AAPS PharmSciTech (2022) 23: 186
https://doi.org/10.1208/s12249-022-02335-4

RESEARCH ARTICLE

Applications of Machine Learning and A.I. in Pharmaceutical Development and Technology

Visible Particle Identification Using Raman Spectroscopy and Machine Learning

Han Sheng[1] · Yinping Zhao[1] · Xiangan Long[1] · Liwen Chen[2,3] · Bei Li[4] · Yiyan Fei[2] · Lan Mi[2] · Jiong Ma[1,2,5]

Received: 20 March 2022 / Accepted: 13 June 2022 / Published online: 6 July 2022

Abstract

Visible particle identification is a crucial prerequisite step for process improvement and control during the manufacturing of injectable biotherapeutic drug products. Raman spectroscopy is a technology with several advantages for particle identification including high chemical sensitivity, minimal sample manipulation, and applicability to aqueous solutions. However, considerable effort and experience are required to extract and interpret Raman spectral data. In this study, we applied machine learning algorithms to analyze Raman spectral data for visible particle identification in order to minimize expert support and improve data analysis accuracy. We manually prepared ten types of particle standard solutions to simulate the particle types typically observed during manufacturing and established a Raman spectral library with accurate peak assignments for the visible particles. Five classification algorithms were trained using visible particle Raman spectral data. All models had high prediction accuracy of >98% for all types of visible particles. Our results demonstrate that the combination of Raman spectroscopy and machine learning can provide a simple and accurate data analysis approach for visible particle identification.

Keywords Raman spectroscopy · Machine learning · Processing · Injectable · Particle identification

✉ Lan Mi
lanmi@fudan.edu.cn

✉ Jiong Ma
jiongma@fudan.edu.cn

1 Institute of Biomedical Engineering and Technology, Academy for Engineer and Technology, Fudan University, 220 Handan Road, Shanghai 200433, China

2 Shanghai Engineering Research Center of Ultra-precision Optical Manufacturing, Key Laboratory of Micro and Nano Photonic Structures (Ministry of Education), Green Photoelectron Platform, Department of Optical Science and Engineering, Fudan University, 220 Handan Road, Shanghai 200433, China

3 Ruidge Biotech Co. Ltd., No. 888, Huanhu West 2nd Road, Lin-Gang Special Area, China (Shanghai) Pilot Free Trade Zone, Shanghai 200131, China

4 State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, No. 3888 Dong Nanhu Road, Changchun, Jilin 130033, China

5 Shanghai Engineering Research Center of Industrial Microorganisms, The Multiscale Research Institute of Complex Systems (MRICS), School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China

Introduction

During the manufacturing of injectable pharmaceutical products, it is necessary to control and monitor particulate matter in accordance with current Good Manufacturing Practice regulations [1, 2]. The US Pharmacopeia <790> requires products to be essentially free of visible particulate matter [3]. Particulate contaminants, such as protein aggregates or foreign materials (e.g., glass, stainless steel, silicone oil), in the final drug product may result in a failure of sterility assurance and severely harm patient safety by generating an adverse immunological response [4, 5]. Typically, high efficiency particulate air filter systems and aseptic process operations are implemented during manufacturing to prevent particle and microbiological contamination [1, 2]. After filling, stoppering, and capping filled vials, 100% visual inspection (manual, semi-automated or automated) can efficiently detect visible particle defects according to the filling site procedure [3, 6]. These defect vials are then rejected. During 100% visual inspection, action limitations on typical defect rates should be established to identify atypical batches [6, 7]. If a limit

is exceeded, it should trigger an investigation, including forensic classification/identification of the particle and examination of the manufacturing processes [6].

According to the particle source used in the injectable drug product manufacturing process, visible particles are divided into extrinsic, intrinsic, and inherent particles [6]. Extrinsic particles are foreign to the manufacturing process and arise from the facility environment (e.g., non-process-related fibers, insect parts, inorganic and organic materials); intrinsic particles emerge from contact with product equipment train or materials (e.g., stainless steel, seals, gaskets, packaging glass, fluid transport tubing, and silicone lubricant); and inherent particles are generated from the product itself due to a specific stress force (e.g., protein aggregation). Since particles formed during manufacturing have complex sources, particle identification is crucial to particle characterization. This is also a prerequisite for investigating the root cause of particle contamination and for developing a suitable and effective contamination control strategy for future batch manufacturing.

Currently, multiple prevailing technologies for particle identification are used in pharmaceutical companies, such as scanning electron microscopy/energy-dispersive X-ray spectroscopy (SEM-EDX) [8, 9], attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) [10, 11], inductively coupled plasma mass spectrometry (ICP-MS) [12], and Raman spectroscopy [13]. SEM-EDX can detect the chemical elements of particles and provide relevant ratios to classify the chemical types of particles, such as protein-like particles with the elements C, O, N, and S. However, SEM-EDX cannot distinguish between two particles' chemical structure if they have similar elements and relevant ratios. ATR-FTIR can identify the chemical composition and structure of infrared active particles by comparing them with the chemical groups and vibrations from an infrared spectroscopy library. ATR-FTIR has some limitations: first, due to the strong water absorption in infrared spectroscopy, it is mainly used for solid samples and is not directly applicable to aqueous samples. Second, sample manipulation (e.g., particle filtration on a membrane prior to the infrared test) is time-consuming. ICP-MS has become the technique of choice for providing information on nanoparticle size and elemental composition (e.g. trace metal) to characterize nanoparticles in solution. However, it has several disadvantages, including a limited group of detectable elements, complicated sample preparation that uses a set of well-defined reference materials for accurate calibration, experience-dependent accurate testing, and many different types of interferences during testing [12, 14]. Raman spectroscopy has consequently emerged as an effective technique for particle identification. This technology has many advantages over other techniques, including high chemical sensitivity, minimal sample manipulation, and rapid and accurate testing. In addition, it can be applied to aqueous solutions due to its reduced water vibration in the fingerprint region [15]. Raman spectroscopy can directly identify foreign particles inside glass containers [16, 17], characterize sub-visible particles with particle sizes as low as 0.5 μm [18], and distinguish surfactant degradation particles in biopharmaceutical formulations [19]. However, Raman spectroscopy still has some limitations that need to be improved. These include limited Raman spectral libraries due to the lack of a thorough understanding of band assignments, and a high fluorescence background *vs* a weak Raman signal, which might be addressed using several different fluorescence suppression techniques [20, 21].

