Peer Review File

Manuscript Title: Digital Colloids-Enhanced Raman Spectroscopy by Single Molecule Counting

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

The manuscript describes the use of "digital SERS" which addresses the problem of detecting low concentrations of target molecules by sampling a large number (typically > 1000) independent voxels in a liquid sample. The data are then analysed to find voxels which show a signal that is larger than the background i.e. the ratio of positive voxels (RPV). The proportion of positive pixels is then used to infer the concentration using calibration plots of RPV versus sample concentration. This is intended to reduce the effect of the fluctuations in SERS intensity that occur when single molecule measurements are being recorded. The manuscript shows data for crystal violet down to 10—15 mol dm-3, along with results for a DNA analogue, an aromatic thiol, glucose and haemoglobin. In addition to these test analytes, the manuscript also show data illustrating potential applications of the approach for detection thiram and paraquat.

I do have some concerns about the originality of the approach. Brolo et al. published a paper in 2018 which also centred on digital SERS and used a similar approach. To be fair to the authors of the current manuscript they did include that paper as ref 19 but the main point is that Brolo's paper does anticipate most of the concepts in the new manuscript. It shows the use of digital SERS for detecting low concentrations of analytes and explicitly describes the concept of working in the single molecule regime, imaging an area and counting pixels which contain the appropriate spectral features as way of quantifying the concentration of the sample. The main difference from the current study is that Brolo used a solid substrate rather than a colloid but apart from that the concept is extremely similar. The submitted manuscript is more comprehensive but I can't see how it significantly advances the concepts beyond what was published in 2018.

If that reservation is set to one side, there are some puzzling aspects to the data in the manuscript. In particular, for the calibrations shown in Figure 1 I would like to understand how the authors account for the observation that the RPV is is not simply proportional to the analyte concentration, at least in the low count region. I would have expected that increasing concentration by factor of 10 would increase number of positive voxels directly in proportion to the concentration of analyte i.e. 10x but in the data in Figure 1 it is clear that changing the concentration gives much smaller increases in RPV. In the cases of D-glucose and haemoglobin for example, increases in concentration by 3 orders of magnitude only causes the proportion of positive events to increase by less than 10x. This is the most extreme example but all the other analytes show similar effects. Of course, this relatively gradual increase allows a concentration range of several orders of magnitude to be analysed without the system saturating, even though the signals are only normally shown in the paper as ranging from CPV 1- 40 %. However, it does seem to be a real anomaly which does cast doubt on the simple mechanism proposed in the manuscript. It is also useful to note that this effect also increases the uncertainty in the concentration derived from the calibration plots since an uncertainty in the RPV of 20% will correspond to a much larger uncertainty in the concentration value.

With regard to the improvement which the dSERS approach gives over simply summing all the pixels, the manuscript shows significantly lower detection limits for CV, although I would argue that the error bars on the signals shown in Figure 2a do not really represent the uncertainty which is relevant to the measurements. It is clear in that figure that the signal recorded even at 10 10 mol dm-3 shows relatively modest differences even between batches while the errors bars are huge. If the error in the total intensity, which is the one which is relevant here, was this large then the batch-to-batch variation would be also be very much larger than the experiments show. In fact, the data appear to show that simply summing the spectra would give a rational linear calibration down to 10-10 mol dm-3 and possibly even lower, although there is no data shown for 10-11 mol dm-3. That is not to say that the digital method is not superior, it is just that it is important not to artificially increase the apparent improvement by downplaying the performance of the standard approach. A similar effect is also apparent in extended data Figure 6a where the plot of the Raman signal falls systematically with the reduction in concentration while the error bars are dramatically larger than the observed changes in signal.

It might also be useful to compare the digital approach with that advanced by Bin Ren et al (Single-Molecule Level Rare Events Revealed by Dynamic Surface-Enhanced Raman Spectroscopy, Anal. Chem. 2020, 92, 24, 15806–15810) https://doi.org/10.1021/acs.analchem.0c02936) who carried out SERS measurements in the single molecule regime and improved the sensitivity by identifying the spectra from an extended time series that contained signals and then summing only including those in the summation. This approach could be applied to the data shown in the current manuscript.

Finally, I would note that the comparison between the dSERS and analogue SERS carried out in the same experiment is not really a fair one. The conditions in the manuscript are set to optimise the digital SERS method and then this compared with analogue data obtained under these conditions. It would be fairer to compare the results with other studies where analogue SERS has been optimised for the analytes has been used. The data for CV is impressive but that for the other molecules less so. In particular, the two examples included to illustrate applications of dSERS were paraquat and thiram. Paraquat was detected down to 10-10 mol dm-3, which is 3 orders of magnitude better than the maximum residual level. However, it is not dramatically better than other much simpler SERS approaches such as Choo et al's method using standard microfluidics which had a detection limit of 2 x 10-9 mol dm-3 (Highly sensitive trace analysis of paraquat using a surface-enhanced Raman scattering microdroplet sensor, Analytica Chimica Acta 2010, 681(1-2):87-91 DOI: 10.1016/j.aca.2010.09.036).

Similarly, SERS studies of thiram based on paper substrates gave a detection limit of 0.5 X 10-9 mol dm-3, which is better than the 0.001 mg/kg (4x 10-9 mol dm-3) in the current paper (Subnanomolar Sensitivity of Filter Paper-Based SERS Sensor for Pesticide Detection by Hydrophobicity Change of Paper Surface ACS Sens. 2018, 3, 1, 151–159 https://doi.org/10.1021/acssensors.7b00782). Indeed, a recent paper has claimed detection limits of 10-14 mol dm-3 and also lists several other SERS studies with sub-nanomolar detection limits (Femtomolar detection of thiram via SERS using silver nanocubes as an efficient substrate, Environ. Sci.: Nano, 2020,7, 3999-4009, https://doi.org/10.1039/D0EN01049A).

Overall, I am not confident that this paper would have a significant impact in the field partly due to the concept being discussed and demonstrated previously and partly because the potential improvements in detection limits are offset by the limitations of the conditions in which the measurements need to be conducted. This means that the detection limits are not strikingly superior

to results obtained using existing SERS approaches.

Referee #2 (Remarks to the Author):

The concept described within the manuscript 'Digital Colloids-Enhanced Raman Spectroscopy by Single-Molecule Counting' is very interesting. Through the digitalization of collected SERS spectra into either detect (1) or non-detect (0) the project team is able to achieve detection limits as low as 1 fM for a range of analytes. While the approach is novel and the results potentially exciting, the manuscript raises a number of concerns and questions that require response if the approach is to achieve its potential. I have listed below both general concerns as well as specific questions that should be answered.

General Concerns

1) How do differences in analyte affinity for the nanoparticle surface impact the dCERS measurements? The work described in the manuscript is focused primarily on analytes that readily associate with noble metal surfaces either via covalent linkages (4-nitrobenzenethiol, thiram), electrostatic interactions between negatively charged colloids and positively charged nitrogen groups (paraquat, crystal violet, adenine within A12) or via hydrophobic interactions (hemoglobin). The only analyte that does not bind strongly with the noble metal surface (D-glucose) exhibited the worst dCERS performance (LOD ~ 10-9 M vs. LOD <10-11 M for the other analytes). The extent that an analyte interacts with the surface and with SERS hotspots should dictate dCERS performance, however, this point is never discussed nor considered. [Interestingly, L-cysteine is included as an analyte in the methods and data on its dCERS detection is found within the extended data, however, it is not mentioned as an analyte within the main body of the manuscript.] Do the observed differences in the slopes of the RPV vs. concentration plots (Figure 1; Extended Data Figs. 6,7) reflect differences in surface affinity or do they reflect something else? It would seem possible to utilize the adapted Poisson formula defined within the Supplementary Information to somewhat quantify this relationship.

2) The project team utilized three different colloids within the work: hydroxylamine-reduced silver (Hya-Ag), citrate-reduced silver (citrate-Ag), and citrate-reduced gold (citrate-Au). They note that the best performance was achieved using Hya-Ag, but do not fully explain how such a conclusion was reached. While there is some data presented on the optical properties and single-molecule sensitivity in Extended Data Figure 1 (and explained in the Supplementary Information) it is not clear why Hya-Ag is preferred. Does Hya-Ag exhibit a different surface chemistry than citrate-Ag which makes it a better colloid for dCERS? What was the surface charge (i.e, zeta potential) for Hya-Ag vs. citrate-Ag vs. citrate-Au under the experimental conditions? Does a difference in surface charge differentially impact the surface affinity of the analytes and thus the dCERS sensitivity?

3) The colloids and solutions used in the experiments are woefully under-characterized. While single TEM images of the three colloids are presented within Extended Data Figure 1 there is minimal to no information provided on TEM determined colloid diameter or surface charge. Looking at the TEM images in Extended Data Figure 1 it would seem that the citrate-Ag particles are much larger than the Hya-Ag or citrate-Au particles. The impacts of this size difference on the dCERS results are not explained. Furthermore, only one type of particle was characterized by DLS (citrate-Ag) – why were the others not subjected to this simple analysis? [On a related note – why is there a disconnect

between the ~100 nm diameter for Ag-citrate as determined by TEM and the DLS diameter of ~40 nm? Typically the TEM diameter is less than that determined by DLS due to particle aggregation. Unfortunately, the methodology and instrumentation used for the DLS measurements is not described.] Simply stating that 'lake water was fetched from a city park' is insufficient. What was the background electrolyte concentration? What macromolecules were present?

4) The impacts of nanoparticle aggregation on the dCERS results are not addressed. For SERS hotspots to form it is necessary for the nanoparticles to aggregate to some extent. Such aggregation may happen due to the interaction between the analyte and the colloids, but aggregation will also be affected by the background electrolyte concentration (with higher electrolyte concentrations enhancing aggregation) as well as the presence of macromolecules (which may hinder aggregation). Under the simple solution conditions (either pure water or ethanol/water) used to establish the dCERS approach aggregation should primarily reflect analyte-colloid interactions, however, under the 'Field' deployment conditions presented in Figure 3 the extent of aggregation will presumably reflect these background constituents. Such an outcome is not considered within the manuscript. Furthermore, if aggregation is too extensive then the colloids will sediment and no longer be detectable by their approach.

5) What is the minimal number of voxels required to be sampled for the approach to work? The authors suggest it is possible to calibrate this term for each analyte, but it is not clear whether the lower limits reflect differences in surface affinity (see #1).

Specific Points (w/ page numbers containing the pertinent text indicated)

A. Pg. 3. The authors note that variable molecular orientations on a substrate impact SERS quantitation. Wouldn't this same phenomenon affect the dCERS results? A different molecular orientation could potentially result in a signal with lower intensity that falls below the threshold defined by the dCERS approach. How can dCERS account for multiple different molecular orientations on the surface?

