Raman spectra-based deep learning – A tool to identify microbial contamination in the pharmaceutical industry

Murali Maruthamuthu[1], Amir Raffiee[1], Denilson de Oliveira[1], Arezoo Ardekani[1], and Mohit Verma[1]

[1]Purdue University

June 22, 2020

Abstract

Deep learning has the potential to revolutionize process analytical technology in the pharmaceutical industry. Here, we used Raman spectroscopy-based deep learning strategies to develop a tool for detecting microbial contamination. We built a Raman dataset for microorganisms that are common contaminants in the pharmaceutical industry for Chinese Hamster Ovary (CHO) cells, which are often used in the production of biologics. Using a convolution neural network (CNN), we classified the different samples comprising individual microbes and microbes mixed with CHO cells with an accuracy of 95-100%. The set of 12 microbes spans across Gram-positive and Gram-negative bacteria as well as fungi. We also created an attention map for different microbes and CHO cells to highlight which segments of the Raman spectra contribute the most to help discriminate between different species. This dataset and algorithm provide a route for implementing Raman spectroscopy for detecting microbial contamination in the pharmaceutical industry.

Keywords: Raman spectroscopy, deep learning, process analytical technology, microbial contamination, convolution neural network, biologics, CHO cells

Real-time release of pharmaceuticals (small molecules and biologics) requires the ability to use in-process data to evaluate and ensure the quality of the final product [1]. Within biologics, determining sterility and measuring microbial contamination is especially important [2]. Traditional United States Pharmacopeia microbial testing methods depend primarily on culturing of microorganisms to determine bioburden and sterility [1, 3]. Since culturing and culture-dependent methods are slow (1-21 days), they cannot be used for real-time release testing. Nucleic acid-based technologies (polymerase chain reaction, next generation sequencing) have reduced the time for analysis to the order of hours but they still require sample preparation and thus, remain invasive methods of detection. Spectroscopic methods, such as Raman spectroscopy, on the other hand are non-invasive, rapid (minutes), and versatile (can detect a variety of microorganisms).

Although incidents of microbial contamination are rare, they can be extremely costly. For example, bioreactors can be operated at scales of about 15,000 L scale with media costs of $65/L and thus, a single contamination could lead to a loss of around $975,000 [1]. Thus, detecting contamination in a timely manner and monitoring critical control points is essential for real-time release. Recently, a proof-of-concept rapid microbiological screening system was able to detect *Escherichia coli* spiked into Chinese Hamster Ovary (CHO) cell line culture within three hours by using filtration (to separate CHO cells), microfluidics (to generate nanoliter-sized droplets), and an indicator dye (to measure the doubling time of bacteria) [4]. Since the method requires filtration and growth of bacteria, it is still limited to at-line or off-line use.

Raman spectroscopy measures the inelastic scattering of light due to molecular vibrations. It is possible to distinguish phenotypes of microorganisms based on their molecular composition [5]. Since the differences in the Raman spectra of different microbes can be subtle, the use of deep learning algorithms is essential to separate signal from noise. A recent demonstration of this approach on human pathogens achieved an accuracy of about 82% for distinguishing isolates of microbes [5].

In the current work, we apply Raman spectroscopy and deep learning to pharmaceutical contaminants and demonstrate detection and discrimination of 12 different microorganisms (encompassing Gram-positive bacteria, Gram-negative bacteria, and fungi listed in Table 1). We have used a Teflon$^{\mathrm{TM}}$-coated polished stainless steel substrate (Figure 1) to obtain high signal-to-noise ratios. We also demonstrate discrimination of bacterial contamination in a mixture with CHO cells. We achieve accuracies in the range of 95-100% for determining microbial identity (Figure 2).

