COMMUNICATIONS PHYSICS

ARTICLE

DOI: 10.1038/s42005-018-0060-1 OPEN

Machine learning classification for field distributions of photonic modes

Carlo Barth[1] & Christiane Becker[1]

Machine learning techniques can reveal hidden structures in large amounts of data and have the potential to replace analytical scientific methods. Electromagnetic simulations of photonic nanostructures often produce data in significant amounts, particularly when three-dimensional field distributions are calculated. An optimisation task, aiming at increased light yield from emitters interacting with photonic nanostructures, enforces systematic analysis of these data. Here we present a method that combines finite element simulations and clustering for the identification of photonic modes with large local field energies and specific spatial properties. For illustration, we use an experimental–numerical data set of quantum dot fluorescence on a photonic crystal surface. The application of Gaussian mixture model-based clustering allows to reduce the electric field distributions to a minimal subset of prototypes and the identification of characteristic spatial mode profiles. The presented clustering method potentially enables systematic optimisation of nanostructures for biosensing, bioimaging, and photon upconversion applications.

[1] Helmholtz-Zentrum Berlin für Materialien und Energie, Albert-Einstein-Str. 16, 12489 Berlin, Germany. Correspondence and requests for materials should be addressed to C.B. (email: christiane.becker@helmholtz-berlin.de)

Machine learning is a rapidly developing discipline which uses statistical approaches to learn from data without explicitly rule-based programming. Driven by today's massive increase in data amounts, the related techniques are extended and improving at a rapid pace[1]. Machine learning is currently applied to all aspects of science, from not only health sciences and psychology[2,3], biology[4–6] and environmental[7] and material sciences[8,9], but also to matters of everyday life from online security to finance and insurance. While supervised learning has led to breakthroughs in computer vision[10] and speech recognition[11], unsupervised learning is expected to become far more important in the future[12]. The latter techniques, such as clustering[13–15], allow for the recognition of patterns in unlabelled data and can therefore reveal a hidden structure. They have been successfully applied to, e.g., anomaly detection[16,17] or genetics[18].

In the field of nanophotonics, increasing computer power, storage space and data throughput, as well as improvements in modelling techniques greatly accelerated all-numerical system design. For nanostructures three typical optimisation tasks are met: (1) simple design, in which scalar parameters are optimised for a scalar output (e.g., lengths/refractive indices to reflectivity); (2) inverse design, which deals with multivariate parameters to optimise a scalar output (e.g., permittivity distribution to reflectivity); and finally (3) qualitative design, where scalar parameters are varied to optimise multivariate outputs (e.g., lengths/refractive indices to three-dimensional (3D) field distribution).

Simple design tasks are generally solved by simulating the system for many different parameter combinations (i.e., grid search) or by applying function minimisation routines. More sophisticated techniques such as the reduced basis method[19] for finite element method (FEM) simulations have successfully been applied to speed up this optimisation process for large parameter spaces. Inverse design tasks introduce a high-dimensional input parameter space, typically by allowing for arbitrary changes in the permittivity distribution $\epsilon(\mathbf{r})$ of the nanostructure. Machine learning techniques have successfully been applied for this purpose, mainly using genetic algorithms[20–26]. Simple and inverse tasks have in common that they possess a scalar measure of success, i.e., they can be seen as minimisation problems. The machine learning approach in inverse design therefore belongs to the field of supervised learning (more specifically regression). The third design task introduced above substantially differs in the way that the system should be optimised for a high-dimensional output. Due to the inaccessibility of a scalar success metric, we denote this problem as qualitative design. This is for example the case if the 3D spatial distribution of the electromagnetic fields has to be taken into account. Usually, such problems are solved by appropriate visualisations. However, since any change in the input parameters leads to a change in the high-dimensional output, the data amounts quickly become extremely large. We will demonstrate below that clustering techniques of the field of machine learning are able to overcome these issues by reducing the output dimensionality.

As indicated before, an example of qualitative design is to optimise a photonic nanostructure, e.g., a photonic crystal (PhC), for an appropriate spatial field distribution. This is of high relevance whenever an interaction of the field with a (potentially vague) particle distribution is present, e.g., for emitters on nanophotonic surfaces or emitters embedded into the nanostructure. PhC slabs exhibit a phenomenon called leaky modes: resonances that can be excited using external radiation[27–31]. Leaky modes have been used to improve various applications (e.g., light trapping in photovoltaic devices[32–36], light-emitting diodes[37,38]), but can also affect near-surface emitters, such as quantum dots (QDs), atoms, or molecules. Especially in the life sciences, the applications range from PhC enhanced microscopy and single molecule detection to enhanced live cell imaging, DNA sequencing, and gene expression analysis[39–42]. Besides the rather well-investigated extraction enhancement effect[28–30,37,43–46], the excitation enhancement effect[41,47–52] increases the stimulated emission rate of the emitters by enhanced near-field energy densities of leaky modes in the absorption wavelength range. To optimise photonic nanostructures for excitation enhancement, it is therefore inevitable to take the 3D spatial electromagnetic field distribution into account.

In this study we present a powerful technique based on clustering for the classification of 3D electromagnetic field distribution data. We directly apply the technique to a specific data set of our previous publication on fluorescence enhancement of lead sulphide (PbS) QDs on a silicon PhC slab surface[53] with the electric field distribution data generated by a commercial finite element Maxwell solver[54] (see Methods section and our previous studies[19,55,56]). We first reconsider the experimental and numerical results of the previous study[53], highlighting aspects which were left unexplained by the prior analysis technique. Afterwards, we introduce the clustering technique and apply it to systematically analyse the 3D field energy distribution properties. The distributions are classified by assigning them to distribution prototypes which are consulted as representative solutions to fully explain the effects observed in the experiment. We further consider a mathematical method based on silhouette coefficients[57] to assess the clustering result. Based on these analyses we explain how the method enables to solve complex optimisation tasks with high-dimensional output. Further, the practicability of the characteristic spatial mode profiles is discussed referring to possible emitter geometries in biosensing and bioimaging applications.

Results

Description of sample geometry and underlying data set. The field enhancement effect on photonic nanostructures is sketched in Fig. 1a, depicting emitters (black dots) that interact with the electric field $\mathbf{E}(\mathbf{r})$ of a leaky PhC mode (redish colours) excited by an external laser source. The illumination conditions introduce four parameters: the laser wavelength λ, the laser polarisation $\mathcal{P}$ (transverse-electric (TE) or transverse-magnetic (TM)), the polar angle θ with the plane normal, and the azimuthal angle ϕ used to define the high-symmetry direction ($\Gamma - M$ or $\Gamma - K$). The latter is also indicated in the scanning electron microscope image of the silicon PhC slab, a hexagonal nanohole array with 600 nm periodicity and slab thickness of 115 nm (Fig. 1b). The experimental fluorescence enhancement data underlying this computational clustering study have been measured on such a PhC slab coated with PbS emitters. A 3D schematic of the unit cell used for simulation of the corresponding numerical data set is depicted in Fig. 1c. It includes the symmetry *xy*, *xz,* and *yz* planes used for electric field export.

As mentioned, the energy density of the electric field of the leaky modes, $w_{\mathrm{lm}}(\mathbf{r})$, can be larger compared to the energy density of the incident plane wave, w_{pw}, known as field energy enhancement ($w_{\mathrm{lm}}(\mathbf{r})/w_{\mathrm{pw}} > 1$). To study this effect in large parameter spaces we define the volume-integrated field energy enhancement

$$E_{+} = \frac{1}{w_{\mathrm{pw}} V_{\mathrm{sup}}} \int_{V_{\mathrm{sup}}} w_{\mathrm{lm}}(\mathbf{r}) \mathrm{d}V_{\mathrm{sup}}, \qquad (1)$$

where V_{sup} is the volume of interest. In our case V_{sup} is the superspace of the computational domain and defined by the photonic crystal hole and a layer above the silicon with height $h_{\mathrm{sup}} = 250$ nm, as indicated by the yellow dashed line in Fig. 1a. The energy density of the plane wave has no spatial dependence

