Supporting Information for:

A Neural Network-based Approach to Predicting Absorption in Nanostructured, Disordered Photoelectrodes

Robert H. Coridan[1,2]

[1]Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, AR 72701

[2]Materials Science and Engineering Program, University of Arkansas, Fayetteville, AR 72701

Finite-difference Time Domain (FDTD) simulations: Simulations were performed using the FDTD method[1] via the Python interface (pymeep) for the open source MEEP package.[2] The simulations performed on a grid with a 20 nm voxel size in a two-dimensional box of area 1.5 µm x 8 µm. The FDTD simulations used periodic boundary conditions along the x-direction and perfectly matched layers at either end of the y-axis to allow for infinite propagation away from the electrode. The illumination source was defined as a uniform plane beneath the electrode with the Poynting vector oriented along the y-axis and the polarization along the z-axis, parallel with axes of the cylinders. The cylinders were defined by the geometric object primitives in MEEP assigned with complex refractive index calculated from literature measurements for GaAs[3] and SiO_2.[4] The dielectric properties of GaAs varies significantly over the wavelengths of interest here, so a unique FDTD simulation was performed for each wavelength. The calculations are normalized to the intensity of the incident illumination and are therefore comparable in absolute quantities even though FDTD simulations are unitless. Simulations were performed on a desktop computer, with each taking roughly 4 seconds to complete.

Definition of error metric: We defined the metric $\boldsymbol{\sigma}_{\mathbf{pp}}$, or the per-pixel variance of a prediction to the true, FDTD-derived value:

$$\boldsymbol{\sigma}_{\mathbf{pp}} = \langle \frac{\sum_{\mathbf{i,j}}\left[\mathbf{A}_{\mathbf{pred}}(\mathbf{i},\mathbf{j}) - \mathbf{A}_{\mathbf{FDTD}}(\mathbf{i},\mathbf{j})\right]^2}{\sum_{\mathbf{i,j}}[\mathbf{A}_{\mathbf{FDTD}}(\mathbf{i},\mathbf{j})]^2} \rangle,$$

where $\mathbf{A}_{\mathbf{pred}}$ is the MLP-predicted absorption profile and $\mathbf{A}_{\mathbf{FDTD}}$ is the FDTD-determined, 'true' intensity profile for a given omission glass configuration. The sum is taken over each pixel (*i,j*) and the average is taken over every prediction in the test data set. $\boldsymbol{\sigma}_{\mathbf{pp}}$ includes a weight term to normalize variance to the integrated absorption of a given configuration to avoid biasing towards configurations with lower total absorption. $\boldsymbol{\sigma}_{\mathbf{pp}} = \mathbf{0}$ represents perfect numerical agreement between the prediction and the true profile, while $\boldsymbol{\sigma}_{\mathbf{pp}} = \mathbf{1}$ for the uniform prediction of a flat, zero-valued profile. $\boldsymbol{\sigma}_{\mathbf{pp}}$ is a more exact measure of the accuracy of the prediction than the differences in integrated intensity, which may artificially indicate agreement for drastically different absorption profiles.

Characterization of the close-packed k = 0 photoelectrode: The steady-state electric field amplitude, $|\mathrm{E}(\mathrm{x},\mathrm{y})|$, from FDTD simulations for the *k* = 0 (no scatterers omitted) and *k* = 41 (all scatterers omitted) simulations are shown in Figures S1a and S1b, respectively. The absorption profile $\mathrm{A}(\mathrm{x},\mathrm{y}) = \varepsilon_{\mathrm{i}}(\lambda)|\mathrm{E}(\mathrm{x},\mathrm{y})|^2$ in the GaAs cylinder for the *k* = 0 and *k* = 41 simulations are shown in Figure S1c for 600 nm, 700 nm, and 800 nm illumination. $\varepsilon_{\mathrm{i}}(\lambda) = 2\, n(\lambda)k(\lambda)$ is the imaginary part of the wavelength-dependent, complex dielectric constant. $n(\lambda)$ and $k(\lambda)$ are the wavelength-dependent index of refraction and extinction coefficient of the material, respectively. The absorption spectrum for an electrode was computed by integrating $\mathrm{A(x,y)}$ over the volume of the absorber, as shown in Figure S1d. The absorption spectrum of a continuous, 250 nm slab of GaAs without included scatterers is shown for comparison. A quadrupole-like resonance was observed at 800 nm, possibly due to coupling through periodic boundary conditions along the x-direction.[5] The addition of the SiO_2 scatterers increases the absorption at smaller wavelengths than this resonance, including a 119% increase at 700 nm and a 25% increase at 600 nm. A k = 0 photoelectrode would enhance the broadband absorption in the GaAs absorber.

