Inverse Design of Complex Nanoparticle Heterostructures via Deep Learning on Heterogeneous Graphs

Eric Sivonxay[1], Lucas Attia[1,2], Evan Walter Clark Spotte-Smith[3],
Benjamin Sanchez-Lengeling[4], Xiaojing Xia[5], Daniel Barter[1],
Emory M. Chan[5*], Samuel M. Blau[1*]

[1]Energy Storage and Distributed Resources Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA.
[2]Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA.
[3]Department of Materials Science and Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA.
[4]Google DeepMind, Cambridge, MA, USA.
[5]The Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA.

*Corresponding author(s). E-mail(s): emchan@lbl.gov; smblau@lbl.gov;
Contributing authors: esivonxay@lbl.gov; lattia@lbl.gov;
ewcspottesmith@cmu.edu; beangoben@gmail.com; xxia@lbl.gov;
danielbarter@gmail.com;

Abstract

Applications of deep learning (DL) to design nanomaterials are hampered by a lack of suitable data representations and training data. We report efforts to overcome these limitations and leverage DL to optimize the nonlinear optical properties of core-shell upconverting nanoparticles (UCNPs). UCNPs, which have applications in e.g., biosensing, super-resolution microscopy, and 3D printing, can emit visible and ultraviolet light from near-infrared excitations. We report the first large-scale dataset of UCNP emission spectra based on accurate but expensive kinetic Monte Carlo simulations ($N > 6{,}000$) and use this data to train a heterogeneous graph neural network (GNN) using a novel representation of UCNP nanostructure. Applying gradient-based optimization on the trained GNN, we identify structures with 6.5x higher predicted emission under 800nm

https://doi.org/10.26434/chemrxiv-2024-1dw4q **ORCID:** https://orcid.org/0000-0003-1554-197X

illumination than any UCNP in our training set. Our work reveals new design principles for UCNPs and presents a roadmap for DL-based inverse design of nanomaterials.

Keywords: Upconversion, Nanoparticle, Machine Learning, Graph Neural Networks, Inverse Design

Introduction

Applications in clean energy, advanced manufacturing, biomedicine, photonics, and microelectronics increasingly demand new materials with complex structures and heterogeneous composition.[1] In principle, machine learning (ML) offers a strategy to accelerate the discovery of such materials, as it has emerged as a transformational tool for the design of small molecules, bulk inorganic crystals, and even single-component nanomaterials.[2–4] Deep learning (DL) approaches are particularly well suited to model the behavior of systems with large numbers of parameters, but several obstacles hinder DL from being used to guide the discovery of the complex materials often needed for real-world applications, including nanostructures and composites.

First, state-of-the-art approaches for representing materials often fail to capture the structural complexity of nanomaterials, such as multi-shell nanoparticles[5–12] (Figure 1a) and nanowire heterojunctions, in which nanostructure controls energy transport (Figure 1b).[1] Nanomaterials exhibit distinct and often superior properties compared to their bulk counterparts, driven by nanoscale confinement and high surface areas relative to their volume.[13–15] However, the large number of features required to adequately describe a nanoscale material (e.g., the morphology, dimensions, composition, heterogeneity, doping, defects in each domain; internal interfaces, surface ligands) make training on naive tabular representations computationally inefficient, in part because they neglect physical relationships between features.[16] While bulk crystals can be represented by their unit cell coordinates (e.g., CIF) and small organic molecules by strings (SMILES,[17] SELFIES,[18]), graphs, or atomic coordinates, such atomistic representations are impractical for complex nanomaterials because their critical features often span length scales of one to $>10^6$ atoms and cannot necessarily be reduced to periodic subunits.[19, 20] More recent DL approaches encode spatial information as pixels or voxels, but these fixed-resolution representations cannot efficiently capture the structural hierarchy of a wide range of nanomaterials, e.g., those of different sizes.

Beyond the challenge of representing nanomaterials, it is also challenging to generate datasets sufficiently large to train DL models that can accurately predict the properties of heterogeneous, multi-component nanostructures.[21] Although high-throughput experimental and computational approaches are growing in their availability and utility,[22–24] the synthesis and simulation of complex heterostructures is often time-consuming, limiting the scale of available datasets[25] and constraining campaigns to the "small data" regime where DL techniques often struggle. Modern DL models can also have difficulty extrapolating outside of the envelope of their training data, which is necessary for the discovery of novel materials with enhanced properties. Finally, the discovery of fundamentally new materials is complicated by the rough response surfaces of material properties with respect to their composition, necessitating dense and tedious ”needle-in-a-haystack” searches across a parameter space. The prediction of materials with targeted properties, or inverse design, would be significantly accelerated by surrogate models that are differentiable so that gradient-based optimization techniques can be used to direct efficient searches.[3] Thus, DL-guided inverse design of complex materials would benefit from the development of large structure-property datasets of complex nanomaterials, new methods to represent them across lengths scales, and differentiable models that are accurate out of their training distribution.

In this work, we develop a heterogeneous graph representation for nanomaterials with a variable number of spatial domains, each containing multiple components that can interact within the same domain and across interfaces. We demonstrate that physics-informed graph neural networks (GNNs) built atop such representations can accurately predict properties of nanostructures which are far more complex than any contained in the training dataset. As a model system, we center our investigation on lanthanide-doped upconverting nanoparticle (UCNP) heterostructures (Figure 1a), whose unique nonlinear optical properties have diverse applications in biological and super-resolution imaging,[26] optogenetics, sensing, photonics,[27] scintillators, secure labeling, 3D printing,[28] and photovoltaics.[29] These applications leverage the ability of UCNPs to absorb multiple near-infrared (NIR) photons and convert them into a single photon of higher frequency, e.g., in the visible and ultraviolet spectrum. Such nonlinear processes are the result of complex networks of energy transfer (ET) interactions between different lanthanide ions (e.g., Yb^{3+}, Er^{3+}, and Nd^{3+}, as in Figure 1b). To promote advantageous ET interactions and inhibit those that quench emission, nearly all practical implementations of UCNPs use doped heterostructures in which a spherical core is surrounded by one to four concentric shells, with each domain having a distinct combination and composition of lanthanide ions (Figure 1c). Due to the large numbers of tunable structural and compositional parameters, and the complex network of energy transfer interactions between dopants, optimizing the intensity and wavelength for such complex heterostructures is extremely challenging.[30] Thus, multi-shell UCNP heterostructures present a stringent test for any new DL model and representation.