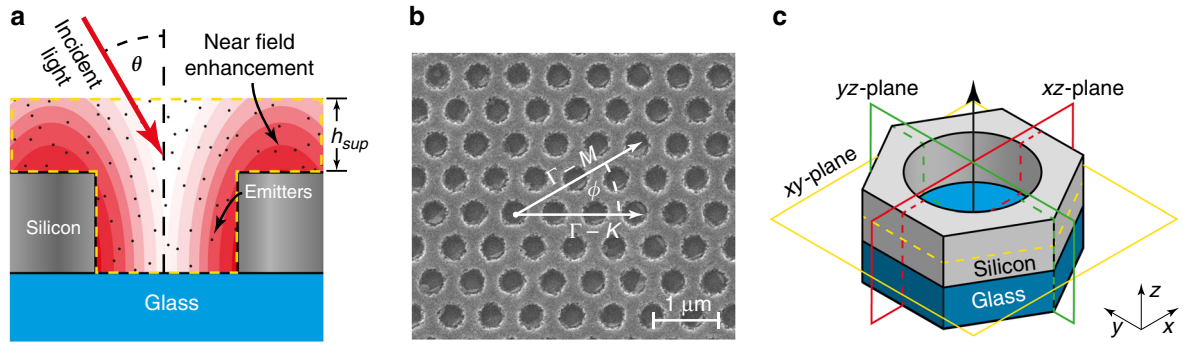

Fig. 1 Overview of the photonic nanostructure. **a** Light incident on a silicon photonic crystal (PhC, grey) on glass (cyan) excites a leaky mode that exhibits enhanced electromagnetic near-field energies (redish colours) in the superspace volume (marked by the yellow dashed line). Emitters (black dots) in the vicinity of the PhC surface interact with the local electric field distribution. **b** Scanning electron microscopy image of the silicon PhC sample with denoted high-symmetry directions $\Gamma - K$ and $\Gamma - M$. **c** A unit cell of the PhC system as used in the simulation. Yellow, green, and red rectangles mark the planes used for the field export

and is proportional to the amplitude of the electric field, $\mathbf{E}_{\mathrm{pw},0}$, and the refractive index n of the surrounding medium, i.e.

$$w_{\mathrm{pw}} = \frac{\epsilon_0}{4} n^2 \left\| \mathbf{E}_{\mathrm{pw},0} \right\|^2. \quad (2)$$

In Fig. 1a uniform random distribution of emitters is shown as an example. However, depending on the coating process, emitters might have a very specific spatial distribution in a real application, e.g., a monolayer attached to the surface, or a higher concentration inside the holes, or at the plateaus between the holes. Consequently, the spatial distribution of the energy density $w_{\mathrm{lm}}(\mathbf{r})$ becomes a determining factor and, therefore, the integrated field energy enhancement E_+ is not sufficient to quantify the effect on the emitters. An optimised design for an application as sketched in Fig. 1a can hence be achieved by identifying a mode which has: (i) a large volume-integrated field energy enhancement E_+ and (ii) an appropriate spatial field energy density distribution overlapping the locations of the emitters, at the same time. Task (i) is a simple design task, as defined in the Introduction, while (ii) is a qualitative task, i.e., an optimisation of a multivariate output.

Figure 2a, b repeat the main findings of the prior fluorescence enhancement study[53]. Figure 2a shows the fluorescence enhancement (F_+) maps obtained by tilting the QD-coated PhC sample along the respective high-symmetry directions of the irreducible Brillouin zone ($\Gamma - M$ or $\Gamma - K$, adjusted using ϕ), and by using TE or TM polarisation $\mathcal{P}$ of the incident laser radiation. Each measured spectrum (for a single incident angle) was first integrated over the fluorescence peak from $\lambda = 1200$ nm to $\lambda = 1700$ nm and normalised to the measured incident laser power and the absorption profile of the QDs, yielding the fluorescence F. A minimum estimate for the fluorescence enhancement F_+ is obtained from dividing by the minimal value in each of the maps. The maps feature regions of enhanced fluorescence caused by increased energy densities of the fields at the emitter positions.

Figure 2b maps the electric field energy enhancement E_+ integrated over the simulated superspace volume V_{sup}, which contains the QDs (see Eq. (1)). The E_+ maps exhibit clearly visible bands of strong field energy enhancement which partly correspond to regions of high measured fluorescence F_+. Some deviations are caused by a Q-factor mismatch between the spectral bandwidths of the leaky modes and the excitation laser source. However, a few features of the measured F_+ maps remain unexplained, for example, the declining band after the anticrossing point, which is visible in the E_+ map for the $\Gamma - K$, TE configuration, but missing in the corresponding experimental fluorescence enhancement F_+.

The E_+ results solve task (i), as described above. Task (ii) potentially enforces to take into account 3D field distribution data of all combinations of the illumination condition parameters λ, $\mathcal{P}$, θ, and ϕ. If the number of considered wavelengths N_λ and the number of angles N_θ becomes large, it is no longer feasible to directly visualise all the 3D field distributions for all points in the λ–θ maps shown in Fig. 2. It is hence necessary to reduce the amount of field distribution data in an appropriate way. One possibility to achieve this reduction is to pitch on specific wavelengths and incident angles for which the field distribution is evaluated, as it was done in the previous study[53]. This way, however, information is mainly gained at random, so that general trends might be overseen. A more systematic approach is to cluster field distributions which are similar, and to therefore derive typical distributions (i.e., distribution fingerprints). It is known that a certain undisturbed photonic band in the leaky-mode regime will not significantly change its symmetry properties when crossing the λ–θ space[31,58], as will be explained in more detail below. As a result, the entirety of field distributions is composed of a finite set of patterns which are basically caused by the finite number of bands. This feature space can efficiently be partitioned into the typical patterns using clustering techniques.

Introduction and justification of the clustering technique. The E_+ maps given in Fig. 2b only provide information about the volume-integrated field enhancement over a characteristic volume V_{sup}, marked by the yellow dashed line in Fig. 1a. Therefore, regions of high E_+ can be regarded as a necessary condition for fluorescence enhancement, but not as a sufficient one. A high E_+ without a corresponding fluorescence enhancement F_+ hence indicates a lack in the spatial overlap of the emitters with the regions of enhanced field energy density.

However, it is known that bands of the photonic crystal have well-behaved spatial properties when varying the k-vector between two high-symmetry points of the irreducible Brillouin zone[31,58]. More specifically, the modes belong to the same symmetry point group as the system seen from the point in k-space, i.e., they exhibit the same spatial symmetry. Consequently, it is theoretically justified to expect that the spatial properties of the bands only change smoothly with θ. For a fixed high-symmetry direction, e.g., $\Gamma - K$, we only expect two types of solutions which are either regions that correspond to leaky-mode bands or regions that are off any photonic band, and therefore correspond to the continuum of radiation modes. The regions