When applying fast and easy-to-handle Raman spectroscopy for particle identification, complex Raman spectral data might be generated during sample testing due to the similarity of some particle spectroscopies and limited Raman spectral libraries. This makes spectral interpretation and accurate information extraction difficult and dependent on expert experience. Machine learning is a rapidly growing data mining tool that can build a Raman spectral data prediction model for particle identification using machine learning algorithms based on interpretations from complex available datasets which humans would likely miss. Such models could be directly used in classification predictions of new samples and would require very little expert access and interpretation after end users are fully trained. Machine learning has been successfully applied in other fields such as food analysis [22, 23], cancer classification [24, 25], microbial identification [26], and protein classification [27, 28]. For example, Le et al. [27] demonstrated the improved predictive ability of four monoclonal antibodies (combined error of 2.4% *versus* 14.6%) using a linear approach. Zhang et al. [28] developed a support vector machine (SVM)–based regression model that can quickly and accurately predict protein aggregation. In addition, using the machine learning algorithms of principal component analysis (PCA)–discriminant function analysis (DFA) and PCA-SVM in combination with Raman spectroscopy can provide a rapid method to distinguish normal breast cells from breast cancer cells with greater than 97% accuracy [25].

In this study, we applied machine learning algorithms to analyze Raman spectra and identify common visible particles that might be observed during biopharmaceutical injectable drug product manufacturing, including cellulose, wool, polypropylene (PP), polyvinylidene fluoride (PVDF), polyether sulfone (PES), polytetrafluoroethylene (PTFE), silicone oil, silicone tubing, glass, and protein. The prediction models for visible particle identification were trained and validated using Raman

spectra obtained from manually prepared visible particle standard solutions. This work demonstrates the application of an accurate, fast, and easy-to-handle Raman spectroscopy–based visible particle detection method combined with an accurate and efficient machine learning algorithm for classification prediction of visible particles.

Materials and Methods

Materials

The following ten particle types were selected for this study: cellulose, wool, PP, PVDF, PES, PTFE, silicone oil, silicone tubing, glass, and protein. Cellulose and wool fibers are extrinsic particles that usually emerge from the facility environment, while PVDF, PES, PP, PTFE, silicone tubing, silicone oil, and glass particles are representative intrinsic particles' sources from process and product contact materials, and protein particles are inherent in biopharmaceutical products. These particle standard solutions were manually prepared in aqueous solutions to simulate the visible particles observed in biotherapeutic injectable drug products.

Polymer (Extrinsic and Intrinsic) Particles

Polymer particle solutions were prepared using a procedure similar to that described by Vollratch et al. [29]. Cellulose fiber is a typical extrinsic particle that is mainly sourced from the environment or autoclave packaging material during injectable liquid product manufacturing. Cellulose particles were cut into small pieces from the surgical grade paper of a self-sealing sterilization pouch autoclave bag (Cat# 89140-804, VWR, GA, USA). Wool fiber is a common material used in the textile industry, and it can be extrinsically introduced to the filled drug product solution via environmental contamination during the operator's gowning and filling operations. In this study, wool fiber particles from a woolen sweater purchased from Migaino (Shenzhen, China) were cut into small pieces.

Modified hydrophilic PVDF and PES are common sterile filter membrane materials used in biotherapeutic injectable liquid product manufacturing owning to their low protein adsorption and fast filtration flow throughput. PVDF and PES particles may enter final drug products by leaching from filter membranes. PVDF and PES particles were respectively prepared from 0.22-µm PVDF Durapore membrane filter (Cat# GVWP04700) and 0.22-µm PES Express PLUS membrane filter (Cat# GPWP04700) purchased from Millipore (Merck, Burlington, USA).

PP biotainers or sampling tools, PTFE valve membranes, and silicone tubing are commonly used in filling line systems. They can become corroded or be spalled into particles under certain stress conditions and subsequently be filled into the final product. PP, PTFE, and silicone tubing particles were respectively crushed into particles from raw materials using a stainless-steel shredder (FSJ-A03D1, Bear, China). The platinum-cured silicone tubing (Cat# 96410-16) was purchased from Masterflex (IL, USA), PTFE diaphragm valves were purchased from Ningci (Shanghai, China), and 1.5 mL PP microcentrifuge tubes (Cat# 509-GRD-Q) were purchased from Quality Scientific Plastics (New Hampshire, USA).

Silicone oil is usually used as a lubricant for primary containers such as vials, pre-filled syringes, cartridges, and plungers. If silicone oil particles leach from the surface of primary containers, they can then be observed in the drug product. Silicone oil (Cat# PMX-200, Aladdin, Shanghai, China) particle solution was diluted 1:100 in purified water and mixed homogenously.

Glass is widely used as a primary container in the injectable pharmaceutical industry, and glass particles are sometimes observed in filled vials due to vial breakage during filling. Glass particles were pestled into a fine glass powder from 2R vials (Cat# V002711080D, Schott, Suzhou China).

The micro-sized (< 500 µm) particles derived above were suspended in purified water, and these standard solutions were stored at 2–8°C for Raman spectroscopy.

Protein Particles

In a typical biotherapeutic injectable drug product manufacturing process, there are multiple sources of stress conditions including unexpected high-temperature exposure during production, shear stress during the freeze-thawing process of the drug substance, mixing of compounded drug product bulk, and bulk transfer via a peristaltic pump or gas pressure, filtration, and filling. Therefore, protein aggregation and protein particles may be generated under unfavorable stress conditions during manufacturing. These protein particles, which are formed after sterile filtration, get filled into the final container (e.g., vials, pre-filled syringes, or cartridges) and impact product quality and patient safety. In this study, to simulate unexpected high-temperature exposure during manufacturing, protein particles were generated by placing a 2R vial filled with IgG1 antibody (Mab1) formulation solution (25 mg mL^{-1} protein in formulation buffer, pH 6.2, donated by a local biopharmaceutical company) on a heating plate at 90°C for 2 h (only the bottom of the vial in contact with the heating plate). The protein solution was initially transparent and colorless, but its color gradually changed until it was translucent white at the end

of the heat stress treatment. Protein particles (approximately 2–100 µm in size, as measured by Raman spectroscopy) were observed under 1200 lx with a black background. The prepared protein particle standard solution was stored at 2–8°C until Raman spectroscopy.

Sample Manipulation

The particle samples for the Raman spectra measurements were transferred to a glass chip under a visual inspection station with a black background. After gentle swirling of the container, the liquid portion containing visible particles in the prepared particle standard solution (polymer or protein particles) was directly pipetted using a 1-mL pipette. After a visual check to confirm the location of the particles, the solution was transferred from the container to a low-background glass chip (HOOKE Instruments Ltd., China). No dilution or filtration was performed during the sample manipulation of the particle standard solutions. As this study mainly focused on the Raman data of a single visible particle of a certain size in the particle standard solutions, the particle count (not tested but sufficient for Raman measurement) had a negligible impact on the experimental procedure and study conclusion.