B. Pg. 4. The authors suggest that '…Ag colloids can be distributed homogeneously throughout the entire solution'. How is such homogeneity ensured given that the particles will aggregate?

C. Pg. 7. What is meant by 'pre-calibration of the prepared colloids with each batch'?

D. Pg. 7. The authors list a number of factors that affect the practical concentration ranges over which dCERS works for different analytes. How do they disentangle the relative enhancement factor of the hotspots from the binding affinity of the target analytes? Won't those two parameters be intricately intermixed?

E. Pg. 9. If the authors are going to suggest that thiram uptake is higher than expected within bean sprouts then they should refer to the established literature on this topic. Stating that '…a much lower usage of this chemical might be considered' is an over-reach given the simplicity of their experimental system.

F. Figure 3. It is unclear in panel b why one of the X-axes increases in concentration from right to left while the other decreases. At first I thought that this result reflected competition between the analytes for hotspots (and thus the NB signal decreases with an increase in CV), but that is not correct. Finally, what is the blue arrow in panel d meant to indicate?

Author Rebuttals to Initial Comments:

Point to Point Response to Reviewers' Comments and Summary of the Revision

Comments from Reviewer 1:

Comment 1.1: I do have some concerns about the originality of the approach. Brolo et al. published a paper in 2018 which also centred on digital SERS and used a similar approach. To be fair to the authors of the current manuscript they did include that paper as ref 19 but the main point is that Brolo's paper does anticipate most of the concepts in the new manuscript. It shows the use of digital SERS for detecting low concentrations of analytes and explicitly describes the concept of working in the single molecule regime, imaging an area and counting pixels which contain the appropriate spectral features as way of quantifying the concentration of the sample. The main difference from the current study is that Brolo used a solid substrate rather than a colloid but apart from that the concept is extremely similar. The submitted manuscript is more comprehensive but I can't see how it significantly advances the concepts beyond what was published in 2018.

Response:

We fully understand the point made by the reviewer. There is no argument that digital concept was first proposed by Brolo et al. and we have clearly and unambiguously acknowledged this fact. However, the digital concept of Brolo et al. has not been reduced to practice since its inception in 2018. The main reason is that solid based systems have so far failed to demonstrate quantitative reproducibility and are refractory to calibrations owing to the uncontrollable heterogeneity both within and across individual solid chips caused by the highly variable distributions of hotspots and ill controlled molecular depositions. In other words, side by side comparisons between solid chips are not numerically consistent regardless how hard one tries. This fact also makes the determination of measurement error dubious at best. Perhaps, it is not unreasonable to suggest that to reduce a concept into applications often requires additional breakthroughs that can be equally challenging and often unappreciated. Being able to detect single molecule signals may be rather common nowadays, but it does not automatically translate into quantitative measurements. Such difficulties have also been recognized by several other investigators (Chem. Soc. Rev., 2008,37, 1052-1060; J. Am. Chem. Soc. 2015, 137, 42, 13698–13705; ACS Nano 2020, 14, 28−117). In fact, statistically defined quantitative measurements at high sensitivity is what required but not demonstrated until our work.

Here, our work shows that with a colloid-based single-molecule detection approach, the above obstacles inherent to solid chips can be overcome with confidence. Because hotspots and

target analytes are randomly and statistically uniform-distributed in the colloidal system, switching to a solution-based system affords two important advances: (1) calibration of any colloidal preparation can be pre-performed, which establishes a quantitative correspondence between analyte concentration and detected single-molecule events under pre-set conditions; (2) as we have shown, with the solution colloidal system, single-molecule events do follow the well-known Poisson distribution, therefore, desired reproducibility and measurement accuracy can be achieved at almost any concentration which is largely determined by the accumulated single-molecule counts, i.e., $error(\%) = \frac{1}{\sqrt{N}} \times 100\%$ (*N* is the number of detected single-molecule events). **Figure R1** (also **Figure 2** in the revised manuscript) demonstrates the experimental verification of this Poisson relationship between the error and the positive counts. These results clearly show that dCERS quantification of target molecules at ultralow concentrations is reproducible and can meet practical demands on accuracy, which has never been accomplished in the past, to our best knowledge.

To make the above points better appreciated, we have revised the manuscript with additional data and explanations in the main text (**Line 146-171 on Page 8-9**) and in the supplementary notes (**Supplementary Note 7**) to make the advantages of dCERS clearer to the readers.

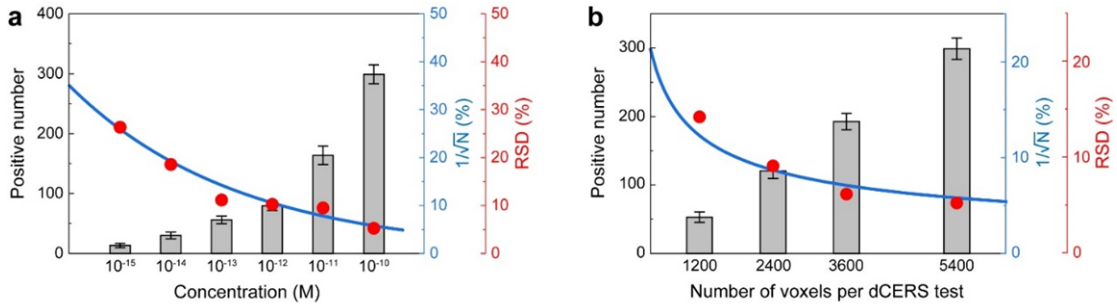

Figure R1. Reproducibility of quantitative dCERS (Figure 2 in the revised manuscript). (a) Concentration-dependent quantitative accuracy under a fixed number of acquired voxels per dCERS test (*i.e.*, 5,400 voxels) for CV using Hya-Ag colloids. (b) Total voxel number-dependent quantitative accuracy at 10^{-10} M CV concentration. The gray bars show the mean number of positive voxels with the error bars indicating the standard deviations (n=3). In both (a) and (b), the red dots show the relative standard deviations (RSD) for the measurements, which are in good agreement with the Poisson rule (blue curve).

Comment 1.2: For the calibrations shown in Figure 1, I would like to understand how the authors account for the observation that the RPV is not simply proportional to the analyte concentration, at least in the low count region. I would have expected that increasing concentration by factor of 10

would increase number of positive voxels directly in proportion to the concentration of analyte, i.e., 10x, but in the data in Figure 1, it is clear that changing the concentration gives much smaller increases in RPV. In the cases of D-glucose and hemoglobin, for example, increases in concentration by 3 orders of magnitude only causes the proportion of positive events to increase by less than 10x. This is the most extreme example, but all the other analytes show similar effects. Of course, this relatively gradual increase allows a concentration range of several orders of magnitude to be analyzed without the system saturating, even though the signals are only normally shown in the paper as ranging from CPV 1-40 %. However, it does seem to be a real anomaly which does cast doubt on the simple mechanism proposed in the manuscript. It is also useful to note that this effect also increases the uncertainty in the concentration derived from the calibration plots since an uncertainty in the RPV of 20% will correspond to a much larger uncertainty in the concentration value.

Response:

Thank you for pointing this out. Indeed, we were also expecting a linear relationship at least for very low concentrations. However, when we reanalyzed our data in detail, we realized that the digitized SERS resulted in a *linear* correspondence between the positive counts and the concentration *only* on the log-log scale (**Figure R2**). We could not find a simple mathematical description of this correspondence. So far, this relationship remains true for all the molecules we have tested with dCERS, and is likely a general characteristic for dCERS. The physical/chemical processes resulting in this highly nonlinear behavior (on linear scale) are not understood at the present, but it is consistent with the Freundlich equation (Chemosphere 2020, 258, 127279), which is a phenomenological theory developed to describe molecular binding to solid surfaces. Moreover, dELISA also showed a similar correspondence, thus, such a relationship might be a universal property of most single molecule based quantitative measurements. This point is now explicitly presented in the revised manuscript (**Line 121-145 on Page 7-8**)

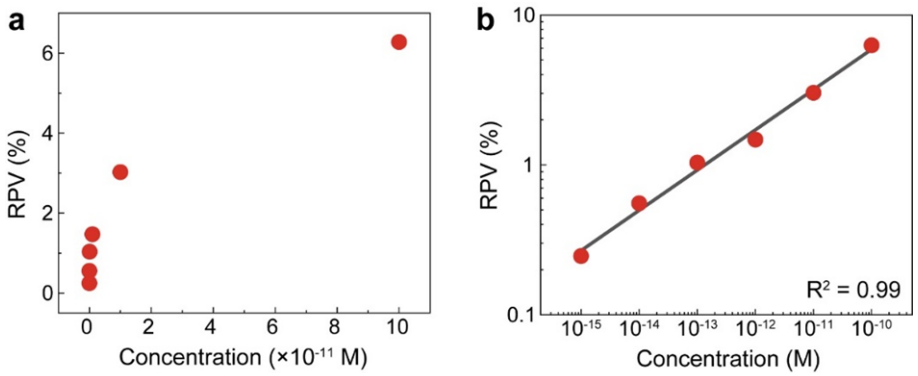

Figure R2. RPV versus analyte (CV) concentration (a) on the linear scale and (b) on the log scale using CV/Hya-Ag colloids as a case for demonstration. In (b), the data points are fitted by a straight line on the log scale ($\log RPV = k \log M + b$). It is noted that the curve shown in (a) cannot be described by a simple mathematical formula. On the linear scale, we have $RPV=B\times M^{k}$, where $\log B = b$, which is the same as the Freundlich equation. At much higher concentrations, the correspondence curve starts to deviate from the linear relationship.

In terms of the measurement error, we have shown in the revised manuscript and the paragraphs above that it is solely determined by the total positive counts, regardless at what concentration. It is true that the same absolute error would correspond to rather different errors in concentration at different concentrations owing to the nonlinear correspondence. But one can always increase the total positive count to mend this problem. In other words, if one wishes to achieve the same relative error at different concentrations, the only thing one should do is to accumulate the same number of positive events at each data point. Of course, for lower concentrations, a longer measurement time (more voxels) will have to be acquired. In fact, this is one of the "nice" properties of the digital approach. Here, the only absolute limitation (or the limit of detection, LOD) is the digitization error of the background, which sets the practical "limit" of detection. In the revision, we have reorganized our presentation to clearly demonstrate this very important point (**Line 146-171 on Page 8-9**). We thank the reviewer for helping us clarifying our thoughts on this issue.