To train these DL models, we generated a dataset of Simulated Upconverting Nanoparticle Spectra for Emission Tuning (SUNSET), consisting of results from ∼6,000 kinetic Monte Carlo (kMC) simulations of nanoparticle photophysics (Figure 1d). Models trained on SUNSET aim to predict photon emission within a specified wavelength band as a function of UCNP heterostructure. By training on simulations of UCNPs with up to three shells and evaluating on simulations with four shells, we can quantify the capacity for models to extrapolate to larger and more complex heterostructures. We find that our heterogeneous graph representation, informed by UCNP physics and geometry, allows DL models to achieve far higher prediction accuracy than tabular, image, and homogeneous graph representations (Figure 1e), especially when extrapolating beyond the training data. The differentiability of our heterogeneous GNN also yields gradients of emission intensity with respect to layer thicknesses and dopant concentrations (Figure 1e), which are not accessible from kMC. Our trained model thus facilitates inverse design of UCNP heterostructure via gradient-based optimization (Figure 1f), identifying novel superior UCNPs with a range of sizes and up to ten shells. When excited at 980 or 800 nm, these optimized UCNP heterostructures exhibit exceptionally high emission between 300-450 nm, a spectral range useful for inducing photochemistry for optogenetic, catalytic, therapeutic, and 3D printing applications. To validate these predictions, we perform additional months-long kMC simulations which indicate that our model possesses considerable ability to extrapolate far out of distribution and can suggest never-before-seen structures with high accuracy, further revealing novel design principles. These findings demonstrate a path forward for the optimization and discovery of technologically useful UCNPs and offer inspiration for the development of novel DL representations and models which enable inverse design for a broad range of optical nanomaterials.

Results

Dataset construction

To develop and train DL models that can predict core-shell UCNP photophysics and ultimately enable inverse design of UCNPs with complex heterostructures that exhibit efficient UV and blue emission (300-450 nm), we generated SUNSET: a dataset of over 30,000 multi-shell UCNP spectra calculated with a high-performance kMC simulation package (RNMC)[31] optimized for chemical reaction networks and UCNP photophysics[9] (Figure 2). SUNSET consists of four sub-collections (SUNSET-[1, 2, 3, 4]) that include different dopant ion combinations and surface effects (see Figure 2a). While each of the sub-collections provides utility for model development and testing, we focus exclusively on SUNSET-1 in our main narrative; discussion of the SUNSET-[2–4] collections can be found in the Supplementary Information. We focus on SUNSET-1 because this sub-collection includes nanoparticles of variable size and a variable number of shells, which is necessary for training models that have the possibility of extrapolating to more complex heterostructures and thus facilitating impactful inverse design. Further, the nanoparticle structures in SUNSET-1 are substantially larger than in SUNSET-[2–4] (given that they contain multiple layers, and each layer must be at minimum 1 nm thick in order to be synthesizable), and thus the vast majority of the computational cost of SUNSET went towards the ∼6,000 simulations in SUNSET-1. We note that individual kMC trajectories often took weeks to complete, necessitating the use of high-throughput self-checkpointing workflows. Further details of our workflow infrastructure are given in Section S1.

SUNSET-1 utilizes a dopant set of Er^{3+}, Nd^{3+}, Yb^{3+} because this combination of dopants has been used to sensitize upconversion[6] and optogenetic activity[32] with 800 nm excitation, a wavelength that lies in the NIR-I biological imaging window. Segregation of these dopants into different shells of UCNP heterostructures has been shown to dramatically enhance emission.[6] In these systems, Nd is typically included to sensitize the absorption of the 800 nm excitation, Er to upconvert absorbed energy and emit UV/visible light, and Yb to act as a conduit to transfer energy between Nd

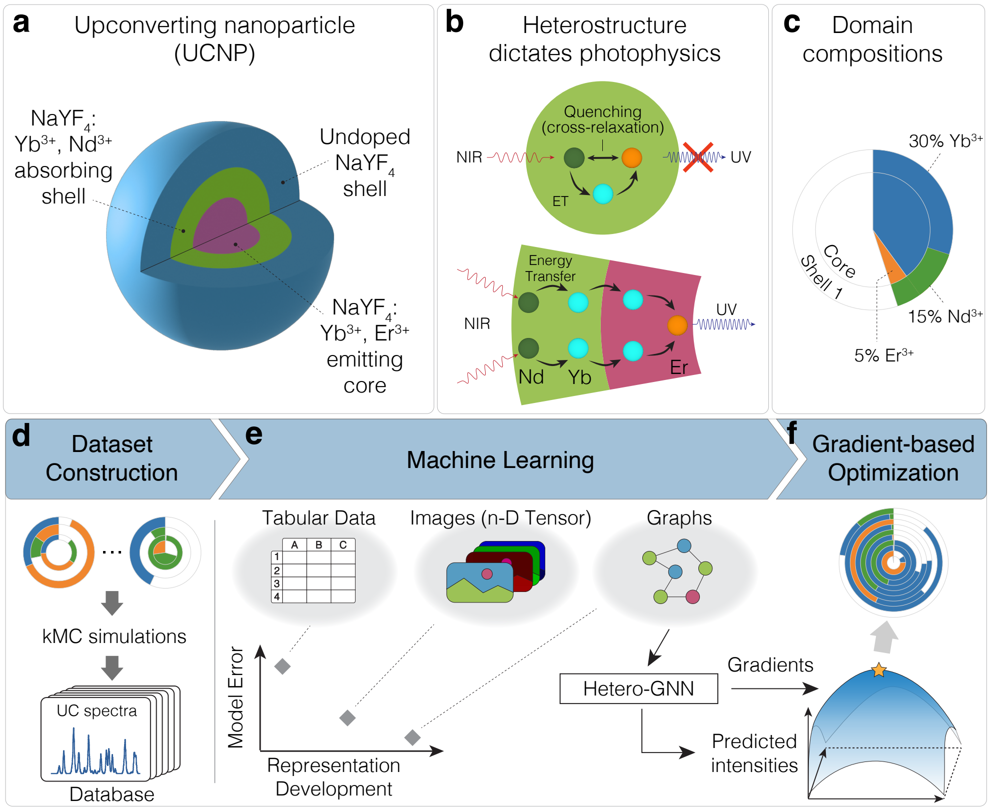
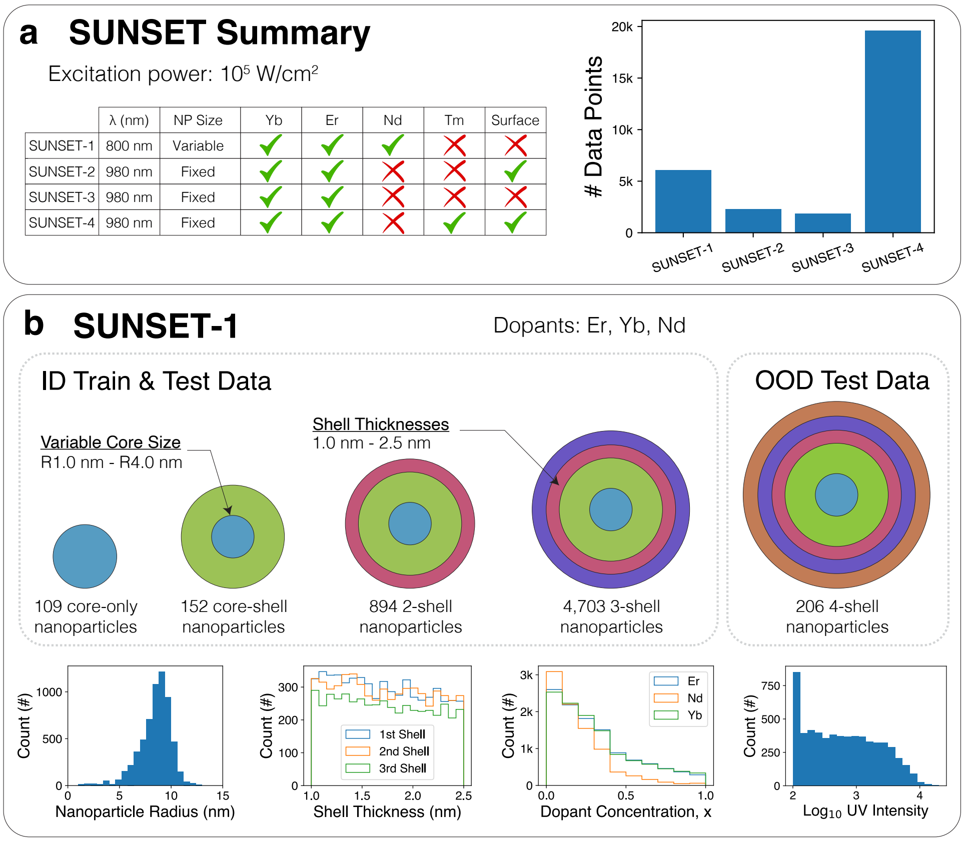