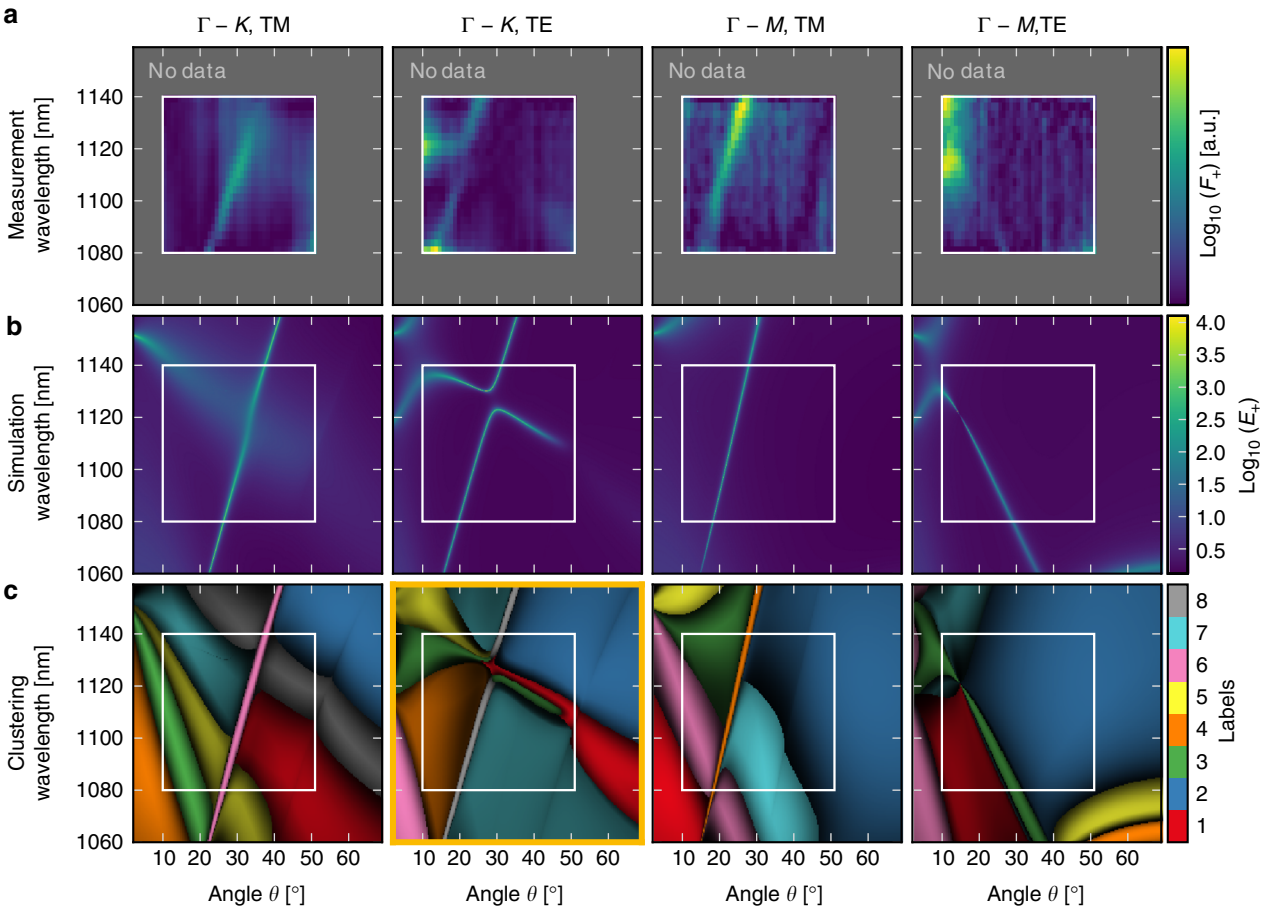

Fig. 2 Comparison of measurement, simulation, and clustering results. **a** Measured fluorescence enhancement F_+ of lead sulphide quantum dots on a silicon photonic crystal slab, a 600 nm-periodic hexagonal nanohole array, as a function of vacuum wavelength and incident angle θ of the laser source (logarithmic colour scale). The columns correspond to the four combinations of sample orientation ($\Gamma - M$ and $\Gamma - K$) and source polarisation (TE and TM). **b** Simulated volume-integrated electric field energy enhancement E_+ for the same conditions as for the measured fluorescence enhancement F_+. The volume V_{sup} is defined by the hole and a 250 nm layer above the silicon photonic crystal. The white lines mark the experimental data limits. **c** Classification maps depicting the cluster assignments (labels) using different colours independently for each plot, and the respective silhouette coefficients using alpha-blending with a black background (colour bar omitted). More saturated colours denote larger silhouette coefficients. The classification map for the $\Gamma - K$, TE configuration is highlighted by an orange frame as it is most detailed analysed in the study. Note that (**a**, **b**) repeat the same results as already shown in ref. [53] for a larger angle and wavelength range. Reproduced from ref. [53], with the permission of AIP Publishing

of the radiation modes are expected to exhibit solutions that resemble plane waves, i.e., show oscillatory behaviour in the the exterior domain. All things considered, only a small number of different spatial symmetry types is expected, which is of the order of the number of bands that cross the parameter scan window.

This is where machine learning comes into play. If we consider the specific 3D electric field distribution for a single illumination setting as a sample, and the electric field values of each point in the considered volume as features, then clustering techniques are able to subdivide the entirety of field distributions into a finite number of field distribution prototypes. This approach is reasonable because the data range is expected to contain a finite number of typical field patterns, and each of the real field patterns can be identified with one of those prototypes. Moreover, these prototypes have a sufficient uniqueness, e.g., they considerably differ in their symmetry properties. The 'Clustering of electric field data' in the Methods section gives a detailed description of how the clustering is performed. In a nutshell, for each illumination condition, i.e., a set of $(\mathcal{P}, \phi, \theta, \lambda)$, the electric field strength $E = (E_x, E_y, E_z)$ is derived from an FEM simulation. It is sufficient to export the fields on symmetry planes to reduce the data volume, for which we use the *xy*, *xz* and *yz* planes marked in Fig. 1c. The validity of this approach was tested using a comparison to full-3D exports using a smaller data set. Note that, in contrast to the the volume-integrated field energy enhancement E_+ which is calculated in the volume V_{sup} (Fig. 1a), the fields for the clustering are also considered in the dielectric materials (silicon PhC and glass substrate, Fig. 1c). To account for the different cluster sizes (narrow bands) and unknown cluster shapes in the data set, the flexible Gaussian mixture model (GMM) clustering technique is used (see 'Gaussian mixture model clustering' in the Methods section for details), implemented in the Python library Scikit-learn[59]. We compared the GMM clustering technique to various other clustering algorithms implemented in Scikit-learn, namely *k*-means, density-based spatial clustering of applications with noise (DBSCAN), and spectral clustering. The characteristics of these techniques are very different and cannot be covered in the scope of this study. While *k*-means tends to ignore the narrow bands, DBSCAN and spectral clustering did not result in classifications in-line with the physical expectation at all. We suppose that the latter two approaches do not adapt well to these kind of data or may require very careful parameter tuning. Note that the data and the code are published and allow to compare different techniques and

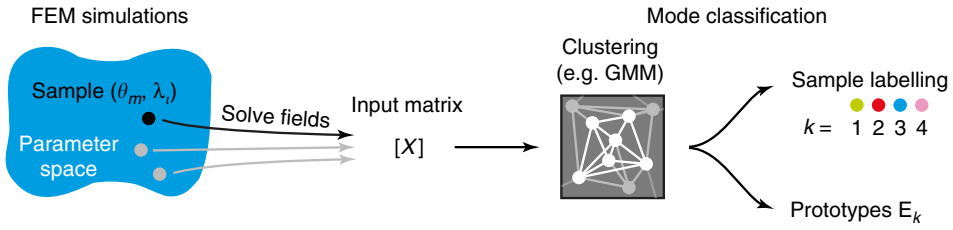

Fig. 3 Flow diagram of the mode classification process. For specific samples of a parameter space (cyan area; spanned by illumination conditions, geometrical parameters, ...), the electric (or magnetic) fields are solved using finite element method (FEM) simulations. From these solutions an input matrix for the clustering algorithm is composed and used to perform the clustering itself. From the resulting model the classification (i.e., sample labelling) and the field prototypes $\mathbf{E}_k$ for each cluster k are gained

parameter settings, so that these results are omitted here (see the data and computer code availability statements below).

From the clustering itself two characteristics can directly be gained: the classification, which labels each observation with a cluster index i, and the distribution prototypes, usually denoted as cluster centres in the general clustering literature. The latter are the average of the electric field distributions (on the chosen planes) of all samples that belong to a specific cluster i. We note that we averaged the normalised input data, i.e., the exact data used for the clustering, to calculate the prototypes. The prototypes therefore represent the actual mathematical cluster centres, with the trade-off that the absolute field amplitude information is lost, because the samples are normalised individually. Another possibility would be to average the unnormalised fields, so that the amplitude information would be conserved, with the trade-off that the prototypes derived that way are not exactly the cluster centres. We settled for the normalised fields, as the amplitude information is essentially included in the E_+ maps. Figure 3 illustrates the complete mode classification process in a flow diagram. For specific samples of a parameter space (cyan area), the electric fields are solved using FEM simulations to construct the clustering input matrix. Performing the clustering algorithm results in the classification (i.e., sample labelling) and the field prototypes $\mathbf{E}_k$ for each cluster k.

As in most clustering techniques, the number of clusters must be specified in GMM clustering, so that the appropriateness of this choice has to be validated. This aspect will be covered shortly.

Classification maps. The classification can be visualised by assigning each point (θ_j, λ_k) to a different colour that corresponds to its label i. Recall that the clustering is carried out individually for each combination of polarisation $\mathcal{P}$ and azimuthal angle ϕ (=high-symmetry direction $\Gamma - M$ or $\Gamma - K$). Plotted in the same fashion as the E_+ maps of Fig. 2b, we denote the resulting figures as classification maps. These classification maps are shown in Fig. 2c. The colour scale relates the colours to the labels and, hence, identify the corresponding cluster. Note that the classification maps cannot be compared among each other, although the same colours have been used. The clusterings for the $\Gamma - K$ cases used 8 clusters, while the $\Gamma - M$ cases only required 7 (i.e., there is no grey region in these maps). The procedure of determining the number of clusters will be explained in the next subsection.