Methods

Raman Measurements

Raman spectra were measured on a HOOKE P300 confocal micro-Raman spectrometer (HOOKE Instruments Ltd., China) equipped with a semiconductor-cooled (−75°C) charge-coupled device (CCD) detector (1340 × 100 pixels, PIXIS 100 B, Princeton Instruments, USA) to achieve a high signal-to-noise ratio (quantum efficiency >90% at 550 nm). Sample excitation occurred at a laser wavelength of 532 nm from a solid-state, fiber-coupled laser (50 mW, 1 MHz) with a 600-g mm^{-1} grating and an objective lens of 100× (LMPlan FLN 100×, Olympus, Japan). This provided a lateral resolution of <1 µm and was used to collect the spectra from 286 to 3745 cm^{-1} with a spectral resolution of 3 cm^{-1}. The Rayleigh-scattered photons were removed using a notch filter to transmit backscattered Raman signals. The laser intensity was maintained at 6 mW with an exposure time of 6 s for Raman spectra measurement on a glass chip. The spectra were calibrated with a silica band at 520.7 cm^{-1} before sample testing.

Data Pre-Processing

Data pre-processing was performed using Origin Pro 9.1 (OriginLab, Northampton, USA). A total of 50 spectra were acquired for each particle standard solution. The mean spectra were calculated, and all spectra were baseline subtracted (2nd derivative with 0.05 threshold) to remove the fluorescence background and [0,1] normalized for Raman intensity. The peak assignment was analyzed from 400 to 3745 cm^{-1} to cover all the characteristic peaks for the investigated visible particle types. Raman shifts from 400 to 1800 cm^{-1} were selected for Raman spectra processing and machine learning classification prediction, as this range is the fingerprint region of skeletal vibrations for most molecules.

Principal Component Analysis Visualization

Independent PCA [30] visualization of the particle standard solutions was performed using MATLAB R2021a (MathWorks, Natick, MA, USA). The 1st, 2nd, and 3rd principal components were selected and plotted in MATLAB R2021a.

Classification of Machine Learning Analysis

Classification of the machine learning analysis was performed using the Classification Learner App in MATLAB R2021a (MathWorks, Natick, MA, USA). Among the 50 spectra derived for each particle standard, 90% (45 spectra) were used as a training set for algorithms to learn, whereas 10% (five spectra) were used as a test set to determine the quality of the model predictions. Numeric components of ten for the PCA analysis and cross-validation folds of ten were implemented for classification model training. The five prevalent classifiers of machine learning algorithms in this study were decision tree [31], discriminant analysis [32], SVM [33], K-nearest neighbor (KNN) classifiers [34], and ensemble classifier [35].

Results and Discussion

Raman Spectroscopy Characterization

Raman spectra of ten types of visible particle standard samples were obtained. The Raman spectra in the fingerprint region of 400–1800 cm^{-1} were selected for data pre-processing with baseline subtraction and normalization, which were subsequently used in the prediction model creation using machine learning algorithms. Figure 1 shows the mean intensity of the acquired Raman spectra for each particle standard (n = 50) in the region of 400–1800 cm^{-1}, plotted with the standard deviation in gray shade overlay. The results show that a small method variation (standard deviation) of the Raman spectra was observed for polypropylene, PES, PVDF, PTFE, silicone oil, silicone tubing, and glass particles. However, relatively more variability

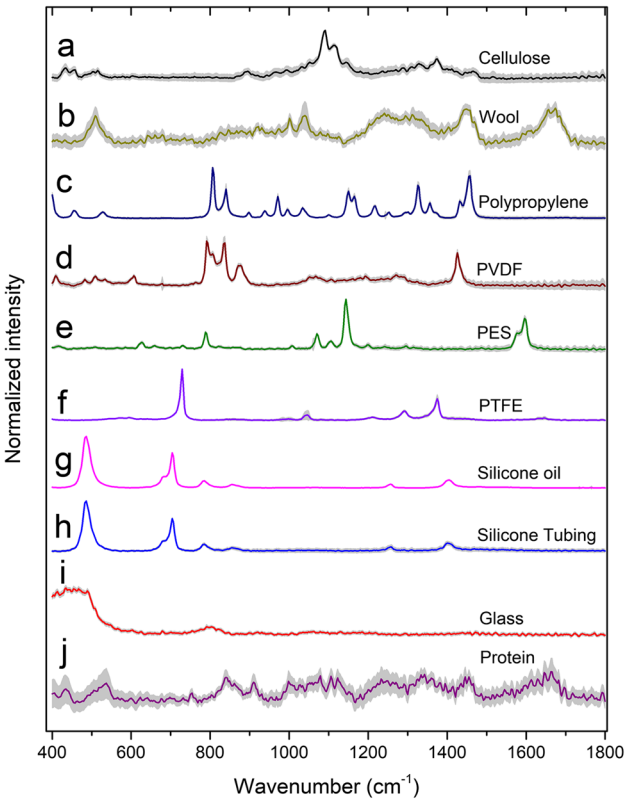

Fig. 1 Raman spectra of typical visible particles in particle standard solutions. **a** Cellulose; **b** wool; **c** polypropylene; **d** PVDF; **e** PES; **f** PTFE; **g** silicone oil; **h** silicone tubing; **i** glass; **j** protein (IgG1 antibody). Raman spectra were measured at 532 nm laser-excited wavelength with a laser intensity at 6 mW and a exposure time of 6 s. Mean spectra plotted in the figure were calculated from $n = 50$ spectra per visible particle type, and standard deviation was added in grey shade for each mean spectra

was observed for cellulose fiber, wool fiber, and protein particles, which may be due to the relatively strong fluorescent background causing a low signal/noise ratio. Additional experiments showed that we could effectively suppress the fluorescence, increase the signal/noise ratio, and minimize the detection variation in several ways: (1) by using the excitation laser to "photobleach" the samples over an extended period of time (typically several minutes) to destroy the fluorescent chromophores [20]; (2) by increasing the laser intensity (e.g., to 10 mW); and/or (3) by extending the exposure time (e.g. by 10s) (data not shown here).

To better understand the main chemical group vibration modes and compare the characteristic peaks for each particle molecule, we analyzed the band assignments for each derived mean spectrum from the particle standard solutions in the detection region of 400–3745 cm^{-1} (summarized in Table 1). The main band assignments in Table 1 demonstrate the consistency and accuracy of the acquired Raman spectra for each particle type in this study compared to the published Raman spectra band assignments [36–46]. Furthermore, this Raman spectral library served as a benchmark for the machine learning prediction model creation in this study. The details of the main Raman vibration mode interpretation for each type of particle are provided in the 18.