Multi-level Perceptron (MLP) emulation from libraries of FDTD simulations: The artificial neural network used to generate the model of an ensemble was built with the MLPRegressor function implemented in the Python machine learning library Scikit-Learn.[6] The MLP emulator accepted a 41-element array of binary values representing the presence (1) or absence (0) of an SiO_2 cylinder at a given position in the close-packed array (Figure S1). The MLP emulator predicted the corresponding 2D absorption profile in the GaAs cylinder as a 16x16 real-valued array. Voxels with an imaginary dielectric constant value $\varepsilon_{\mathrm{i}}(\lambda) = 0$ (ie, outside of the GaAs volume) were omitted from the training and prediction of the MLP emulator. The parametric structure of the MLPRegressor network was composed of three hidden layers, each of 2000 nodes with rectifier ('ReLU') activation and a tolerance of 10^{-7}. This structure was chosen empirically as it demonstrated the best σ_{pp} measurements on its own training data ('self-score') of all the network structures we evaluated. We observed a strong relationship between the self-score of a trained MLP emulator and the absolute variance scores σ_{pp} on unique test data. While the absolute values of σ_{pp} improved, the qualitative conclusions of this work were consistent over the range of MLP network structures we evaluated.

Error analysis for MLP emulator predictions with full ensembles: We chose a number of simulations from the k = 3 or the k = 4 to measure the relationship between the training set size and the accuracy of prediction. A number of FDTD simulations, N_{sim}, were chosen at random from the noted ensemble as the training set, and the accuracy of the trained emulator was tested on the entire ensemble. This procedure was repeated eight times for each N_{sim}, generating a new training set at random and MLP emulator at each iteration. The data for λ =

600 nm is shown in Figure S5. An MLP emulator trained exclusively with data from *k* = 3 or from *k* = 4 FDTD simulations was the most accurate approach for making predictions on the respective ensemble for a given N_{sim}. Additionally, σ_{pp} decreased monotonically with increasingly large training sets. σ_{pp} is biased for large N_{sim}, as a significant portion of the ensemble is included in the training set, though this quantifies the limit of achievable statistical error in the MLP emulator when the training set is the entire ensemble. σ_{pp} increased for the *k* = 3-trained and *k* = 4-trained MLP emulators for predictions in the *k* = 5 and *k* = 6 ensembles, though increasing N_{sim} resulted in lower values of σ_{pp}. A training set comprised of simulations from a single ensemble made predictions with the highest accuracy for that ensemble but showed diminished accuracy for other ensembles.

We used a similar procedure to compare the effect that training an MLP emulator with simulations sampled from multiple ensembles has on prediction accuracy. We evaluated MLP emulators that where the training set was composed of a total of N_{sim} simulations: the entire *k* = 0-2 ensemble FDTD data set (862 simulations total) included with a combination of simulations from the k = 3 and *k* = 4 ensembles (added in a ratio of ten *k* = 4 simulations per *k* = 3 simulation). The mixed training set yielded nearly identical σ_{pp} values and dependence on N_{sim} compared to the k = 3-trained and *k* = 4-trained models on their respective ensembles. The mixed training set also improved the prediction accuracy for *k* = 5 and *k* = 6 ensembles for the larger values of N_{sim} tested. Similar results were observed for the *λ* = 700 nm (Figure S6) and *λ* = 800 nm (Figure S7) cases. These experiments showed that mixing ensembles into the training set resulted in prediction accuracy that was virtually equal to independent models trained, though with high accuracy for multiple ensembles. Mixing ensembles in the training set also improved the prediction accuracy for ensembles outside of the training set. For a fixed total number of simulations, there is a notable accuracy improvement to train an MLP emulator with data from many ensembles.

1 A. Taflove and S. C. Hagness, *Computationai Electrodynamics*, Artech House, Boston, MA, 3rd edn., 2005.

2 A. F. Oskooi, D. Roundy, M. Ibanescu, P. Bermel, J. D. Joannopoulos and S. G. Johnson, *Computer Physics Communications*, 2010, 181, 687–702.

3 D. E. Aspnes and A. A. Studna, *Phys. Rev. B*, 1983, 27, 985–1009.

4 I. H. Malitson, *J. Opt. Soc. Am., JOSA*, 1965, 55, 1205–1209.

5 V. E. Babicheva and A. B. Evlyukhin, *Phys. Rev. B*, 2019, 99, 195444.

6 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, 12, 2825–2830.