Neural network-based microbial contamination classification: We used the convolution neural network (CNN) as a deep learning strategy to classify the microbial contaminants (and CHO cells) relevant to the pharmaceutical industry. The CNN consists of multiple hidden convolutional layers. In each layer, a certain number of filters convolve over the input map and abstract it into the feature map, which is passed to the next layer. Each layer extracts a pattern (which is determined during the optimization process) in the input data and passes the resulting feature maps to the next layer to search for higher-level patterns. The final output is passed into a fully-connected layer that converts the extracted feature maps into the probability distribution over various classes [18]. In our study, the input layer is Raman spectroscopy data obtained from different samples, and the output is the probability distribution over the 16 classes of samples (12 microbes, 1 CHO cell, 3 mixtures of CHO cells and microbes to represent Gram-positive bacteria, Gram-negative bacteria, and fungi). To evaluate the multi-class classification model, we use a confusion matrix shown in Figure 2. In this matrix, the vertical axis denotes the actual classes of samples, and the horizontal axis represents the predicted classes. In this study, we classified the samples into 16 categories. Using the confusion matrix, we can evaluate the performance of the model on every single class and learn about the type of microbe where the model has the weakest capability in recognition. In our study, the model has the lowest accuracy for*Staphylococcus epidermis* that is misclassified as*Propionibacterium acnes* in 5% of the cases. On the other hand, the model has very high accuracy in detecting the difference between microbes and microbes mixed with CHO cells. According to the confusion matrix, the average accuracy of the model is 98.19 ± 0.55% (the standard deviation is calculated over the 5 splits of training and validation sets in the LOOCV approach).

Attention map for classification: To explain the internal functionalities of proposed CNN, we use the recently developed Grad-Cam++ method [17]. This method uses a linear combination of positive partial derivatives of class scores with respect to last convolutional layers features as weights to provide the attention map of particular class labels. The resulting attention map helps us understand the regions that are important for CNN to predict the class of input data. In this case, we can identify the range of wavenumbers in spectral data of species that are significant in categorizing them, as shown in Figure 3. According to the attention maps for various species, we notice that any patterns after the largest peak in spectral data (2850-3050 cm^{-1}) do not have any significance for the model, and CNN focuses on a range of wavenumbers before the largest peak, which is around (400 - 3050 cm^{-1}) in our study. The results imply that the wavenumbers in the aforementioned wavenumbers are important in identifying the species.

Important features of Raman spectra for distinguishing microbial contaminants: The Raman spectra were collected in the wide range of 100 - 6000 cm^{-1} to avoid missing any minute variations within the different microbes. We collected 10 technical replicates by measuring the same dried sample from different points on the substrate (with 200 scans per point) and three biological replicates by repeating the experiment on three different days for each species of interest. The average (bold lines) of 6000 spectra/sample class of all the microbes/cells are depicted in Figure 3 where shaded regions indicate standard deviations.

The Raman spectra of all the microbes and CHO cells have prominent peaks of nucleic acids (1575, 1481, 812, 783 cm^{-1}), proteins (1002 cm^{-1}), and lipids (1658, 1448 cm^{-1}) [19, 20]. A strong Raman shift found in all the microbes/CHO cells is around 2850-3050 cm^{-1}. This region is found to be a non-specific organic $>CH_2$ and $-CH_3$ stretching modes [21]. Though a subtle difference can be observed between the spectra visually, high-throughput analysis requires an automated tool for discrimination [22]. Thus, CNN helped to classify the microbes and the CHO cells and to highlight which parts of the spectra had the most impact on discrimination between classes.

Although Raman spectroscopy typically suffers from low signal-to-noise ratios, here, the use of a polished stainless-steel substrate (Figure 4) has enabled concentration of the bacteria and reduction of background noise. The same substrate has been used in the past for detecting proteins at levels as low as 1 fmol [7].

The use of Raman spectroscopy has the following four advantages over other rapid microbial testing methods in the pharmaceutical industry: i) it can distinguish between several different types of microbes (spanning over Gram-positive bacteria, Gram-negative bacteria, and fungi), ii) it can provide a signal even in the presence of CHO cells and thus, does not require physical separation or filtration of the cell types before detection, iii) when a small number of scans are used, it is non-destructive and thus, the samples could be used for culturing or sequencing if needed for tracing the contaminant, and iv) collecting spectra requires less than a minute and thus, the technique can be used at-line in the production plant.