Comment 1.3: With regard to the improvement which the dSERS approach gives over simply summing all the pixels, the manuscript shows significantly lower detection limits for CV, although I would argue that the error bars on the signals shown in Figure 2a do not really represent the uncertainty which is relevant to the measurements. It is clear in that figure that the signal recorded even at 10^{-10} mol dm^{-3} shows relatively modest differences even between batches while the errors

bars are huge. If the error in the total intensity, which is the one which is relevant here, was this large then the batch-to-batch variation would be also be very much larger than the experiments show. In fact, the data appear to show that simply summing the spectra would give a rational linear calibration down to 10^{-10} mol dm^{-3} and possibly even lower, although there is no data shown for 10^{-11} mol dm^{-3}. That is not to say that the digital method is not superior, it is just that it is important not to artificially increase the apparent improvement by downplaying the performance of the standard approach. A similar effect is also apparent in extended data Figure 6a where the plot of the Raman signal falls systematically with the reduction in concentration while the error bars are dramatically larger than the observed changes in signal.

Response:

Good point and well taken. In retrospect, we should have noted this as well. In **Figure 2a** (original submission), we followed the method deployed by A.G. Brolo, et al. (Anal. Chem. 2018, 90, 1248−1254) where the error bars were the standard deviation of the SERS signals computed from individual voxels. However, this approach may not be appropriate in the present setting and had resulted in the puzzling observation indicated by the reviewer. For the present purpose, we should have used multiple measurements of the summed spectra over all voxels for each acquisition and compute the error from these measurements (Details of this calculation is found in the revised **Supplementary Note 1**). In this way, as shown in **Figure R3a**, with the summed spectra, the spectra in the CV-specific window of 780 to 820 cm^{-1} was still observable with a weak peak at the concentration of 10^{-10} M and 10^{-12} M, but was essentially flat at concentrations lower than 10^{-13} M with no spectral features, which is indistinguishable from the null control under the same conditions. As shown in **Figure R3c**, when this analog data is plotted against CV concentration, no concentration below 10^{-12} M could be measured. In this figure, the Raman signal was obtained by subtracting the null spectra in the same window. On the contrary, when compared with the individual spectrum from each voxel, the spectral features in this range are certainly recognizable (**Figure R3b**) and the RPV decreases as the concentration with the error bar following the Poisson rule (**Figure R3d**). For the above reasons, we have now replotted the results in **Figure R3c-d**. Here, the advantage of the digital approach is easily appreciated. Furthermore, with the digital approach, the error bar increases at lower concentrations owing to the reduced number of positive voxels for the same number of total voxels acquired, as expected with Poisson statistics. Therefore, accuracy at very low concentrations can be further improved by acquiring more voxels per measurement. By the same token, much lower concentrations should still be accessible by the digital approach and should be solely limited by the acceptable measurement duration and the false positives within the

measurement time (digitization error of the background). This will not be possible with the analog method. This point has been discussed in the **Supplementary Note 1** to illustrate the merit of dCERS for quantification at low concentrations. Again, we must thank the reviewer for his insight and knowledge, and this is perhaps one of the most important corrections/clarifications that we have made in the revision.

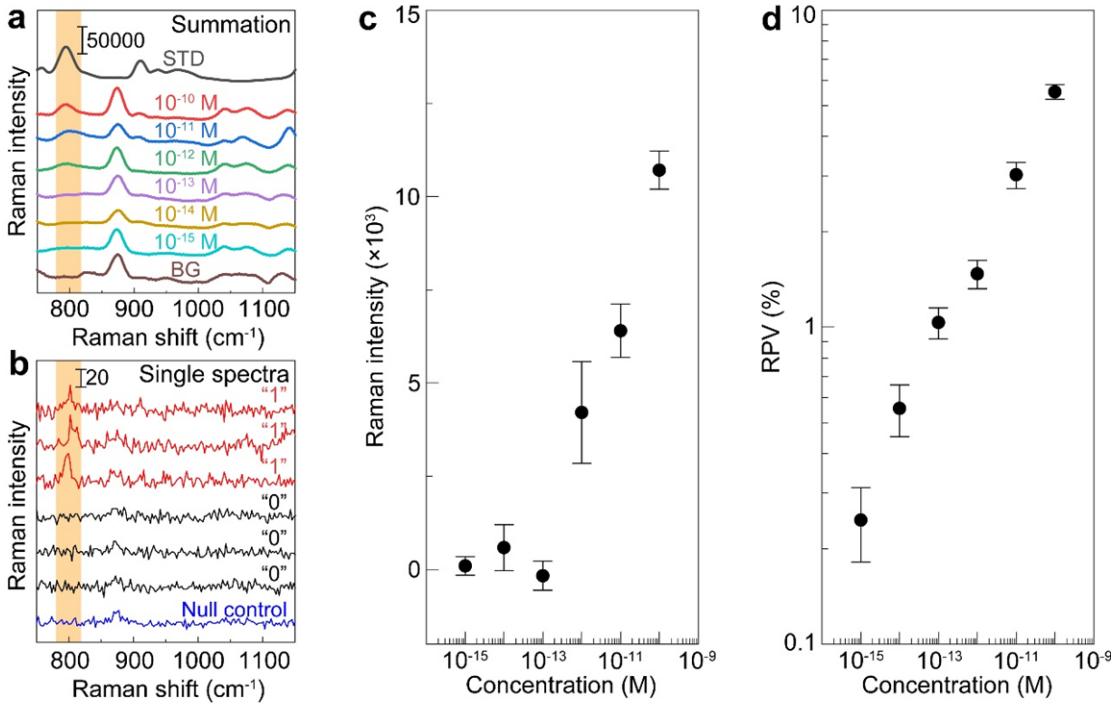

Figure R3. (a) The summation of all acquired spectra at different concentrations of CV (10^{-15} M to 10^{-10} M) and the background (BG) reference (null control). The standard SERS spectra (STD) of crystal violet (CV)-ethanol solution is provided after scaling for clarity. (b) Typical single molecule spectra acquired from the voxels at 10^{-15} M CV. Red: positive spectra with discernable CV-specific peak (assigned 1 for that voxel); black: negative spectra without the targeted signals (assigned 0); blue: spectrum acquired without CV (null control). The specific SERS band corresponding to CV (780-820 cm^{-1}) is indicated by the orange stripes in (a) and (b). (c) The peak intensity from the summed spectra over all voxels and (d) the ratio of positive voxels (RPV) versus CV concentration by dCERS. The standard deviations are calculated from 3 independent measurements (5,400 voxels/mapping) for (c) and (d).

Second, the inter-batch variations are dependent on the enhancement capability of the SERS colloids from each batch, including colloidal concentration, colloidal particle size and morphology, etc. Now using the proper method of quantification (explained above), we found that the variations between batches were in fact larger than that within each batch (**Figure R4**), consistent with the conventional wisdom. Nevertheless, each batch of colloids can be fully pre-calibrated by taking a

small fraction from the stock to establish the correspondence curve specific to each batch, therefore inter-batch differences can be fully accounted for in the practical sense.

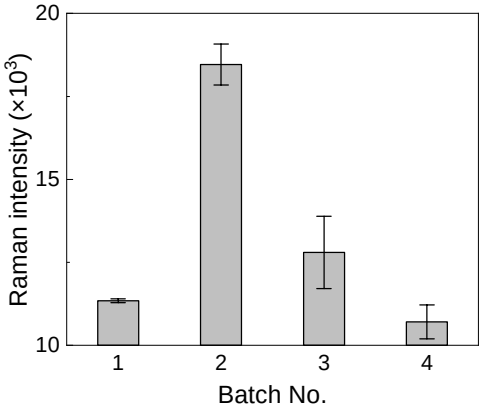

Figure R4. The signal intensity within the CV-specific window of the summed spectra (5,400 total voxels) after subtraction of the null control for 4 separately synthesized batches of Hya-Ag colloids at 10^{-10} M CV. The data shown are the average and the standard deviation (n = 3) for each batch.

Comment 1.4: It might also be useful to compare the digital approach with that advanced by Bin Ren et al (Single-Molecule Level Rare Events Revealed by Dynamic Surface-Enhanced Raman Spectroscopy, Anal. Chem. 2020, 92, 24, 15806–15810) https://doi.org/10.1021/acs.analchem.0c02936) who carried out SERS measurements in the single molecule regime and improved the sensitivity by identifying the spectra from an extended time series that contained signals and then summing only including those in the summation. This approach could be applied to the data shown in the current manuscript.

Response:

Thank you very much for your suggestion. We understand that the reviewer might be thinking the possibility of counting the number of positive events within an extended period of time with a fixed acquisition spot, as shown in the work of Bin Ren *et al*., instead of using a scanning spot to acquire positive events over space as used in the present work. Since the colloidal system is solution based and both the analyte and the colloids diffuse freely in the measurement chamber, the results from the two modes of acquisition should be equivalent. But from the point of designing practical instruments, the fixed spot mode of acquisition does have its own merit, which can simplify the design of the instrument. Thanks for the reviewer's suggestion, we have made some preliminary measurements with a fixed acquisition spot and obtained comparable results with the scanning

mode. Since the current manuscript is aiming at the proof of principle, we would prefer not to include this new mode of measurements for the lack of space and the lack of time to acquire comprehensive datasets.

Comment 1.5: Finally, I would note that the comparison between the dSERS and analogue SERS carried out in the same experiment is not really a fair one. The conditions in the manuscript are set to optimise the digital SERS method and then this compared with analogue data obtained under these conditions. It would be fairer to compare the results with other studies where analogue SERS has been optimised for the analytes has been used. The data for CV is impressive but that for the other molecules less so. In particular, the two examples included to illustrate applications of dSERS were paraquat and thiram. Paraquat was detected down to 10^{-10} mol dm^{-3}, which is 3 orders of magnitude better than the maximum residual level. However, it is not dramatically better than other much simpler SERS approaches such as Choo et al's method using standard microfluidics which had a detection limit of 2 x 10^{-9} mol dm^{-3} (Highly sensitive trace analysis of paraquat using a surface-enhanced Raman scattering microdroplet sensor, Analytica Chimica Acta 2010, 681(1-2):87-91 DOI: 10.1016/j.aca.2010.09.036). Similarly, SERS studies of thiram based on paper substrates gave a detection limit of 0.5 X 10^{-9} mol dm^{-3}, which is better than the 0.001 mg/kg (4x 10^{-9} mol dm^{-3}) in the current paper (Subnanomolar Sensitivity of Filter Paper-Based SERS Sensor for Pesticide Detection by Hydrophobicity Change of Paper Surface ACS Sens. 2018, 3, 1, 151–159 https://doi.org/10.1021/acssensors.7b00782). Indeed, a recent paper has claimed detection limits of 10^{-14} mol dm^{-3} and also lists several other SERS studies with sub-nanomolar detection limits (Femtomolar detection of thiram via SERS using silver nanocubes as an efficient substrate, Environ. Sci.: Nano, 2020,7, 3999-4009, https://doi.org/10.1039/D0EN01049A).