When comparing the classification maps to the E_+ maps above, a striking accordance can be observed. The narrow bands of high field enhancement in the E_+ maps correspond to narrow areas at the same positions in the classification maps. Note that the E_+ maps and the classification maps are based on very different data sets: the former are derived from a spatial integration over the electric field energy density distribution $w_{\mathrm{lm}}(\mathbf{r})$ in the superspace volume V_{sub} only (Eq. (1)), while the latter uses electric field patterns $\mathbf{E}(\mathbf{r})$ on planes that include the PhC and glass domains. When observing the regions off the leaky-mode bands, i.e., the domains of the radiation modes, it is seen that these regions are multiply subdivided in some cases; e.g., $\Gamma - K$, TM bottom left. In contrast, other parts are homogeneous over large ranges, such as $\Gamma - M$, TE top right.

Another detail of these plots are the different levels of saturation used for each point, obtained by alpha-blending with a black background. This additional layer of information illustrates the representation quality of the local solution by the assigned cluster, as determined using so-called silhouette coefficients[57]. The silhouette coefficients provide a way to assess the initial choice of the number of clusters, and how well the samples lie in their respective clusters, at the same time. The silhouette coefficient rates how well a sample fits into its own cluster. If it is far away from all other clusters and very close to the cluster centre (i.e., prototype), the sample gets a positive rating. If the distances to a different cluster and its own cluster are comparable, it is rated with values close to zero. Finally, if it is much closer to a different cluster, a negative rating is assigned. See 'Solution quality rating using silhouette coefficients' in the Methods section for a severe definition.

In all cases, we observe that the saturation decreases at the border of two clusters. This is expected, as silhouette scores close to 0 indicate a sample which is in fact close to the border of the neighbouring cluster. It is apparent from this phenomenon, and important to stress here, that the clustering technique is a tool. The field distribution data are not categorical, and so we expect superposed solutions which are badly represented by pure modes. That said, these intermediate parts are small, as it is seen from the saturation distributions, so that the clustering is still a valid and effective approach.

Silhouette analysis and the number of clusters. Before we investigate the field distribution prototypes, the quality of the clustering itself is evaluated using a mathematical analysis. Without loss of generality, we focus in the following analyses on the $\Gamma - K$, TE configuration and refer to ref. [60] for the other ϕ, $\mathcal{P}$ combinations. We calculate the silhouette coefficients using a scheme known as silhouette analysis[57]. Figure 4 depicts a so-called silhouette plot. The silhouette coefficients for each sample are plotted as a bar in x-direction with a length corresponding to its value (negative values point into the $-x$-direction). The samples are sorted by their silhouette coefficients, with smaller values being located at smaller y-positions. In addition, the samples are grouped for each cluster k and colour-coded using the same colours as in the classification map highlighted with an orange frame in Fig. 2c. The red dashed line marks the average of all silhouette coefficients, which is a measure for the absolute quality of the representation denoted as silhouette score. The results are the typical sails or shark fins. The width of each fin in the silhouette plots is proportional to the area of the correspondingly labelled points in the classification maps.

Considering the distribution of the silhouette coefficients, fins which are not too sharp are observed, i.e., having broad plateaus of high silhouette coefficients. There is only a minimum number of values with negative coefficients. Both arguments together give a validation for the fact that the number of clusters is not underestimated: negative values would occur if there were too few clusters, leaving back samples which do not fit in one of the classes ($s \sim -1$). Too many clusters could be identified by a large fluctuation in the fin widths. However, this does not fully apply here, as the areas occupied by the bands and the residual parts are unequal. Therefore, equally broad classes are not expected. A slightly too large number of clusters can be seen as unproblematic, because it would basically subdivide the radiation mode regions further, which are of limited relevance for the interpretation. Another point that suggests a good representation is that there are few clusters with below average silhouette scores. In summary, the optimum number of clusters equals to the minimum number of clusters for which all the bands visible in the field enhancement maps are distinguished from flat regions, and no extensive presence of negative silhouette coefficients occurs. Since this criterion is not completely rigorous, it is necessary to compare the results for different choices of the number of clusters. Using this reasoning, the optimum number of clusters was determined for each case (results omitted). We note that subdivisions of the radiation mode regions do not imply a real physical effect, but are to be considered as a numerical artefact of the approach. Especially if sharp bands are present in the parameter window, the number of clusters may have to be increased in order to detect these bands, as they take a small fraction of the entirety of samples. We tried to minimise the unphysical subdivisions by choosing a clustering technique that adapts well to the problem and by choosing proper parameters of this algorithm (especially for the covariance type in Gaussian mixture model clustering).

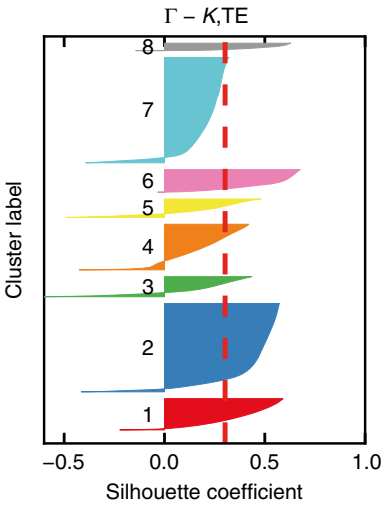

Fig. 4 Silhouette analysis plots for the for the $\Gamma - K$, TE configuration. The silhouette coefficients for each sample are plotted as a bar in x-direction with a length corresponding to its value. The samples are sorted by their silhouette coefficients, with smaller values being located at smaller y-positions. The red dashed lines mark the silhouette score

Field distribution prototypes. As the second essential outcome, the clustering procedure yields the field distribution prototypes. As the input data for the GMM algorithm have been electric field values on three planes, namely *xy*, *xz*, and *yz*, the prototype data are available on these planes as well. The prototypes for all clusters of the $\Gamma - K$, TE configuration are depicted in Fig. 5 on all three planes. For the other ϕ, $\mathcal{P}$ combinations, please see ref. [60]. The number of columns accounts for the number of clusters, and each column has a coloured edge in the top-most row that corresponds to the colour used for that label in the classification maps shown in Fig. 2c. The cluster label is further given in the title of the *xz* row. Each distribution plot depicts the electric field energy distribution $\|\mathbf{E}\|^2$ in the respective plane. The distribution plots further feature semitransparent markings for the glass superstrate (blue) and the silicon of the PhC (grey) in the case of *xz* and *yz*; and a white circle indicating the hole circumference in the case of *xy*. Recall that the colour scales do not give absolute values, as the prototypes are based on normalised data and, therefore, cannot be compared with respect to their absolute amplitudes.

For each prototype, the field energy plots on the three planes give a notion of the 3D field energy distribution. The solutions with the same label (colour) in the classification maps of Fig. 2c, which is highlighted by the orange frame, all share this distribution type. Lower saturations quantify how much the individual solutions deviate from the prototype. Clusters that correspond to leaky-mode bands with strong field enhancement, such as cluster 8 (grey), 3 (green) and 6 (pinkish), have strongly localised energy distributions as shown in Fig. 5. These three specific field distributions are also denoted as Mode A, B and C and have a maximum field enhancement at the plateau, at the

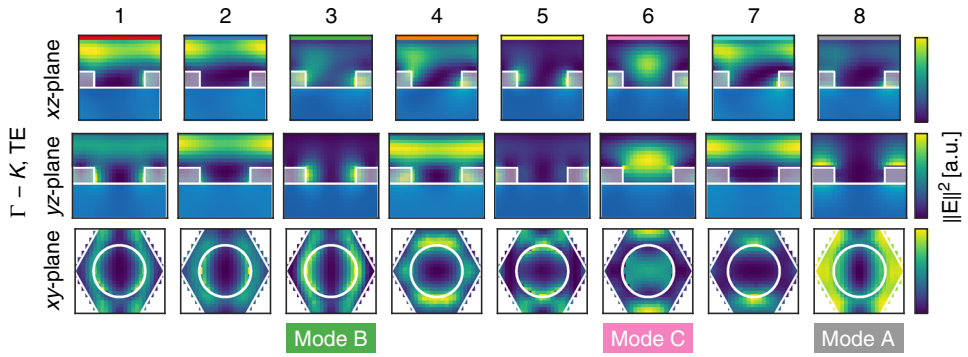

Fig. 5 Prototypes/Cluster centres for the $\Gamma - K$, TE configuration. The figure consists of 3 rows for the *xz*, *yz*, and *xy* planes, respectively (as indicated). Each of these rows has eight columns accounting for eight clusters found during clustering. Each distribution plot depicts the electric field energy distribution $\|\mathbf{E}\|^2$ in the respective plane. The distribution plots feature semitransparent markings for the glass superstrate (blue) and the silicon of the PhC (grey) in the case of *xz* and *yz*; and a white circle indicating the hole circumference in the case of *xy*. Three of the prototypes are highlighted and exhibit a characteristic field enhancement at the plateau (mode A), the flanks of the holes (mode B), and inside the holes (mode C) of the PhC slab. (Colour scales do not give absolute values, as the prototypes are based on normalised data.)