Fig. 1 a) (left) Schematic of doping heterostructured UCNPs. b) Illustration of the importance of heterostructuring to physically separate dopants. In the core-only particle, all dopants interact to a high degree, leading to cross-relaxation and consequently low upconversion efficiency, while separation of Er and Nd into different regions via heterostructuring prevents cross-relaxation and facilitates UV emission. c) 2D visualization of a heterostructured UCNP with core and shell doping. d) Overview of physics-infused deep learning for the inverse design of UCNPs. We first construct a large dataset of simulated UCNPs using high-throughput kMC simulations. Then, we investigate different representations of heterostructured UCNPs and find best performance with a heterogeneous graph representation in a GNN. The resulting trained model can provide the gradient of emission intensity with respect to each structural parameter, enabling inverse design via gradient-based optimization of UCNP heterostructure to maximize UV emission.

and Er dopants that would otherwise quench each other via cross-relaxation energy transfer. The nanoparticle heterostructures sampled in SUNSET-1 are variable, with core radii ranging from 1–4 nm and up to 3 shells, each measuring between 1–2.5 nm in thickness, as depicted in Figure 2b. To probe the extrapolatory power of developed models, we simulate explicit 4-shell nanoparticles and hold them out of training data to use as an out-of-distribution (OOD) test set. Thus, SUNSET-1 as a whole has nanoparticle radii spanning from 1–13.6 nm, and the brightest particle has an intensity of $\sim$ 20,000 cps. I_{vis-UV} most closely follows an exponential distribution, so we use the $log(I_{vis-UV})$ as the target label for model training.



	λ (nm)	NP Size	Yb	Er	Nd	Tm	Surface
SUNSET-1	800 nm	Variable	✓	✓	✓	✗	✗
SUNSET-2	980 nm	Fixed	✓	✓	✗	✗	✓
SUNSET-3	980 nm	Fixed	✓	✓	✗	✗	✗
SUNSET-4	980 nm	Fixed	✓	✓	✗	✓	✓

Fig. 2 a) Summary of the SUNSET dataset, including excitation wavelength, fixed versus variable nanoparticle size, dopants used, presence or absence of luminescence-quenching surface ligands, and scale in terms of number of data points. b) Graphical depiction of SUNSET-1 heterostructures, ID versus OOD split, and statistical distributions of structural features.

Representing Nanoparticle Structure for Machine Learning

To train ML models on the SUNSET data, we initially investigated several existing representations for encoding the compositional and dimensional features of each UCNP. As we summarize in Table 1 and discuss in more detail in the "Model Performance" section, we found that standard ML models (e.g., random forest regressors and convolutional neural networks) utilizing tabular and image-based representations exhibited poor ability to extrapolate, with 3- to 10-fold lower accuracy during out-of-distribution testing (OOD) than during in-distribution testing (ID).

Seeking representations and models with greater ability to extrapolate to more complex nanostructures, we explored the use of graph-structured representations, which have recently gained prominence due to their ability to effectively capture complex relationships (edges) between entities (nodes).[33] The simplest graph representation of a UCNP is a homogeneous graph in which each node is labeled with the identity and concentrations of a single type of dopant (e.g., Er) in a specific domain of the UCNP (e.g., the core), while edges encode interactions between dopants represented by those nodes, i.e., energy transfer processes. However, we found that GNNs utilizing these homogeneous graph representations exhibited equally poor accuracy for both ID and OOD testing (Table 1).

Reasoning that the poor performance of homogeneous GNNs was related to inadequate representation of the physical interactions between dopants, we developed a UCNP representation based on a directed heterogeneous graph (Figure 3a). Unlike the homogeneous graphs, dopant-dopant interactions in our heterogeneous graphs are represented by interaction nodes that connect dopant nodes (via edges), allowing the encoding of additional physical features of the interactions. Two different types of interaction nodes are used, intra-layer and trans-layer, to delineate interactions between dopants within the same geometric region (i.e., core or shell domain) and those in different regions, respectively. It is important to note that dopant nodes are never connected to other dopant nodes, and every interaction node connects exactly two dopant nodes. A self-interaction node, describing interactions between different dopant ions of the same type in the same region, has edges both from and to the same dopant node. The use of a directed graph introduces asymmetry in energy transfer between two dopants (e.g. Yb → Er as compared to Er → Yb). This is important for energy transfer processes that are not reversible, such as non-resonant, phonon-assisted energy transfer that results in irreversible heat dissipation.[34]

When establishing features encoded in the nodes, we chose a minimal set of descriptors that are most relevant for UCNPs. Dopant node features include dopant type, dopant concentration (within the respective region), and geometric bounds (inner and outer radii of the core/shell domain they reside in). The interaction nodes contain the interaction type (*e.g.*, Yb-Er, Er-Yb, Er-Er, . . .) and features derived from the pair of connected dopant nodes — dopant concentrations and geometric bounds. Since we explore only spherical nanoparticles with multiple concentric shells, layer radii fully specify heterostructure geometry.

