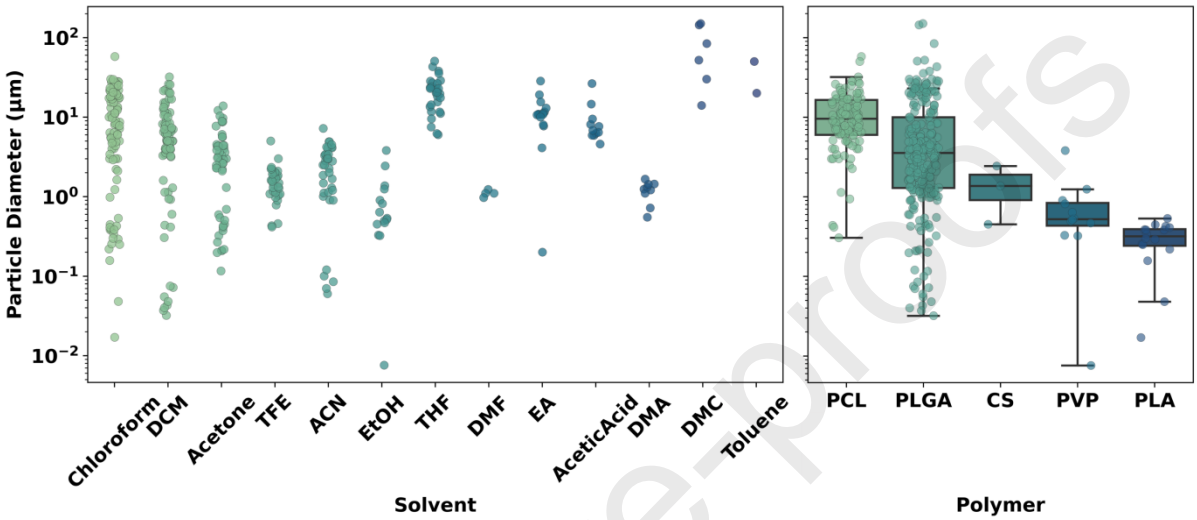



**Figure 4**. Strip and box plots of particle diameter distribution of solvents and polymers.

## 3.2. Preliminary Modelling

In this study, several feature engineering techniques were implemented, including different solvent featurization to determine the optimal means to compute solvent information. Furthermore, data imputation methods were tested to address the lack of some datapoints. Therefore, the study carried out preliminary modelling to determine a good combination of feature engineering options before commencing ML model benchmarking. As a starting point, XGBoost was selected due to its ability to handle small dataset. A baseline model using XGBoost was trained on features with *one-hot* solvent featurization, feature standardization, *Mean* data imputation and no label logarithm transformation. This model reached an $R^2$ of 0.81 and an RMSE of 6.09 in 5-fold CV. With this baseline, a series of experiments using different solvent featurization techniques, feature standardization, data imputation, and label logarithm transformation was done to choose a preliminary "best" combination of feature engineering techniques (data not included). Finally, the combination of *EHDProperties* solvent featurization, feature standardization, *kNN* data imputation, and label logarithm transformation achieved the best result with an $R^2$ of 0.87 and an RMSE of 4.57 under 5-fold CV. Hence, such set of feature engineering techniques were chosen as the default in the following model benchmarking and hyperparameter optimizing process.

## 3.3. Model Benchmarking

Seven ML models were tested for the task of predicting particle diameters, including SVR, KRR, kNN, MLP, RF, XGBoost, and LGBM. These models were trained and evaluated through 5-fold CV with the feature engineering techniques determined in the preliminary study. Firstly,

hyperparameter optimization of each model was conducted with randomized searching in the parameter space as shown in **Table S2**. The optimized hyperparameters were listed in **Table S4**. With the optimized model hyperparameter, we were able to compare the performance of the model at their best condition (**Figure 5**). XGBoost model had the best performance with an $R^2$ of 0.91, an RMSE of 3.91 μm and a MAPE of 0.50. Other tree-based models also had comparable performance. For example, RF had an $R^2$ of 0.84, an RMSE of 6.19 μm and an MAPE of 0.61. KRR and kNN produced less satisfying results indicated by higher RMSE and lower $R^2$ values. Interestingly, SVR possessed the best MAPE of 0.42, but performed weaker in other metrics. Considering all three metrics, XGBoost possessed the best performance, and was selected for further optimization.

In this study, linear modelling techniques, like multiple linear regression, were not examined since previous studies already suggested poor performances of linear models on similar tasks [62]. Of seven ML models studied here, SVR and KRR are regression models based on kernel methods, kNN is a local regression model based on similar cases in the training set, MLP is a neural-network-based model, and RF, XGBoost, and LGBM are all tree-based models. With different learning mechanisms, these models have their own specializations and perform differently in tasks. As shown in **Figure 5**, tree-based models had the best performance when compared with other models. The better performance of tree-based algorithms agrees well with previous research of ML in other areas where limited amount of data is available [63]. These tree-based models own various merits including shorter training time, better interpretability, and the capability of dealing with missing values (only for XGBoost and LGBM). These advantages made them widely used in other areas [32,64,65]. However, previous studies listed in **Table S3** only chose a handful of algorithms like MLP and SVR to set up their ML model. According to these observations, it is recommended future modelling studies of ES and other fabrication techniques to incorporate and evaluate tree-based models like XGBoost and RF.

Moreover, the result in **Figure 5** also emphasized the importance of using CV for model evaluation. The high variance between folds in our study could be resulted from anomalies and outliers in the data. CV is known to address the problem of data containing outliers, where outliers are detrimental to model learning [58]. When CV was included in the evaluation phase, all data was portioned equally into several parts after random shuffling and one part was treated as the test/evaluation set and the remaining parts were used as the training set at a time. The model training and evaluation was carried out several times after all data had been treated as the test set. It was noted that by including CV the effect of outliers in the evaluation procedure was reduced by averaging the performance on different test sets. It is commonly recommended to implement CV for small data ML tasks, particularly in material science and medical data [66,67]. In addition to benchmarking various ML methods, it is also highly recommended to perform CV to better evaluate the model performance for ML practices for ES and other small data learning scenarios [67].