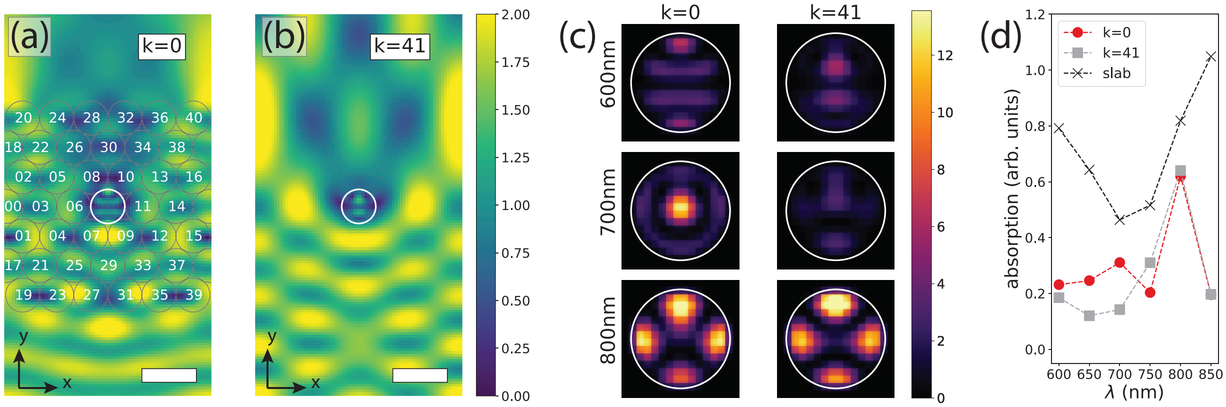

Figure S1 – The geometry and steady-state |*E*(*x*,*y*)| for (a) the *k* = 0 and (b) the *k* = 41 omission glass electrode structure at *λ* = 600 nm (scale bar = 400 nm). The planar illumination is incident from the bottom, propagating from below along the y-direction. (c) The absorption profiles $A_{FDTD}(x,y)$ for the *k* = 0 and *k* = 41 electrodes, calculated from FDTD simulations for *λ* = 600 nm, *λ* = 700 nm, and *λ* = 800 nm (circle = boundary of 250 nm-diameter GaAs cylinder). (d) The integrated absorption spectra for the *k* = 0 (red circles) and *k* = 41 (gray squares) photoelectrodes. The FDTD-simulated integrated absorption spectrum of a 250 nm GaAs thin film is shown for comparison (black '*x*').