Response:

The first question here is whether analog method under optimized conditions, such as based on uniform solid chips, could achieve a comparable or even better performance than dCERS (as presented above). To address this directly, we performed additional experiments with the commercial Karite® chips (considered the most advanced and best performing) in the appropriate concentration range of the target molecule (i.e., CV). We acquired spectra in the pointwise scanning mode under the same parameters as in dCERS, and on each chip, 3 independent areas were measured. As shown in **Figure R5**, according to the analog method, by using the signal in the signature window (780-820 cm^{-1}) with the summed spectra of each scanned area, the Karite® chips

could only produce useful results down to about 10^{-7} M with a large uncertainty (also see **Figure S2** in the **Supplementary Note 2** for details). It is clear that even under the optimized conditions, analog method might be inadequate for quantifying low concentrations, which is still quantifiable by the digital method based on the ratio of positive voxels, as discussed in Response 1.3.

The issue of detectability for dCERS is quite different from the analog based approaches. The reason for this has been discussed above in much detail. Although several SERS methods have demonstrated detectability of ultra-low concentrations even down to single molecules, reproducible quantification, or limit of detection in the quantitative sense, is quite a different story. What we have demonstrated in this work was not intended to imply the limit of detection for dCERS, a critical point not well presented in the original submission. As we have discussed in Response 1.1 (and in the main text, **Line 165-169 on Page 9**), the true detection limit is determined by the digitization error of the background. As long as the error count from the background is much lower than the positive counts for the acquisition period, one can always increase the number of total voxels in order to detect lower concentrations. Although we have not pursued this absolute limit, an estimate could be made for this particular case of CV: 3.6×10^{-17} M (see **Supplementary Note 7.1**). If one likes to talk about absolute limits, this value must be somewhat impressive.

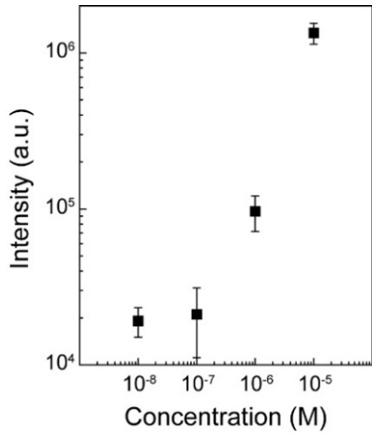

Figure R5. The peak intensity from the summed spectra of CV (concentration: 10^{-5}, 10^{-6}, 10^{-7} and 10^{-8} M) enhanced by Karite® chips. The standard deviation for each concentration was calculated from 3 mapped areas.

Comment from Reviewer 2:

Comment 2.1: How do differences in analyte affinity for the nanoparticle surface impact the dCERS measurements? The work described in the manuscript is focused primarily on analytes that readily associate with noble metal surfaces either via covalent linkages (4-nitrobenzenethiol, thiram), electrostatic interactions between negatively charged colloids and positively charged nitrogen groups (paraquat, crystal violet, adenine within A12) or via hydrophobic interactions (hemoglobin). The only analyte that does not bind strongly with the noble metal surface (D-glucose) exhibited the worst dCERS performance (LOD ~ 10^{-9} M vs. LOD <10^{-11} M for the other analytes). The extent that an analyte interacts with the surface and with SERS hotspots should dictate dCERS performance, however, this point is never discussed nor considered. [Interestingly, L-cysteine is included as an analyte in the methods and data on its dCERS detection is found within the extended data, however, it is not mentioned as an analyte within the main body of the manuscript.] Do the observed differences in the slopes of the RPV vs. concentration plots (Figure 1; Extended Data Figs. 6,7) reflect differences in surface affinity or do they reflect something else? It would seem possible to utilize the adapted Poisson formula defined within the Supplementary Information to somewhat quantify this relationship.

Response:

Thank you for raising these important issues that we did not present well in the original submission. Perhaps, the above critique could be addressed in two parts. The first deals with the relationship between the number of detected single molecule events vs. the analyte concentration. We have discussed this issue in our response to Reviewer 1, Comment 1.3. With all the analyte molecules that we have characterized, they all follow an intriguing relationship that can be phenomenologically described as $\log RPV = k \log M + b$. The parameter *k* and *b* are determined by the combined effect of all those factors as indicated by the reviewer, including the association with the metal surfaces, molecular cross sections and colloidal enhancements of the spectra and others at each concentration. Although this surprising linear relationship on the log-log plot is similar to the Freundlich equation in form, the physics and chemistry behind this phenomenon remain to be delineated. To deconvolute the contributions of these processes for their effect on *k* and *b* should be an enormous undertaking and technically challenging. However, because these correspondence relations are highly reproducible for a given batch of colloids, a pre-characterization should be sufficient in practical applications. Of course, this does not mean that an understanding of the underlying principles behind this nearly universal effect (also found with dELISA) is unimportant. It does certainly warrant systematic and detailed analysis in the future by us and perhaps other

investigators as well. In the revision, we have rewritten our discussion to make this point better appreciated.

The second question is the LOD. We have to apologize that in the original submission, we have not discussed this important issue with clarity. As indicated above, when a measuring process follows Poisson distribution, the actual LOD will only be determined by the null control (*i.e*., the false positives detected in the background over the measurement time period). We have now performed this measurement and found that the false positives are extremely low (see revised **Supplementary Note 7.1**, "Positive determination and null controls" for details). In this sense, even for an analyte-colloidal system that is not very sensitive, one can still use a longer acquisition time to accumulate sufficient positive voxels to succeed, provided that the accumulated positive counts far exceed the false count in null control for the same acquisition time.

We also thank the reviewer for the comment on L-cystine. It was our oversight and this mistake has been corrected in the revision. The calibration curve of L-cysteine is moved to **Figure S12** and included in the statement "all other target molecules, from the relatively simple glucose to poly-nucleic acids and proteins, exhibited the same linear relationship on the log-log plot" (**Page 7 Line 127-128** in the main text).

Comment 2.2: The project team utilized three different colloids within the work: hydroxylamine-reduced silver (Hya-Ag), citrate-reduced silver (citrate-Ag), and citrate-reduced gold (citrate-Au). They note that the best performance was achieved using Hya-Ag, but do not fully explain how such a conclusion was reached. While there is some data presented on the optical properties and single-molecule sensitivity in Extended Data Figure 1 (and explained in the Supplementary Information) it is not clear why Hya-Ag is preferred. Does Hya-Ag exhibit a different surface chemistry than citrate-Ag which makes it a better colloid for dCERS? What was the surface charge (i.e, zeta potential) for Hya-Ag vs. citrate-Ag vs. citrate-Au under the experimental conditions? Does a difference in surface charge differentially impact the surface affinity of the analytes and thus the dCERS sensitivity?

Response:

We totally agree with the reviewer here. Generally speaking, the best colloidal system should have the following properties: (1) for a particular analyte at a specific concentration, the detected positive events should be sufficiently high while the background false positives remain very low; (2) the colloidal system should be suitable for a wide range of different analytes, a property very

important for practical applications; (3) the long-term stability of the system (uniformly dispensed) is another critical factor that has significant value in practice.

To provide further support for our choice of Hya-Ag, we compared 3 types of colloids and re-performed experiments under the same colloidal concentration (0.5 nM) in the revised manuscript (**Extended Data Figure 2**). Using CV as a test, we found that Hya-Ag colloids can provide a higher detection efficiency than citrate-Ag colloids and citrate-Au colloids (**Extended Data Table 1** and **Supplementary Note 7.3**). Additionally, Hya-Ag colloids are also suitable for a wide range of analyte molecules as demonstrated in the manuscript and remain stable over long periods of time without noticeable aggregation or sedimentation. The latter was validated by the unchanged hydrodynamic diameter for the whole duration of the measurement used in the work (or even longer) by DLS (0 min: 39.9±16.34 nm, PDI = 0.323; 1 hour: 40.1±15.85 nm, PDI = 0.272; 3 hours: 41.3±17.93 nm, PDI = 0.228). We also attempted to have some understanding for the better performance of Hya-Ag over the other two colloids. We first tested the zeta potential as suggested by the reviewer but found no significant differences between them (Hya-Ag colloids: - 35.6 mV; Citrate-Ag colloids: - 33.8 mV; Citrate-Au colloids: - 36.0 mV). When comparing Ag and Au colloids, we found that Ag colloids often produced more intense plasmonic effects than that of Au, making the signals stronger thus more likely to be detected. While for citrate-Ag and Hya-Ag colloids, we speculated that citrate-Ag colloids are covered by a layer of citrate residues which may cause steric hinderance to impede CV from getting close to the metallic surface to some extent, thus reducing the probability of single-molecule detection, but Hya-Ag colloids are mainly covered by chloride ions, which enables CV to get closer the metallic surface.

To clarify this important point and strengthen our reasoning, we have included more experimental details, *i.e.*, zeta potential of the SERS colloids in **Table S3** in **Supplementary Note 8** with the instrumental information in the section of **Method** and the corrected calibration curves of CV enhanced by different SERS colloids in **Extended Data Figure 2** with the relevant explanation on the influence of colloids on the dCERS measurement in **Supplementary Note 7.3.**

Comment 2.3: The colloids and solutions used in the experiments are woefully under-characterized. While single TEM images of the three colloids are presented within Extended Data Figure 1 there is minimal to no information provided on TEM determined colloid diameter or surface charge. Looking at the TEM images in Extended Data Figure 1 it would seem that the citrate-Ag particles are much larger than the Hya-Ag or citrate-Au particles. The impacts of this size difference on the dCERS

results are not explained. Furthermore, only one type of particle was characterized by DLS (citrate-Ag) – why were the others not subjected to this simple analysis? [On a related note – why is there a disconnect between the ~100 nm diameter for Ag-citrate as determined by TEM and the DLS diameter of ~40 nm? Typically the TEM diameter is less than that determined by DLS due to particle aggregation. Unfortunately, the methodology and instrumentation used for the DLS measurements is not described.] Simply stating that 'lake water was fetched from a city park' is insufficient. What was the background electrolyte concentration? What macromolecules were present?