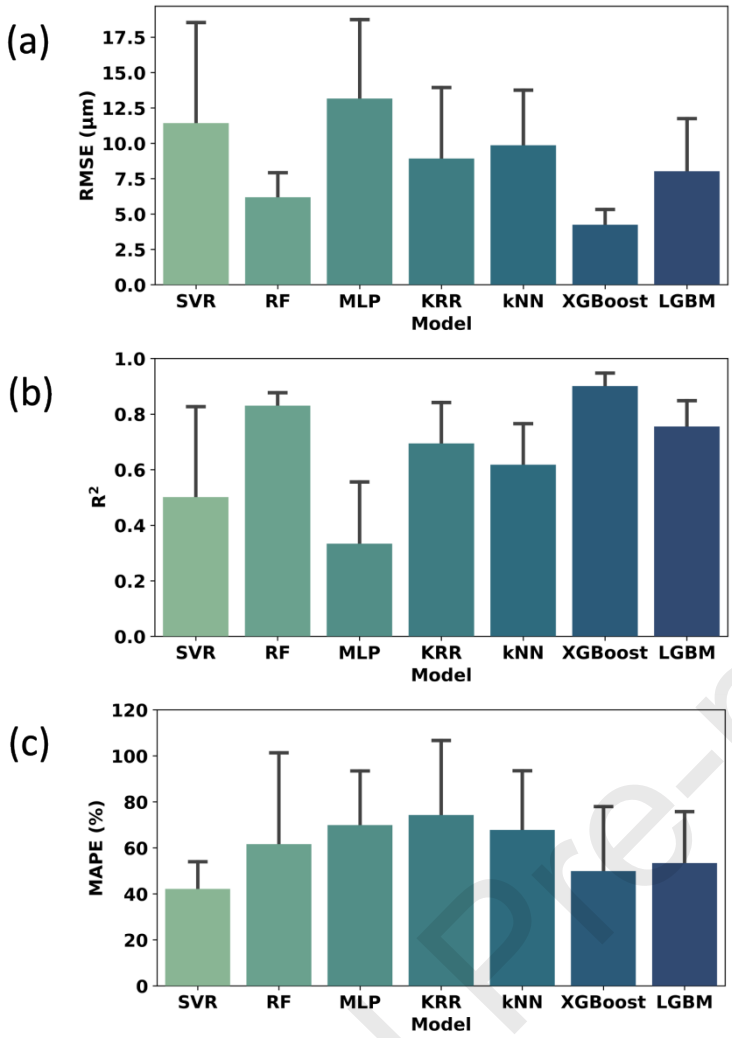



**Figure 5**. Model performance evaluated with 5-fold CV by (a) RMSE, (b) R2, and (c) MAPE. Error bars obtained from 5-fold CV.

## 3.4. Reviewing Feature Engineering Techniques

Albeit nominal feature engineering techniques were determined at the beginning of the study, after selecting the best ML model with XGBoost, a review of these parameters and feature engineering techniques was necessary. In **Figure 6**, performances of different data imputation and solvent featurization methods were plotted. Regarding data imputation, it can be seen from the plot that data imputation did affect model performance. Since XGBoost model could handle data with missing values, the *leave_empty* method was applied here for comparison. A reduction in performance, as indicated by increased RMSE and decreased in $R^2$, could be found in the *None* group where all data records with missing values were deleted. This result highlighted the importance of data imputation. For other ML methods like SVR, RF, KRR, kNN, and MLP which are not able to handle missing values, the only two choices to treat these blanks were either to use imputation methods or directly delete all records with missing values. The present research demonstrated that using data imputation methods could help with the model performance.

Molecular featurization techniques were introduced in this study to determine the optimal means to representing solvents for ML applications. Six different featurization approaches were investigated. *One-hot* featurization essentially records the solvents by their name, and thus lacks generalizability. Herein, there were 13 solvents in the database. Therefore, *one-hot* featurization represented each solvent as a vector with 13 elements (a "1" and twelve "0"s). However, if the user wishes to apply the model to predict for a new solvent system outside these 13 solvents (e.g., chloroethane), this 13-element vector will not be able to represent new solvents. This greatly reduced the capability of the model to generalize to new formulations. From a chemistry viewpoint, chloroform is similar to DCM than it is to water, both in molecular structure and physico-chemical properties. Hence, other featurization methods were investigated to allow the model to achieve generalizability by capturing solvent chemical and physico-chemical properties [68]. The features *Mol2Vec*, *Mordred*, *RDKit*, *ECFP,* and *EHDProperties* represents the solvents by their chemical structure and/or physico-chemical properties, using information such as molecular weight, number of acid groups and number of rings.

It was revealed that the six solvent featurization methods had comparable performance (**Figure 6**). *One-hot* featurization did not result in any noticeable effect on model performance. One potential explanation is that the model automatically assigned weights to these categorically represented solvents, and these weights compensated for the differences of solvents, which was confirmed in the feature importance analysis below. Such differences in weights when using *one-hot* featurization is believed to help models differentiate between molecules (which were actually treated as separate "features" in model inputs). Thus, although *one-hot* featurization is not chemically meaningful, it still can generate comparable results as other featurization techniques, as can be seen in this study and also in other research [69]. More importantly, however, is that the use of molecular featurization methods did not diminish model performance in comparison to *One-hot* featurization*.* Therefore, using molecular features allows for both high prediction performance and generalizability.