Machine Learning Analysis for Visible Particle Investigation

Principal Component Analysis Feature Dimension Reduction

PCA is a classical method for reducing high-dimensional data while retaining most of the variation in a dataset [30]. Using a few components, the sample data can be plotted, and the similarities and differences between samples can be visually assessed [30]. Most importantly, it can speed up machine-learning algorithms and apply them to the entire dataset. In this study, principal components were analyzed for the entire Raman spectral data set in the region of 400–1800 cm^{-1} after baseline subtraction and normalization. PC1 (56%), PC2 (14%), and PC3 (7%) were selected for 3D visualization. Figure 2 shows all ten types of visible particles: four (glass, PTFE, PES, cellulose) are displayed as fully isolated and six (wool and protein, PVDF and polypropylene, silicone oil and silicone tubing) as partially overlapping.

Training and Testing of Machine Learning Models

In this study, we used the classification learner app for machine learning analysis in MATLAB to build a model that could be applied to the prediction test set. The dataset consisted of ten types of typical visible particles with 50 Raman spectra acquired from each particle standard ($n = 500$ total). PCA was initiated for all machine learning algorithms. Raman spectra in the fingerprint region of 400–1800 cm^{-1} were used in machine learning analysis after baseline subtraction and normalization. The particle name was imported as a response, and the band frequency (cm^{-1}) was imported as a predictor. During model training, the cross-validation fold was $k = 10$ to overcome the insufficient data package and to avoid model overfitting. The top ten components (96.9% of the total variance: 55.5%, 14.4%, 6.9%, 5.7%, 3.8%, 3.6%, 2.5%, 2.0%, 1.8%, and 0.7%) were selected to reduce the data dimension number from 473 to 10, but remained as highly representative as possible of the data characteristics. The five machine learning algorithms in this study included ensemble classifiers, SVM, KNN, discriminant analysis, and decision tree. All the machine

Table 1 Raman Data of Different Types of Visible Particles and Band Assignment

Assignment	Band frequency (cm^{-1}) and intensity								
	Cellulose	Wool	Polypropylene	PVDF	PES	PTFE	Silicone tubing/ oil	Glass	Protein
υCC, υCO [36]	432m	-	-	-	-	-	-	438s	438w
4-fold silicate ring [45]	-	-	-	-	-	-	-	491s	-
υSiO sym, δCF_2 [40, 41]	-	-	-	486w	-	-	486s	-	-
υS-S, cysteine [37, 38]	-	~509s	-	-	-	-	-	-	537m
υC-Si-C sym. [40]	-	-	-	-	-	-	704s	-	-
υCF_2 sym. [44]	-	-	-	-	-	735vs	-	-	-
υC-Si-C asym [40].	-	-	-	-	-	-	785m	-	-
C–C aromatic ring [37, 42]	-	-	-	-	788s	-	-	-	-
SiO_4 with zero bridging oxygen [45]	-	-	-	-	-	-	-	797s	-
ρCH_2, υCCb, $\upsilon C\text{-}CH_3$, ρCH_3 [39–41]	-	-	811s, 842m	797s, 809s-sh, 837s, 878s	-	-	858w	-	-
δHCC, δHCO, (C–O–C) skeletal [36, 37]	895m	828-849w	-	-	-	-	-	-	843m
υCCb, ρCH_3, δCH, ωCH_2, C–C aromatic ring, CH deformation, phenylalanine [36, 37, 39]	995m-sh	1001m	994w	-	1009w	-	-	-	1001m-sh
υCO, υCC; υCOC, υCC asym., υCF_3 sym. [36, 37, 41, 43, 44]	1052s-sh, 1091vs	1079w	-	1052w	1073s	1043w	-	-	1080m
υSO [43]	-	-	-	-	1109m	-	-	-	-
υCCb, $\upsilon C\text{-}CH_3$, δCH,ρCH_3, τCH_2, $\upsilon C\text{-}C_6H_5$, υCF_2 sym. [36, 37, 39, 44]	1115s-sh, 1144s-sh	-	1152m, 1165m-sh, 1220m	1197w	1203w	1218w	-	-	1106m, 1121m
υO=S=O sym [43].	-	-	-	-	1145vs	-	-	-	-
Amide III, umbrella [37, 38]	-	1221–1282m	-	-	-	-	-	-	1190–1291m
υO=S=O asym., υCF_2 asym. [41]	-	-	-	1273w	1297w	1300m	-	-	-
δHCC, δHCO, δHOC, ωCH_2, τCH_2, δCH, τCH_2, δCH_3 sym., δCH [36, 37, 39]	1337m, 1377s	1294m, 1308m-sh	1305w, 1329s, 1359m-sh	-	-	-	-	-	1337m
υCF [44]	-	-	-	-	-	1378s	-	-	-
δHCH, δHOC, δCH_2, ωCH_2, CH_2/CH_3 deformation; δCH_3 asym. [36, 37, 39, 41]	1446~1472m	1449s	1433m, 1458s	1426s	-	-	-	-	1452m
υC=C aromatic ring chain vibrations, tryptophan, C=N [37, 42, 43]	-	1593w, 1607w-sh	-	-	1580s-sh, 1597vs	-	-	-	1618w
Amide I (α Helix) [37]	-	~1657s	-	-	-	-	-	-	-
Amide I (β-Pleated sheet) + disordered [46]	-	~1677s	-	-	-	-	-	-	1666m
Amide I ($CO\text{-}NH_2$) [37]	-	~1685m-sh	-	-	-	-	-	-	-
υCH_2 sym. [39, 41]	-	-	2836s	2976vs	-	-	-	-	-
υCH_2 asym., υCH_3 sym. [36, 37, 39, 41]	2889vs	-	2880vs	3019s	-	-	-	-	-
υCH sym. [36, 37, 39, 40, 43]	2901vs-sh	-	2903s-sh	-	3072s	-	-	-	-

Table 1 (continued)

Assignment	Band frequency (cm^{-1}) and intensity								
	Cellulose	Wool	Polypropylene	PVDF	PES	PTFE	Silicone tubing/ oil	Glass	Protein
υCH_2, υCH_2 asym.,υCH_3 sym. [37, 44]	-	2922vs	2921s-sh	-	-	2929w	2901vs	-	-
υCH_3 asym., υCH asym. [36, 37, 39, 40]	2954s-sh	-	2959s	-	-	-	2964s	-	2937vs
υOH [36, 37]	3147–3549m	3138–3445m	-	-	-	-	-	-	3019–3669s

υ stretching, *ω* wagging, *δ* bending, *τ* twisting, *ρ* rocking, *b* backbone, *vs* very strong, *s* strong, *m* medium, *w* weak, *sh* shoulder, *PVDF* polyvinylidene fluoride, *PES* polyether sulfone, *PTFE* polytetrafluoroethylene