Response:

We apologize for our sloppiness and we have now performed comprehensive and systematic characterizations of the colloidal systems used in the experiments. Now the colloidal sizes have been determined to be about 24, 60 and 16 nm by TEM and about 31, 90 and 28 nm by DLS, corresponding to Hya-Ag colloids, citrate-Ag colloids and citrate-Au colloids, respectively. As expected, the values measured by DLS are typically larger than TEM since the former is measuring the hydrodynamic diameter of aqueous colloids. Detailed information about instrumentations, methods and related results of characterization are now provided in the section of **Methods** and **Figure S10** in the **supplementary file** in the revision.

As for the lake water, we indeed should have provided more information. This deficiency has been addressed in the revision. Briefly, the collected lake water was centrifugated and filtered with by 0.8 μm, 0.22 μm membrane filter units and 3 kDa-cutoff ultrafiltration centrifugal tube (Millipore) to remove particulates. By DLS, there was no identifiable particulates after this filtering step. No further processing of the lake water was performed. The conductivity of the filtered water was 154.4 μS/cm and the pH at 7.92. These values are well within the range established by US Environmental Protection Agency for normal lake water. These details are now added in the method section "Materials and instrumentation" with a brief discussion in the main text (**Line 199-203 on Page 10**).

Comment 2.4: The impacts of nanoparticle aggregation on the dCERS results are not addressed. For SERS hotspots to form it is necessary for the nanoparticles to aggregate to some extent. Such aggregation may happen due to the interaction between the analyte and the colloids, but aggregation will also be affected by the background electrolyte concentration (with higher electrolyte concentrations enhancing aggregation) as well as the presence of macromolecules (which may hinder aggregation). Under the simple solution conditions (either pure water or ethanol/water)

used to establish the dCERS approach aggregation should primarily reflect analyte-colloid interactions, however, under the 'Field' deployment conditions presented in Figure 3 the extent of aggregation will presumably reflect these background constituents. Such an outcome is not considered within the manuscript. Furthermore, if aggregation is too extensive then the colloids will sediment and no longer be detectable by their approach.

Response:

We agree that the problem of aggregation is important and should be better addressed. As now demonstrated in the revised manuscript, accurate quantification can be achieved by performing a dilution series, so that after adequate dilution, problems associated with field applications can be significantly reduced or becomes negligible. Two examples are shown in **Figure 3**. Here we show that for paraquat in the filtered lake water and thiram in the filtered bean sprouts extract, the same linear relationship as the calibration can be recovered upon adequate dilutions, and the concentration of the target analyte in the complex background can be accurately determined. Although not demonstrated in this work, both the ionic strength and the pH of a sample can always be adjusted accordingly into the acceptable range and, if required, constant stirring could be used to further reduce aggregation. Since dCERS can detect much lower concentrations of the analyte (see response to Comment 1.5), a serial dilution should be able to recover the calibrated response curve, thus, to achieve accurate quantification.

Comment 2.5: What is the minimal number of voxels required to be sampled for the approach to work? The authors suggest it is possible to calibrate this term for each analyte, but it is not clear whether the lower limits reflect differences in surface affinity (see #1).

Response:

As discussed above in much detail, given the Poisson-dominated nature of our approach, the minimal number of voxels required is determined by the desired accuracy regardless the actual concentration of the analyte, while the detection limit is determined by the digitization error of the background (which can be different for different analytes even for the same colloidal system, thus, needs to be determined for each). As for the calibration, it provides a monotonic relationship between the number of positive voxels and the analyte concentrations, which reflects the combined effects of all processes involved in detection and is not directly related to the limit of detection. We have now comprehensively revised this part of the work in order to make this point better appreciated.

Comment 2.6: Pg. 3. The authors note that variable molecular orientations on a substrate impact SERS quantitation. Wouldn't this same phenomenon affect the dCERS results? A different molecular orientation could potentially result in a signal with lower intensity that falls below the threshold defined by the dCERS approach. How can dCERS account for multiple different molecular orientations on the surface?

Response:

Indeed, different molecular orientations will result in a different signal enhancement, and as such, some orientations may not be detectable. However, if identical criterion is applied to the actual measurements and the calibration, this should not lead to serious errors, especially when the accumulated positive events are relatively abundant. The important is the reproducibility of the correspondence between the positive events and the analyte concentration. Fortunately, this is held up with all the analyte molecules that we have tested.

Comment 2.7: Pg. 4. The authors suggest that '…Ag colloids can be distributed homogeneously throughout the entire solution'. How is such homogeneity ensured given that the particles will aggregate?

Response:

We have monitored our experimental system by DLS and found that during our measurements, no apparent aggregation was detectable, as discussed in response to Comment 2.2. Therefore, at least for the data presented, this has not been a problem. However, we do agree with the reviewer that for a practical system, such aggregation must be controlled in order to achieve robust and reliable measurements. For that purpose, dilution and constant agitation with minute magnetic bars could be used to minimize potential aggregations and consequent colloidal sedimentation. Accordingly, we have revised this statement with more explanations in the revision (see **Line 205-210 on Page 11** and **Line 229-233 on Page 12** in the revised manuscript).

Comment 2.8: C. Pg. 7. What is meant by 'pre-calibration of the prepared colloids with each batch'?

Response:

Pre-calibration is done as follows. For each freshly made batch of colloids, a small fraction was drawn and dCERS was performed for various analytes over an appropriate concentration range. These calibration curves will be used to convert measured positive counts into quantitative concentrations for the target analyte. Because there are uncontrollable differences for each fabrication process, such a calibration is necessary with every batch of colloids for quantitative measurements.

Comment 2.9: Pg. 7. The authors list a number of factors that affect the practical concentration ranges over which dCERS works for different analytes. How do they disentangle the relative enhancement factor of the hotspots from the binding affinity of the target analytes? Won't those two parameters be intricately intermixed?

Response:

This is indeed one of the most challenging mechanistic questions. Honestly speaking, there are so many processes and factors involved which are difficult to resolve. Although it is technically difficult to deconvolute the above-mentioned factors, it is rather straightforward to assess the overall effect in dCERS measurements with pre-performed calibrations for each analyte. The advantage of single molecule counting is that the absolute sensitivity is less critical as only yes or no is required for each event. As long as a proper calibration can be achieved in the measurement concentration range, one does not need to be concerned with the efficiencies for each target analyte.

Comment 2.10: Pg. 9. If the authors are going to suggest that thiram uptake is higher than expected within bean sprouts then they should refer to the established literature on this topic. Stating that '…a much lower usage of this chemical might be considered' is an over-reach given the simplicity of their experimental system.

Response:

Point well taken. The mentioned statement is indeed an over reach at present. This sentence has been removed in the revision (**Line 242 on Page 12**).

Comment 2.11: Figure 3. It is unclear in panel b why one of the X-axes increases in concentration from right to left while the other decreases. At first I thought that this result reflected competition

between the analytes for hotspots (and thus the NB signal decreases with an increase in CV), but that is not correct. Finally, what is the blue arrow in panel d meant to indicate?

Response:

We agree with the reviewer that Fig 3 is indeed confusing. We apologize for that. In the revision, we have rearranged the results for this figure with additional data so that it is easier to follow. Specifically in the revised **Figure 3b (also Figure R8b)**, we separated the data points of CV and NB into two panels and for each, the points from left to right correspond to sample No.1 to No.3 with different analyte proportion which has been described in an additional table (**Table S4** in the supplementary file) for clarity. As for the blue arrow in **Figure 3d**, it was used to indicate a particular set of experiments. But in the end, these experiments were not included in the final version, and it has escaped our notice when submitted. **Figure 3d** has now been revised without this arrow.

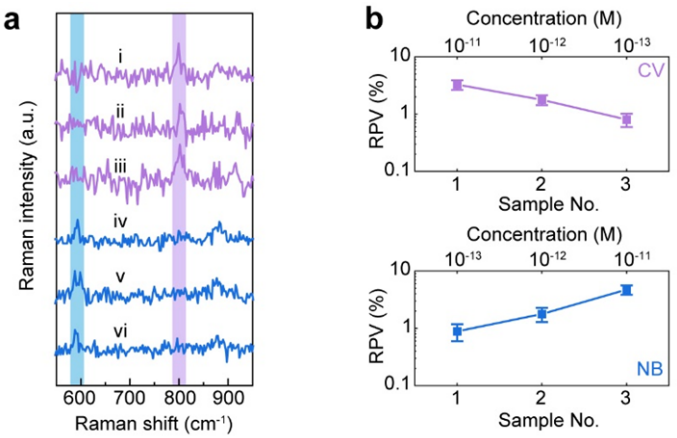

Figure R8. Quantification of CV and NB when mixed at different concentrations. Samples 1, 2 and 3 contained different concentration ratios, *i.e.*, sample 1: 10^{-11} M CV and 10^{-13} M NB; sample 2: 10^{-12} M CV and 10^{-12} M NB; sample 3: 10^{-13} M CV and 10^{-11} M NB. (a) Typical SM spectra corresponding to CV (i-iii) (purple, "0" for NB and "1" for CV) and NB (iv-vi) (blue, "1" for NB and "0" for CV) in a mixed sample. The characteristic peaks for CV (purple shaded) and NB (blue shaded) are clearly identified and adequate for quantification. (b) RPV versus the concentrations of CV (top) and NB (bottom) in the 3 samples with different concentration ratios.

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

The authors have significantly improved the presentation of the work in this version of the manuscript and have clarified many of the issues that I raised in the initial review. For the sake of brevity maybe it is easiest to say that in points 1.1 and 1.5 in the Response I feel that the changes which have been made do answer the queries which I had raised before.

A minor clarification on my part is around 1.4, what I was suggesting was not associated with taking different experimental data, as the response is centered on but instead was a suggestion that only the spectra which show a signal above the threshold are included in a summation of the pixels for the analogue comparison. This would be as a better alternative to simply summing all the pixels since it will avoid adding the noise from the pixels that have no signal into the sum.

For response 1.3, the Figure R3 now clearly shows the significant improvement which is possible using the digital method but I don't really understand how it comes about. In the digital spectra the %RPV decreases by less than a factor of 2x between 10^{-12} M and 10^{-13} M. I would therefore have expected the summed spectra over all voxels to decrease approximately 2x as well but in fact it drops dramatically. Why does this happen? Is the average intensity of the signals in the positive voxels also falling? That would seem to be the only way to account for the observed difference but it then raises the question of why that would be happening. This is also relevant to the question in 1.2, which is why the %RPV does not increase linearly. The authors have now very clearly pointed this out as a question but I am not sure that the answer is very satisfactory. To say the data fit a log -log plot and this is the same as fitting to a Freundlich isotherm is true but it is a circular argument. As the authors themselves point out, the Freundlich isotherm doesn't really allow for any physical interpretation. This means that there is no microscopic picture for what is being measured in these experiments. I do sympathize with the authors because the data are very puzzling but I am also very aware that publication in Nature is normally on the basis of finding an important problem and then advancing the solution. In the current draft of the paper the data are really excellent. They convincingly show how the signals change with concentration and how this allows for low concentration measurements but the manuscript still lacks an explanation for the most important part of the work, which is why the signals don't change the way one would expect. This is the fundamental question that lies at the heart of the work and I would have thought the answer would have been the reason for justifying publication in Nature, particularly since the authors point out that it seems to be a general phenomenon and is observed other types of digital analysis methods.