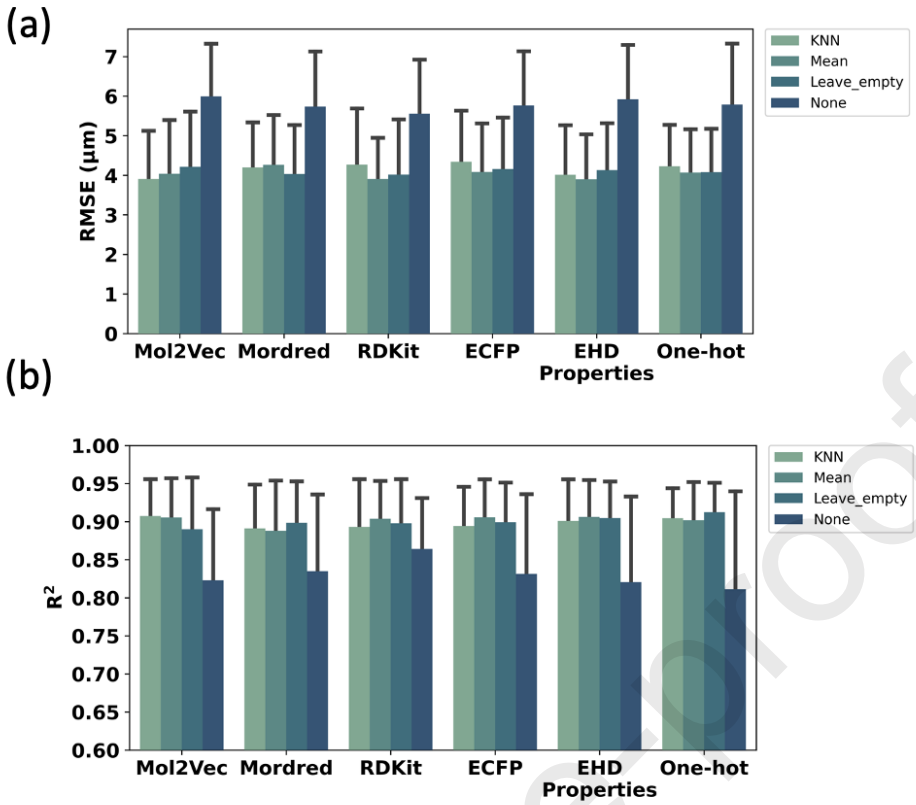



**Figure 6.** Model performances of XGBoost model with different molecular featurization techniques and data imputation methods evaluated by (a) RMSE and (b) $R^2$. Error bars obtained from 5-fold CV.

## 3.5. Model Validation

To validate the model, 13 new formulations were tested with our in-house ES setup. Nine of these formulations were carried out with varying polymer concentration, flow rate, applied voltage, and solvent type. An additional four experiments were conducted to test the generalizability of the model, by testing formulations at voltages below values previously used in the literature for the solvents examined. As the best performing model, XGBoost was used for model validation, using *kNN Imputation*, *EHDProperties* for the solvent features and label logarithm transformation. **Figure 7(a)** presents the results of XGBoost when applied to the literature database, where the training data and test data were split in to an 80/20 ratio from the original dataset as a comparison with the in-house experimental validation results, which are given in **Figure 7(b)**. XGBoost successfully predicted both nanoand micro particle sizes for the training set (**Figure 7(a)**); whereas the test set was marginally less accurate. Nevertheless, this provided confidence to apply the models to the in-house experiments. For experiments No. 1-9, the XGBoost model achieved an RMSE and MAPE of 1.30 µm and 0.33, respectively. For experiments No. 10-13, the values were 3.94 µm and 0.43, respectively. This revealed that the accuracy decreased when extrapolating the prediction to new voltage values, inferring model overfitting or poor generalizability. The assumption was confirmed by the observation of a decreased RMSE (**Table S5**) by raising the regularization hyperparameter "reg_lambda" to 1.4 (previously was 1.1 in the optimized model) and reducing the "max_depth" to 3 (previously was 5 in the optimized model).

These modifications were suggested by the XGBoost documentation to help reduce the overfitting of the XGBoost model [56].

To further test the hypothesis, the predictions of RF to the experimental ground truth were also compared (**Figure 7(b)**), where not RF is known to avoid overfitting [70]. RF was found to perform worse than XGBoost in experiments No. 1-9 but surprisingly better in experiments No. 10-13, which highlighted that RF was capturing the underlying mechanisms instead of noises (e.g., human errors and random fluctuations in experiments) in the data. The results suggest that avoiding overfitting yields better predictions when extrapolating to new processing parameters. Overall, the in-house experimental validation results revealed that ML models trained on previously published data can be readily applied to in-house ES setups without tuning or prior empirical experiments. Most predictions of experiments fell in the range of 25% relative error range. Only two experiments, No. 8 and No. 9 both with DMA as the solvent, had relatively high deviation. This might be due to limited training data for the solvent and could be improved with a larger dataset. For commonly used solvents DCM and acetone, the XGBoost model gave satisfying predictions with very low RMSE (0.71 µm for DCM group and 0.23 µm for acetone group).

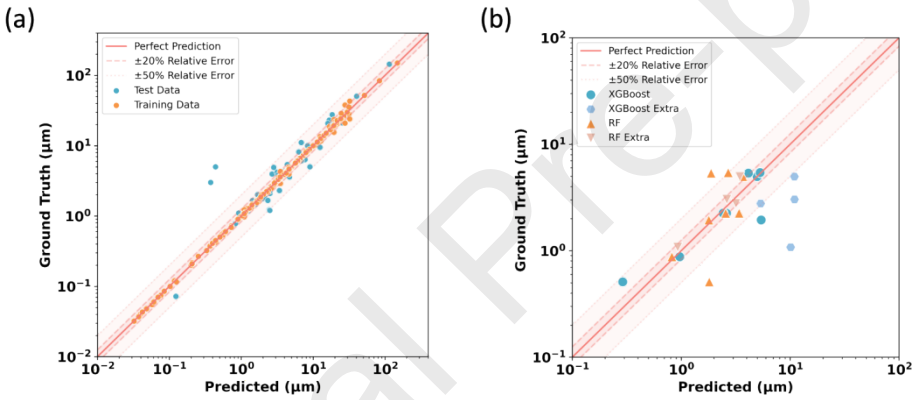



**Figure 7.** (a) Predictions given by XGBoost model based on the data records in the literature database. (b) Experiment results and predictions given by XGBoost and RF model. Dots and triangles pointing upwards are predictions for experiment No. 1-9 with XGBoost and RF model, respectively. Hexagons and triangles pointing downwards are predictions for extra experiments No. 10-13 with XGBoost and RF model, respectively. In both figures, solid line shows perfect perditions where the ground truth values equal to predictions. Dashed lines and the coloured area indicate relative error range (±20% and ±50%) for the predictions.