The use of convolution neural network and attention mapping enables the following three advances: i) high-accuracy classification despite only subtle differences between different classes, ii) when a training set has been incorporated, classification is rapid (in seconds), and iii) highlighting which parts of the spectra are relevant to classification helps understand the reasoning behind the classification (instead of using a completely black box approach).

The key limitations of the current study are: i) we used a high concentration of cells (10^5 cells/mL) to show proof-of-concept, ii) we dried the cells down before detection, and iii) we fixed the cells using glutaraldehyde before detection (due to concerns of biosafety).

In future studies, we aim to improve the sensitivity of Raman spectroscopy by using microfluidics and acoustic concentration. We also aim to detect the cells directly in a liquid sample to simplify the process. Our work serves as stepping stones for developing sensors for PAT and enabling real-time release of biologics.

Materials and Methods

Microorganisms and growth conditions: The list of microbes contaminating the pharmaceutical industry was identified from the FDA's manual of pharmaceutical microbiology and we also included few environmental microbial source found in pharmaceutical industry [6–10] The list of microbes/cells used in the study and media used for culturing these strains are listed in Table 1.

Raman substrate fabrication and sample preparation: The substrates (21 mm x 21 mm) were made from polished stainless steel with alumina and were coated with a thin layer (50 nm) of Teflon using spin coater as described previously [11] (Figure 1). The surface characteristics of the substrates were performed with the Hitachi S-4800 field emission scanning electron microscope (SEM). The microbes were cultured overnight to obtain 10^8 cells/mL (as measured by optical density at 600 nm of 0.1 for bacteria, and 0.6 for fungi. The overnight grown cultures were fixed with 2.5 % of glutaraldehyde and washed with water to remove the debris and diluted to a concentration of 10^5 cells/mL for Raman dataset development. The CHO cells were cultured up to 80% confluent inT75 cell culture flask and the cells were trypsinized and processed for Raman spectroscopy as mentioned by Rangan *et.al* [12]. 10^5cells/mL of CHO cells were measured using the Invitrogen Countess Automated Cell Counter. The prepared cells/microbes were placed in the substrate using a micropipette (5 µl) on the substrate and dried for 5 mins. Once dry, the sample forms a circular spot on the substrate with a diameter of about 2 mm. The dried cells on the substrate are used to collect the Raman spectra for individual species of microbes/cells. Raman measurements were performed with a customized, micro-Raman system with an argon-ion laser (532nm, 20 mW power at the sample) with thermoelectrically cooled charge-coupled device detector (1,340 pixels x 4000 pixels) mounted on a 300-mm focal length imaging with a working distance of 20 mm as described previously [13]. The spectra were collected on three different days (biological replicates) and 10 different points (technical replicates) on the 2 mm spot. At each point, 200 scans were obtained; a total of 2000 scans were obtained for each microbe/cell every day (10 points x 200 scans/point = 2000 scans). These 6000 spectra were used for the deep learning-based analyses. The Raman spectra signal to noise ratio was 1,000:1 and there is almost no interference of the background (Figure S1).

Deep learning-based classification between the potential microbial contaminants: The architecture for deep learning is composed of the following three layers: 1) initial convolution layer, 2) eight residual blocks, and 3) fully connected layer [14]. The convolution layer is composed of a kernel size of 7 and stride of 2. All the residual blocks consist of kernels with a size of 3 and strides of 1 and 2 [14]. The convolution layer proceeds with the batch normalization layer [15], and ReLU (Rectified Linear Unit) is used as a non-linear function. The residual blocks contain a shortcut connection between input and output, which enhances the training stability and addresses the problem of degradation in the deep neural network [14].