I wonder if it would be helpful to look at the distribution of the signal intensities at various concentrations? The digital approach assumes that the discrimination is binary but are all positive voxels the same (or more accurately do they fit to the expected distribution) and are the blank pixels actually zero? It would be very useful to have a plot of the distribution of the intensities of the signals in the voxels, since moving away from a binary molecule/no molecule interpretation may allow the data to be rationalized.

In the absence of a meaningful physical model I do think that in my judgement the balance would be

against publication at this point but the margin is small.

Referee #2 (Remarks to the Author):

The authors' have done a decent job at responding to both my and the other reviewer's prior comments. However, they have been lazy in implementing our suggestions and/or addressing our comments within the manuscript. Instead of revising the order of materials in the Supplemental section to keep things in a sequential order (i.e, Note 1 is referred to first followed by Notes 2, 3, etc.., they instead chose to add a new Supplemental Note 7 (that is actually subdivided into sub-notes 7.1-7.4) and Supplemental Note 8 and refer to them at varied points throughout the manuscript. This is confusing and does not reflect typical practice for referring to materials in sequential order. This issue needs to be fixed to address the readability and the usability of the manuscript.

Another challenge that readers of this manuscript will face is how to repeat the methods in order to replicate the results. While the authors' have improved the description of their overall system (i.e., including greater detail about the Ag colloid characteristics) there still remain a number of important details that are poorly explained. Simply making statements like (lines 80-82): 'The optimal concentration of the colloids was determined based on the criterion at which a sufficiently high concentration of hotspots was achieved but with an acceptable background scattering within the excitation volume of the probe' does not suffice. What defines a 'high concentration of hotspots'? What is 'acceptable background scattering'? Supplemental note 3 has some illustration of the effect of colloid concentration on the SERS signal, but there is nothing within that section that truly illustrates either hotspot density or background scattering - instead the reader is given two plots that suggest a concentration of 0.5 mM colloids is ideal, but have no idea as to WHY it is is ideal. As the authors' later note, this concentration is expected to vary as a function of the analyte identity and the background chemistry. Given this fact, if an interested reader wanted to replicate these experiments it is not clear that they readily could.

Another challenge of the dCERS approach relates to its broad applicability to detect analytes. In their response to the reviewers, the authors clearly state that not all analytes are amenable to the dCERS approach - this point, however, is only weakly stated within the actual manuscript. There is some effort to address it within Supplemental Section 7.3, but there is a need to add text to the manuscript that not only states that this approach is not universally applicable but that also summarizes its limitations. Following this line of thought - the authors want to suggest that their approach has the capacity to detect analytes in a range of different media. However, to achieve such measurements requires extensive dilution and thus removal of the media background with the extent of dilution needing to be experimentally evaluated for each system. Such an approach is not practical. Instead of suggesting their method, as currently presented, is useful for real-world application they should instead highlight that this is a proof-of-concept effort and use the experiments depicted in Figure 3 (and elsewhere) to test its capabilities under challenging conditions.

Some specific comments:

Abstract. The abstract is missing key information regarding what digital (nano)colloids are. There is discussion of their large scale fabrication later in the abstract, but no real information regarding what they are.

Line 29. Suspension not solution.
Line 39. Poisons not poisoning.
Line 48. Replace 'with their' with 'based upon their'
Line 80. Why Is Supplementary Note 7.3 referred to prior to any other notes? Why is the numbering system different for Supplementary Note 3? Shouldn't the Supplemental materials be presented and referred to sequentially in the main manuscript?
Lines 81-82. Does this optimal concentration vary for each target? Water quality tested? If so, how can this approach be extended to unknown samples or environmental samples with widely variant background chemistries?
Line 83. How is the homogeneity of the hotspot distribution determined? Won't the nanoparticles continue to aggregate with time?
Lines 93-94. How does the hotspot distribution change following addition of the analyze to the suspension? How is the distribution affected by changes in the analyte identity?
Line115. The authors set a fairly high threshold (> 10$\boldsymbol{\sigma}$), but do not provide any detailed analysis as to how changes in the threshold (either higher or lower) impact their digital SERS results. What happens if the threshold is set lower? What exactly is the tradeoff between not detecting something that is there vs. minimization of false positives? There is some analysis presented in Table S2 using a lower threshold (> 5$\boldsymbol{\sigma}$), but there is no discussion of what this change in threshold means with respect to the data.

Supplemental Notes. What is a Bernoulli trial? It is not defined either in the Supplemental information or in the main text.

Author Rebuttals to First Revision:

Point to Point Response to Reviewers' Comments

Response to Reviewer 1:

Comments 1: The authors have significantly improved the presentation of the work in this version of the manuscript and have clarified many of the issues that I raised in the initial review. For the sake of brevity maybe it is easiest to say that in points 1.1 and 1.5 in the Response I feel that the changes which have been made do answer the queries which I had raised before.

Response:

We thank this reviewer for his/her positive comments of our work.

Critique 1: A minor clarification on my part is around 1.4, what I was suggesting was not associated with taking different experimental data, as the response is centered on but instead was a suggestion that only the spectra which show a signal above the threshold are included in a summation of the pixels for the analogue comparison. This would be as a better alternative to simply summing all the pixels since it will avoid adding the noise from the pixels that have no signal into the sum.

Response:

As we understand, most analog measurements are essentially equivalent to a summation of all of the pixels in our system, and so we compared this summation to our digital approach in the paper. We now realize that this was not what was asked by the reviewer. For the sake of the comparison that was suggested by the reviewer (while, we note that such an approach would also have to invoke an idea of "digitization" with a preset judgment based on some criteria to determine what to include), we have calculated the results as shown in **Figure R1**. Although the detectability might improve, the uncertainty of the quantification varies in a broad range, owing to the well-documented single molecule intensity fluctuations at very low concentrations where single/few molecule events dominate. Therefore, even under this more "optimal" condition, reliable quantification at very low concentrations still could not be achieved with the analog mode, in contrast to that of dCERS.

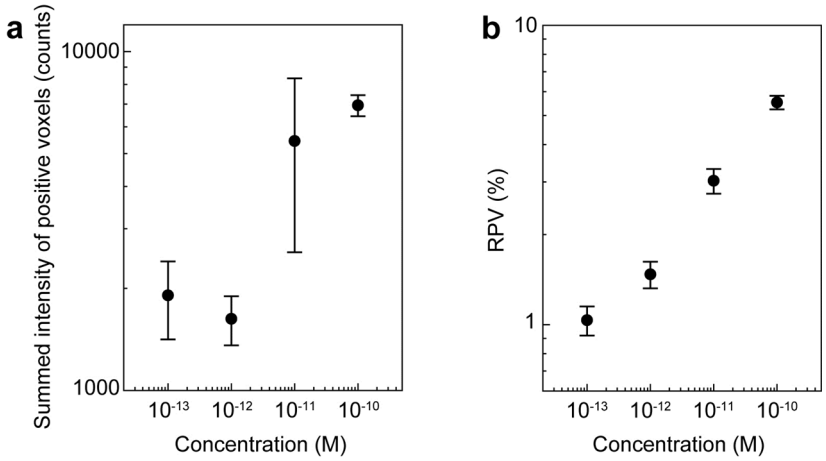

Figure R1. (a) The summed intensity of positive voxels and (b) the ratio of positive voxels (RPV) versus CV concentration by dCERS. The standard deviations are calculated from 3 independent measurements (5,400 voxels in total per measurement) at each concentration.

Critique 2: For response 1.3, the Figure R3 now clearly shows the significant improvement which is possible using the digital method but I don't really understand how it comes about. In the digital spectra the %RPV decreases by less than a factor of 2x between 10^{-12} M and 10^{-13} M. I would therefore have expected the summed spectra over all voxels to decrease approximately 2x as well but in fact it drops dramatically. Why does this happen? Is the average intensity of the signals in the positive voxels also falling? That would seem to be the only way to account for the observed difference but it then raises the question of why that would be happening.

Response:

This is a good point – we should have provided more description of the situation. Simply, the measurement at 10^{-13} M is below the analogue detection limit (with all voxels summed up). The noise is too significant – it drowns out the signal. Summing over all voxels, the apparent "signals" in the absence of any analyte give values of 0.2 ± 319.4 counts. Thus, we would consider all such measurements below roughly 900 counts (3 standard deviations) as unreliable. As for the average intensity in the positive voxels, they indeed do not change with concentration, as shown in **Figure R2** below.

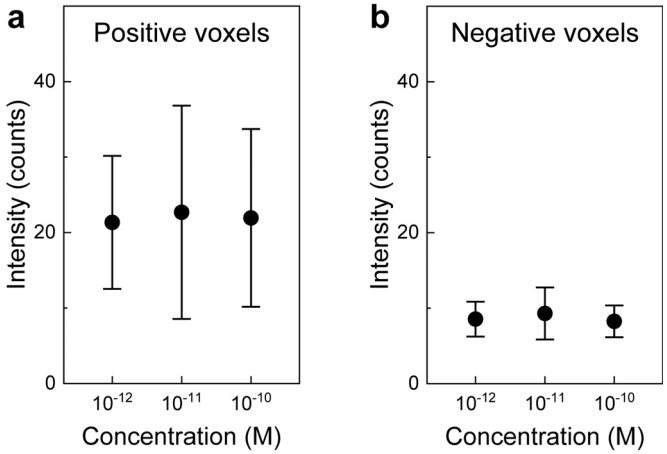

Figure R2. Distribution of the intensities of (a) 100 positive voxels and (b) 100 negative voxels at each concentration. Error bar: standard deviation of the 100 positive/negative voxel intensities.

Critique 3: This is also relevant to the question in 1.2, which is why the %RPV does not increase linearly. The authors have now very clearly pointed this out as a question but I am not sure that the answer is very satisfactory. To say the data fit a log-log plot and this is the same as fitting to a Freundlich isotherm is true but it is a circular argument. As the authors themselves point out, the Freundlich isotherm doesn't really allow for any physical interpretation. This means that there is no microscopic picture for what is being measured in these experiments. I do sympathize with the authors because the data are very puzzling but I am also very aware that publication in Nature is normally on the basis of finding an important problem and then advancing the solution. In the current draft of the paper the data are really excellent. They convincingly show how the signals change with concentration and how this allows for low concentration measurements but the manuscript still lacks an explanation for the most important part of the work, which is why the signals don't change the way one would expect. This is the fundamental question that lies at the heart of the work and I would have thought the answer would have been the reason for justifying publication in Nature, particularly since the authors point out that it seems to be a general phenomenon and is observed other types of digital analysis methods.