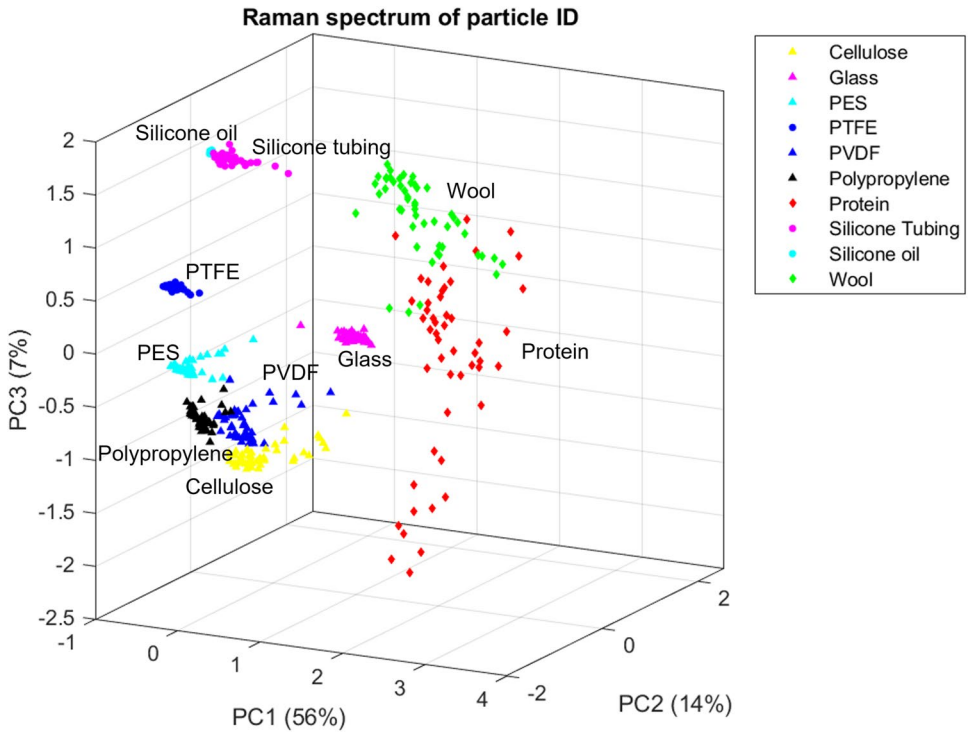

Fig. 2 PCA 3D visualization with 3 PCs

learning algorithms were trained using 90% of the visible particle Raman spectra; the remaining 10% were used as the test dataset.

The confusion matrix generated from the classification learner app in MATLAB was used to evaluate the best model with the highest prediction accuracy for visible particle identification using the Raman spectra. Table 2 lists the training (validation) accuracy and test accuracy of the five algorithm models to predict visible particle identification using Raman spectral data from particle standard solutions. Figures 3 and 4 show the confusion matrices (number of observations) obtained using these five algorithms for the training and test datasets of the particle standard solutions, respectively.

In Table 2, among the five investigated machine learning algorithms, the ensemble classifier (subspace KNN) shows

Table 2 Training and Test Accuracy for Visible Particle Identification with Raman Spectra Data from Particle Standard Solutions Using Five Machine Learning Algorithms

No.	Model	Particle standard solutions	
		Training (validation) accuracy	Test accuracy
1	1.23 ensemble — subspace KNN	100.0%	100.0%
2	1.10 SVM — cubic SVM	99.8%	100.0%
3	1.14 KNN — fine KNN	99.8%	100.0%
4	1.5 quadratic discriminant — quadratic discriminant	99.3%	98.0%
5	1.1 tree — fine tree	98.4%	100.0%

KNN K-nearest neighbor, *SVM* support vector machine

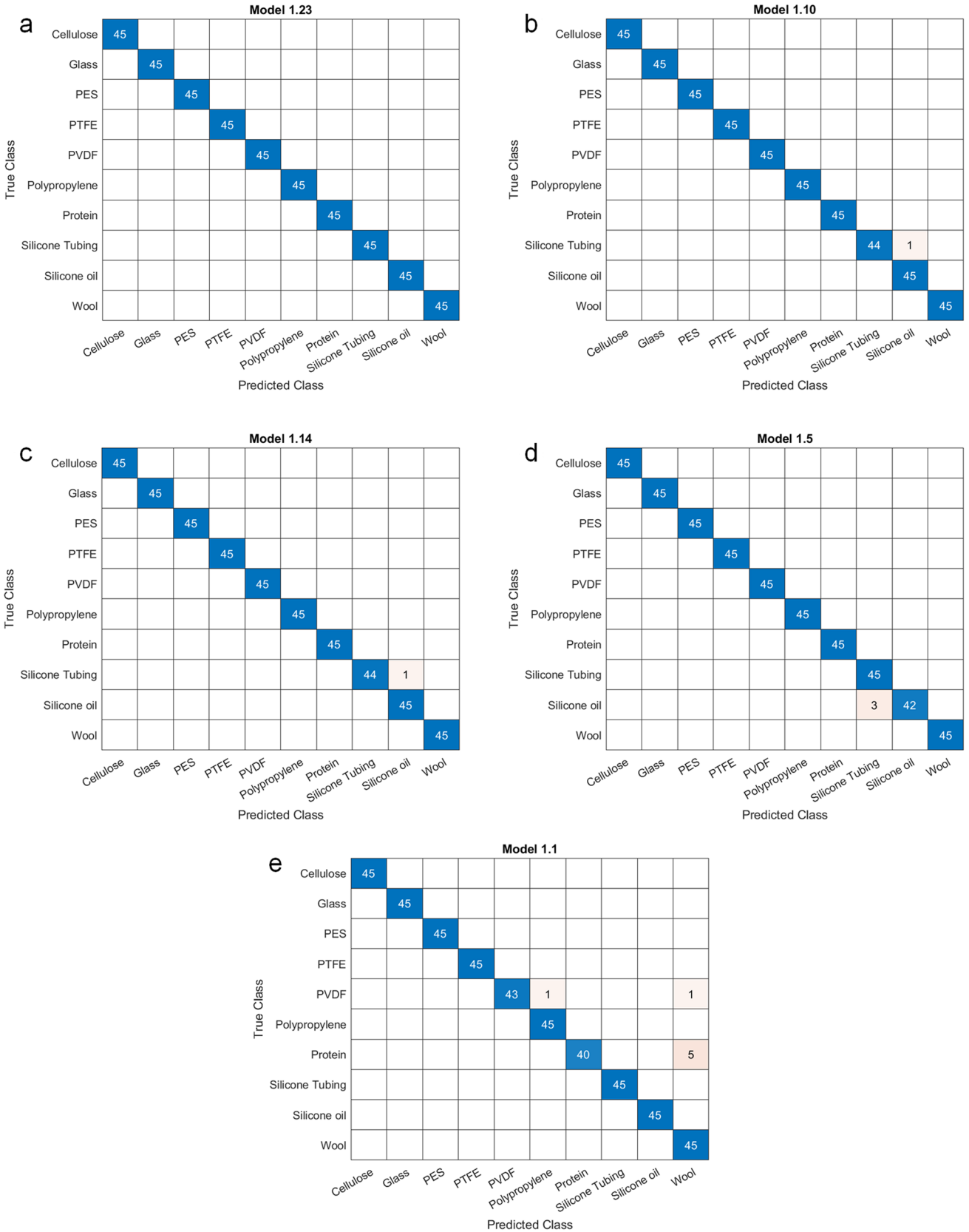

Fig. 3 Training (validation) confusion matrix for variant visible particles in particle standard solutions using different machine learning algorithms. **a** 1.23 ensemble — subspace KNN; **b** 1.10 SVM — cubic SVM; **c** 1.14 KNN — fine KNN; **d** 1.5 quadratic discriminant; **e** 1.1 tree — fine tree

