We would like to thank the respected reviewer for their valuable comments that helped us to improve the quality of this article. We would also like to state that we were unable to add all the requested details directly to the paper due to the page limitation. However, in this reply, we have tried to address all points that are raised by the reviewers to the best of our ability. Also due to small changes in the manuscript, reference numbers have shifted. We have used the reference numbers as discussed by the reviewers and offered the new reference number at the end of each reply.

Reviewer 1

> Page 3, Line 13, Right Column- Why you have specifically used 5 hidden layers? What will be the impact on performance when using less or more hidden layers and neurons?

During our initial experiments, we have studied the effect of the network architecture on the accuracy. During these studies we have found 5 layers are necessary for ANN to be able estimate the loss curve. Specifically, in SPR dataset, this requirement is evident. Higher number of layers increases both the risk of overfitting and the required number of samples to train the network, causing further problems. Similar situation exists for the number of neurons. However, it is possible that there might be a better combination of layers and neurons as we have not performed any exhaustive optimization on the neural network.

A summary of this reply has been added to the paper in Section II-B at the end of the first paragraph.

> Page 4, Line 43, Right Column- The paper cited [18] doesn't have 7 inputs as looking from the dataset file (from github) and the mauscript itself. It looks to me that there are 5 inputs. Can you recheck this?

We want to thank the reviewer for his keen eyes. Indeed the number of features were wrongly listed in the paper even though there were 5 features listed in that sentence. The mistake is fixed and highlighted in the revised manuscript.

> Also, have you used the dataset of [18] along with your own data making it finally as 1117+432 data samples. If yes, how did you merge both datasets as proposed manuscript (d1,d2,d3,etc) data sets has different inputs parameters than [18]. What happens to non-common columns like number of rings input feature as mentioned in [18].

Both datasets are treated separately during experiments. Input neurons of the ANN systems are adjusted depending on the dataset used in a particular experiment without altering the rest of the architecture.

> Figure 4c, Figure 6- Give more details about the paramaters (d1,d2,d3,refrative index,etc) at which this particular actual blue line graph is obtained.

Requested information is added to the respective figures.

> Page 5, Line 43, Left Column- You have stated that [18]'s model has been failing in Figure 4c. It looks to me that the blue line in Figure 4c is for your proposed PCF structure design, while [18]'s PCF design was a simple hexagonal structure with varying rings. Also, there was no analyte (but you have used refractive index 1.35 as shown in this manuscript) in [18]. So if you compare ML models (your's and [18]) only on the criteria of wavelength as stated in Figure 4c, then [18] model is supposed to fail. This comparison doesn't make much sense as both designs and input features are different. If you disagree, then explain this comparison in more detail and mention the design specification values. The similar comparison (your's and [18]) explanation is required for Figure 6.

As the respected reviewer pointed out, we have used our SPR dataset in Figure 4c as well as Figure 4a which includes all configurations. The aim here is to show that the changes in the state-of-the-art is necessary to enable the ANN architecture to handle more complex systems. We should point out in here that this improved network is not the focus of our research, it is merely a prerequisite to the more important contribution of GAN phase, which in turn allows us to tackle this more difficult problem. In Figure 6, the comparison is drawn using the proposed architecture with and without GAN phases. The method in [18] is excluded in this graph as the improvement that can obtained by applying GAN phase to the network proposed in [18] does not yield any net benefit. In the text below we would like to explain our rationale for modifying the state-of-the-art to the proposed architecture.