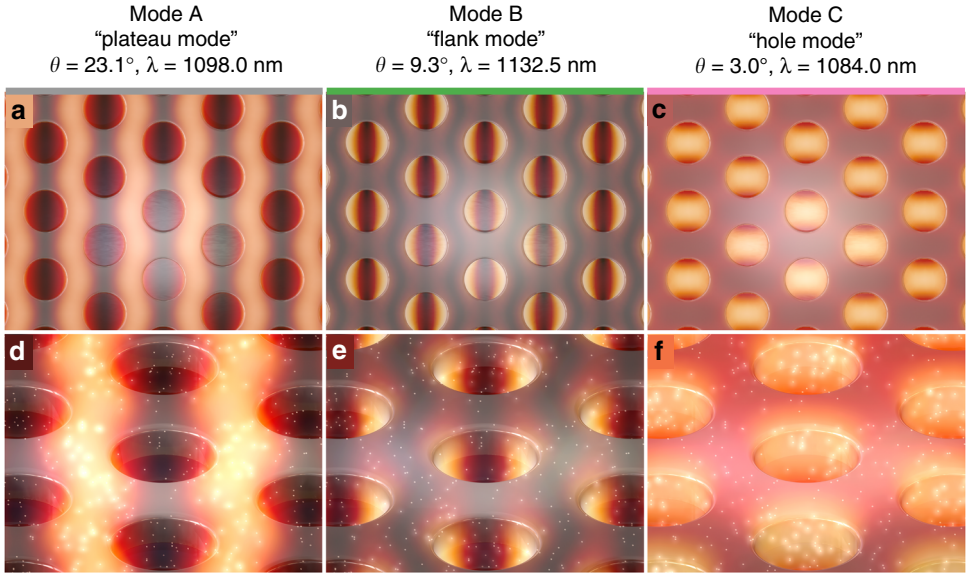

Fig. 6 Full-3D volume renderings of selected modes for the $\Gamma - K$, TE configuration. Semi-artistic ray tracing images depicting multiple periods of the photonic crystal as a greyish material. **a**–**c** Top views of the full-3D *E*-field energy density, colour-coded using a heat map (not comparable between the figures). **d**–**f** Closer views indicate the same random distribution of quantum dots (bright small spheres), emitting white light with an intensity proportional to the field energy density at their specific positions. The pairs (**a**, **d**), (**b**, **e**), and (**c**, **f**) relate to three different modes of the $\Gamma - K$/TE case, denoted as plateau mode (mode A), flank mode (mode B), and hole mode (mode C), respectively. The modes are the actual solutions from the finite element solver that have the smallest deviations from the assigned prototype (i.e., cluster centre), determined using the silhouette coefficients. Incident angle θ and wavelength λ for each mode are given in the headings. The figures use real physical proportions

flanks of the hole and inside the hole of the PhC, respectively. In contrast, clusters that belong to radiation modes have energy distributions that increase away from the PhC, e.g. cluster 2 (dark blue).

Combining field enhancement and clustering results. To explain the measured fluorescence enhancement effects shown in Fig. 2a, it is necessary to combine all information gained from the numerical analysis. This is, the volume-integrated field energy enhancement maps (E_+, Fig. 2b), the classification maps (Fig. 2c), and the prototype maps (Fig. 5). A guide on how the different aspects of the results can be connected to yield a complete interpretation may read as follows. First, select a feature in the volume-integrated field energy enhancement (E_+) map. For these features the simulation suggests a possible excitation enhancement effect. Second, check whether there is an according feature in the experimental fluorescence enhancement (F_+) map. Third, observe the corresponding region in the classification maps and determine the cluster label from the colour using the colour bar. Fourth, using this label or colour, locate the related column in the prototype map. Check if the field energy distributions on the three planes can explain the observed fluorescence enhancement.

We will analyse the results for the $\Gamma - K$, TE configuration as an example. There are two very clear bands that show anticrossing, a steeper band crossing the complete wavelength range from roughly 20° to 40°, and a shallower one coming from top left. The former is very clearly seen in the clustering by the grey region with label 8. From the prototype map it is observed that this band has a node along the *x*-direction and concentrates its energy at the flanks and the plateaus in *y*-direction. The shallower mentioned band undergoes a transition from the green cluster (label 3) to the red cluster (label 1) in the classification maps. For the green parts the energy is strongly localised at the flanks, while the red one can be identified with a radiation mode. The energy is therefore less well confined to the surface in the red case, which is exactly seen in the fluorescence maps, where only at the location of a broad green region a fluorescence enhancement can be observed, but no enhancement is seen at the prosecution of the mode at higher incident angles.

To give a clear idea of the 3D energy density distribution for three selected modes, and also to show how well the clusters match the actual physical fields, Fig. 6 shows full-3D renderings[61]. The images depict multiple periods of the photonic crystal as a greyish metal-like material, without showing a superspace material. Figure 6a–c shows top views of the volume-rendered electric field energy density colour-coded using a heat map, which is not comparable between the figures. Figure 6d–f shows a closer view and indicates a random distribution of QDs as bright small spheres, emitting white light with an intensity proportional to the field energy density at their specific positions. The QD distribution is the same for all three images. The columns relate to three different modes of the $\Gamma - K$, TE case, denoted as A, B, and C. They correspond to clusters 8, 3, and 6, as also marked in Fig. 5. The modes are the actual solutions from the finite element solver that have the smallest deviations from the assigned prototype (i.e., cluster centre), determined using the silhouette coefficients. Incident angle θ and wavelength λ for each mode are given in the headings. Note that these images have an illustrative character, but can be very helpful to imagine the actual physical situation. Modes A and B are the ones which have been discussed recently, and it is clearly seen that the former concentrates its energy at the plateaus, while the latter has high energy densities at the flanks of the holes. A third type is shown with mode C, which focusses the energy directly inside the holes. The illustrations in Fig. 6d–f give a notion of how these modes activate different QDs, depending on their position. Only a small density of QDs is used for the images for purposes of visibility, and they are randomly distributed in a layer that fills the holes and extents 100 nm in *z*-direction. Mode A very efficiently excites QDs at the plateaus, just as expected, while modes B and C do the same at the flanks and inside the holes, respectively. Consequently, these renderings completely confirm the results of the clustering approach.

Discussion

The aim of the numerical approach presented here is the systematic identification of suitable leaky modes of nanophotonic structures for interaction with near-surface emitters. For instance, (a) a monolayer of emitting species attached at the surface of the nanophotonic structure is expected to strongly interact with a leaky mode with shallow field distribution. In contrast, in an experiment with (b) emitters in a coating on top of the photonic nanostructure, the interaction with a shallow leaky mode will be rather small due to the limited spatial overlap of the mode volume with the emitting material. Here, a leaky mode with an energy density enhancement in a large volume outside the photonic nanostructure would be better suited. In a third scenario, (c), emitters fill the voids of the photonic nanostructure, for example if they are solved in a liquid solution and dropped onto the structure. In that case, leaky modes with strong field enhancement inside these voids are expected to cause the strongest effects. In the chosen data set, the fluorescence enhancement experiment of PbS quantum dots on a silicon PhC slab in nanohole geometry, the distribution of QDs resembles a mixture of cases (b) and (c). The clustering technique revealed that the modes which have the best spatial overlap with the QD distribution effectively cause the strongest fluorescence enhancement effects in the measurements.