## 3.6. Model Interpretation

Interpreting the model provides valuable insights into its learning characteristics. Feature importance learnt by the RF model was plotted to represent the ML's interpretation of the ES (**Figure 8**). Three solvent feature sets were examined for now, which were the *Mordred*, *one-hot encoding* and *EHDProperties*. PCA dimensionality reduction was used to reduce the *Mordred* solvent features, which were over 1600 features, into one single feature called "solvent" in **Figure 8(a)**. Since *ECFP* and *Mordred*, *RDKit* and *Mol2Vec* methods yielded identical feature importance results, here feature importance given by *Mordred* was displayed. In addition, the feature

importance when *EHDProperties* was used as the solvent featurization was plotted in **Figure 8(b)**. And the feature importance of the RF model trained on *one-hot* featurization was plotted in **Figure 8(c)**. All these plots promoted flow rate as the dominating factor in ES, followed by polymer concentration, and needle diameter.

Interestingly, the feature importance plot from **Figure 8(a)** suggested the limited influence of solvents on the particle diameter produced from ES. Indeed, Faramarzi and colleagues revealed that the type of solvents had a more significant effect on particle morphology, while the size of the particle was dominated by the flow rate [71]. In **Figure 8(b)**, the importance of different solvents varied from each other. It confirmed the hypothesis previously discussed herein about molecular featurizations that the model assigned different weights for the *one-hot* featurization method automatically to distinguish different solvents. Thus, it performed as well as other featurization methods herein. In **Figure 8(c)**, the plot demonstrated the dominating effect of flow rate over solvent properties. This result could be confirmed by research into scaling laws for ES. Gañán-Calvo *et al.* proposed a scaling law of droplets produced by ES in 1997 and Hartman *et al.* reported a similar scaling law in 1999 [72,73]. These laws related droplet diameter with various experiment parameters. Coincidentally, in both laws, the flow rate had the power of 1/2 whereas solution properties like conductivity, density, and surface tension all owned the power of 1/6. This evidently demonstrated the importance of flow rate on droplet diameter and confirmed the RF were learning the correct attributes established by these laws.

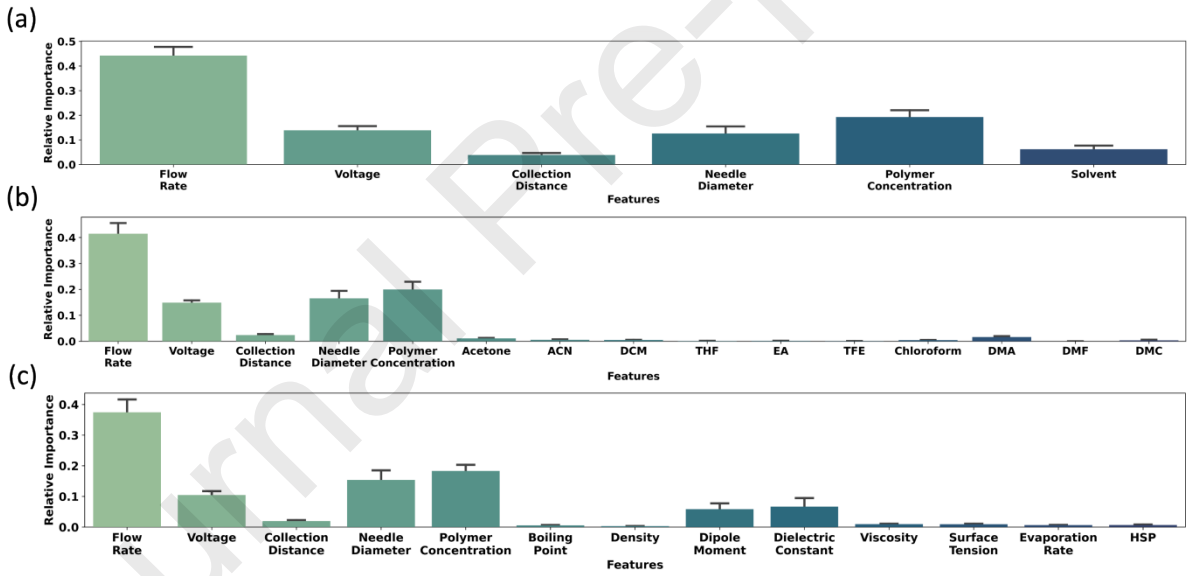



**Figure 8.** Feature Importance assigned by RF with (a) *Mordred*, (b) One-hot, and (c) *EHDProperties* as the molecular featurization technique. Error bars obtained from 5-fold CV.

The dipole moment and dielectric constant of the solvent stood out as two important solvent factors (**Figure 8(c)**). A study by Luo *et al.* on the choice of electrospinning solvents noted a strong correlation between solvent dielectric constant and the fibre produced through electrospinning [74].

## 3.7. Study Implications

A salient finding of this study was the successful application of literature-extracted data for building effective ML models to predict electrosprayed-particle sizes. Data availability in material science, and other fields, has been a key issue, hence the recent concerted effort to find ML algorithms that require a small dataset [67]. While research into ML algorithms for low datasets

continues, the present study pragmatically presented a different approach to developing predictive tools utilising well-established ML algorithms but placing emphasis on the means of data generation. Herein, the data generation method led to 248 PLGA formulations, which is a data size greater than previous applications of ML for ES (**Table S3**). Thus, trained models were provided with more instances to learn and predict PLGA particle sizes. ML models are known to be more effective as the size of the data grows [75]. Moreover, collecting information from the literature and from different research teams is expected to result in less bias, as different research teams use different ES machines and setups. Furthermore, literature-extracted data is both a cost- and resource-effective process in comparison to experimentally generating the formulation data.