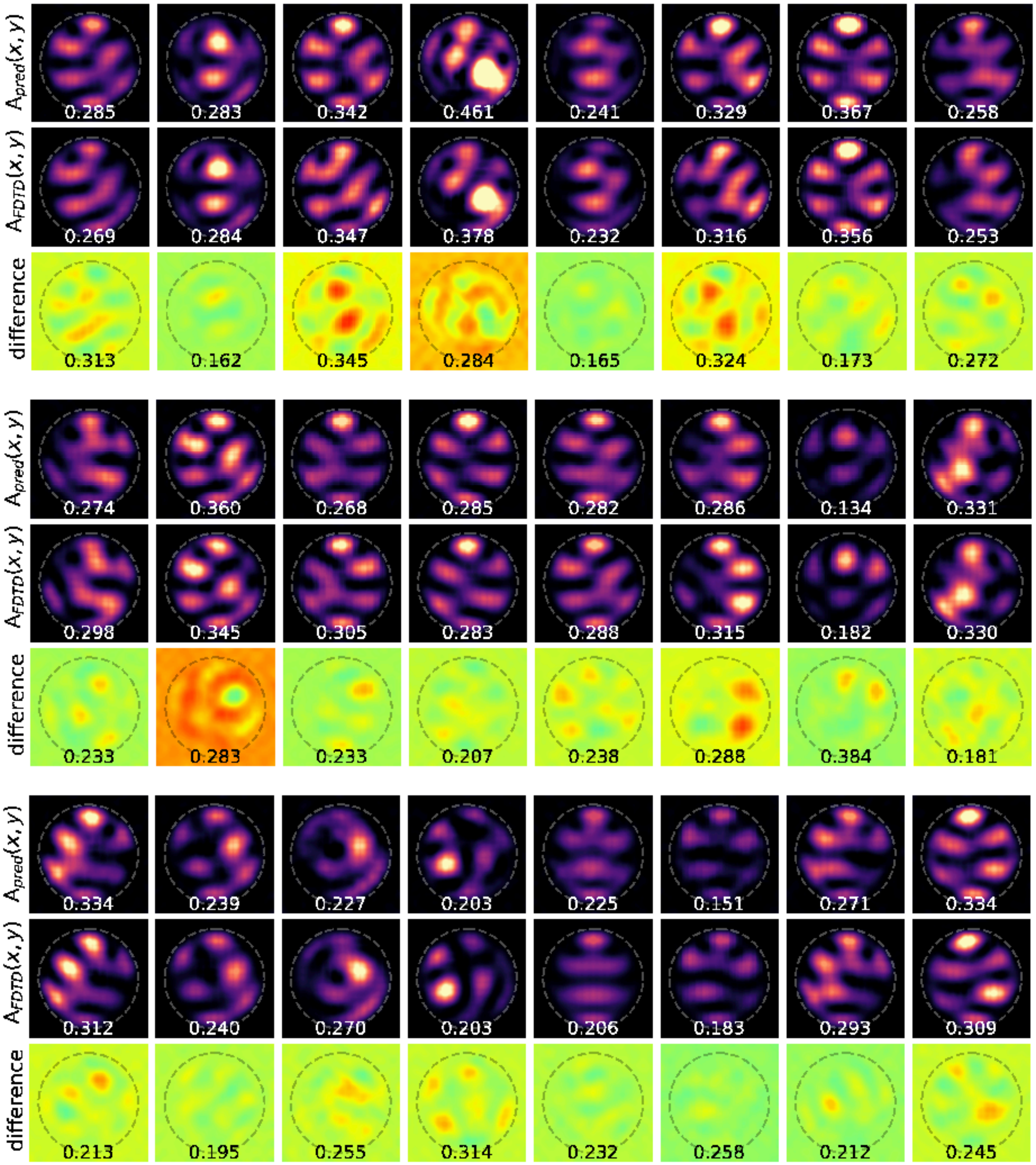

Figure S2 – Randomly chosen examples of the MLP emulator-predicted profile A_{pred}, the corresponding true profile A_{FDTD}, and the difference between the two profiles (*difference*). The integrated absorption for each profile is included at the bottom of each panel. The $\boldsymbol{\sigma_{pp}}$ metric for each pair is included at the bottom of the *difference* panel. λ = 600 nm for these examples.

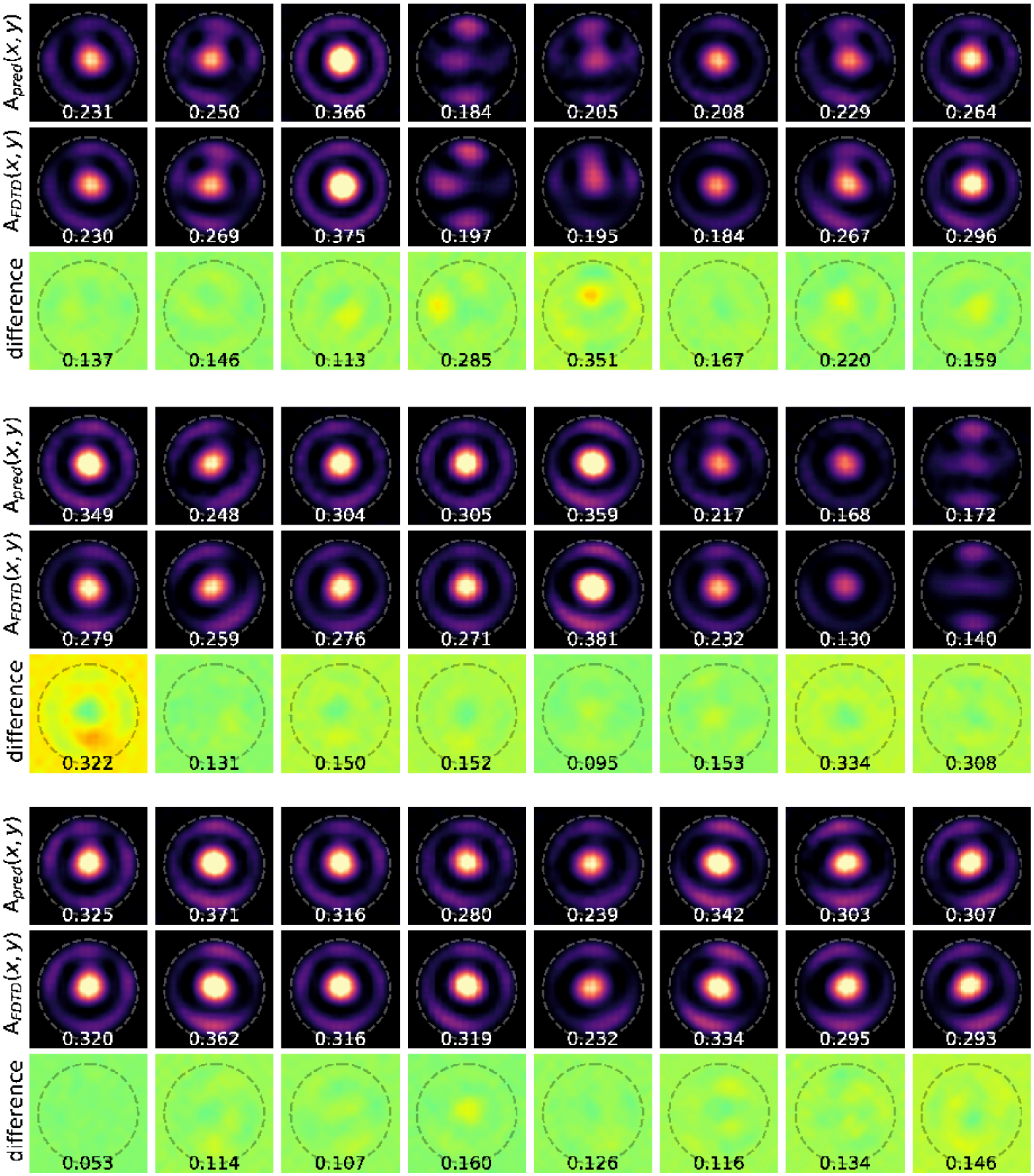

Figure S3 – Randomly chosen examples of the MLP emulator-predicted profile A_{pred}, the corresponding true profile A_{FDTD}, and the difference between the two profiles (*difference*). The integrated absorption for each profile is included at the bottom of each panel. The $\boldsymbol{\sigma_{pp}}$ metric for each pair is included at the bottom of the *difference* panel. λ = 700 nm for these examples.