In the previous study[53] we used a selection of a small number of points for which the field energy distributions were analysed. The clustering technique not only confirmed the results that were achieved this way, but also helped to explain complicated details, e.g., as caused by the superposition of two modes. Therefore, the clustering approach gave a much more coherent and detailed explication of the underlying physical phenomena. It emphasises the interesting parts automatically and systematically, e.g., by revealing regions of rapidly changing field distributions through individual clusters, or through large deviations from the assigned prototype. Moreover, the clustering technique seems to be applicable to even more complicated cases, e.g., in windows with more bands for which an analysis using selected points is not reasonable any more.

The presented technique composed of (i) the field energy enhancement maps and (ii) the 3D electric field distribution clustering provides a versatile tool for the analysis and design of photonic nanostructures for applications that utilise near-field enhancement effects for increased emission. For any known distribution of near-surface emitters that should be affected by leaky modes, optimum values for all relevant parameters can in principle be determined. It is, e.g., possible to define a wavelength range for the excitation of the emitters by considering their absorption properties, and to numerically calculate the field energy enhancement E_+ and field values in 3D for clustering (as provided in Fig. 2b, c, here). By choosing the mode with the largest spatial overlap of high field energy with the emitter distribution from the prototypes (as in Fig. 5), an optimum mode can systematically be determined. This process can moreover be repeated for possible geometrical parameters of the photonic nanostructure, e.g., the lattice constant, slab thickness, or hole radius. Alternatively, if the geometrical parameters should be varied extensively, the technique could be applied for an initial set of geometrical parameters to select a potential mode and to reduce the wavelength and angle window. Successively, only the field energy enhancement E_+ may be calculated in the scan over the possible geometrical parameters to determine the absolute maximum of the enhancement.

The clustering technique is extremely flexible. It is not limited to uniformly sampled feature spaces as shown in our example application. It would also have been possible to choose arbitrary snapshot points in the θ–λ space, e.g., with a higher density in regions of high field energy enhancement E_+. It is further not limited to the shown number of feature parameters; i.e., we could have added a variation of the hole diameter or other geometrical parameters as well. However, the method is even more powerful, because the trained classifier can be used to classify field distributions that it has not seen yet, known as prediction. In contrast to the clustering itself, this is a computationally cheap process, and the classifier can even be persistently stored on disk for later use. To make these considerations more clear, it would have been possible to choose a smaller number of possibly non-uniformly sampled points in the θ–λ space for efficient clustering. The silhouette analysis can be used to make sure that the number of samples is sufficient to reach an appropriate clustering result. From this clustering the prototype field distributions can be derived and the classifier can be stored to disk. Afterwards, a uniform scan over θ, λ, and other parameters that are expected to not change the field distributions considerably (e.g., hole diameter, slab thickness, refractive indices, …) could be performed. The resulting new solutions could then be assigned to the prototypes using the classifier from disk with minimal computational effort.

Numerous applications could benefit from these optimisation abilities. In the field of biosensing, photonic nanostructures have become an important platform for, e.g., label-free biosensing or for the enhancement of the output of photon emitting tags used in the life sciences and in vitro diagnostics. A recent review article[39] shows that nanophotonic-enhanced biosensors are yet extremely relevant, even commercially and potentially on industrial scale. Exploiting leaky modes with large Q-factors enables for narrow bandwidths (<1 nm) and extremely high sensitivities, e.g., for detection of disease biomarkers in serum with concentrations of ~1 pg ml^{-1}. The numerous applications that are described in the mentioned review article have in common that the nanophotonic structure is designed for a very specific mode, i.e., a specific illumination condition and a determinable distribution of the molecules/cells/virus particles in question. This is where the technique presented here could be utilised for a systematic optimisation in the design process, and hence to further increase the sensitivities of related sensors. Photon upconversion[62,63] in biomedical imaging and solar energy is another application that could benefit from the discussed all-numerical design abilities. Recent publications[64,65] demonstrate upconversion using thin emitter layers, which as well could potentially be improved using specifically tailored nanophotonic structures.

In summary, we have developed a numerical method that allows to systematically optimise nanophotonic structures pertaining to the 3D field distribution and field energy enhancement of modes. The method applies a combination of FEM simulations and post-processing using clustering. We showcased the modelling power of the method by explaining experimentally measured fluorescence enhancement of QDs on a photonic crystal slab surface. The method yielded information that was not easily accessible using, e.g., a visualisation-based analysis for selected parameter combinations, and which allowed to fully explain the experimental results. Consequently, the presented technique could be very useful for applications that utilise effects that depend on the spatial field distribution of nanophotonic modes, such as in the fields of biosensing[39,63] or spectral conversion in solar energy[62,64–66].

Methods

Finite element simulations. For simulations we use a time-harmonic FEM Maxwell solver (JCMsuite[54]) on a 2D periodic unstructured, prismatoidal mesh of the unit cell. Each simulation uses a distinct plane wave excitation corresponding to the direction of incidence, wavelength, and polarisation. We assure numerical convergence by appropriate tests, guaranteeing an accuracy of 1, by comparing the derived quantities to those calculated in a highly accurate reference solution. Quantities such as the electric or magnetic field distribution, and therefore field

energy distributions, are derived from the near-field solutions. Appropriate post-processes are directly implemented in the used commercial finite element Maxwell solver. Please consult our previous publication[53] for details of the numerical model.

Clustering of electric field data. The clustering is executed on an input matrix X of shape $N_s \times N_f$, where N_s is the number of samples and N_f the number of features. A sample is the solution for a specific set of input parameters, in our case incident angle θ and wavelength λ. The features, in the present case, are absolute values of the electric field components E_j with $j \in \{x, y, z\}$ for a number of points $\mathbf{r}_i \in \mathbb{R}^3$, i.e., of the form $\left|E_j(\mathbf{r}_i)\right|$. Consequently, if the field is evaluated at N_p points, these are $N_f = 3N_p$ features. To avoid exporting the electric field on a full Cartesian grid in 3D, which would cause huge amounts of data when trying to achieve a reasonable resolution, data are only exported on the symmetry planes marked in Fig. 1c, respectively. More symmetry planes could be used as well, but based on these three planes a reasonable classification can be reached, as tested using smaller data sets and compared to a full-3D field output. A field pattern of a single simulation holds data for each of the three spatial directions, and for each component j of the electric field (altogether a 4D data set). As each sample X_i must be a 1D row vector with observations of single scalar values $x_0, \ldots, x_{N_f-1}$, it is necessary to flatten these data sets in always the same way, yielding 1D representations of the fields. The data are moreover normalised by scaling each sample to unit norm individually. The field export is performed for each point in each map of Fig. 2b, so that the samples are unique simulations for a given direction/polarisation combination, wavelength λ, and incident angle θ. The number of samples for a single map is given by $N_s = N_\lambda \cdot N_\theta$. To give an expression for the complete input matrix X we abbreviate $\hat{E}_j^{i,m,l} = \left|E_j(\mathbf{r}_i, \theta_m, \lambda_l)\right|$, where the additional indices $m = 0 \ldots N_\theta$ and $l = 0 \ldots N_\lambda$ have been introduced, and where the hat denotes the absolute value and normalisation. The input matrix then reads

$$X = \begin{pmatrix} \hat{E}_x^{0,0,0} & \cdots & \hat{E}_x^{N_p,0,0} & \cdots & \hat{E}_z^{0,0,0} & \cdots & \hat{E}_z^{N_p,0,0} \\ & & & \vdots & & & \\ \hat{E}_x^{0,N_\theta,0} & \cdots & \hat{E}_x^{N_p,N_\theta,0} & \cdots & \hat{E}_z^{0,N_\theta,0} & \cdots & \hat{E}_z^{N_p,N_\theta,0} \\ & & & \vdots & & & \\ \hat{E}_x^{0,N_\theta,N_\lambda} & \cdots & \hat{E}_x^{N_p,N_\theta,N_\lambda} & \cdots & \hat{E}_z^{0,N_\theta,N_\lambda} & \cdots & \hat{E}_z^{N_p,N_\theta,N_\lambda} \end{pmatrix}. \quad (3)$$

For the wavelength and angle resolution, values of 0.5 nm and 0.3° have been used, respectively. For each clustering procedure the input matrix X had a size of $N_s \times N_f = 47{,}034 \times 8616$. This is a comparably large problem size, especially because of the large feature dimensionality (N_f), so that the procedure took more than 10 h on a hexa-core workstation with roughly 40 GB of memory consumption.