An additional impact of the study will be realized in the ES domain, as well as the wider electrohydrodynamic processes (EHDP). While the experiment setup of ES is relatively simple, there has been a recent effort to expand the versatility of the system through the incorporation of, for example, multi-axial nozzles [76]. Moreover, EHDP are being integrated with other technologies, such as additive manufacturing, to enrich the latter's product performance [77–79]. While such systems are warranted, they will undoubtedly expand the processing search space, consequently prolonging development. Hence, there will be a need for ML to reduce the number of experiments needed to optimize these contemporary hybrid technologies.

# 4. Conclusion

In this study, the modelling relationship between ES product diameter and key experiment parameters were established. Several predictive ML models were trained on data extracted from previous publications. Different molecular featurization, data imputation, and ML models were benchmarked to identify the model with the best model performance evaluated by 5-fold CV. Notably, these models trained from published data were validated with wet-lab experiments. The study demonstrated outstanding prediction performance for particles at both micron and submicron scales without any prior tuning or experience. It was determined that the XGBoost model using *kNN* data imputation and *EHDProperties* molecular featurization stood out with the RMSE of 4.24 μm and an $R^2$ of 0.91 evaluated with CV on the published data. In addition, it was observed that the RF model performed better model generalizability in broader experiment parameter ranges. Furthermore, model interpretation revealed ML learns similarly to previously established mechanistic models. The study demonstrated ML as a promising tool for integrating with ES manufacturing, to enable precise control over process parameters and improve the manufacturability of ES technologies through automation.

## Acknowledgement

M.E. would like to acknowledge Engineering and Physical Sciences Research Council (EPSRC) [Grant Numbers: EP/S009000/1].

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**REFERENCES**


[1] B. Arauzo, M.P. Lobera, A. Monzon, J. Santamaria, Dry powder formulation for pulmonary infections: Ciprofloxacin loaded in chitosan sub-micron particles generated by electrospray, Carbohydrate Polymers. 273 (2021) 118543. https://doi.org/10.1016/j.carbpol.2021.118543.

[2] M.C. Castrovilli, P. Bolognesi, J. Chiarinelli, L. Avaldi, A. Cartoni, P. Calandra, E. Tempesta, M.T. Giardi, A. Antonacci, F. Arduini, V. Scognamiglio, Electrospray deposition as a smart technique for laccase immobilisation on carbon black-nanomodified screen-printed electrodes, Biosensors and Bioelectronics. 163 (2020) 112299. https://doi.org/10.1016/j.bios.2020.112299.

[3] S. Kavadiya, D.M. Niedzwiedzki, S. Huang, P. Biswas, Electrospray-Assisted Fabrication of Moisture-Resistant and Highly Stable Perovskite Solar Cells at Ambient Conditions, Adv. Energy Mater. 7 (2017) 1700210. https://doi.org/10.1002/aenm.201700210.

[4] S.J. Lee, S.M. Park, S.J. Han, D.S. Kim, Electrolyte solution-assisted electrospray deposition for direct coating and patterning of polymeric nanoparticles on non-conductive surfaces, Chemical Engineering Journal. 379 (2020) 122318. https://doi.org/10.1016/j.cej.2019.122318.

[5] M. Rasekh, Z. Ahmad, R. Cross, J. Hernández-Gil, J.D.E.T. Wilton-Ely, P.W. Miller, Facile Preparation of Drug-Loaded Tristearin Encapsulated Superparamagnetic Iron Oxide Nanoparticles Using Coaxial Electrospray Processing, Mol. Pharmaceutics. 14 (2017) 2010–2023. https://doi.org/10.1021/acs.molpharmaceut.7b00109.

[6] L. Lan, J. Xiong, D. Gao, Y. Li, J. Chen, J. Lv, J. Ping, Y. Ying, P.S. Lee, Breathable Nanogenerators for an On-Plant Self-Powered Sustainable Agriculture System, ACS Nano. 15 (2021) 5307–5315. https://doi.org/10.1021/acsnano.0c10817.

[7] Y. Du, M. Han, K. Cao, Q. Li, J. Pang, L. Dou, S. Liu, Z. Shi, F. Yan, S. Feng, Gold Nanorods Exhibit Intrinsic Therapeutic Activity via Controlling *N* 6-Methyladenosine-Based Epitranscriptomics in Acute Myeloid Leukemia, ACS Nano. 15 (2021) 17689–17704. https://doi.org/10.1021/acsnano.1c05547.

[8] P. Fantuzzi, L. Martini, A. Candini, V. Corradini, U. del Pennino, Y. Hu, X. Feng, K. Müllen, A. Narita, M. Affronte, Fabrication of three terminal devices by ElectroSpray deposition of graphene nanoribbons, Carbon. 104 (2016) 112–118. https://doi.org/10.1016/j.carbon.2016.03.052.

[9] Y. Xue, L. Qi, Y. Niu, H. Huang, F. Huang, T. Si, Y. Zhao, R.X. Xu, Integration of Electrospray and Digital Light Processing for Freeform Patterning of Porous Microstructures, Adv. Mater. Technol. 5 (2020) 2000578. https://doi.org/10.1002/admt.202000578.

[10] S. Patil, J. Kulkarni, K. Mahadik, Exploring the Potential of Electrospray Technology in Cocrystal Synthesis, Ind. Eng. Chem. Res. 55 (2016) 8409–8414. https://doi.org/10.1021/acs.iecr.6b01938.

[11] S.C. Hong, G. Lee, K. Ha, J. Yoon, N. Ahn, W. Cho, M. Park, M. Choi, Precise Morphology Control and Continuous Fabrication of Perovskite Solar Cells Using Droplet-Controllable Electrospray Coating System, ACS Appl. Mater. Interfaces. 9 (2017) 7879–7884. https://doi.org/10.1021/acsami.6b15095.

[12] H. Hu, S. Rangou, M. Kim, P. Gopalan, V. Filiz, A. Avgeropoulos, C.O. Osuji, Continuous Equilibrated Growth of Ordered Block Copolymer Thin Films by Electrospray Deposition, ACS Nano. 7 (2013) 2960–2970. https://doi.org/10.1021/nn400279a.