The newly added average MSE values in the Figure 4 shows that the method in [18] is actually better at minimizing the error in these particular cases. Additionally, in the detailed performance analysis table (Table II), it is evident that the method at [18] and proposed method without the GAN phase has comparable results. However, Figure 4c shows that the error rate of both methods are accumulated differently. Even with the limited samples, the method at [18] is a good predictor of the real curve, except the shape of the curve does not actually model the real output; but a very rough estimation of it. The proposed ANN structure predicts the shape of the curve much better, however, it clearly requires more samples to fit that curve to the real output. Therefore, method at [18] cannot be refined further with less-than-ideal generated samples, on the other hand, the proposed architecture can benefit from increased number of samples. This was our aim when designing the proposed architecture. In fact, in our experiments, we have found that GAN phase has nearly no positive effect on the method proposed in [18]. We have excluded these experiments from the paper as it is already over the page limit. One last point we would like to show is that in Table II, in the SPR dataset, the reduction in error achieved

by changing the network is less than 10%, however, when GAN phase is also included, reduction reaches 68%.

> Page 5, Line 46, Left Column- "This can be linked to the fact that this method uses full-batch learning and therefore under-fits the model[46]"- There is no surity of this reason if the above comment/query in line 43 is logical. So reframe/reconsider this statement.

We have tried both full-batch and mini-batch learning models during our experiments. Mini-batch training generally had an edge over the full-batch training during these experiments. However, we agree with the reviewer that without a comprehensive experiment and its reported results, this claim has no basis. Therefore, we have decided to retract this statement.

> Page 7, Line 45, Left Column- "In our experiments, due to optimized dataset the accuracy of the method proposed in Ref. [18] is also improved."- Following from the previous comments- if the comparison is not for the similar designs then you can't state this improvement. So please elaborate on this.

We thank the respected reviewer for catching this sentence. We believe the meaning of this sentence has been completely altered during language editing. We have fixed this mix up.

> Page 7, Line 40, Right Column- Recurrent Neural Networks (RNN) generally rely memory to recognise patterns like in speech recognition or text generation and tells what should/will come after. How do you state/propose here than RNN can be used to discover optimal geometrical shapes?- because all the geometrical shapes might be independent of each other.

After an internal debate, we have understood the concern of the respected reviewer and decided to drop this claim.

> Page 7, Line 43, Right Column- Github link for the code is broken. I hope you will correct it. Looking forward to check your code along with data samples you have used to obtain results in this manuscript.

The given link is our current working repository that includes the paper. We will alter and publish the repository after the acceptance according to the requirements of the journal preprint policies.

Reviewer 2

> Please, reread the paper to correct some misspelling words. For example, Abstract (line 3): should it be "researchers"? Another example in "Introduction", page 1, column 2, line 35: "Snell's Total" instead of "Snell?s Total".

We would like to thank the respected reviewer for his help in finding these problems. We have carefully reread the manuscript and corrected spelling and grammatical errors

> The authors wrote, "The most important contribution of this research is the use of the GAN phase, where the available data is expanded to be used in the training phase". Why use a Generative adversarial network? What would be the advantages and disadvantages of a GAN approach over other possible data generators, such as evolutionary mutation functions? The authors should provide a comparison methodology and discuss differences between GAN and other data generators.

Respected reviewer poses a very important remark. Our primary aim in this paper is to introduce data augmentation to this field as data augmentation has never used in this field before. This also means that there are no prior experiments to draw comparisons to. We have selected GAN as it is shown to be effective in many fields. We do believe the alternative generative methods may have similar results and should be explored. However, each of these methods should be explained, implemented and optimized for this problem. This will increase the scope of this study substantially and will require us to expand the paper further. Even though the former is possible; due to the page limit put forth by the journal, the latter is not.

> How did the authors find the ANN architectures? How long did it take to find the ANN designs suitable for the problem being addressed?

We have used the networks described in [18] as a starting point and applied changes that are known to solve the issues that are surfaced when this method is applied to a more difficult dataset. While we cannot give an exact time frame to the solution, we have spent at least several weeks adapting the ANN for this difficult problem.

> Which kind of performance metric the authors are measuring in Table II? Is it MSE? Please, define it.

We would like to thank the reviewer for noticing the missing metric in the table, it is indeed MSE. We have updated the table title to include the metric.

> Please, define which is RI (page 4, column 1, line 38).