The output of the model is a 1-d (R^d,$R \in [0, 1]$) vector containing the probability distribution over all the classes of bacteria. To train the model, we used Adam optimizer with betas = (0.9, 0.999), and the learning rate is set to 0.001. The factor of 0.1 decays the learning rate if the accuracy on the validation set reaches a plateau during training [16]. In order to train the model, we use 5-fold Leave-One-Out Cross-Validation (LOOCV) method to split the collected data set into training and validation sets. In this method, the reference data set is randomly split into five groups, and in each round of training, one group is held out to be used as the validation set and the remaining data is used as the training set. This process is repeated five times to ensure that all the samples fall into the validation set once. The performance of the model was evaluated on the individual class scale to form a confusion matrix. Furthermore, using Grad-Cam++, we developed a saliency map for each sample that shows the attention map of each microbe/cell with the Raman spectra [17]. With this feature, we can explain how the deep learning model chooses a class for an arbitrary input by providing the corresponding attention map.

Acknowledgment: This work was partially funded by the grant “Continuous Manufacturing of Biologics,” funded by Purdue College of Engineering’s Faculty Conversations (EFC) and by the National Science Foundation (CBET 1700961).

References:

1. Shintani, H (2016). Validation Study of Rapid Assays of Bioburden, Endotoxins and Other Contamination. *Biocontrol Sciences* , 21, 63–72 . https://doi.org/10.4265/bio.21.63

2. Jiang, M., Severson, KA., Love, JC., Madden, H., Swann, P., Zang, L., Braatz, RD. (2017). Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing. *Biotechnology Bioengineering* , 114, 2445–2456 . https://doi.org/10.1002/bit.26383

3. England, MR., Stock, F., Gebo, JET., Frank, KM., Lau, AF. (2019). Comprehensive Evaluation of Compendial USP<71>, BacT/Alert Dual-T, and Bactec FX for Detection of Product Sterility Testing Contaminants. *Journal of Clinical Microbiology* , 57, 1548-18. https://doi.org/10.1128/JCM.01548-18

4. Surrette, C., Scherer, B., Corwin, A., Grossmann, G., Kaushik, AM., Hsieh, K., Zhang, P., Liao, JC., Wong, PK., Wang, TH., Puleo, CM. (2018). Rapid Microbiology Screening in Pharmaceutical Workflows.*SLAS Technol* , 23. 387–394 . https://doi.org/10.1177/2472630318779758

5. Ho, C-S., Jean, N., Hogan, CA., Blackmon, L., Jeffrey, SS., Holodniy, M., Banaei, N., Saleh, AAE., Ermon, S., Dionne, J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature Communications* , 10. 4927. https://doi.org/10.1038/s41467-019-12898-9

6. Pharmaceutical Microbiology Manual | FDA. https://www.fda.gov/media/88801. Accessed 2 Jan 2020

7. Deal, A., Klein, D., Lopolito, P., Schwarz, JS. (2016). Cleaning and Disinfection of Bacillus cereus Biofilm. PDA *Journal of Pharmaceutical Science and Technology* , 70. 208–217. https://doi.org/10.5731/pdajpst.2014.005165

8. Pacheco, FLC., Pinto, TDJA. (2010). The bacterial diversity of pharmaceutical clean rooms analyzed by the Fatty Acid methyl ester technique. PDA *Journal of Pharmaceutical Science and Technology* , 64.156–166 https://pubmed.ncbi.nlm.nih.gov/21502015/

9. Salaman-Byron, AL. (2019). Probable Scenarios of Process Contamination with Cutibacterium (Propionibacterium) acnes in Mammalian Cell Bioreactor. PDA *Journal of Pharmaceutical Science and Technology* . pdajpst.2019.010710. https://doi.org/10.5731/pdajpst.2019.010710

10. Cobo, F., Concha, Á. (2007). Environmental microbial contamination in a stem cell bank. Letters in *Applied Microbiology* , 44, 379–386 . https://doi.org/10.1111/j.1472-765X.2006.02095.x

11. Zhang, D., Xie, Y., Mrozek, MF., Ortiz, C., Davisson, VJ., Ben-Amotz, D. (2003) Raman Detection of Proteomic Analytes.*Analytical Chemistry* , 75, 5703–5709 . https://doi.org/10.1021/ac0345087

12. Rangan, S., Kamal, S., Konorov, SO., Schulze, HG., Blades, MW., Turner, RFB., Piret, JM. (2018). Types of cell death and apoptotic stages in Chinese Hamster Ovary cells distinguished by Raman spectroscopy. *Biotechnology Bioengineering* , 115:401–412 . https://doi.org/10.1002/bit.26476