[13] Z. Gu, T.T. Dang, M. Ma, B.C. Tang, H. Cheng, S. Jiang, Y. Dong, Y. Zhang, D.G. Anderson, Glucose-Responsive Microgels Integrated with Enzyme Nanocapsules for Closed-Loop Insulin Delivery, ACS Nano. 7 (2013) 6758–6766. https://doi.org/10.1021/nn401617u.

[14] L. Fei, S.H. Yoo, R.A.R. Villamayor, B.P. Williams, S.Y. Gong, S. Park, K. Shin, Y.L. Joo, Graphene Oxide Involved Air-Controlled Electrospray for Uniform, Fast, Instantly Dry, and Binder-Free Electrode Fabrication, ACS Appl. Mater. Interfaces. 9 (2017) 9738–9746. https://doi.org/10.1021/acsami.7b00087.

[15] M. Parhizkar, P.J.T. Reardon, J.C. Knowles, R.J. Browning, E. Stride, R.B. Pedley, T. Grego, M. Edirisinghe, Performance of novel high throughput multi electrospray systems for forming of polymeric micro/nanoparticles, Materials & Design. 126 (2017) 73–84. https://doi.org/10.1016/j.matdes.2017.04.029.

[16] A. Ali, A. Zaman, E. Sayed, D. Evans, S. Morgan, C. Samwell, J. Hall, M.S. Arshad, N. Singh, O. Qutachi, M.-W. Chang, Z. Ahmad, Electrohydrodynamic atomisation driven design and engineering of opportunistic particulate systems for applications in drug delivery, therapeutics and pharmaceutics, Advanced Drug Delivery Reviews. 176 (2021) 113788. https://doi.org/10.1016/j.addr.2021.04.026.

[17] Y. Wu, L. Li, Y. Mao, L.J. Lee, Static Micromixer–Coaxial Electrospray Synthesis of Theranostic Lipoplexes, ACS Nano. 6 (2012) 2245–2252. https://doi.org/10.1021/nn204300s.

[18] Y. Wang, R. Zhang, W. Qin, J. Dai, Q. Zhang, K. Lee, Y. Liu, Physicochemical properties of gelatin films containing tea polyphenol-loaded chitosan nanoparticles generated by electrospray, Materials & Design. 185 (2020) 108277. https://doi.org/10.1016/j.matdes.2019.108277.

[19] Y. Luo, Y. Li, X. Feng, Y. Pei, Z. Zhang, L. Wang, Y. Zhao, B. Lu, B. Zhu, Triboelectric nanogenerators with porous and hierarchically structured silk fibroin films via water electrospray-etching technology, Nano Energy. 75 (2020) 104974. https://doi.org/10.1016/j.nanoen.2020.104974.

[20] S. Moschetto, M. Bolognesi, F. Prescimone, M. Brucale, A. Mezzi, L. Ortolani, M. Caporali, P. Pingue, M. Serrano-Ruiz, D. Pisignano, M. Peruzzini, L. Persano, S. Toffanin, Large-Area Oxidized Phosphorene Nanoflakes Obtained by Electrospray for Energy-Harvesting Applications, ACS Appl. Nano Mater. 4 (2021) 3476–3485. https://doi.org/10.1021/acsanm.0c03465.

[21] P. Jayaraman, C. Gandhimathi, J.R. Venugopal, D.L. Becker, S. Ramakrishna, D.K. Srinivasan, Controlled release of drugs in electrosprayed nanoparticles for bone tissue engineering, Advanced Drug Delivery Reviews. 94 (2015) 77–95. https://doi.org/10.1016/j.addr.2015.09.007.

[22] A. Jaworek, A.T. Sobczyk, A. Krupa, Electrospray application to powder production and surface coating, Journal of Aerosol Science. 125 (2018) 57–92. https://doi.org/10.1016/j.jaerosci.2018.04.006.

[23] H. Wang, Z. Zhao, Y. Liu, C. Shao, F. Bian, Y. Zhao, Biomimetic enzyme cascade reaction system in microfluidic electrospray microcapsules, Sci. Adv. 4 (2018) eaat2816. https://doi.org/10.1126/sciadv.aat2816.

[24] J. Xie, J. Jiang, P. Davoodi, M.P. Srinivasan, C.-H. Wang, Electrohydrodynamic atomization: A two-decade effort to produce and process micro-/nanoparticulate materials, Chemical Engineering Science. 125 (2015) 32–57. https://doi.org/10.1016/j.ces.2014.08.061.

[25] N. Bock, M.A. Woodruff, D.W. Hutmacher, T.R. Dargaville, Electrospraying, a Reproducible Method for Production of Polymeric Microspheres for Biomedical Applications, Polymers. 3 (2011) 131–149. https://doi.org/10.3390/polym3010131.

[26] B. Almería, A. Gomez, Electrospray synthesis of monodisperse polymer particles in a broad (60nm–2μm) diameter range: guiding principles and formulation recipes, Journal of Colloid and Interface Science. 417 (2014) 121–130. https://doi.org/10.1016/j.jcis.2013.11.037.

[27] A.Í.S. Morais, E.G. Vieira, S. Afewerki, R.B. Sousa, L.M.C. Honorio, A.N.C.O. Cambrussi, J.A. Santos, R.D.S. Bezerra, J.A.O. Furtini, E.C. Silva-Filho, T.J. Webster, A.O. Lobo, Fabrication of Polymeric Microparticles by Electrospray: The Impact of Experimental Parameters, Journal of Functional Biomaterials. 11 (2020) 4. https://doi.org/10.3390/jfb11010004.

[28] T.J. Struble, J.C. Alvarez, S.P. Brown, M. Chytil, J. Cisar, R.L. DesJarlais, O. Engkvist, S.A. Frank, D.R. Greve, D.J. Griffin, X. Hou, J.W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C.A. Nicolaou, A.D. Palmer, D.J. Price, R.I. Robinson, S. Salentin, L. Xing, T. Jaakkola, William.H. Green, R. Barzilay, C.W. Coley, K.F. Jensen, Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis, J. Med. Chem. 63 (2020) 8667–8682. https://doi.org/10.1021/acs.jmedchem.9b02120.