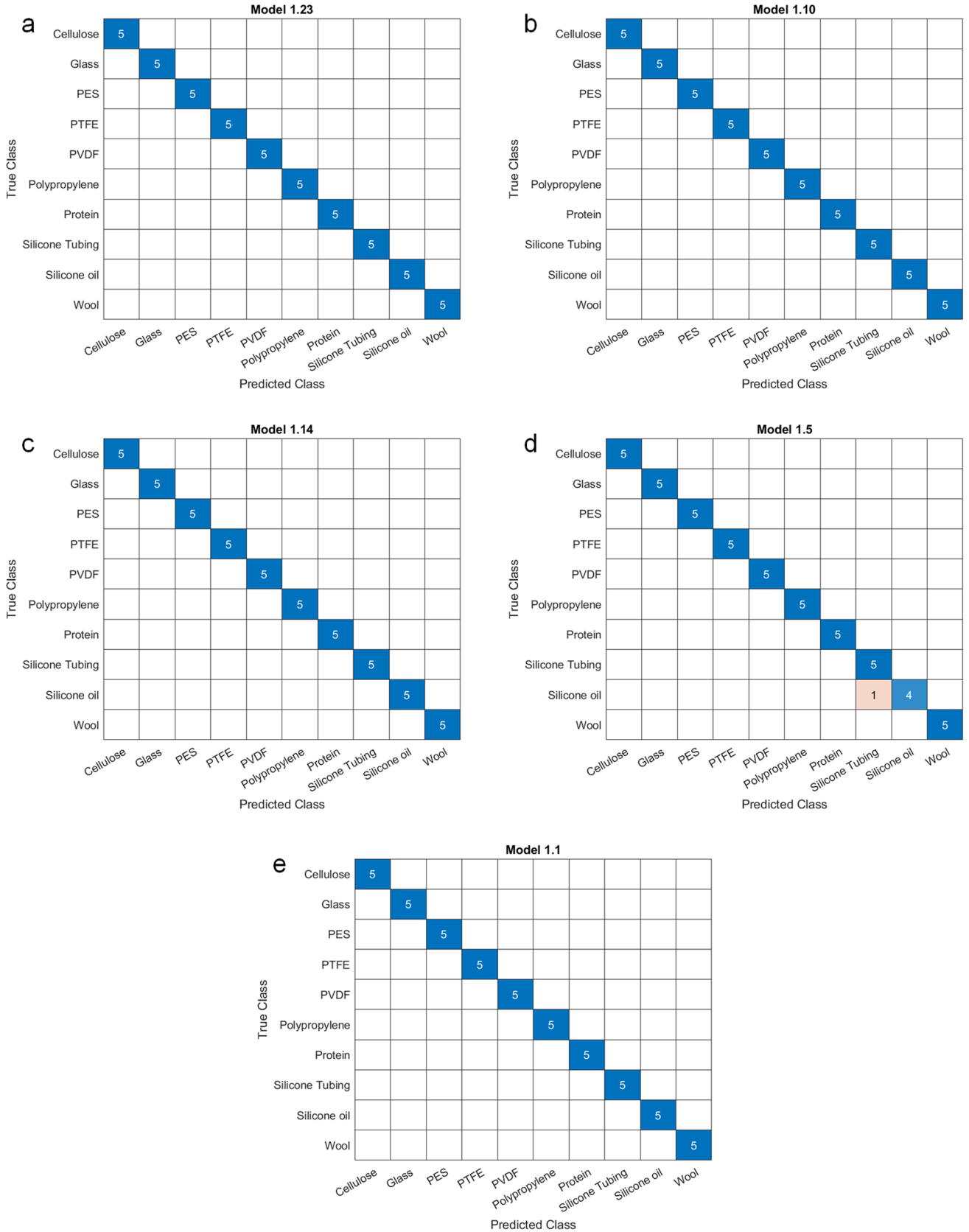

Fig. 4 Test confusion matrix for variant visible particles in particle standard solutions using different machine learning algorithms. **a** 1.23 ensemble — subspace KNN; **b** 1.10 SVM — cubic SVM; **c** 1.14 KNN — fine KNN; **d** 1.5 quadratic discriminant; **e** 1.1 tree — fine tree

a relatively good model with the highest prediction accuracy of 100% for both the training and test sets. The other four algorithms also achieved a high prediction accuracy of >98% for both training and test sets.

In the ensemble classifier model (subspace KNN), all Raman spectra from the particle standard solutions were successfully classified into the correct particle type (Figures 3a and 4a). This model was able to distinguish correctly between particle types (e.g., silicone oil and silicone tubing) with very similar spectral characteristics and could also recognize potential slight differences that cannot be identified visually.

In the validation confusion matrices of Figure 3b and 3c, one (1/45) silicone tubing particle Raman data point (true class) was mispredicted as silicone oil for both cubic SVM and fine KNN classifiers. In contrast, three (3/45) silicone oil Raman data points were misclassified as silicone tubing for the quadratic discriminant classifier (Figure 3d). For the fine tree classifier (Figure 3e), one (1/45) PVDF Raman data point was misclassified as polypropylene, and one (1/45) PVDF Raman data point was misclassified as wool. An additional five (5/45) protein Raman data points were misclassified as wool. Unsurprisingly, the misprediction pairs observed in different machine learning algorithms coincided with the partially overlapping particle groups shown in the PCA 3D visualization plot (Figure 2) (e.g., silicone oil and silicone tubing, wool and protein, PVDF, and polypropylene), which is mainly due to the partial or high similarity of Raman spectra among these groups (Figure 1).

Compared to other algorithms, the fine tree classifier (one of the decision tree algorithms in MATLAB) gave a relatively high misprediction number (5/45) for protein particles when used in model training. This might be because the classification tree is built through a binary recursive partitioning process from the root node to each leaf note, and the predictors for each node are generated and optimized step-by-step using each individual data. By contrast, other classifiers are primarily built through a distribution or fitting function optimized from the overall dataset; thus, the individual data have a greater impact on the prediction accuracy of the decision tree classifier compared to that on accuracy of other algorithms.