13. Davis, JG., Gierszal, KP., Wang, P., Ben-Amotz D. (2012) Water structural transformation at molecular hydrophobic interfaces.*Nature* 491:582–585 . https://doi.org/10.1038/nature11570

14. He, Kaiming., Zhang, Xiangyu., Ren, Shaoqing., Sun, Jian. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). pp 770-778. https://doi.org/10.1109/CVPR.2016.90

15. Ioffe, S., (2017). Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. Conference on *Advances in neural information processing systems* , 1945-1953, http://papers.nips.cc/paper/6790-batch-renormalization-towards-reducing-minibatch-dependence-in-batch-normalized-models.pdf.

16. Kingma, DP., Ba, J. (2015) Adam: A Method for Stochastic Optimization.3rd *International conference on learning representations* , ICLR, Conference track proceedings. https://arxiv.org/pdf/1412.6980.pdf

17. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks - IEEE Conference Publication. https://ieeexplore.ieee.org/document/8354201. Accessed 3 Jan 2020

18. Krizhevsky, A., Sutskever, I., Hinton, GE. (2012). ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Curran Associates Inc., Lake Tahoe, Nevada, pp 1097–1105

19. Teng, L., Wang, X., Wang, X., Gou, H., Ren, L., Wang, T., Wang, Y., Ji, Y., Huang, WE., Xu, J. (2016). Label-free, rapid and quantitative phenotyping of stress response in E. coli via ramanome. *Scientific Reports* , 6. https://doi.org/10.1038/srep34359

20. Ren, Y., Ji, Y., Teng, L., Zhang, H. (2017). Using Raman spectroscopy and chemometrics to identify the growth phase of Lactobacillus casei Zhang during batch culture at the single-cell level.*Microbial Cell Factories* , 16. https://doi.org/10.1186/s12934-017-0849-8

21. Naja, G., Bouvrette, P., Hrapovic, S., Luong, JHT. (2007). Raman-based detection of bacteria using silver nanoparticles conjugated with antibodies. *Analyst* 132, 679–686. https://doi.org/10.1039/B701160A

22. Razek, SA., Ayoub, AB., Swillam, MA., (2019) One Step Fabrication of Highly Absorptive and Surface Enhanced Raman Scattering (SERS) Silver Nano-trees on Silicon Substrate. *Scientific Reports* 9, 1–8 . https://doi.org/10.1038/s41598-019-49896-2

Table 1: List of microbes/cells used in this study.

No	Name	Source	Growth media	Condition
1.	*Aspergillus brasiliensis*	ATCC 16404	Potato dextrose broth	Aerobic, 25 °C
2.	*Bacillus cereus*	ATCC 10876	Nutrient broth	Aerobic, 30 °C
3.	*Bacillus subtilis*	ATCC 6633	Brain heart infusion broth	Aerobic, 37 °C
4.	*Candida albicans*	ATCC 10231	Yeast extract peptone dextrose (YPD media)	Aerobic, 25 °C

No	Name	Source	Growth media	Condition
5.	*Clostridium sporogenes*	ATCC 19404	Trypticase Soy Broth with defibrinated sheep blood	Anaerobic, 37 °C
6.	*Escherichia coli*	ATCC 8739	Nutrient broth	Aerobic, 37 °C
7.	*Micrococcus luteus*	ATCC 10240	Trypticase Soy Broth	Aerobic, 30 °C
8.	*Propionibacterium acnes*	ATCC 29399	Tryptone Yeast glucose media (TYG)	Anaerobic, 37 °C
9.	*Pseudomonas aeruginosa*	ATCC 9027	Trypticase Soy Broth	Aerobic, 37 °C
10.	*Salmonella enterica*	ATCC 14028	Trypticase Soy Broth	Aerobic, 37 °C
11.	*Staphylococcus aureus*	ATCC 6538	Trypticase Soy Broth	Aerobic, 37 °C
12.	*Staphylococcus epidermis*	ATCC 35984	Trypticase Soy Broth	Aerobic, 37 °C
13.	CHO cells	ATCC CCL-61	F-12K medium with 10% Fetal bovine serum (FBS)	Aerobic, 37 °C