[29] B. Aramide, A. Kothandaraman, M. Edirisinghe, S.N. Jayasinghe, Y. Ventikos, General Computational Methodology for Modeling Electrohydrodynamic Flows: Prediction and Optimization Capability for the Generation of Bubbles and Fibers, Langmuir. 35 (2019) 10203–10212. https://doi.org/10.1021/acs.langmuir.8b03763.

[30] G.R. Mirams, C.J. Arthurs, M.O. Bernabeu, R. Bordas, J. Cooper, A. Corrias, Y. Davit, S.-J. Dunn, A.G. Fletcher, D.G. Harvey, M.E. Marsh, J.M. Osborne, P. Pathmanathan, J. Pitt-Francis, J. Southern, N. Zemzemi, D.J. Gavaghan, Chaste: An Open Source C++ Library for Computational Physiology and Biology, PLOS Computational Biology. 9 (2013) e1002970. https://doi.org/10.1371/journal.pcbi.1002970.

[31] X. Liu, Y. Shao, T. Lu, D. Chang, M. Li, W. Lu, Accelerating the discovery of high-performance donor/acceptor pairs in photovoltaic materials via machine learning and density functional theory, Materials & Design. 216 (2022) 110561. https://doi.org/10.1016/j.matdes.2022.110561.

[32] M. Elbadawi, L.E. McCoubrey, F.K.H. Gavins, J.J. Ong, A. Goyanes, S. Gaisford, A.W. Basit, Harnessing artificial intelligence for the next generation of 3D printed medicines, Advanced Drug Delivery Reviews. 175 (2021) 113805. https://doi.org/10.1016/j.addr.2021.05.015.

[33] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, Counting on natural products for drug design, Nature Chem. 8 (2016) 531–541. https://doi.org/10.1038/nchem.2479.

[34] D. Reker, Y. Shi, A.R. Kirtane, K. Hess, G.J. Zhong, E. Crane, C.-H. Lin, R. Langer, G. Traverso, Machine Learning Uncovers Food- and Excipient-Drug Interactions, Cell Reports. 30 (2020) 3710-3716.e4. https://doi.org/10.1016/j.celrep.2020.02.094.

[35] J. Janai, F. Güney, A. Behl, A. Geiger, Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art, CGV. 12 (2020) 1–308. https://doi.org/10.1561/0600000079.

[36] B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A general-purpose face recognition library with mobile applications, (n.d.) 20.

[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv:1810.04805 [Cs]. (2019). http://arxiv.org/abs/1810.04805 (accessed February 18, 2021).

[38] M. Elbadawi, B. Muñiz Castro, F.K.H. Gavins, J.J. Ong, S. Gaisford, G. Pérez, A.W. Basit, P. Cabalar, A. Goyanes, M3DISEEN: A novel machine learning approach for predicting the 3D printability of medicines, International Journal of Pharmaceutics. 590 (2020) 119837. https://doi.org/10.1016/j.ijpharm.2020.119837.

[39] J.D. Toscano, Z. Li, L.J. Segura, H. Sun, A Machine Learning Approach to Model the Electrospinning Process of Biocompatible Materials, in: American Society of Mechanical Engineers Digital Collection, 2021. https://doi.org/10.1115/MSEC2020-8362.

[40] F. Wang, M. Elbadawi, S.L. Tsilova, S. Gaisford, A.W. Basit, M. Parhizkar, Machine learning to empower electrohydrodynamic processing, Materials Science and Engineering: C. 132 (2022) 112553. https://doi.org/10.1016/j.msec.2021.112553.

[41] S. Tsai, Y. Ting, Synthesize of alginate/chitosan bilayer nanocarrier by CCD-RSM guided co-axial electrospray: A novel and versatile approach, Food Research International. 116 (2019) 1163–1172. https://doi.org/10.1016/j.foodres.2018.11.047.

[42] F. Esmaeili, M. Aghajani, A. Rashti, M. Abdollahi, R. Faridi-Majidi, H. Ghanbari, A. Amani, Parameters influencing size of electrosprayed chitosan/HPMC/TPP nanoparticles containing alendronate by an artificial neural networks model, Journal of Electrostatics. 112 (2021) 103598. https://doi.org/10.1016/j.elstat.2021.103598.

[43] B. Muñiz Castro, M. Elbadawi, J.J. Ong, T. Pollard, Z. Song, S. Gaisford, G. Pérez, A.W. Basit, P. Cabalar, A. Goyanes, Machine learning predicts 3D printing performance of over 900 drug delivery systems, Journal of Controlled Release. 337 (2021) 530–545. https://doi.org/10.1016/j.jconrel.2021.07.046.

[44] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature. 559 (2018) 547–555. https://doi.org/10.1038/s41586-018-0337-2.

[45] G.B. Goh, N.O. Hodas, A. Vishnu, Deep Learning for Computational Chemistry, ArXiv:1701.04503 [Physics, Stat]. (2017). http://arxiv.org/abs/1701.04503 (accessed September 7, 2020).

[46] L. Pattanaik, C.W. Coley, Molecular Representation: Going Long on Fingerprints, Chem. 6 (2020) 1204–1207. https://doi.org/10.1016/j.chempr.2020.05.002.

[47] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug Discovery Today. 23 (2018) 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

[48] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, Chem. Sci. 9 (2018) 513–530. https://doi.org/10.1039/C7SC02664A.

[49] S. Jaeger, S. Fulle, S. Turk, Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition, J. Chem. Inf. Model. 58 (2018) 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

[50] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, Journal of Cheminformatics. 10 (2018) 4. https://doi.org/10.1186/s13321-018-0258-y.

[51] D. Rogers, M. Hahn, Extended-Connectivity Fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754. https://doi.org/10.1021/ci100050t.

[52] C.M. Hansen, Hansen solubility parameters: a user's handbook, 2nd ed, CRC Press, Boca Raton, 2007.

[53] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Research. 49 (2021) D1388–D1395. https://doi.org/10.1093/nar/gkaa971.

[54] A. Zheng, A. Casari, Feature engineering for machine learning: principles and techniques for data scientists, First edition, O'Reilly, Beijing : Boston, 2018.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 12 (2011) 2825--2830.