To further account for the effect of distance and region geometry on dopant energy transfers, we introduce a quantity that we call the "integrated interaction." This quantity is derived by integrating a Gaussian function, denoted as $\mathcal{N}(s; 0, \sigma)$, over all pairwise distances, s, between interacting regions, V_i and V_j, as illustrated in Fig. 3b:

$$II = x_i * x_j \iint \mathcal{N}(s; 0, \sigma) dV_i dV_j, \tag{1}$$

$$s(r_1, \theta_1, \phi_1, r_2, \theta_2, \phi_2) = \sqrt{r_1^2 + r_2^2 - 2r_1r_2\left(\sin(\theta_1)\sin(\theta_2)\cos(\phi_1 - \phi_2) + \cos(\theta_1)\cos(\theta_2)\right)} \tag{2}$$

where x_i and x_j are the doping concentrations in regions i and j, respectively and s is written in spherical coordinates. While the probability of energy transfer between two dopants is actually proportional to $distance^{-6}$, we chose to represent this probability as a sum of Gaussian functions because they are continuous at $x = 0$ and can be

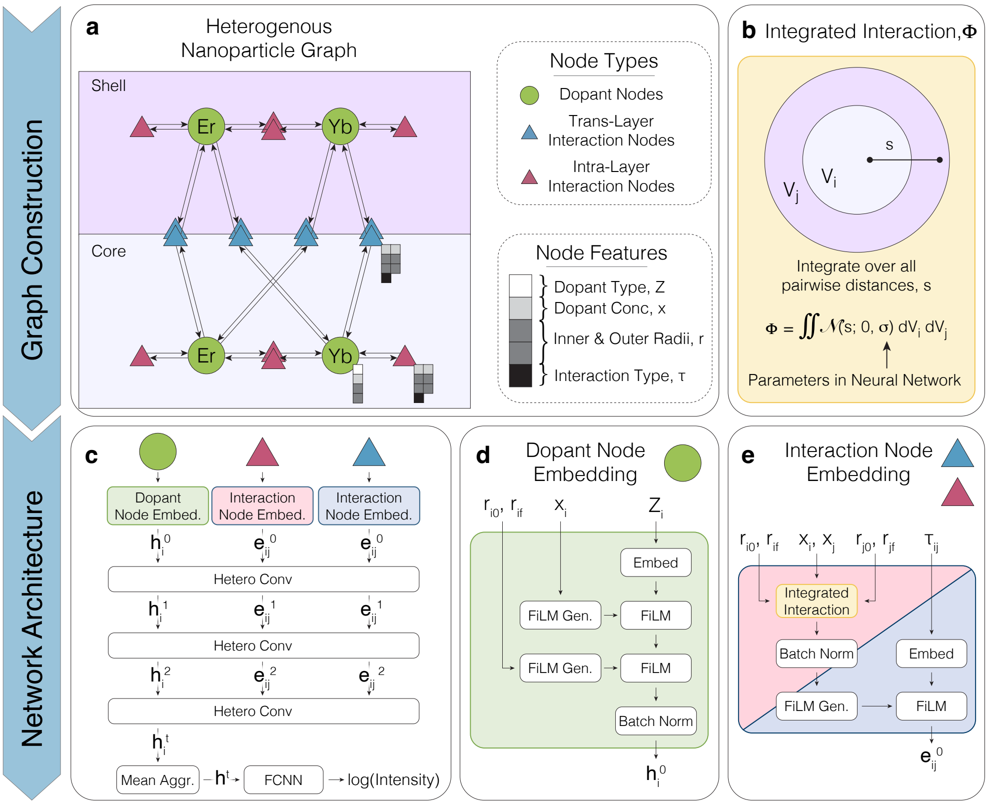


Fig. 3 a) Heterogeneous graph structure of an example two-region (core plus one shell) nanoparticle with Yb and Er dopants. b) Depiction of the Integrated Interaction, which captures the geometric interaction between two regions. c) The architecture of the heterogeneous graph model. d) Schematic of the operations for embedding dopant features to obtain the initial dopant attribute. e) Schematic of the embedding process for interaction nodes.

integrated multiple times while still capturing the decaying nature of energy transfer with increasing distance between two ions. Adjusting the σ parameters within the Gaussians allows for the modulation of the effective interaction distance of dopants. When used in the ML model, the integrated interaction module is parameterized by n learnable weights (here $n = 5$) each corresponding to a σ value of one of the Gaussians in the sum.

The proposed heterogeneous graph structure lends naturally to the use of GNNs for DL. In our heterogeneous GNN (Figure 3c), we first embed information from each node into a continuous vector space. We construct the dopant node embeddings by passing the dopant type (Z_i) through an embedding layer, then contextualizing the initial embedding on the dopant concentration (x_i) and the radii (r_{i0} and r_{if}) using feature-wise linear modulation (FiLM) layers, as shown in Figure 3d. Likewise, to obtain the embedding vector for the interaction nodes (Figure 3e), we pass the interaction type (τ_{ij}) through an embedding layer and then condition on the integrated interaction using a FiLM layer. A batch normalization is applied prior to the FiLM layer to shift the distribution of integrated interaction values. Note that ”contextualizing” or ”conditioning” a vector (i.e., an initial embedding) on another value (e.g. dopant concentration, layer radii, or integrated interaction) with a FiLM layer is a way of combining the information contained in each via an operation that is controlled by many learnable parameters, which often provides better expressivity and performance than a simple concatenation or addition.[35] The resulting dopant and interaction embeddings are then used as the inputs for three iterations of message passing (MP) based on the heterogeneous graph's directional edges, where each MP iteration employs graph attention via the GATv2 operator,[36] after which we use mean aggregation to obtain the global latent representation. Finally, a fully connected neural network (FCNN) is used for label prediction (i.e. predicting the log of the emission intensity over the specified wavelength band) from the global latent representation.

Model Performance

To assess the performance of the heterogeneous GNN (“hetero-GNN”) described above with respect to other models and representations, we train each on the SUNSET-1 dataset, where our target label is emission intensity from the UV-blue (300-450 nm) wavelength band, and training data include 800-nm excited UCNPs with 0-3 shells. We evaluate the mean squared errors (MSEs) of each model when predicting the intensities of held-out ID samples as well as for OOD nanoparticles with 4 shells (Table 1). We compare the hetero-GNN to four well-established supervised learning models: two models (a random forest regressor and a FCNN multi-layer perceptron) using a tabular representation; a CNN using an image representation; and the homogeneous GNN described above. Model hyperparameters are provided in Section S8. Additional model, representation, and feature details are provided in Sections S4, S5, and S6.