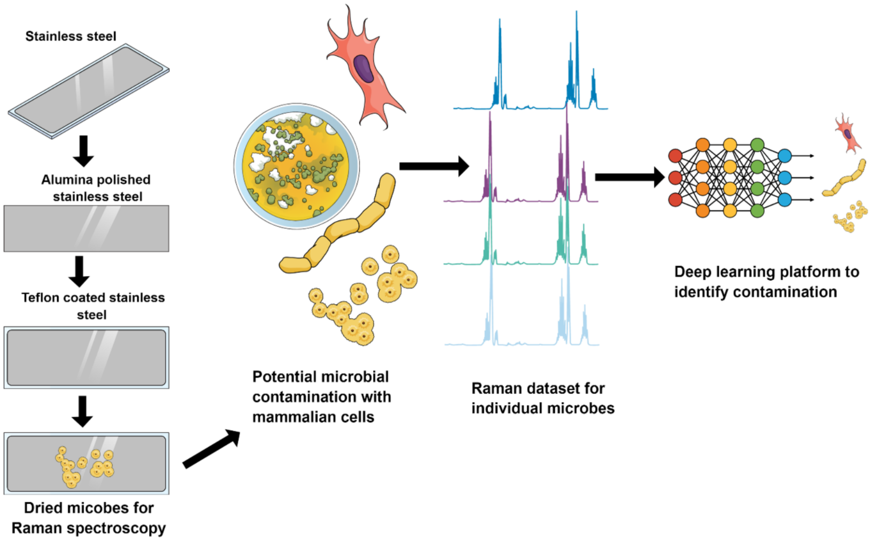

Stainless steel
Alumina polished stainless steel
Teflon coated stainless steel
Dried micobes for Raman spectroscopy
Potential microbial contamination with mammalian cells
Raman dataset for individual microbes
Deep learning platform to identify contamination

Figure 1: Schematic of workflow to identify contamination using deep learning strategy.

Figure 2: Confusion matrix from the developed neural network for classification of microbes using the Raman dataset.