Gaussian mixture model clustering. Simple clustering techniques, such as the k-means algorithm[67], can be extremely robust, but also have their disadvantages. For example, k-means assumes that the clusters are circular, i.e., representable by a (hyper-)sphere in feature space. The centre of this sphere defines the cluster centre (i.e., prototype), while the radius acts as a hard boundary used to decide which samples belong to the cluster. In contrast, the GMM[67,68] is a so-called soft method. That is, a score for each cluster is assigned to the samples which account for the probability that the sample belongs to a specific cluster. In GMM clustering, the clusters are represented by Gaussian distributions of the dimensionality of the features space (i.e., N_f).

In general, a superposition of N multivariate Gaussian distributions of the form

$$p(\mathbf{x}) = \sum_{i=1}^{N} c_i \mathcal{N}_i(\mathbf{x}) \quad (4)$$

can be used to approximate almost any continuous density to arbitrary accuracy (this is intuitive with 1D Gaussians, which can fit almost any 1D signal if enough Gaussians are superimposed). Here, the $\mathcal{N}_i(\mathbf{x})$ are multivariate Gaussian distributions of the form[67]

$$\mathcal{N}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\left|\tilde{\Sigma}\right|^{1/2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T}{2}\tilde{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (5)$$

for a D-dimensional vector $\mathbf{x}$, the D-dimensional mean-vector $\boldsymbol{\mu}$, and the $D \times D$ covariance matrix $\tilde{\Sigma}$ with determinant $\left|\tilde{\Sigma}\right|$. Equation (4) is called a Gaussian mixture, the $\mathcal{N}_i(\mathbf{x})$ are called components of the mixture, and the c_i are weight factors. Loosely speaking, the distribution of sample points is fitted using a set of high-dimensional Gaussians. A GMM can therefore represent much more complex data sets and can be seen as a generalisation of the k-means algorithm for non-circular clusters. One can imagine that it would be straightforward to fit the multivariate Gaussians to a data set for which the labels are known. With unlabelled data the case is more difficult, and enforces to take into account another step. In the literature, this problem is commonly denoted as to find out which (latent) component is responsible for a certain sample, which is somehow a different way of asking to which cluster the sample belongs. However, it underlines that the GMM clustering is a probabilistic approach, because it calculates the probability that the sample was generated by cluster i for all clusters. These probabilities, which are also called responsibilities, are simply the weight factors c_i of Eq. (4). In the implementation that was utilised here, the cluster assignment is solved using a method known as expectation-maximisation[69,70]. This algorithm starts with a random Gaussian mixture (i.e., random components),which is typically initialised using a prior application of k-means to improve the convergence. In the next step it determines for each sample the probability of being generated by each component of the mixture. Based on these probabilities, the parameters of the Gaussian distributions are fitted to give the best approximation of the data by maximising their likelihood[67]. This process is executed iteratively and is guaranteed to converge to a local optimum.

Solution quality rating using silhouette coefficients. To give a definition of the silhouette coefficient, let X_i^k be a sample that was assigned to the cluster k and $a(i)$ be the average dissimilarity of X_i^k to all other members $X_{j\neq i}^k$ of this cluster. The measure for the dissimilarity is usually the Euclidean distance. Let $d(i, m)$ be the average dissimilarity of X_i^k to all members of the cluster $m \neq k$ and $b(i)$ be the minimum of $d(i, m)$ for these clusters, i.e.,

$$b(i) = \min_{m \neq k} d(i, m).$$

The cluster m for which this minimum is obtained is called the neighbouring cluster of X_i. If the number of clusters N_k is >1, we can define the silhouette coefficient $s(i)$ for the sample X_i by

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i)<b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i)>b(i) \end{cases}. \quad (6)$$

From this definition it is seen that the silhouette coefficient s is in the range $-1 \leq s \leq 1$. Values near $+1$ indicate that the sample is far away from the neighbouring cluster and accordingly fits well into its own cluster. A value of 0 indicates that the sample is on or very close to the boundary between its own and the neighbouring cluster, and negative values indicate that it might have been assigned to the wrong cluster. A sorted diagram of all silhouette coefficients can thus be used to visualise the representation quality of a clustering. In addition, the average silhouette coefficient for all samples—usually denoted as silhouette score—can be used to compare the representation quality for different clusterings, e.g., using different N_k values. It hence even provides a single numeric value for solution quality assessment.

Code availability. The complete and documented Python code to generate the presented results is available via https://doi.org/10.5281/zenodo.1234734. The code can automatically download the published data set and comes with an additional small data set for testing purposes. In addition, interactive Jupyter notebooks are provided with the code base in order to illustrate the usage and how to change parameters (including different clustering algorithms).

Data availability

The complete field distribution data generated using FEM simulations is available via https://doi.org/10.5442/ND000002.

Received: 7 March 2018 Accepted: 3 September 2018
Published online: 28 September 2018

References

1. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
2. Just, M. A. et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.* **1**, 911–919 (2017).
3. Gunčar, G. et al. An application of machine learning to haematological diagnosis. *Sci. Rep.* **8**, 411 (2018).
4. Steinegger, M. & Soding, J. Clustering huge protein sequence sets in linear time. *bioRxiv* **1**, 104034 (2018).
5. Chen, C.-C., Juan, H.-H., Tsai, M.-Y. & Lu, H. H.-S. Unsupervised learning and pattern recognition of biological data structures with density functional theory and machine learning. *Sci. Rep.* **8**, 557 (2018).
6. Kan, A. Machine learning applications in cell image analysis. *Immunol. Cell Biol.* **95**, 525–530 (2017).
7. Exbrayat, J. F., Liu, Y. Y. & Williams, M. Impact of deforestation and climate on the Amazon Basin's above-ground biomass during. *Sci. Rep.* **7**, 1–7 (2017).

8. Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. A bridge for accelerating materials by design. *npj Comput. Mater.* **1**, 15008 (2015).
9. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
10. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision.* **115**, 211–252 (2015).
11. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
13. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Comput. Surv.* **31**, 264–323 (1999).
14. Xu, R. & WunschII, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**, 645–678 (2005).
15. Aghabozorgi, S., Shirkhorshidi, A. Seyed & Wah, T. Ying Time-series clustering - a decade review. *Inf. Syst.* **53**, 16–38 (2015).
16. Bhuyan, M. H., Bhattacharyya, D. K. & Kalita, J. K. Network anomaly detection: methods, systems and tools. *IEEE Commun. Surv. Tutor.* **16**, 303–336 (2014).
17. Pimentel, M. A., Clifton, D. A., Clifton, L. & Tarassenko, L. A review of novelty detection. *Signal Process.* **99**, 215–249 (2014).
18. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
19. Hammerschmidt, M. et al. Reconstruction of photonic crystal geometries using a reduced basis method for nonlinear outputs. *Proc. SPIE* **9756**, 97561R (2016).
20. Smajic, J., Hafner, C. & Erni, D. Optimization of photonic crystal structures. *J. Opt. Soc. Am. A* **21**, 2223 (2004).
21. Hakansson, A., Sanchez-Deh, J. & Sanchis, L. Inverse design of photonic crystal devices. *IEEE J. Sel. Areas Commun.* **23**, 1365–1371 (2005).
22. J. Lu. *Nanophotonic Computational Design*, Dissertation, Stanford University (2013)
23. Lu, J. & Vučković, J. Nanophotonic computational design. *Opt. Express* **21**, 13351 (2013).
24. Piggott, A. Y. et al. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nat. Photonics* **9**, 374–377 (2015).
25. Piggott, A. Y., Petykiewicz, J., Su, L. & Vučković, J. Fabrication-constrained nanophotonic inverse design. *Sci. Rep.* **7**, 1786 (2017).
26. Michaels, A. & Yablonovitch, E. Inverse design of near unity efficiency perfectly vertical grating couplers. *Opt. Express* **26**, 4766–4779 (2018).
27. Rosenblatt, D., Sharon, A. & Friesem, A. Resonant grating waveguide structures. *IEEE J. Quantum Electron.* **33**, 2038–2059 (1997).
28. Astratov, V. N. et al. Photonic band-structure effects in the reflectivity of periodically patterned waveguides. *Phys. Rev. B* **60**, R16255–R16258 (1999).
29. Astratov, V. N. et al. Resonant coupling of near-infrared radiation to photonic band structure waveguides. *J. Light. Technol.* **17**, 2050–2057 (1999).
30. Erchak, A. A. et al. Enhanced coupling to vertical radiation using a two-dimensional photonic crystal in a semiconductor light-emitting diode. *Appl. Phys. Lett.* **78**, 563–565 (2001).
31. Ochiai, T. & Sakoda, K. Dispersion relation and optical transmittance of a hexagonal photonic crystal slab. *Phys. Rev. B* **63**, 125107 (2001).
32. Chutinan, A. & John, S. Light trapping and absorption optimization in certain thin-film photonic crystal architectures. *Phys. Rev. A* **78**, 023825 (2008).
33. Han, S. E. & Chen, G. Toward the Lambertian limit of light trapping in thin nanostructured silicon solar cells. *Nano Lett.* **10**, 4692–4696 (2010).
34. John, S. Why trap light? *Nat. Mater.* **11**, 997–999 (2012).
35. Mellor, A. et al. Nanoimprinted diffraction gratings for crystalline silicon solar cells: implementation, characterization and simulation. *Opt. Express* **21**, A295–304 (2013).
36. Branham, M. S. et al. 15.7% Efficient 10-μm-thick crystalline silicon solar cells using periodic nanostructures. *Adv. Mater.* **27**, 2182–2188 (2015).
37. Fan, S., Villeneuve, P. R., Joannopoulos, J. D. & Schubert, E. F. High extraction efficiency of spontaneous emission from slabs of photonic crystals. *Phys. Rev. Lett.* **78**, 3294–3297 (1997).
38. Wiesmann, C., Bergenek, K., Linder, N. & Schwarz, U. Photonic crystal LEDs - designing light extraction. *Laser Photonics Rev.* **3**, 262–286 (2009).
39. Cunningham, B. T., Zhang, M., Zhuo, Y., Kwon, L. & Race, C. Recent advances in biosensing with photonic crystal surfaces: a review. *IEEE Sens. J.* **16**, 3349–3366 (2016).
40. Block, I. D. et al. A detection instrument for enhanced-fluorescence and label-free imaging on photonic crystal surfaces. *Opt. Express* **17**, 13222 (2009).
41. Ganesh, N., Mathias, P. C., Zhang, W. & Cunningham, B. T. Distance dependence of fluorescence enhancement from photonic crystal surfaces. *J. Appl. Phys.* **103**, 083104 (2008).
42. Threm, D., Nazirizadeh, Y. & Gerken, M. Photonic crystal biosensors towards on-chip integration. *J. Biophotonics* **5**, 601–616 (2012).
43. Boroditsky, M. et al. Spontaneous emission extraction and Purcell enhancement from thin-film 2-D photonic crystals. *J. Light Technol.* **17**, 2096–2112 (1999).
44. Ganesh, N. et al. Leaky-mode assisted fluorescence extraction: application to fluorescence enhancement biosensors. *Opt. Express* **16**, 21626–21640 (2008b).
45. Ondič, L. et al. Diamond photonic crystal slab: Leaky modes and modified photoluminescence emission of surface-deposited quantum dots. *Sci. Rep.* **2**, 914 (2012).
46. Ondič, L. et al. Two-dimensional photonic crystal slab with embedded silicon nanocrystals: Efficient photoluminescence extraction. *Appl. Phys. Lett.* **102**, 251111 (2013).
47. Adachi, M. M. et al. Broadband solar absorption enhancement via periodic nanostructuring of electrodes. *Sci. Rep.* **3**, 2928 (2013).
48. Kim, S. et al. Lead sulfide nanocrystal quantum dot solar cells with trenched ZnO fabricated via nanoimprinting. *ACS Appl. Mater. Interfaces* **5**, 3803–3808 (2013).
49. Su, L. T. et al. Photon upconversion in hetero-nanostructured photoanodes for enhanced near-infrared light harvesting. *Adv. Mater.* **25**, 1603–1607 (2013).
50. Zhang, F., Deng, Y., Shi, Y., Zhang, R. & Zhao, D. Photoluminescence modification in upconversion rare-earth fluoride nanocrystal array constructed photonic crystals. *J. Mater. Chem.* **20**, 3895 (2010).
51. Hofmann, C. L. M., Herter, B., Fischer, S., Gutmann, J. & Goldschmidt, J. C. Upconversion in a Bragg structure: photonic effects of a modified local density of states and irradiance on luminescence and upconversion quantum yield. *Opt. Express* **24**, 14895 (2016).
52. Ganesh, N. et al. Enhanced fluorescence emission from quantum dots on a photonic crystal surface. *Nat. Nanotechnol.* **2**, 515–520 (2007).
53. Barth, C. et al. Increased fluorescence of PbS quantum dots in photonic crystals by excitation enhancement. *Appl. Phys. Lett.* **111**, 031111 (2017).
54. Pomplun, J., Burger, S., Zschiedrich, L. & Schmidt, F. Adaptive finite element method for simulation of optical nano structures. *Phys. Status Solidi b* **244**, 3419–3434 (2007).
55. Becker, C. et al. 5×5 cm^2 silicon photonic crystal slabs on glass and plastic foil exhibiting broadband absorption and high-intensity near-fields. *Sci. Rep.* **4**, 5886 (2014).
56. Barth, C., Burger, S. & Becker, C. Symmetry-dependency of anticrossing phenomena in slab-type photonic crystals. *Opt. Express* **24**, 10931 (2016).
57. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
58. K. Sakoda. *Optical Properties of Photonic Crystals*, Springer Series in Optical Sciences, Vol. 80 (Springer-Verlag, Berlin/Heidelberg, 2005)
59. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).
60. C. Barth. *Analysis of photonic crystals for interaction with near-surface emitters*. PhD thesis, Technische Universität Berlin (2018).
61. Software, Persistence of Vision (TM) Raytracer (POV-Ray), Version 3.7, Persistence of Vision Pty. Ltd. (2013)
62. Schulze, T. F. & Schmidt, T. W. Photochemical upconversion: present status and prospects for its application to solar energy conversion. *Energy Environ. Sci.* **8**, 103–125 (2015).
63. Park, W., Lu, D. & Ahn, S. Plasmon enhancement of luminescence upconversion. *Chem. Soc. Rev.* **44**, 2940–2962 (2015).
64. Wu, M. et al. Solid-state infrared-to-visible upconversion sensitized by colloidal nanocrystals. *Nat. Photonics* **10**, 31–34 (2015).
65. Wu, T. C., Congreve, D. N. & Baldo, M. A. Solid state photon upconversion utilizing thermally activated delayed fluorescence molecules as triplet sensitizer. *Appl. Phys. Lett.* **107**, 031103 (2015).
66. Hoang, N.-V. et al. Giant enhancement of luminescence down-shifting by a doubly resonant rare-earth-doped photonic metastructure. *ACS Photonics* **4**, 1705–1712 (2017).
67. C. M. Bishop. *Pattern Recognition and Machine Learning* (Springer, New York, 2006)
68. T. Hastie, R. Tibshirani, & J. Friedman. *The Elements of Statistical Learning*, 2nd edn, Springer Series in Statistics, Vol. 27 (Springer New York, New York, NY, 2009)
69. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**, 1–38 (1977).
70. G. J. McLachlan & T. Krishnan. *The EM Algorithm and Extensions*, 2nd edn (John Wiley & Sons, Inc., Hoboken, 1997)

Acknowledgements

The authors thank Klaus Jäger from Helmholtz-Zentrum Berlin for useful discussions. The German Federal Ministry of Education and Research is acknowledged for funding the research activities of the Nano-SIPPE group within the program NanoMatFutur (No. 03X5520). We further acknowledge support by the Einstein Foundation Berlin through ECMath within subproject SE6. Parts of the results were obtained at the Berlin Joint Lab for Optical Simulations for Energy Research (BerOSE) of

Helmholtz-Zentrum Berlin für Materialien und Energie, Zuse Institute Berlin, and Freie Universität Berlin.

Author contributions

C.Be. and C.Ba. designed the study. C.Ba. performed the simulations, developed the analysis technique, and wrote the manuscript. C.Be. supervised the project. Both authors discussed the results and reviewed the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Electronic supplementary material