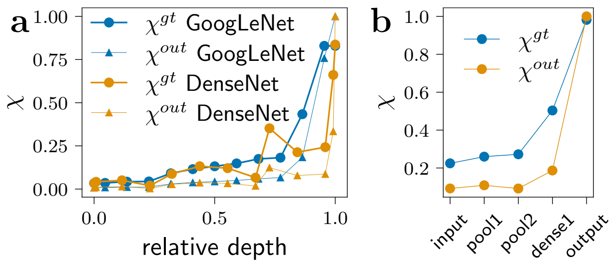

We thank all the referees for their helpful comments and the constructive feedbacks. All the referees agree that the paper is a relevant and meaningful contribution. Even the most negative referee (**R3**) recognizes that the results we obtain are "intriguing". **Our main result is the** observation that the hidden representations become ordered via a **sharp transition near the end of the network** which is sharper for deeper networks and for more complex datasets. **R2** recognizes the novelty of this result, but then s/he claims that *"it is unclear how unusual it might be"*. This is presented as a weakness, but if we are the first to make such an observation and present quantitative evidence supporting it, we believe this should be considered a strong point, even if the results are compliant with the intuition of the referee. Importantly, to address the concerns raised by **R2** and **R4** about the usefulness of our results, **we point out at least three possible practical applications of our findings**. **1)** The analysis of the density peaks hierarchy can allow an optimal truncation of the network in transfer learning schemes, exploiting the semantic hierarchy. **2)** The profiles of χ^{gt} can be used to check if the architecture is oversized in relation to the complexity of the classification task. **3)** Our observation on the existence of a sharp transition and our tools to characterise it could help in designing better performing architectures and training schemes. For instance the transition could be facilitated by "seeding" it using a metric loss function similar to that used for Siamese and Triplet networks [SIMBAD, V 9370, pp 84-92, 2015]. We will clarify these (overlooked) implications in the revised manuscript.

Following **R1** and **R4 we corroborated our results by performing new tests** on GoogleLeNet and DenseNet121 pre-trained on Imagenet (Fig. a). We also trained a small convnet on the UrbanSound8K dataset reaching a 88% test accuracy (Fig. b). In both cases we found the same characteristic transition curve shown in the manuscript (Fig. 2c). **R1** suggests that more categories could be analysed to improve the robustness of our results. This would be impractical with current HPC infrastructure. However, as also acknowledged by **R4**, the scaling analysis done in A1 makes us very confident that the results would not change significantly using more points. **R1** and **R3** mentioned the possibility of linking more strongly to existing literature in the field. Thanks to the useful direction indicated by **R1**, we found two very revant works: the network dissection analysis of [CVPR, pp. 3319-3327, 2017] and the linear probe analysis of [arXiv:1610.01644, 2016]. These works tackle the challenging problem of understanding the hidden representations of deep CNNs, but our analysis is complementary to theirs because **1)** the analysis of data probability densities has never been performed and, as a consequence, **2)** the results we obtain (lines 65-77 of the manuscript) have not been previously reported. A critical discussion of our work in comparison to the above references will be added to the manuscript.

R3 is particularly critical of our submission. Her/his most important concern is that *"it is not sufficiently clear that the method from [14] is yielding valid results"*. It is important to stress that the main assumption of the Density Peak clustering [Science, V. 344, p. 1492, 2014] (i.e., that *"the density peaks are surrounded by neighbors with lower local density"*) has been verified by several groups independently, and that the method has been tested and used extensively receiving thousands citations. One of the goals of our work is demonstrating that this approach is useful also for analyzing the activations in a DNN. In addition, as indicated by **R1**, **R2** and **R4** our work is fully reproducible since: **1)** we provide a self-contained and easy-to-run jupyter notebook in the SM for generating the main results of the paper and **2)** a detailed explanation of the method can be found in the literature.

In connection to Fig. 4a **R3** also remarks that *"low dimensional embeddings of high dimensional structures can be misleading"*. We stress here that for our analysis **we *do not* require any low dimensional embedding of the data**, and this is one of the greatest advantages of the method we use. In other words, Fig. 4a is *not* used as source of evidence for our claims but only to aid the visualisation of the results which are obtained independently through the analysis of the density peaks. **R3** raises a concern about the invariance properties of the overlap χ. Equation (1) shows that χ is computed from the product of two adjacency matrices which are built using the euclidean distances between images, as such they are invariant to orthogonal transformations but not to any arbitrary linear transformation of the activations.

A third concern of **R3** is that *"Any nucleation seems to be happening only for the last few layers (142-153)[...] it might simply be a consequence of learning more complex features*". We explicitly state in the manuscript that the transition from disordered to ordered neighbourhoods is a consequence of progressive learning. However, what we call "nucleation" **does not happen progressively in all last layers, but as a sudden change** immediately after layer 142. This, in our opinion, is a highly non-trivial result.

R3's concern on Hypothesis 4 seems more like a positive remark than the description of a weakness.

Regarding the question of **R3** and **R4** about Fig. 2a, we chose to display evenly spaced the pooling layers and the outputs of the four ResNet blocks, since these are "architectural milestones" where the networks downsample the images. On the contrary we computed $\chi^{l,l+1}$ between each couple of layers, this is why the number of measurements is different in Fig. 2a. **R4** Fig. 2b is indeed the histogram over all the data instances without the outer summation and division of Eq. (1). We will clarify this in the final version of the manuscript.