Response:

We apologize for a lack of rigor on this important point – we only meant to point out that our observations of a log-log relationship agreed with what has been found in other systems that are explained by the Freundlich equation. But indeed, as the reviewer pointed out, the original reference was empirical and remained so for many decades afterwards. However, after a thorough search of the literature, we have found that there is now a firm theoretical basis for this equation based on classical Gibbs thermodynamics (ACS Omega, 2020, 5, 13130-13135; Philosophical Magazine Letters, 2022, 102(7), 239-253). Indeed, as the latter publication notes, the Gibbs formalism in fact predicts the Freundlich relationship at the lowest analyte concentrations down to single molecule detections. We are particularly grateful to the reviewer for raising this concern.

As a result, we have removed much of the previous discussion in the manuscript and include the following on Line 127-129, Page 7: "…The concentration dependence of RPV for CV is well-described by a straight line with R^2 = 0.99 when plotted on a log-log scale (Figure 1c), consistent with the well-known Freundlich equation that can be explained based on classical Gibbs thermodynamics…"

Critique 4: I wonder if it would be helpful to look at the distribution of the signal intensities at various concentrations? The digital approach assumes that the discrimination is binary but are all positive voxels the same (or more accurately do they fit to the expected distribution) and are the blank pixels actually zero? It would be very useful to have a plot of the distribution of the intensities of the signals in the voxels, since moving away from a binary molecule/no molecule interpretation may allow the data to be rationalized.

Response:

Thank you for these questions. As shown in **Figure R2**, the intensity of positive voxels varies above the threshold, but shows no significant difference at different analyte concentrations, which further indicates the benefit of using the number of positive voxels for quantification.

Comments 2: In the absence of a meaningful physical model, I do think that in my judgement the balance would be against publication at this point but the margin is small.

Response:

We appreciate your insightful suggestions which have broader implications even beyond dCERS. We hope that we have now addressed these concerns and perhaps have moved the "small margin" in the opposite direction?

Response to Reviewer 2:

Comments 1: The authors' have done a decent job at responding to both my and the other reviewer's prior comments.

Response:

We are grateful to this reviewer for their positive comments of our work.

Critique 1: However, they have been lazy in implementing our suggestions and/or addressing our comments within the manuscript. Instead of revising the order of materials in the Supplemental section to keep things in a sequential order (i.e, Note 1 is referred to first followed by Notes 2, 3, etc.., they instead chose to add a new Supplemental Note 7 (that is actually subdivided into sub-notes 7.1-7.4) and Supplemental Note 8 and refer to them at varied points throughout the manuscript. This is confusing and does not reflect typical practice for referring to materials in sequential order. This issue needs to be fixed to address the readability and the usability of the manuscript.

Response:

We apologize for this. Previously, we intended to put all practical considerations in one Supplementary Note, but now we realize that this might have caused poor readability in the main text. Therefore, we have now corrected this and have rearranged the supplementary files in the order with which they are discussed in the main text.

Critique 2: Another challenge that readers of this manuscript will face is how to repeat the methods in order to replicate the results. While the authors' have improved the description of their overall system (i.e., including greater detail about the Ag colloid characteristics) there still remain a number of important details that are poorly explained. Simply making statements like (lines 80-82): 'The optimal concentration of the colloids was determined based on the criterion at which a sufficiently high concentration of hotspots was achieved but with an acceptable background scattering within the excitation volume of the probe' does not suffice. What defines a 'high concentration of hotspots'? What is 'acceptable background scattering'? Supplemental note 3 has some illustration of the effect of colloid concentration on the SERS signal, but there is nothing within that section that truly illustrates either hotspot density or background scattering - instead the reader is given two plots that suggest a concentration of 0.5 nM colloids is ideal, but have no idea as to WHY it is ideal. As the authors' later note, this concentration is expected to vary as a function of the analyte identity and the background chemistry. Given this fact, if an interested reader wanted to replicate these experiments it is not clear that they readily could.

Response:

We apologize for this deficiency. The optimal colloids concentration was empirically determined based on two factors: hotspots concentration and background scattering of the colloids that interfere with both illumination and signal acquisition. To be specific, hotspots concentration clearly increases with the colloidal concentration, but at the same time, high concentration of colloids also causes significant background scattering thus reducing the usable hotspots. Therefore, there should be an optimal concentration of colloids at which the usable hotspots are at a maximum. In the supplementary note, we have now added more explanation on the determination of optimal colloidal concentration.

For the analytes we have tested, we found that a colloidal concentration of 0.5 nM was optimal (**Figure R3**). Though the precise optimal colloidal concentration is likely to vary among different analytes, **Figure R3** also shows that the "optimal range" is relatively broad,

and therefore, a reasonable choice of colloidal concentrations should be adequate for most measurements. In fact, achieving the same accuracy (the same number of accumulated positive events) at suboptimal concentrations would only require an increase to the measurement time, which is a strong advantage of this method.

To remedy this, we have revised the appropriate phrasing in the main text (line 81-84, page 5):

"The most effective concentration of the colloids to use in these measurements was experimentally determined (Figure S5 in Supplementary Note 4) owing to the fact that while the number of hotspots increases with the concentration of the colloids, the background scattering also increases at the same time, leading to a reduction of detectable signals".

In addition, we have included the following text in the Supplementary Note 4:

"For dCERS, the detection efficiency is closely related with the concentration of the SERS colloids via the density of hotspots in the voxels and the colloidal background scattering, which should, therefore, be optimized to balance the two factors. The first issue is the hotspot density. This refers to the areas adjacent to the colloidal surface with high electromagnetic field, where the SERS signals of the molecules can be greatly enhanced and has the potential to generate positive voxels if in the probe volume. Therefore, the colloidal concentration is desired to be higher to increase the hotspot density, making the molecules more likely to be detected. The second issue is the background scattering, which, on the contrary, is preferred to be low. Too much colloids, actually, may reduce the excitation efficiency and scatter out the molecular SERS signals, thus reducing the usable hotspots. Though the precise optimal colloidal concentration is likely to vary among different analytes, **Figure S5** also shows that the "optimal range" is relatively broad, and therefore, a reasonable choice of colloidal concentrations (e.g., 0.5 nM) should be generally adequate for most measurements. In fact, achieving the same accuracy at suboptimal concentrations would only require an increase to the measurement time, which is a strong advantage of this method."

We again thank the reviewer for this important clarification.

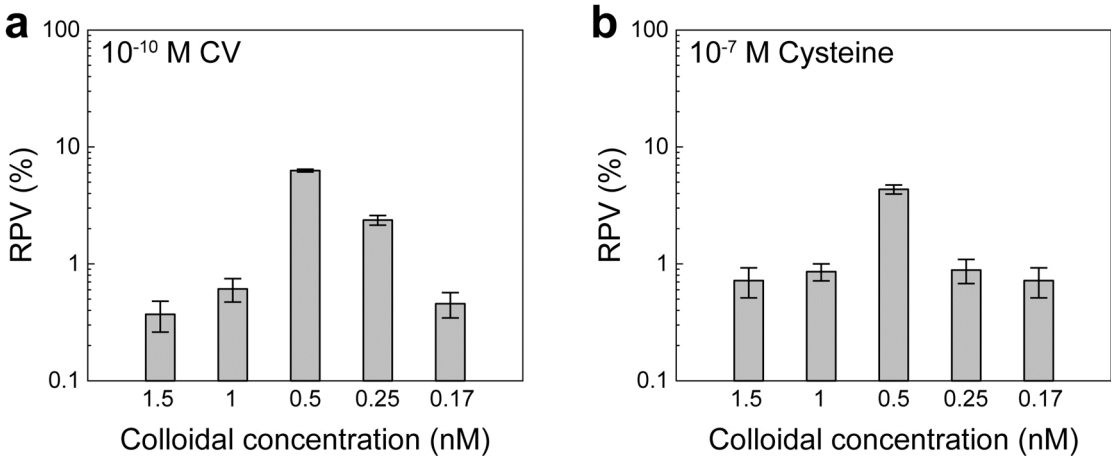

Figure R3. Optimal concentration of the SERS colloids for dCERS (Figure S5 in the supplementary file). (a) 10^{-10} M CV and (b) 10^{-7} M L-cysteine were detected by the Hya-Ag colloids at 0.17, 0.25, 0.5, 1 and 1.5 nM. The optimal concentration for Hya-Ag colloids was found around 0.5 nM.

Critique 3: Another challenge of the dCERS approach relates to its broad applicability to detect analytes. In their response to the reviewers, the authors clearly state that not all analytes are amenable to the dCERS approach - this point, however, is only weakly stated within the actual manuscript. There is some effort to address it within Supplemental Section 7.3, but there is a need to add text to the manuscript that not only states that this approach is not universally applicable but that also summarizes its limitations. Following this line of thought - the authors want to suggest that their approach has the capacity to detect analytes in a range of different media. However, to achieve such measurements requires extensive dilution and thus removal of the media background with the extent of dilution needing to be experimentally evaluated for each system. Such an approach is not practical. Instead of suggesting their method, as currently presented, is useful for real-world application they should instead highlight that this is a proof-of-concept effort and use the experiments depicted in Figure 3 (and elsewhere) to test its capabilities under challenging conditions.

Response:

We agree with the reviewer's point. We have now added more qualifying statements in the conclusion section to indicate the limitations

(Line 122-124, Page 7):

"It should also be noted that for any analyte to be suitable for dCERS detection, well-recognized spectral features must be present without significant complications either from the background or from other analyte species (Supplementary Note 9)."