In the test confusion matrices (Figure 4), only the fine KNN classifier (Figure 4d) mispredicted one (1/5) silicone oil Raman data point as silicone tubing. The other classifiers showed 100% prediction accuracy after training.

Conclusion

In summary, this study generated classification models with a high prediction accuracy of >98%, by applying machine learning to Raman spectral data analysis for visible particle identification of manually prepared particle standard solutions. The highest prediction accuracy was 100% and was achieved using the ensemble (subspace KNN) classifier. Using this model, all Raman spectral data were successfully classified into the correct particle type, even for highly similar Raman spectral data species of silicone oil *versus* silicone tubing.

However, to further explore the applicability of this approach for particle identification in real drug products, it will be necessary to expand the library to include additional sources of particles (e.g., new particle types, composite particles, and stress conditions) and to set up various offline sample manipulation approaches. The proposed particle prediction models could be utilized in different Raman-based process analytical technologies through the development of an online micro-Raman detection probe and automatic software for particle identification. With these efforts, the optimized approach could potentially accelerate visible particle identification by simplifying data analysis, improving result accuracy, and potentially responding in real-time. This will ultimately shorten the investigation into causes of visible particles in injectable liquid drug product manufacturing and advance process improvement and control.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1208/s12249-022-02335-4.

Author Contribution

- Substantial contributions to the conception or design of the work or the acquisition, analysis, or interpretation of data for the work: Jiong Ma, Lan Mi, Han Sheng, Yinping Zhao, and Xiangan Long
- Drafting the work or revising it critically for important intellectual content: Han Sheng, Yinping Zhao, Xiangan Long, Liwen Chen, Bei Li, and Yiyan Fei
- Final approval of the version to be published: Jiong Ma, and Lan Mi
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: Han Sheng, Yinping Zhao, Xiangan Long, Liwen Chen, Bei Li, Yiyan Fei, Lan Mi, and Jiong Ma

Funding This work was supported by the Medical Engineering Fund of Fudan University (yg2021-022), Pioneering Project of Academy for Engineering and Technology of Fudan University (gyy2018-001, gyy2018-002), Shanghai Key Discipline Construction Plan (2020-2022) (Grant No. GWV-10.1-XK01), and National Natural Science Foundation of China (62175034, 62175036).

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. CFR 21 Part 211 Current good manufacturing practice for finished pharmaceuticals. Office of the Federal Register. National Archives and Records Administration. 2020. https://www.acces

sdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=211. Accessed 6 Oct 2021.
2. EU GMP Annex 1 Revision: manufacture of sterile medicinal products (draft). European Commission. 2020. https://www.gmp-compliance.org/files/guidemgr/2020_annex1ps_sterile_medicinal_products_en.pdf. Accessed 6 Oct 2021.
3. General Chapter: USP. <790> Visible particulates in injections. In: USP-NF. Rockville, MD: USP; May 1, 2016. https://doi.org/10.31003/USPNF_M7197_01_01
4. Jiskoot W, Randolph TW, Volkin DB, Middaugh CR, Schöneich C, Winter G, et al. Protein instability and immunogenicity: roadblocks to clinical application of injectable protein delivery systems for sustained release. J Pharm Sci. 2012;101(3):946–54. https://doi.org/10.1002/jps.23018.
5. Jiskoot W, Kijanka G, Randolph TW, Carpenter JF, Koulov AV, Mahler HC, et al. Mouse models for assessing protein immunogenicity: lessons and challenges. J Pharm Sci. 2016;105(5):1567–75.
6. General Chapter: USP. <1790> Visual inspection of injections. In: USP-NF. Rockville, MD: USP; May 1, 2022. https://doi.org/10.31003/USPNF_M7198_06_01
7. ICH. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Q10 Pharmaceutical Quality System (PQS). 2009. [accessed 2021 Oct 6]. https://www.fda.gov/media/71553/download
8. Li GG, Cao S, Jiao N, Wen ZQ. Classification of glass particles in parenteral product vials by visual, microscopic, and spectroscopic methods. PDA J Pharm Sci Technol. 2014;68(4):362–72. https://doi.org/10.5731/pdajpst.2014.00986.
9. Idris AM, El-Zahhar AA. Indicative properties measurements by SEM, SEM-EDX and XRD for initial homogeneity tests of new certified reference materials. Microchem J. 2019;146:429–33.
10. Brückl L, Hahn R, Sergi M, Scheler S. A systematic evaluation of mechanisms, material effects, and protein-dependent differences on friction-related protein particle formation in formulation and filling steps. Int J Pharm. 2016;511(2):931–45.
11. Nashed-Samuel Y, Torraca G, Liu D, Fujimori K, Zhang Z, Wen ZQ, et al. Identification of an extraneous black particle in a glass syringe: extractables/leachables case study. PDA J Pharm Sci Technol. 2010;64(3):242–8.
12. Semenova D, Silina YE. The role of nanoanalytics in the development of organic-inorganic nanohybrids-seeing nanomaterials as they are. Nanomaterials (Basel). 2019;9(12):1673. https://doi.org/10.3390/nano9121673.
13. Stefaniak EA, Worobiec A, Potgieter-Vermaak S, Alsecz A, Van Grieken R. Molecular and elemental characterisation of mineral particles by means of parallel micro-Raman spectrometry and scanning electron microscopy/energy dispersive X-ray analysis. Spectrochim Acta Part B At Spectrosc. 2006;61(7):824–30.
14. Bulska E, Wagner B. Quantitative aspects of inductively coupled plasma mass spectrometry. Philos Trans A Math Phys Eng Sci. 2016;374(2079):20150369. https://doi.org/10.1098/rsta.2015.0369.
15. Benevides JM, Overman SA, Thomas GJ Jr. Raman spectroscopy of proteins. Curr Protoc Protein Sci 2004;Chapter 17. https://doi.org/10.1002/0471140864.ps1708s33
16. Cao X, Wen ZQ, Vance A, Torraca G. Raman microscopic applications in the biopharmaceutical industry: in situ identification of foreign particulates inside glass containers with aqueous formulated solutions. Appl Spectrosc. 2009;63(7):830–4. https://doi.org/10.1366/000370209788701026.
17. Caudron E, Tfayli A, Monnier C, Manfait M, Prognon P, Pradeau D. Identification of hematite particles in sealed glass containers for pharmaceutical uses by Raman microspectroscopy. J Pharm Biomed Anal. 2011;54(4):866–8. https://doi.org/10.1016/j.jpba.2010.10.023.
18. Singh SK, Afonina N, Awwad M, Bechtold-Peters K, Blue JT, Chou D, et al. An industry perspective on the monitoring of subvisible particles as a quality attribute for protein therapeutics. J Pharm Sci. 2010;99(8):3302–21. https://doi.org/10.1002/jps.22097.
19. Saggu M, Liu J, Patel A. Identification of subvisible particles in biopharmaceutical formulations using Raman spectroscopy provides insight into polysorbate 20 degradation pathway. Pharm Res. 2015;32(9):2877–88. https://doi.org/10.1007/s11095-015-1670-x.
20. Kostamovaara J, Tenhunen J, Kögler M, Nissinen I, Nissinen J, Keränen P. Fluorescence suppression in Raman spectroscopy using a time-gated CMOS SPAD. Opt. 2013;21(25):31632–45. https://doi.org/10.1364/OE.21.031632.
21. Wei D, Chen S, Liu Q. Review of fluorescence suppression techniques in Raman spectroscopy. Appl Spectrosc. 2015;50(5):387–406. https://doi.org/10.1080/05704928.2014.999936.
22. Berghian-Grosan C, Magdas DA. Raman spectroscopy and machine-learning for edible oils evaluation. Talanta. 2020;218:121176. https://doi.org/10.1016/j.talanta.2020.121176.
23. Berghian-Grosan C, Magdas DA. Application of Raman spectroscopy and machine learning algorithms for fruit distillates discrimination. Sci Rep. 2020;10(1):21152. https://doi.org/10.1038/s41598-020-78159-8.
24. Mandrell CT, Holland TE, Wheeler JF, Esmaeili S, Amar K, Chowdhury F, et al. Machine learning approach to Raman spectrum analysis of MIA PaCa-2 pancreatic cancer tumor repopulating cells for classification and feature analysis. Life (Basel). 2020;10(9):181. https://doi.org/10.3390/life10090181.
25. Zhang L, Li C, Peng D, Yi X, He S, Liu F, et. al. Raman spectroscopy and machine learning for the classification of breast cancers. Spectrochim Acta A Mol Biomol Spectrosc 2022;264:120300. https://doi.org/10.1016/j.saa.2021.120300
26. Lu W, Chen X, Wang L, Li H, Fu YV. Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. Anal Chem. 2020;92(9):6288–96. https://doi.org/10.1021/acs.analchem.9b04946.
27. Le L, Kégl B, Gramfort A, Marini C, Nguyen D, Cherti M, et al. Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach. Talanta. 2018;184:260–5. https://doi.org/10.1016/j.talanta.2018.02.109.
28. Zhang C, Springall JS, Wang X, Barman I. Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy. Anal Chim Acta. 2019;1081:138–45. https://doi.org/10.1016/j.aca.2019.07.007.
29. Vollrath I, Mathaes R, Sediq AS, Jere D, Jörg S, Huwyler J, et al. Subvisible particulate contamination in cell therapy products - can we distinguish? J Pharm Sci. 2020;109(1):216–9. https://doi.org/10.1016/j.xphs.2019.09.002.
30. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26(3):303–4. https://doi.org/10.1038/nbt0308-303.
31. Pandya R, Pandya J. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. Int J Comput Appl. 2015;117:18–21.
32. Sun J, Zhao H. The application of sparse estimation of covariance matrix to quadratic discriminant analysis. BMC Bioinformatics. 2015;16:48. https://doi.org/10.1186/s12859-014-0443-6.
33. Lee Y. Support vector machines for classification: a statistical portrait. Methods Mol Biol. 2010;620:347–68. https://doi.org/10.1007/978-1-60761-580-4_11.
34. Abu Alfeilat HA, Hassanat A, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of distance measure choice on K-nearest neighbor classifier performance: a review. Big Data. 2019;7(4):221–48. https://doi.org/10.1089/big.2018.0175.
35. Gul A, Perperoglou A, Khan Z, Mahmoud O, Miftahuddin M, Adler W, et al. Ensemble of a subset of kNN classifiers. Adv