Of the five models tested, the hetero-GNN exhibited the lowest error for both ID and OOD testing, with MSE values of 13.9 and 22.2 photon counts-per-second (cps), respectively (Table 1). The ID loss is 4-fold lower than that of models utilizing tabular representations (RFR and FCNN) and 21.5% lower than the CNN utilizing an image representation. The fact that the image- and heterogenous graph-based models have the best ID accuracies demonstrates how their representations allow them to leverage spatial information to learn relationships between heterostructure and properties and to connect the common behavior of dopant ions of the same type but located in different regions. It is notable, however, that the least accurate model for ID testing was also based on a graph representation. The homogeneous GNN exhibited 6-fold higher MSE than its hetereogeneous analogue, highlighting that the enhanced accuracy of the hetero-GNN is the result of its incorporation of interactions as nodes in the graphs.

We believe that the promotion of dopant-dopant interactions to a node-level property specifically improves hetero-GNN performance by elevating the prominence of interaction features during message passing. When included as explicit nodes, the interactions are able to alter the content of the passed messages, allowing for the transmission of richer and more physically relevant information. This parallels the photophysics of UCNPs, in which energy transfer interactions between dopants critically determine the excited state populations of donors, acceptors, and

Table 1 An overview of the performance of different models on the SUNSET-1 dataset, assessed using the mean squared errors (MSE, Equation 3) of predicted I_{vis-UV} derived from 10-fold cross-validation both for SUNSET-1 in-distribution (ID) test, containing structures with up to three shells, and for SUNSET-1 out-of-distribution (OOD) test, containing structures with four shells. The first value in each cell is the mean squared error (MSE) in photon counts-per-second (cps) multiplied by 10^{-3} for easier interpretation. The second value in parentheses is the normalized mean squared error (NMSE) which is the MSE normalized by the sum of squares to yield a relative error (Equation 4). The lowest errors for each category, both of which are achieved by the hetero-GNN, are emphasized with bold text.

	RFR	FCNN	CNN	GNN	Hetero GNN
ID Test	57.2 (7.3%)	55.8 (6.6%)	17.6 (2.1%)	84.3 (9.9%)	**13.8 (1.6%)**
OOD Test	365.8 (36.6%)	526.6 (55.5%)	49.0 (5.2%)	89.6 (9.4%)	**22.1 (2.3%)**

their neighbors,[30, 37] driving nonlinear processes such as upconversion, photon avalanching,[5, 38] and quantum cutting.[39] In contrast, interaction features in homo-GNNs have less influence on predictions because edge properties can only contribute to the attention score, influencing the weighting of the messages being passed between dopant nodes rather than the information contained therein. The heterogeneous graph structure also allows for the embedding of interactions between lanthanide dopants. This embedding introduces valuable inductive bias concerning the distinctness of lanthanide interactions, constraining the model to treat the dopant pairs (e.g., Yb-Er) equally, agnostic to which layers they reside in (e.g. the first or second shell), albeit with varying strengths.

The most striking benefit of the hetero-GNN is its ability to extrapolate, in this case to 4-shelled nanostructures not included in its training set. When switching from ID to OOD testing, the MSE for the hetero-GNN model increased by 8.3 cps, or 1.6-fold. This modest increase in loss is in stark contrast to the tabular-representation-based RFR and FCNN models, for which extrapolation resulted in 6- to 10-fold increases in the MSE, respectively. The tabular models lack the geometric and relational information of the graph models and therefore must learn the influence of the dopants in each layer independently. This increases the data demand of these models, making them prone to overfitting and reducing their ability to predict the properties of unseen heterostructures. Even the image-based CNN, which had high ID accuracy, exhibited 2.8-fold greater loss during OOD testing. We ascribe the significantly greater ability of the hetero-GNN to generalize to its graph representation. This conclusion is supported by the fact that the homogeneous GNN also exhibits very little change in loss (+6.5%) when shifting from interpolation to extrapolation. In summary, representing UCNP heterostructures as heterogeneous graphs results in a hetero-GNN that exhibits both high accuracy and extrapolative capacity, in contrast to existing models that are inaccurate (homo-GNN), poor at extrapolation (image CNN), or both (tabular models).

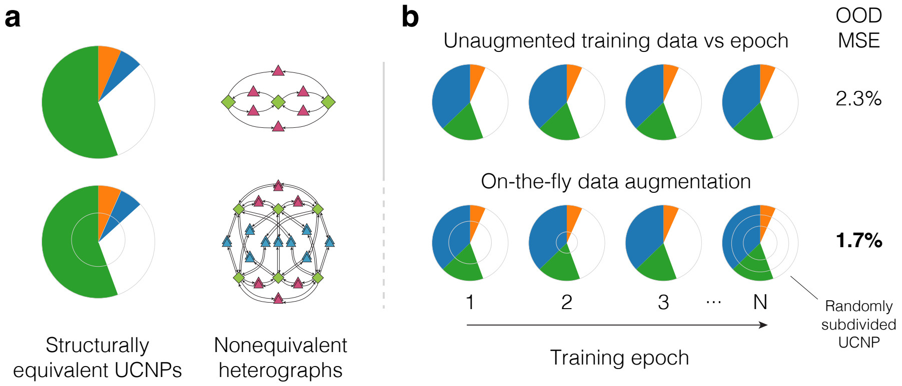


Fig. 4 Arbitrarily subdividing a structural region in a heterostructured nanoparticle leaves the nanoparticle physically unchanged. Such "subdivision invariance" is not reflected in nanoparticle graph representations. a) 2D visualizations of physically equivalent core and arbitrarily subdivided core-shell nanoparticles, along with their non-equivalent heterogeneous graph representations. b) Subdivision invariance is promoted in hetero-GNNs using on-the-fly data augmentation. In contrast to the nonaugmented training procedure (b, top), on-the-fly data augmentation is performed during each training epoch by inserting a random number of subdivisions into UCNPs at random radial coordinates (b, bottom). This process implicitly multiplies the number of unique heterograph representations of UCNPs seen by the model during training by the number of training epochs, and reduces OOD error by 25%

Data Augmentation: Training Subdivision Invariance

While the hetero-GNN exhibits superior performance to the CNN, the image representation has the physically intuitive property that arbitrarily subdividing a given nanoparticle region (e.g. dividing a shell into two smaller shells, where both have the same dopant concentrations as the originally undivided region) has no impact on the model's structural representation or subsequent label prediction. This property, which we call "subdivision invariance", is physically motivated by the fact that region subdivision is arbitrary and leaves the nanoparticle being described completely unchanged. However, neither our heterogeneous graph nor any of the other non-image representations are inherently subdivision invariant. For example, as shown in Figure 4a, subdividing an originally core-only particle into a core and a shell dramatically changes the heterogeneous graph, and thus our hetero-GNN model may predict very different latent representations for physically identical nanoparticles. This is clearly undesirable and may be detrimental to both the learning process and subsequent structural optimization. On the other hand, the voxelization that makes the image representation subdivision invariant simultaneously causes layer dimensions to only be present in the model implicitly, preventing UCNP emission from being differentiated with respect to layer thicknesses and precluding gradient-based optimization of UCNP heterostructure. Thus, any DL model that aims to enable inverse design of nanomaterial heterostructure via gradient-based optimization will need to reckon with the problem of subdivision invariance.