(Line 162-163, Page 9):

"To further demonstrate the validity of dCERS, we first quantified the mixture of two different target molecules dispersed in the same solution, …"

(Line 188-190, Page 10):

"To further determine whether such measurements could also be achieved under complex and uncharacterized backgrounds similar to field situations, …"

(Line 210, Page 11):

"As a final proof of its validity, we further applied dCERS to the detection of a fungicide, thiram, …"

(Line 236-246, Page 12-13):

"In conclusion, we have demonstrated that dCERS based on SM counting is an effective and efficient technology for molecular quantifications at ultra-low concentrations solely based on the intrinsic vibrational signatures of the target molecules as long as they exhibit well-defined, specific spectral signatures as well as adequate interaction with the hotspots. …To further validate our approach, we used thiram and paraquat with complex and unknown backgrounds and found that dCERS can still achieve reproducible quantifications. As demonstrated by these proof-of-concept examples of toxic compound detection in aqueous solution, …"

We have also included the following in the abstract (Line 28-29, Page 2):

"Here, as a proof of concept, we show that, using digital (nano)colloids-enhanced Raman spectroscopy (dCERS), …"

In terms of the use of serial dilutions in actual measurements, we would beg to differ. In fact, serial dilution is an often-adapted strategy for quantitative measurements. An example that readily comes to mind is the quantitative PCR technology where serial dilution is almost a requirement for accurate quantitation. In terms of dCERS, owing to the ultrasensitive nature of our approach, dCERS can quantify very low concentrations of analyte at desired accuracy, and so will allow valid measurements with a broad range of serial dilutions. As we have indicated, only when a linear relationship (on log scale) is achieved, can one be assured of reliable quantification. The original analyte concentration can be recovered by simply multiplying the dilution factor. Since the pre-calibrated colloids are readily available, such dilution measurements do not impose any significant challenge in terms of the materials either, in contrast to solid chip-based methods.

Critique 4: Abstract. The abstract is missing key information regarding what digital (nano)colloids are. There is discussion of their large-scale fabrication later in the abstract, but no real information regarding what they are.

Response:

We appreciate this suggestion. We have now included the following in the abstract (Line 31-33 Page 2):

“Since metallic colloidal nanoparticles that significantly enhance these vibrational signatures, including hydroxylamine-reduced silver (Hya-Ag) colloids, can be fabricated at large scale under routine conditions, …”.

Critique 5: Line 29. Suspension not solution; Line 39. Poisons not poisoning; Line 48. Replace ‘with their’ with ‘based upon their’

Response:

Thank you for the correction. We have changed all of these words accordingly.

Critique 6: Line 80. Why Is Supplementary Note 7.3 referred to prior to any other notes? Why is the numbering system different for Supplementary Note 3? Shouldn’t the Supplemental materials be presented and referred to sequentially in the main manuscript?

Response:

We apologize for this mistake and have now corrected it.

Critique 7: Supplemental Notes. What is a Bernoulli trial? It is not defined either in the Supplemental information or in the main text.

Response:

We have added the definition in the Supplementary Note 5.

Critique 8: Lines 81-82. Does this optimal concentration vary for each target? Water quality tested? If so, how can this approach be extended to unknown samples or environmental samples with widely variant background chemistries?

Response:

Though the optimal concentration may vary among different target molecules and background chemistry, given the fact that the “optimal range” is relatively broad, it is relatively easy to identify a colloidal concentration that is applicable to most measurements. Even at the slightly suboptimal concentration, only more voxels would have to be accumulated for comparable accuracy. Also please see **Figure R3**.

Critique 9: Line 83. How is the homogeneity of the hotspot distribution determined? Won't the nanoparticles continue to aggregate with time?

Response:

The uniformity of the hotspots can be experimentally assessed by the signal variation among voxels with a relatively high concentration of analyte. Using Hya-Ag colloids as an example, we found that signal variation with 10^{-7} M CV among 200 voxels was only about 2.8%, indicating that the hotspots present in each probed voxel was statistically comparable throughout the measurement space (**Figure R4**). We also monitored the colloids by DLS and found no aggregation for the whole duration of the measurement used in the work (or even longer): the hydrodynamic diameter was unchanged (0 min: 39.9 ± 16.3 nm, PDI = 0.323; 1 hour: 40.1 ± 15.8 nm, PDI = 0.272; 3 hours: 41.3 ± 17.9 nm, PDI = 0.228). These data have been now added in the Supplementary Note 5.

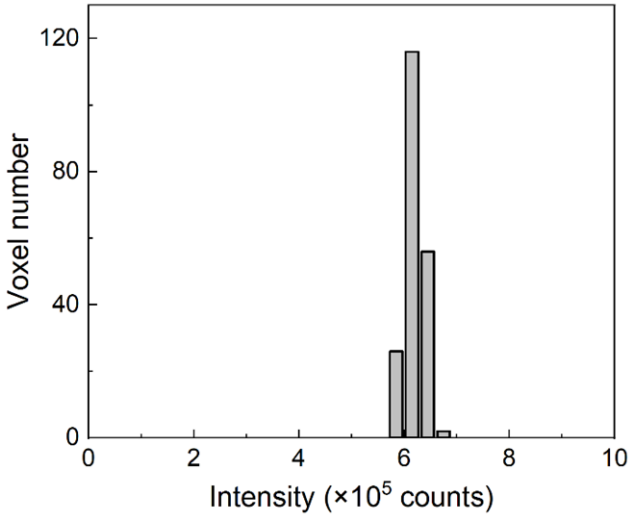

Figure R4. Histogram of the intensity distribution among voxels by 10^{-7} M CV detected by Hya-Ag colloids. The relative standard deviation is 2.8% among the 200 voxels.

Critique 10: Lines 93-94. How does the hotspot distribution change following addition of the analyze to the suspension? How is the distribution affected by changes in the analyte identity?

Response:

If there is no colloidal aggregation, the hotspot density will not change either. At least in the analytes we used, we did not find appreciable changes in DLS so far (10^{-10} M CV: 39.2 ± 15.9 nm, PDI = 0.321; 10^{-8} M paraquat: 43.9 ± 21.2 nm, PDI = 0.274). Actually, the

difference in detection efficiency is largely caused by the molecular cross section and their affinity to metallic colloids.

Critique 11: Line115. The authors set a fairly high threshold (> 10**σ**), but do not provide any detailed analysis as to how changes in the threshold (either higher or lower) impact their digital SERS results. What happens if the threshold is set lower? What exactly is the tradeoff between not detecting something that is there vs. minimization of false positives? There is some analysis presented in Table S2 using a lower threshold (> 5**σ**), but there is no discussion of what this change in threshold means with respect to the data.

Response:

Specifically, with 5,400 total voxels, the number of false positives is 2 when the threshold is set at 10σ, but this increases to 250 when set at 5σ and further to 1,689 if 3σ is used as the threshold as calculated in the control samples. Thus, for the low concentrations that we wish to measure, even increasing to 5σ will introduce an unacceptably high number of false positives. As indicated in the text, this false positive rate will increase the lower bound of the detection limit. We are grateful to this reviewer for mentioning this concern and have now included these details in Supplementary Note 8.

"Actually, a higher threshold can also be used yet at the cost of reducing SM events at a certain analyte concentration, while a lower threshold may include more false positives thus reducing the accuracy as well as increasing the lower bound of the detection limit (i.e., 250 false positives in 5,400 voxels at 5σ threshold and even 1,689 in 5,400 at 3σ threshold)."

Reviewer Reports on the Second Revision:

Referee #1 (Remarks to the Author):

The authors have answered the second set of questions I raised appropriately and as far as I can see have also given reasonable responses to those of the other reviewer. In particular, their successful search for a theoretical basis of the Freundlich isotherm to explain the log/log concentration dependence is convincing. It is a remarkable coincidence of timing that this new treatment, and in particular the 2022 paper, was published so long after the original isotherm but just in time for the current manuscript.
Similarly, I am content that the other questions I asked were dealt with appropriately.
I would now support publication of the manuscript in it current form.

Referee #2 (Remarks to the Author):

The authors' have responded appropriately to all of the prior reviewer comments and my remaining comments are minor:

1) The word 'solution' is used within the paper when 'suspension' is more appropriate. The authors should use a global search to verify that the proper term is used whenever necessary.

2) lines 122-124. The authors should also note that the analyte in question must associate with the colloids. As a surface enabled technique such association is an inherent requirement. Changes in surface functionalization may be required to get a particular analyte to associate the colloids.

Author Rebuttals to Second Revision:

Point to Point Response to Reviewers' Comments

Response to Reviewer 1:

Comment 1: The authors have answered the second set of questions I raised appropriately and as far as I can see have also given reasonable responses to those of the other reviewer. In particular, their successful search for a theoretical basis of the Freundlich isotherm to explain the log/log concentration dependence is convincing. It is a remarkable coincidence of timing that this new treatment, and in particular the 2022 paper, was published so long after the original isotherm but just in time for the current manuscript. Similarly, I am content that the other questions I asked were dealt with appropriately. I would now support publication of the manuscript in it current form.

Response:

We are delighted to see that this reviewer is satisfied with our revised manuscript. We like to thank the reviewer for his/her insightful comments in the revision process which significantly improved the quality of our submission and deepened our understanding of the mechanisms in this fascinating single molecule Raman spectroscopy for sensitive quantifications.

Response to Reviewer 2:

Comment 1: The authors have responded appropriately to all of the prior reviewer comments.

Response:

We thank this reviewer for approving our responses and revisions. We also wish to express our heartfelt appreciation to this reviewer for his/her generosity, tolerance and patience. With his/her critique, we were able to progressively improve the quality and the clarity of our manuscript to reach the end.

Comment 2: The word 'solution' is used within the paper when 'suspension' is more appropriate. The authors should use a global search to verify that the proper term is used whenever necessary.

Response:

We appreciate this advice. We have now made relevant changes so that the implication is clearly conveyed. The following is a summary of these changes (underlined: changed from solution to suspension):

(Line 68-69 Page 4):

"In our system, we used a quartz capillary (I.D.: 1 mm) containing 10 μL of a metallic colloid suspension to generate enhanced Raman spectra from target molecules."

(Line 77-78 Page 4):

"the mono-dispersed Ag colloids exhibit homogeneous distributions throughout the entire suspension…"

(Line 87-88, Page 5):

"Hya-Ag colloids at this concentration remained monodispersed in suspension over time…"

(Caption of Figure 1):

"The SERS spectra are collected with a quartz capillary containing a suspension of metallic colloidal nanoparticles and target molecules."

(Caption of Extended Data Figure 1)

"Shown are the colloidal suspensions contained in the vials…"

Comment 3: lines 122-124. The authors should also note that the analyte in question must associate with the colloids. As a surface enabled technique such association is an inherent requirement. Changes in surface functionalization may be required to get a particular analyte to associate the colloids.

Response:

We agree with the reviewer's point. We have now added few more sentences to make this important point. These additions (underlined) are summarized below:

(Line 111-114, Page 6):

“It should also be noted that for any analyte to be suitable for dCERS detection, well-recognized spectral features must be present without significant complications either from the background or from other analyte species and adequate analyte-colloid interaction should be ensured by a suitable colloidal surface functionalization (Supplementary Note 9).”

In Supplementary Note 9:

“In this sense, the colloidal functionalization can be optimized to increase the probability of molecule-colloid interaction, thus improving the apparent efficiency.”