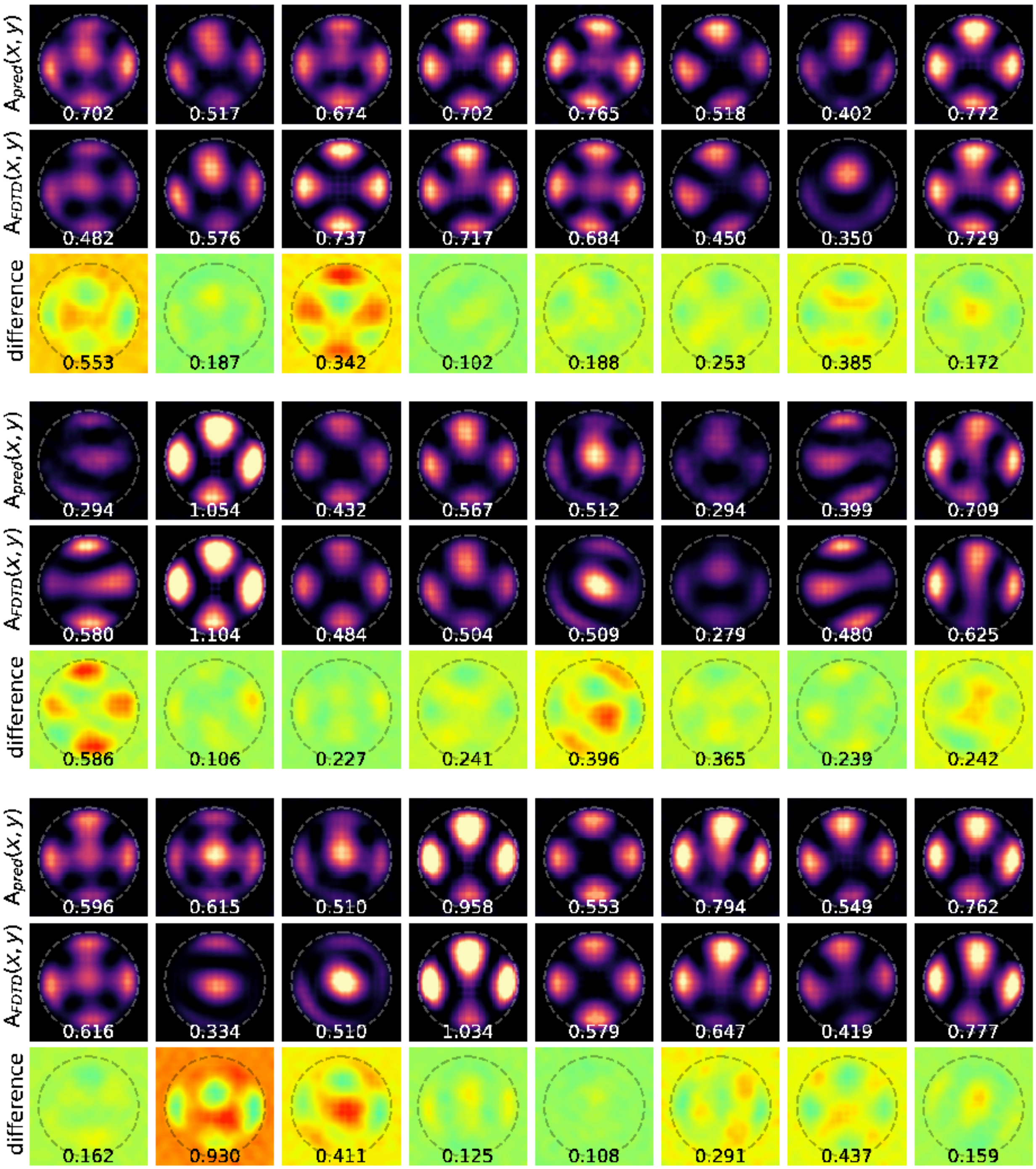

Figure S4 – Randomly chosen examples of the MLP emulator-predicted profile A_{pred}, the corresponding true profile A_{FDTD}, and the difference between the two profiles (*difference*). The integrated absorption for each profile is included at the bottom of each panel. The $\boldsymbol{\sigma_{pp}}$ metric for each pair is included at the bottom of the *difference* panel. λ = 800 nm for these examples.