Even when an input representation is not inherently subdivision invariant, it is possible to design the DL model built atop the representation to explicitly enforce subdivision invariance such that physically identical structures yield identical latent representations. However, in the context of a graph representation, such explicit enforcement is only possible by avoiding the use of any non-linear operations, which dramatically limits model expressivity and performance.

An alternate strategy is to train models to approximate subdivision invariance via data augmentation. Using data augmentation to train approximate invariances in different DL contexts (e.g. image rotation, reflection, etc. in CNNs,[40] molecular rotation and translation in interatomic potentials[41]) is well established and can enhance model prediction accuracy and robustness. We apply this augmentation strategy to our hetero-GNN model by artificially subdividing the UCNP input with the same labels (emission intensities) but with structural representations modified with random subdivisions. This data augmentation is meant to guide the learned latent representation to exhibit approximate subdivision invariance, which should improve model performance. This strategy is implemented on-the-fly (Fig. 4 b), so that data in each batch are augmented as they appear during training, rather than augmenting the entire dataset before training. Thus, this does not explicitly increase the size of the training dataset, but implicitly multiplies the number of unique heterograph representations of UCNPs seen by the model during training by the number of training epochs. Random subdivisions result in a node with r_{inner}, r_{outer}, being split into two nodes with r_{inner}, $r_{\text{subdivision}}$ and $r_{\text{subdivision}}$, r_{outer}, where $r_{\text{inner}} < r_{\text{subdivision}} < r_{\text{outer}}$. Additional details about data augmentation are provided in Section 5.

When the hetero-GNN is trained using on-the-fly-augmentation, its performance improves by 23% on the ID test set (with error falling from 13.8 cps to 10.6 cps) and by 25% on the OOD test set (with error falling from 22.1 cps (2.3%) to 16.5 cps (1.7%)). Additionally, we validated that this augmentation scheme actually trains the model to learn subdivision invariance by evaluating the vector distance between the representation of nanoparticles and their subdivided analogs, observing that the augmented hetero-GNN model more closely represented the subdivided UCNPs in the ID and OOD test sets in embedding space than the nonaugmented model across a range of subdivisions (Figure S6). These results underscore the importance of considering subdivision invariance in model training.

Nanoparticle Optimization

The hetero-GNN is fully differentiable and takes the features that define the UCNP heterostructure (i.e., layer radii and dopant concentrations) as explicit inputs. For optimization, the hetero-GNN acts as a surrogate model for kMC which is not only orders of magnitude faster, but also provides derivatives of a predicted label with respect to structural features (which are inaccessible with kMC), enabling the use of more powerful gradient-based optimizers to identify UCNP structures that minimize or maximize one or multiple properties.

To explore the utility of our differentiable model for inverse design, we use the hetero-GNN trained on SUNSET-1 with augmented data to search for Yb/Er/Nd-codoped UCNPs with the highest UV/blue intensities under 800-nm excitation. To facilitate the discovery of novel heterostructures, we conduct optimizations far beyond the structural distribution spanned by our training data. While the SUNSET-1 training set contains UCNP structures with up to four regions (i.e., a core and three shells) and with a maximum radius of 11.5 nm, our optimizations explore UCNPs with up to ten regions (i.e., a core and nine shells) and a maximum radius of 15 nm. Further, while the core radius is limited to a maximum of 4 nm and shells are limited to a maximum thickness of 2.5 nm in the training data, we remove both of these upper bounds and only limit the overall radius during optimization. However, we do retain a minimum core radius and shell thickness of 1 nm to ensure that the optimized heterostructures are synthesizeable[42].

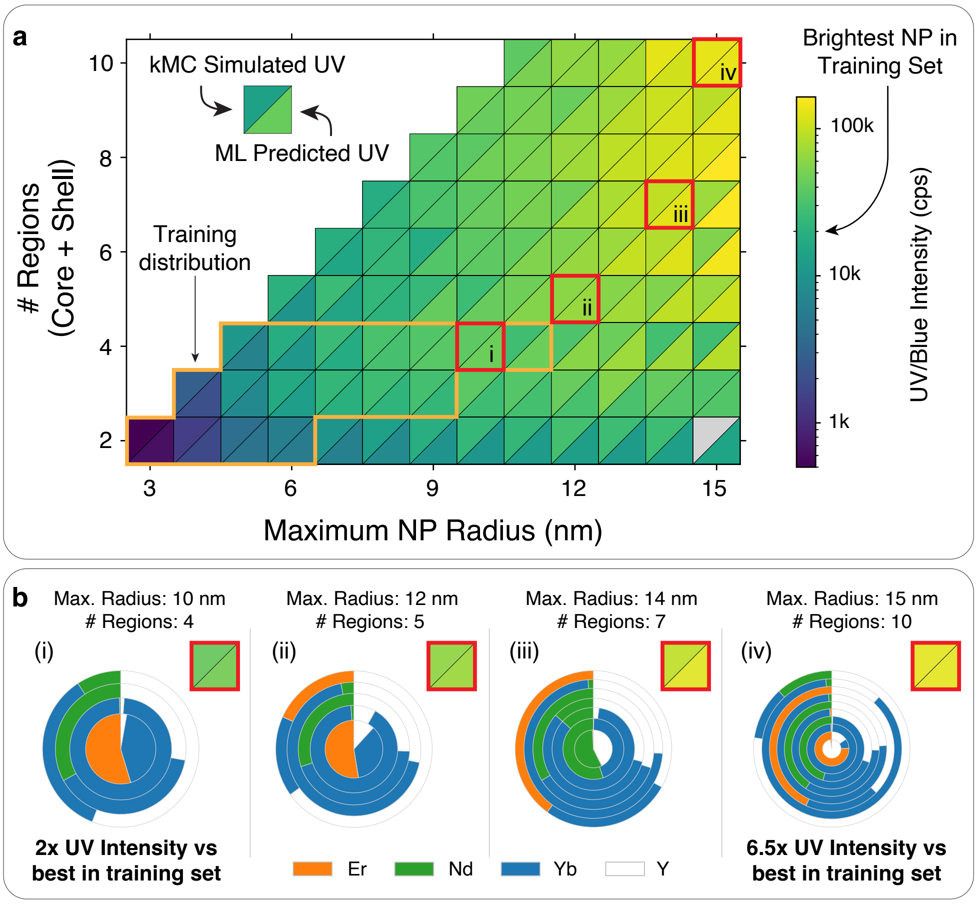


Fig. 5 a) The optimization matrix illustrates I_{vis-UV} intensity predictions for different particle sizes and numbers of layers. Each square represents the total UV and near-UV intensity obtained from validation and predicted by the ensemble of models. b) This panel provides a representative sample of the optimized particles. Note: kMC simultions of the optimized 2-layer 15 nm particle were not possible due to too many Er^{3+} dopants crashing the simulation