We have added the definition of RI in the manuscript.

> How do the authors know the ANN is not overfitting in Fig. 3? Have the authors employed any validation during training? Please, provide the ANN models' performance on the validation by showing the validation MSE during training.

In original manuscript we have omitted the validation graph to save space. We have included the validation curve in the revised manuscript as Fig. 4.

> Please provide the test MSEs for comparison between the proposed method and the ref. [18] (illustrated in Fig. 4).

The requested MSEs are provided on the graphs in the new manuscript.

Reviewer 3

> I suggest that the authors give more information about the stopping criteria of the GAN. As the author pointed out, human judgment has an important role in the generation of images and audio. So, could the known physical laws about PCF and SPR be used to replace the human role in the GAN system to improve the performance? How does the author verify the validity of the augmented samples?

We are thankful for the reviewer's valuable comment. This is the exact reason we have used WGAN. WGAN does not require human judgment, instead, it uses an adaptive stopping criteria. It might be possible to embed simulation approach into GAN training phase to offload judgment duty. This approach requires an extensive study of simulation systems and further experiments to test the viability of this approach. However, authors will consider respected reviewer's suggestion for a future study.

WGAN features its own loss function to determine stopping criteria. During the experiments, we have observed that the loss curve of WGAN converges to zero during the training. In addition to this observation, to ensure the generated data is relevant, we have filtered out the generated samples that fall outside the bounding box of the actual samples.

> In GAN and ANN, how many training samples and test samples are used respectively? How are the test samples selected?

We are sorry for not openly stating the requested information in the manuscript. We have expanded the information regarding the training and test samples in the fourth paragraph of the Experimental setup section. We include this information below for easy access:

For the SPR set: we have used 9 fold testing, we used 336 samples for training, 48 for validation and 48 for testing. Please note that each geometric configuration has 16 data point per analyte making up to 48 for the 3 analytes used. In this experimental setup, each configuration set (48 samples) are tested separately in its own fold and the reported results are the average of these 9 folds.

For the PCF set: we have used 10 fold testing, 112 samples are used for testing, 101 (10% of training) samples are used for validation and the rest is used for training. For this 10 fold testing, we have randomly shuffled the samples into 10 bins, each bin is used for testing over these folds and reported results are the average of these 10 folds.

> In Fig 5, the MSE increase when the augmented samples exceed 1000. Why does more samples cause performance degradation?

Data augmentation is not perfect, the augmented data contains some degree of error. When the number of samples are too low, such is the state in SPR dataset, the performance loss due to low number of samples massively outweighs the error introduced by the augmented samples. However, as the number of generated samples increase, this error accumulates and starts to reduce the performance of the overall system. We would like to point out that while 1000 augmented samples achieve the best result, it requires 3000 or more augmented samples to perform worse than no augmentation. We have also added more data points to Fig 5 to answer this question with surety.

> In Fig 6, in the case of SPR, the calculated loss spectrum has a large deviation under some specific parameters. Could the authors discuss the reasons for the deviation and how to judge the validity of the calculation result?

There are two reasons for the deviation in Fig 6a. The first and most important is the fact that SPR dataset experiments is much more difficult compared to PCF dataset experiments. The most important difference here is the testing strategy. In PCF dataset, the testing samples are selected randomly from all the samples in the dataset. However, in SPR dataset, testing samples are from a completely new geometric configuration, in other words, we never show the testing configuration to the ANN system and expect it to estimate its loss function. Additionally, SPR dataset contains far fewer samples. The second reason for the variation is the range of the experiments, SPR experiment output is constrained in a small region, accentuating the deviation. There is certainly more variation the SPR dataset, however, the difference is not as high as Fig 6 suggests.

Finally, it is possible to verify the results from Fig 6c as these are two of the individual geometric configurations that the ANN systems estimated. Even though there still is a deviation from the ground truth, ANN trained with the help of augmented samples is good at estimating the location of the peak.