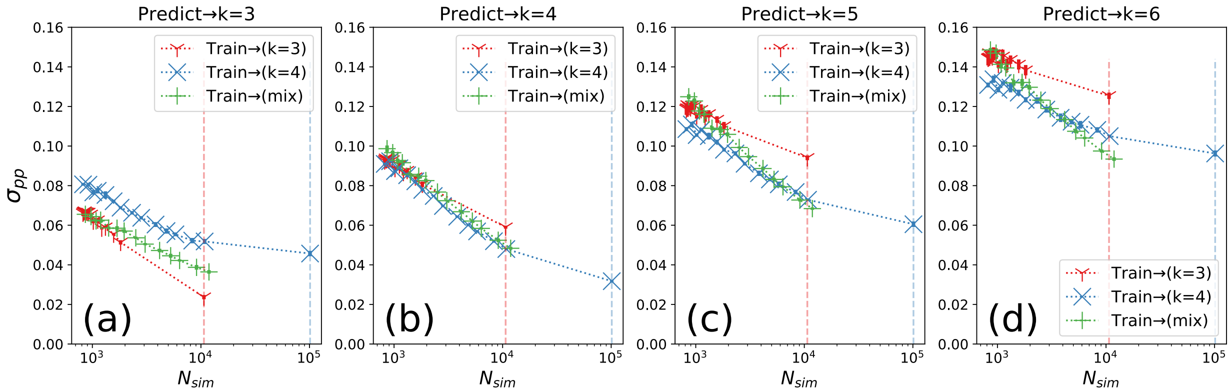

Figure S5 – $\boldsymbol{\sigma_{pp}}$ plotted as a function of the size (N_{sim} = total number of simulations in the training set) and composition of the training set (λ = 600 nm) and the targeted ensembles used for testing. Predictions for an MLP emulator trained on the k = 3 ensemble are shown as the red '*Y*' data. Predictions for an MLP emulator trained on the k = 4 ensemble are shown as the blue '*X*' data. The green '+' data represents predictions made by an MLP emulator trained by the entire k = 0, k = 1, and k = 2 (862 simulations) and the balance made of a random selection of k = 3 and k = 4 included in a 1:10 ratio. The test set for each $\boldsymbol{\sigma_{pp}}$ calculation was (a) the entire k = 3 ensemble (10,660 FDTD simulations), (b) the entire k = 4 ensemble (101,270 simulations), (c) 20,000 FDTD simulations from the k = 5 ensemble, and (d) 20,000 FDTD simulations from the k = 6 ensemble. Each point represents the average $\boldsymbol{\sigma_{pp}}$ over eight independent MLP emulators trained on a random selection of simulations. The dashed lines represent the value of N_{sim} where the MLP emulator is trained on the entire k = 3 (red) and k = 4 (blue) ensembles.

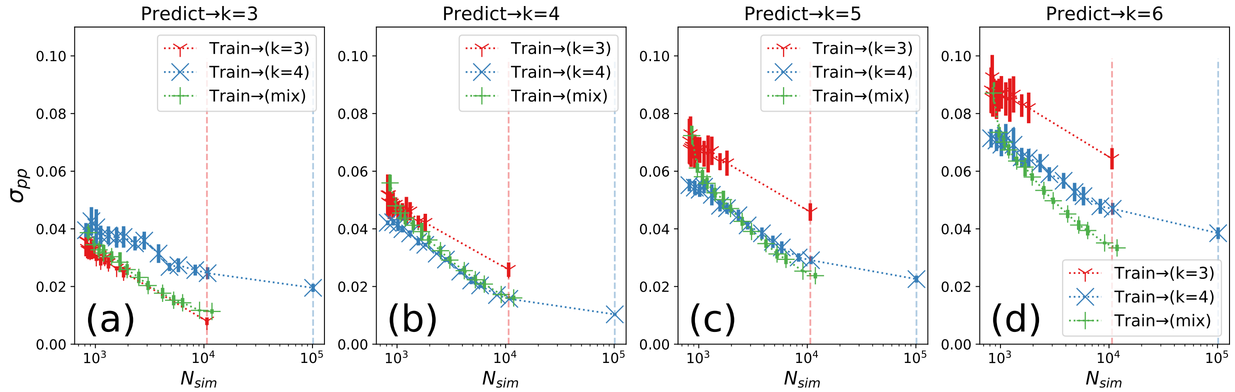

Figure S6 – $\boldsymbol{\sigma_{pp}}$ plotted as a function of the size (N_{sim} = total number of simulations in the training set) and composition of the training set (λ = 700 nm) and the targeted ensembles used for testing. Predictions for an MLP emulator trained on the k = 3 ensemble are shown as the red '*Y*' data. Predictions for an MLP emulator trained on the k = 4 ensemble are shown as the blue '*X*' data. The green '+' data represents predictions made by an MLP emulator trained by the entire k = 0, k = 1, and k = 2 (862 simulations) and the balance made of a random selection of k = 3 and k = 4 included in a 1:10 ratio. The test set for each $\boldsymbol{\sigma_{pp}}$ calculation was (a) the entire k = 3 ensemble (10,660 FDTD simulations), (b) the entire k = 4 ensemble (101,270 simulations), (c) 20,000 FDTD simulations from the k = 5 ensemble, and (d) 20,000 FDTD simulations from the k = 6 ensemble. Each point represents the average $\boldsymbol{\sigma_{pp}}$ over eight independent MLP emulators trained on a random selection of simulations. The dashed lines represent the value of N_{sim} where the MLP emulator is trained on the entire k = 3 (red) and k = 4 (blue) ensembles.