Data Anal Classif. 2018;12(4):827–40. https://doi.org/10.1007/s11634-015-0227-5.
36. Wiley JH, Rajai H, Atalla RH. Band assignments in the Raman spectra of celluloses. Carbohydr Res. 1987;160:113–29. https://doi.org/10.1016/0008-6215(87)80306-3.
37. Movasaghi Z, Rehman S, Rehman IU. Raman spectroscopy of biological tissues. Appl Spectrosc. 2007;42(5):493–541. https://doi.org/10.1080/05704920701551530.
38. Liu H, Yu W. Study of the structure transformation of wool fibers with Raman spectroscopy. J Appl Polym Sci. 2007;103:1–7.
39. Andreassen E. Infrared and Raman spectroscopy of polypropylene. In: Karger-Kocsis J, editor. Polypropylene, Polym Sci Technol Ser, vol. 2; 1999. p. 320–8. https://doi.org/10.1007/978-94-011-4421-6_46.
40. Jayes L, Hard AP, Séné C, Parker SF, Jayasooriya UA. Vibrational spectroscopic analysis of silicones: a Fourier transform-Raman and inelastic neutron scattering investigation. Anal Chem. 2003;75(4):742–6. https://doi.org/10.1021/ac026012f.
41. Nallasamy P, Mohan S. Vibrational spectroscopic characterization of form II poly (vinylidene fluoride). Indian J Pure Appl Phys. 2005;43:821–7.
42. Lobo H, Bonilla JV (Eds.). Handbook of plastics analysis (1st ed.). CRC Press. 2003;265-266. https://doi.org/10.1201/9780203911983
43. Abdelrazek EM, Abdelghany AM, Oraby AH, Morsi MA. Effect of inorganic filler in the structural and optical properties of polyether sulfone. Res J Pharm Biol Chem Sci. 2012;3(4):277–93.
44. Mihály J, Sterkel S, Ortner H, Kocsis L, Hajba L, Furdyga E, et al. FTIR and FT-Raman spectroscopic study on polymer based high pressure digestion vessels. Croat Chem Acta. 2006;79(3):79.
45. Han C, Chen M, Rasch R, Yu Y, Zhao B. Structure studies of silicate glasses by raman spectroscopy. In: Reddy RG, Chaubal P, Pistorius PC, Pal U. (eds) Advances in molten slags, fluxes, and salts: Proceedings of the 10th International Conference on Molten Slags, Fluxes and Salts. 2016;p. 175-182 . https://doi.org/10.1007/978-3-319-48769-4_18
46. Gómez de la Cuesta R, Goodacre R, Ashton L. Monitoring antibody aggregation in early drug development using Raman spectroscopy and perturbation-correlation moving windows. Anal Chem. 2014;86(22):11133–40. https://doi.org/10.1021/ac5038329.