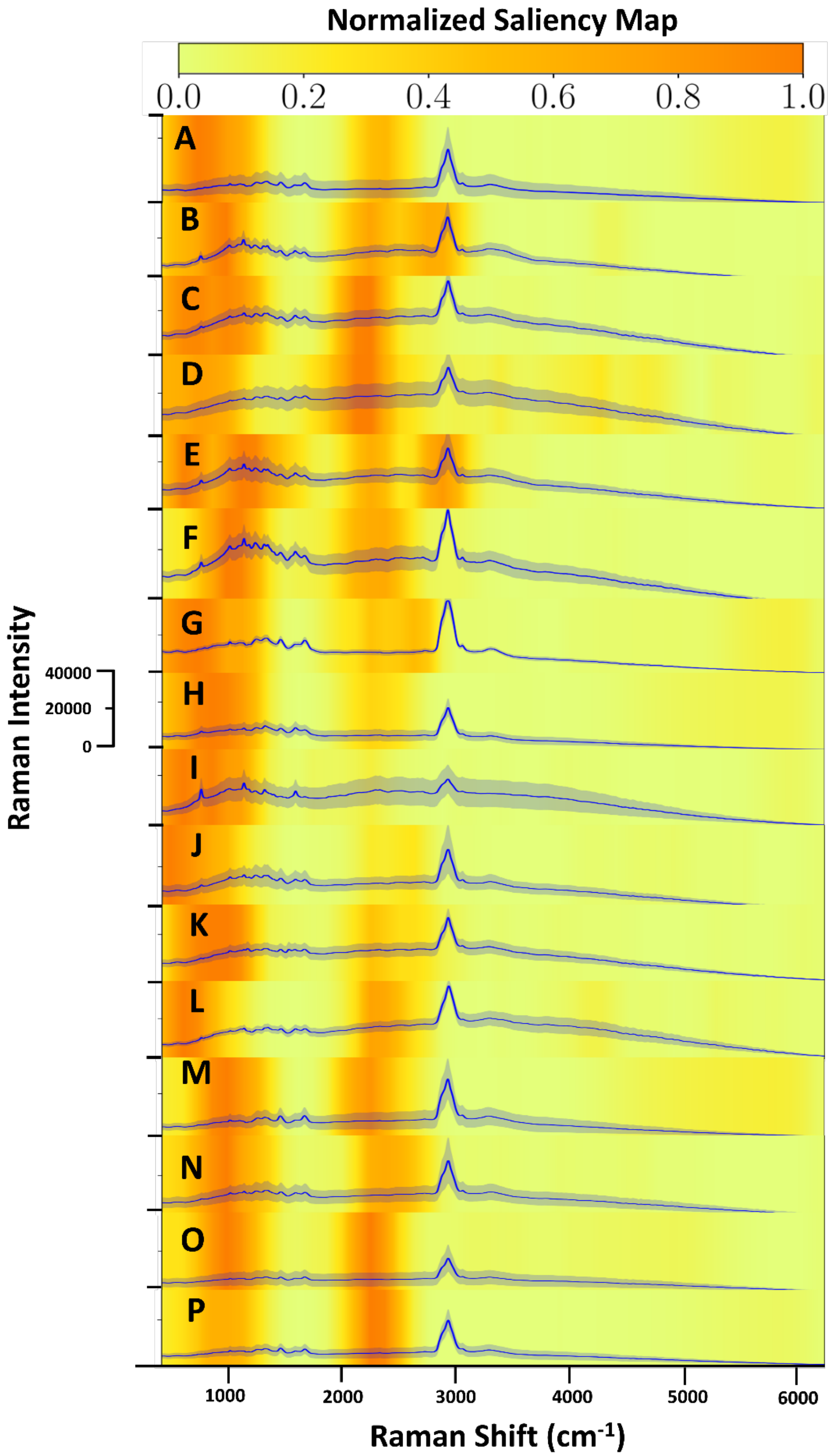
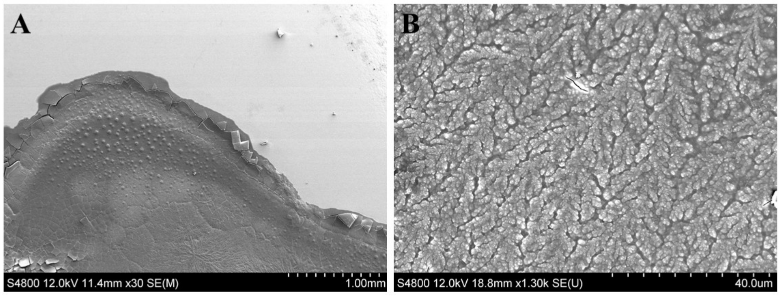


Figure 3*:* Attention map and Raman spectra for classification of microbes, CHO cells, CHO cells with gram-negative, CHO cells with gram-positive and CHO cells with Fungus. The bold blue line indicates

average spectra (6000 scans) and the shaded area around the bold blue line indicates standard deviation. The heatmap (yellow-orange) indicates the importance of the different segments of the spectra according to attention map. A. *Aspergillus brasiliensis* , B. *Bacillus cereus* , C. *Bacillus subtilis* , D. *Candida Albicans* , E. *Clostridium sporogenes* , F.*Escherichia coli* , G. *Micrococcus luteus* , H.*Propionibacterium acnes,* I. *Pseudomonas aeruginosa* , J.*Salmonella enterica* , K. *Staphylococcus aureus* , L.*Staphylococcus epidermis* , M. CHO cells. N. CHO cells and*Aspergillus brasiliensis* , O. CHO cells and *Bacillus cereus* , P. CHO cells and *Staphylococcus aureus* .


A
S4800 12.0kV 11.4mm x30 SE(M)
1.00mm
B
S4800 12.0kV 18.8mm x1.30k SE(U)
40.0um