[56] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016) 785–794. https://doi.org/10.1145/2939672.2939785.

[57] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed August 17, 2021).

[58] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995: pp. 1137–1143.

[59] C.M. Bishop, Pattern recognition and machine learning, Springer, New York, 2006.

[60] J.P. Borra, Y. Tombette, P. Ehouarn, Influence Of Electric Field Profile And Polarity On The Mode Of EHDA Related To Electric Discharge Regimes, Journal of Aerosol Science. 30 (1999) 913–925. https://doi.org/10.1016/S0021-8502(98)00779-4.

[61] S. Zhang, C. Campagne, F. Salaün, Influence of Solvent Selection in the Electrospraying Process of Polycaprolactone, Applied Sciences. 9 (2019) 402. https://doi.org/10.3390/app9030402.

[62] S. Kalantary, A. Jahani, R. Jahani, MLR and ANN Approaches for Prediction of Synthetic/Natural Nanofibers Diameter in the Environmental and Medical Applications, Sci Rep. 10 (2020) 8117. https://doi.org/10.1038/s41598-020-65121-x.

[63] A. Rácz, D. Bajusz, K. Héberger, Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification, Molecules. 26 (2021) 1111. https://doi.org/10.3390/molecules26041111.

[64] W. Li, Y. Yin, X. Quan, H. Zhang, Gene Expression Value Prediction Based on XGBoost Algorithm, Frontiers in Genetics. 10 (2019) 1077. https://doi.org/10.3389/fgene.2019.01077.

[65] A. Ogunleye, Q.-G. Wang, XGBoost Model for Chronic Kidney Disease Diagnosis, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 17 (2020) 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071.

[66] M.J. Willemink, W.A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L.R. Folio, R.M. Summers, D.L. Rubin, M.P. Lungren, Preparing Medical Imaging Data for Machine Learning, Radiology. 295 (2020) 4–15. https://doi.org/10.1148/radiol.2020192224.

[67] Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science, Npj Comput Mater. 4 (2018) 1–8. https://doi.org/10.1038/s41524-018-0081-z.

[68] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, A Structure-Based Platform for Predicting Chemical Reactivity, Chem. 6 (2020) 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.

[69] D.T. Ahneman, J.G. Estrada, S. Lin, S.D. Dreher, A.G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning, Science. 360 (2018) 186–190. https://doi.org/10.1126/science.aar5169.

[70] P. Probst, M.N. Wright, A.-L. Boulesteix, Hyperparameters and tuning strategies for random forest, WIREs Data Mining and Knowledge Discovery. 9 (2019) e1301. https://doi.org/10.1002/widm.1301.

[71] A.-R. Faramarzi, J. Barzin, H. Mobedi, Effect of solution and apparatus parameters on the morphology and size of electrosprayed PLGA microparticles, Fibers Polym. 17 (2016) 1806–1819. https://doi.org/10.1007/s12221-016-6685-3.

[72] A.M. Gañán-Calvo, Cone-Jet Analytical Extension of Taylor's Electrostatic Solution and the Asymptotic Universal Scaling Laws in Electrospraying, Phys. Rev. Lett. 79 (1997) 217–220. https://doi.org/10.1103/PhysRevLett.79.217.

[73] R.P.A. Hartman, D.J. Brunner, D.M.A. Camelot, J.C.M. Marijnissen, B. Scarlett, Electrohydrodynamic Atomization In The Cone–Jet Mode Physical Modeling Of The Liquid Cone And Jet, Journal of Aerosol Science. 30 (1999) 823–849. https://doi.org/10.1016/S0021-8502(99)00033-6.

[74] C.J. Luo, M. Nangrejo, M. Edirisinghe, A novel method of selecting solvents for polymer electrospinning, Polymer. 51 (2010) 1654–1662. https://doi.org/10.1016/j.polymer.2010.01.031.

[75] H. Masood, C.Y. Toe, W.Y. Teoh, V. Sethu, R. Amal, Machine Learning for Accelerated Discovery of Solar Photocatalysts, ACS Catal. 9 (2019) 11774–11787. https://doi.org/10.1021/acscatal.9b02531.

[76] Y. Yuan, N. He, L. Dong, Q. Guo, X. Zhang, B. Li, L. Li, Multiscale Shellac-Based Delivery Systems: From Macro- to Nanoscale, ACS Nano. (2021) acsnano.1c07121. https://doi.org/10.1021/acsnano.1c07121.

[77] F. Chen, G. Hochleitner, T. Woodfield, J. Groll, P.D. Dalton, B.G. Amsden, Additive Manufacturing of a Photo-Cross-Linkable Polymer via Direct Melt Electrospinning Writing for Producing High Strength Structures, Biomacromolecules. 17 (2016) 208–214. https://doi.org/10.1021/acs.biomac.5b01316.

[78] G. Hochleitner, T. Jüngst, T.D. Brown, K. Hahn, C. Moseke, F. Jakob, P.D. Dalton, J. Groll, Additive manufacturing of scaffolds with sub-micron filaments via melt electrospinning writing, Biofabrication. 7 (2015) 035002. https://doi.org/10.1088/1758-5090/7/3/035002.

[79] P.D. Dalton, C. Vaquette, B.L. Farrugia, T.R. Dargaville, T.D. Brown, D.W. Hutmacher, Electrospinning and additive manufacturing: converging technologies, Biomater. Sci. 1 (2013) 171–185. https://doi.org/10.1039/C2BM00039C.

**Declaration of interests**

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

**Dr Maryam Parhizkar**

**17/01/2022**

**Highlights**

- Machine learning models predicting electrospraying particle size were developed from a literature database of 445 records.
- XGBoost and Random Forest (RF) models yielded root-mean-squared errors (RMSE) of 3.91 μm and 6.19 μm evaluated by 5-fold cross-validation (CV).
- In-house experiments validated the models, revealing an accuracy of ± 1.3 μm and providing insight into model generalisation capability.
- Models successfully predicted electrospray processing attributes governing particle size, as previously identified by scaling laws.