Given randomly initialized nanoparticle structures, we employ a combination of trust region constrained local optimization[43, 44] and basin hopping global optimization to identify UCNPs with maximized UV/blue emission as a function of maximum allowed nanoparticle radius and the number of distrinct regions in the heterostructure. Section 5 contains additional details of our optimization approach. The maximum intensities identified for UCNPs of different sizes and different numbers of regions (core + shells) are illustrated in Figure 5a. The optimal structure for each distinct radius + region number was then simulated with kMC, for which larger radius particles often required months-long simulations. The colors of the two triangles in each square indicate the kMC-simulated and ML-predicted intensities for the optimized structure. Remarkably, the model demonstrates accurate predictions for the upconversion luminescence of particles with significantly out-of-distribution radii and numbers of regions, including when the UV emissions approach an order of magnitude higher than those in the training set.

Optimization results display several trends that are well established in the experimental literature on UCNP heterostructures. Optimized UCNPs generally achieve higher absolute brightness at larger diameters, presumably because they are able to host a greater number of absorbing and emitting dopants.[45] The optimized heterostructures for several representative sizes (Fig. 5b) show that the domains of these champion UCNPs are in fact heavily doped – or rather, alloyed,[46] with up to 100% lanthanide substitution – to maximize absorption and emission throughput. Rather than spreading dopants homogeneously through UCNPs, the brightest structures partition Er and Nd dopants into separate shells, reflecting the established knowledge that Er and Nd are prone to quench each other via cross-relaxation.[47]. Since the energy absorbed by Nd must be transferred to Er for upconversion, the optimizer produced structures that separate Nd- and Er-rich domains by a thin shell heavily doped with only Yb.[6]. Such layers transmit the energy absorbed by Nd dopants to the upconverting Er dopants via rapid energy migration through the Yb sublattice. To maximize Nd-Yb and Er-Yb energy transfer, the Nd- and Er-containing shells are also heavily doped with Yb. Many of the GNN-optimized structures, particularly ones with fewer layers, are reminiscent of the 3-layered heterostructure refined by Zhong et al.[6] and others.[37] Nd-rich domains are located in outer shells to maximize absorption by a larger number (volume) of the sensitizers. Meanwhile, an Er-rich core is used to promote upconversion by concentrating absorbed energy into a smaller volume and smaller number of Er activators. The fact that gradient-based optimization of the hetero-GNN surrogate model can rapidly learn design rules that have been developed over decades of UCNP research suggests its potential to discover even more complex and functional nanostructures.

In addition to validating established domain knowledge, the extrapolated results from hetero-GNN optimization provide the opportunity to understand the behavior of complex UCNPs with a greater number of shells than can be readily synthesized or simulated. For example, a major unanswered question is the optimal number of layers for a UCNP, e.g., are more layers better? The optimization matrix in Figure 5a suggests that for smaller particles (below 10 nm radius), moving beyond 2-3 layers does not substantially improve the brightness, most likely because the shell thicknesses would be thinner than characteristic energy transfer distances. However, larger UCNP heterostructures (r ¿10 nm) do see benefit from complex many-shell architectures, with the brightest 14- and 15-nm UCNPs having 7 and 10-shells, respectively. These many-shelled structures also suggest novel strategies to enhance upconversion efficiency. The most striking characteristic of the optimized 12-, 14-, and 15-nm UCNPs (Fig. 5b.ii-iv) is their interleaving of multiple layers of Nd- and Er-rich shells. Rather than converging on one large layer of Nd sensitizer, the optimized 12-nm UCNP sandwiches a layer of Nd sensitizer between two layers of Er activator (with the appropriate Yb buffer layers, as in Figure 5b.ii), while the brightest 15-nm UCNP exhibits the inverse arrangement (Figure 5b.iv). This sandwich shell arrangement allows energy transfer to occur from two sizes, maximizing the number of donors or acceptors within a given distance while minimizing concentration quenching in those outer shells. Curiously, the 14- and 15-nm-radius UCNPs also exhibit motifs in which two Nd-rich shells are separated by an intermediate shell of less concentrated Nd. It is unclear what advantage this motif provides. It is possible that UCNPs may be relatively insensitive to variations near their core (where this motif is observed) since the fraction of dopants is relatively small compared to those in outer shells. This argument may also explain the curious dearth of dopants in the core of the 15-nm UCNP. All of these intriguing structural design motif predictions must be verified and investigated more thoroughly through future experiments (preferably with the aid of precision automated synthesis[39]) and mechanistic analysis of energy transfer pathways from kMC simulations[30] that the hetero-GNN cannot report on.

In a particle utilizing all of these strategies, we find a 6.5x increase in I_{vis-UV}, as compared to the brightest nanoparticle in the training set. Even within the feature distribution, the optimization identifies a particle which utilizes these design rules to achieve a 2x increase in emission intensity as compared to the brightest particle in the training set. These results illustrate that, even within a training distibution but especially far OOD, optimization with a differentiable hetero-GNN can rapidly discover new structures with properties that exceed historical training data and identify novel heterostructure design rules.

Discussion

In order to assess the accuracy of our model prediction during optimization, particularly in the far OOD region, we performed explicit kMC simulations. Particularly for the largest particles, these simulations were extremely expensive, and we terminated many simulations early (after 20%-80% of the requested kMC steps had run) to reduce cost. Overall, the validating kMC simulations took >120,000 CPU-hours on AMD EPYC 7763 and Intel Xeon Gold 6330 CPUs, and individual simulations could take dozens of weeks. All optimizations using our trained hetero-GNN took $\approx$ 2,000 GPU-hours on NVIDIA A100 GPUs. Because optimal particles often emerged early during the optimization process, this GPU-hour figure could probably be reduced by improving our optimization procedure. Coupled with the fact that the kMC simulations could not be directly used for gradient-based optimization, as they are not inherently differentiable, this indicates the massive acceleration in nanomaterial design that can be achieved using DL.

While we have here focused on UCNPs, we believe that the heterogeneous graph representation that we have described and implemented could be suitable to predict heterostructure-dependent properties in other multi-layered nanomaterials. Possible applications include engineering the nanophotonic properties of plasmonic[7, 10] and dielectric nanoparticles,[8] the catalytic properties of polyelemental heterostructures,[48] the optoelectronic properties of complex semiconductor nanoparticle heterostructures,[11] the layer-by-layer assembly of nanoparticles for drug delivery,[49] multilayered magnetic nanospheres,[12] and multilayer graphene sheets for diverse energy and mechanical applications.[50] We also note that, given a DL model which can predict multiple properties controlled by heterostructure, our approach could allow for structural optimization while maximizing or minimizing multiple properties simultaneously.