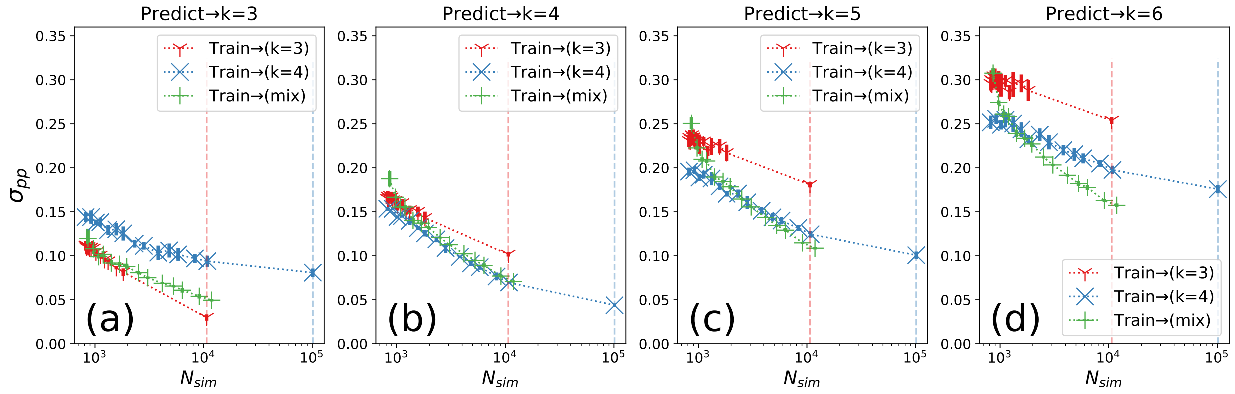

Figure S7 – $\boldsymbol{\sigma_{pp}}$ plotted as a function of the size (N_{sim} = total number of simulations in the training set) and composition of the training set (λ = 800 nm) and the targeted ensembles used for testing. Predictions for an MLP emulator trained on the k = 3 ensemble are shown as the red '*Y*' data. Predictions for an MLP emulator trained on the k = 4 ensemble are shown as the blue '*X*' data. The green '+' data represents predictions made by an MLP emulator trained by the entire k = 0, k = 1, and k = 2 (862 simulations) and the balance made of a random selection of k = 3 and k = 4 included in a 1:10 ratio. The test set for each $\boldsymbol{\sigma_{pp}}$ calculation was (a) the entire k = 3 ensemble (10,660 FDTD simulations), (b) the entire k = 4 ensemble (101,270 simulations), (c) 20,000 FDTD simulations from the k = 5 ensemble, and (d) 20,000 FDTD simulations from the k = 6 ensemble. Each point represents the average $\boldsymbol{\sigma_{pp}}$ over eight independent MLP emulators trained on a random selection of simulations. The dashed lines represent the value of N_{sim} where the MLP emulator is trained on the entire k = 3 (red) and k = 4 (blue) ensembles.

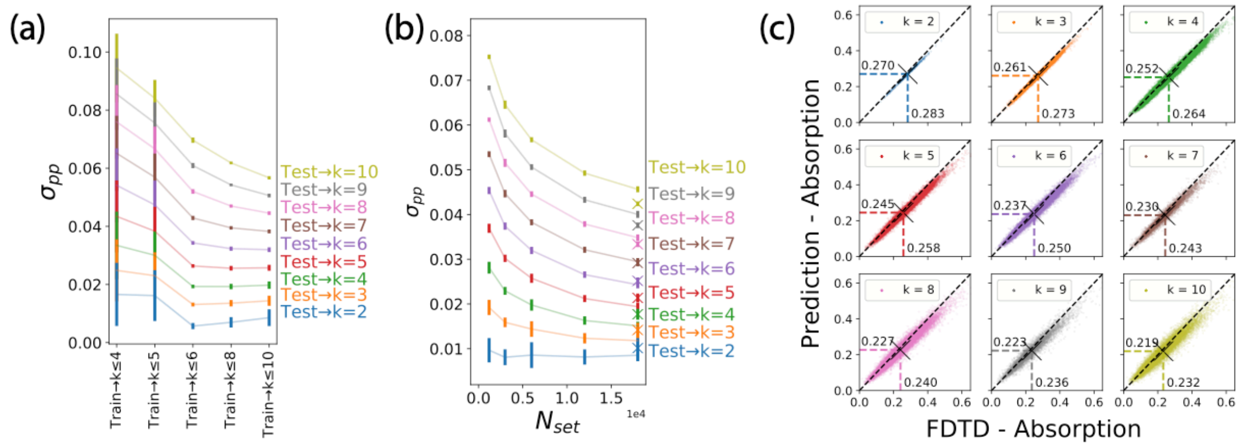

Figure S8 – (a) $\boldsymbol{\sigma_{pp}}$ measurements for MLP emulators trained on with a fixed number of samples (N_{set} = 3000) with varying distribution across the ensembles in a library of λ = 700 nm FDTD simulations. For example, the training set 'k ≤ 4' was composed of the entire set of simulations from the k = 0, k = 1, and k = 2 ensembles, 1500 randomly-sampled k = 3 simulations and 1500 k = 4 simulations. No simulations from the k = 7 or k = 9 ensembles were included in the training set. Each line represents the $\boldsymbol{\sigma_{pp}}$ measurement (averaged over eight unique models) for testing on the complete library of the targeted ensemble. (b) $\boldsymbol{\sigma_{pp}}$ measurements for MLP emulators (solid lines) trained on the 'k ≤ 10' training set as a function of N_{set}. $\boldsymbol{\sigma_{pp}}$ was also calculated for varied contributions from each ensemble. The 'X' data represents $\boldsymbol{\sigma_{pp}}$ measurements for a model trained with the entire set of simulations for k = 0-2 ensembles, 500 simulations from k = 3, 500 from k = 4, 2500 from k = 5, 2500 from k = 6, 6000 from k = 8, and 6000 from k = 10. (c) Scatter plots showing the distribution of integrated absorption values for all of the test data from each ensemble, using emulators trained on the 'X' composition in (b). For each configuration in the ensemble, the x-coordinate of the point represents the FDTD-derived absorption for the configuration while the y-coordinate represents the MLP-predicted absorption. The black 'X' in each plot indicates the average of the predicted (y-coordinate) and the FDTD-determined (x-coordinate) absorption for that ensemble.

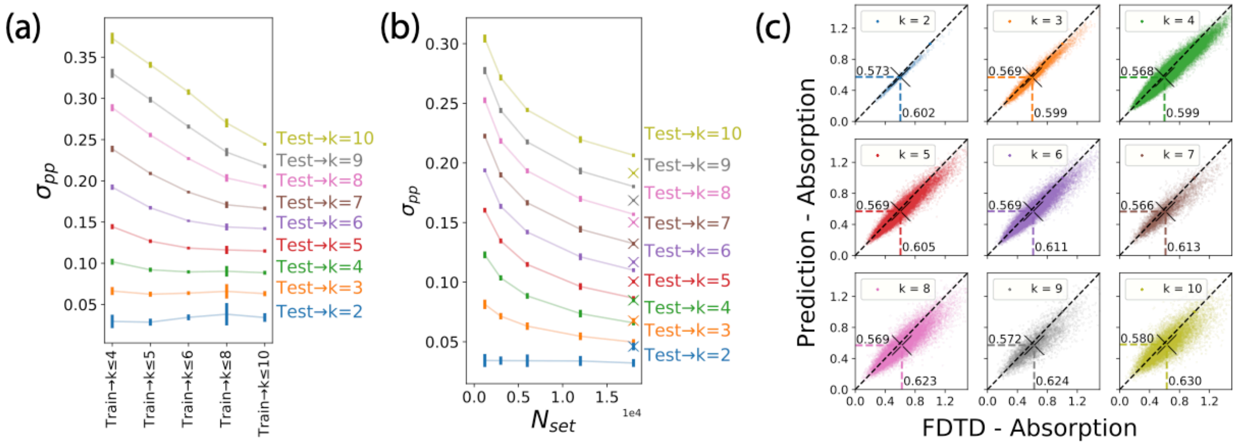

Figure S9 – (a) $\boldsymbol{\sigma_{pp}}$ measurements for MLP emulators trained on with a fixed number of samples (N_{set} = 3000) with varying distribution across the ensembles in a library of λ = 800 nm FDTD simulations. For example, the training set 'k ≤ 4' was composed of the entire set of simulations from the k = 0, k = 1, and k = 2 ensembles, 1500 randomly-sampled k = 3 simulations and 1500 k = 4 simulations. No simulations from the k = 7 or k = 9 ensembles were included in the training set. Each line represents the $\boldsymbol{\sigma_{pp}}$ measurement (averaged over eight unique models) for testing on the complete library of the targeted ensemble. (b) $\boldsymbol{\sigma_{pp}}$ measurements for MLP emulators (solid lines) trained on the 'k ≤ 10' training set as a function of N_{set}. $\boldsymbol{\sigma_{pp}}$ was also calculated for varied contributions from each ensemble. The 'X' data represents $\boldsymbol{\sigma_{pp}}$ measurements for a model trained with the entire set of simulations for k = 0-2 ensembles, 500 simulations from k = 3, 500 from k = 4, 2500 from k = 5, 2500 from k = 6, 6000 from k = 8, and 6000 from k = 10. (c) Scatter plots showing the distribution of integrated absorption values for all of the test data from each ensemble, using emulators trained on the 'X' composition in (b). For each configuration in the ensemble, the x-coordinate of the point represents the FDTD-derived absorption for the configuration while the y-coordinate represents the MLP-predicted absorption. The black 'X' in each plot indicates the average of the predicted (y-coordinate) and the FDTD-determined (x-coordinate) absorption for that ensemble.