Conclusion

Inverse chemical and materials design often requires identifying optimal structures in vast search spaces. DL can dramatically accelerate the optimization and design process, but applications of DL are limited by available data and appropriate representations. In this work, we presented SUNSET, a large dataset of emission spectra for core-shell upconverting nanoparticles simulated using explicit kinetic Monte Carlo simulations. To leverage SUNSET for UCNP design, we developed a new heterogeneous graph representation for nanomaterial heterostructures, which we used to train a heterogeneous graph neural network (GNN). We found that this heterogeneous GNN achieved superior in-distribution (ID) and out-of-distribution (OOD) performance compared to existing representations, including vastly higher accuracy compared with a more traditional GNN using a homogeneous representation. Data augmentation, achieved by artificially partitioning UCNP layers, allowed the heterogeneous GNN to approximately learn the subdivision invariance of UCNP emission, improving both ID and OOD accuracy. Applying gradient-based optimization to the heterogeneous GNN trained with augmented data, we identified new UCNP structures with more than 6.5x higher emission intensity than any UCNP in the training set. Optimized particles further elucidated both previously known and novel heterostructure design rules. Our approach has the potential to considerably improve the rate at which we discover new functional nanomaterials and could provide inspiration for applications of DL to underexplored areas of chemistry and nanoscience.

Methods

We utilize kinetic Monte Carlo (KMC) to simulate the optical response of lanthanide-doped nanoparticles. We use the high-performance C++ implementation, as implemented in the RNMC software package.[31] Input generation is handled by NanoParticleTools. We utilize Jobflow [51] and FireWorks [52] to build a workflow for high-throughput kMC simulations, enabling the generation of datasets for machine learning. The effects of parasitic surface ligands are incorporated by including dopant species which are acceptors, that effectively act as energy sinks.

SUNSET Dataset

We present the dataset titled Simulated Upconverting Nanoparticle Spectra for Emissions Tuning (SUNSET). SUNSET consists of 5 sub-datasets which encompass a range of nanoparticles. Evolution of the energy transfer network via kMC are carried out to 10 ms.

Within each sub-dataset, we provide a training (and validation) split, an in-distribution (ID) test split, and an out-of-distribution (OOD) test split. In SUNSET-2, SUNSET-3, & SUNSET-4, the OOD test split also contains the particles with the top 5% and bottom 5% of core sizes. In the case of SUNSET-2 and SUNSET-3, where many of the structures are obtained from Bayesian Optimization,[9] we also partition the highest emitting particles into the OOD test splits. We set these particles aside to ensure the model performance and therefore the optimization is not biased by these pre-optimized configurations. SUNSET-1 comprises a wider parameter space and is sampled in an unbiased manner, thus we do not follow the same splitting scheme. Instead, we use the 4-shell nanoparticles as the OOD, rather than splitting based on emissions.

- **SUNSET-1** targets Er-Nd-Yb system, commonly used to achieve upconversion with 800 nm light.[6, 47] It is comprised of multi-layered UCNPs with a core and up to 4 shells, illustrated in Fig. 2b. Each nanoparticle consists of a doped core with variable radius r_{core} ranging from 1 to 4 nm and doped shells with variable thickness r_{shell} ranging between 1 to 2.5 nm. The total UCNP size ranges between 1 to 13.6 nm.
- **SUNSET-2** targets the Yb-Er system, a pair that is known to absorb at 980 nm and emit UV light. It is comprised of core-shell UCNPs characterized by a total radius of 4 nm. Each UCNP consists of a doped core with variable radii, r_{core}, ranging from 0 to 3.4 nm, enclosed within a fixed outer shell radius of 3.4 nm. Additionally, the UCNP features a cap shell containing surface species, which mimic the parasitic nature of surface ligands in real UCNP systems. The total UCNP size remains constant at 4 nm. Notably, certain particles lack a shell when $r_{core} = 0$ or 3.4 nm.
- **SUNSET-3** also focuses on the Yb-Er system but excludes the presence of surface species. Removing parasitic surface ligands mirrors the effect of growing an inert external shell of undoped $NaYF_4$ over the synthesized UCNP. With the external shell excluded, the UCNP size is limited to 3.4 nm. Parameters, including the core and shell sizes, dopants, and incident wavelength, remain identical.
- **SUNSET-4** expands on SUNSET-2 by introducing Tm as a possible dopant. Like Er, Tm can also facilitate UV emission. All other parameters in SUNSET-4 remain identical to SUNSET-2. This dataset is intended to be used in conjunction with SUNSET-2.

Model Training

A learning rate of 1e-3 is used, with a warm-up period of 10 epochs, where the learning rate linearly increases from 1e-4 to 1e-3. During training, the Mean Squared

Error (MSE) of the validation set is monitored and the learning rate is reduced on plateau, with a patience of 50 epochs. Early stopping is triggered when the validation MSE has not decreased for 200 epochs. Model performance is reported using the MSE and NMSE, as shown in Equations 3-4 below:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (I_{vis\hat{-}UV} - I_{vis-UV})^2 \tag{3}$$

$$\mathrm{NMSE} = \frac{\sum_{i=1}^{N} (I_{vis\hat{-}UV} - I_{vis-UV})^2}{\sum_{i=1}^{N} (I_{vis-UV})^2} \tag{4}$$

where N is the number of UCNPs, $I_{vis\hat{-}UV}$ is the predicted UV emission intensity, and I_{vis-UV} is the actual emission intensity.

Data Augmentation During training of the hetero-GNN model, we utilize data augmentation as discussed in Section 5. We augment our training data on-the-fly by subdividing the input UCNPs with additional random subdivisions each time an UCNP is seen during training. Subdivided layers retain the dopant composition of their original parent layer. For each UCNP input, we randomly subdivide each parent layer into up to 3 child layers in the augmented UCNP. The subdivision is inserted between 5-90% of the parent layer radii.

Optimization For a nanoparticle with N control volumes, we can define the following bounds.

$$0 \leq x_n^i \leq 1, \quad \text{for } i \in [0\,..\,l] \text{ and } n \in [0\,..\,z] \tag{5}$$

$$0 < r_i \leq 1, \quad \text{for } i \in [0\,..\,l] \tag{6}$$

All dopant concentrations are within the closed interval [0, 1]. The fractional radii is defined on an interval of [0, 1] as a fraction of a the maximum nanoparticle size, r$_{max}$ (which we identify a priori). This is necessary to keep on the same interval as the concentration, since the trust region optimizer defines the same trust region for all independent variables $r_{fraction} = r_{true}/r_{max}$

In addition, we define linear constraints that bound the total concentration within each layer [0, 1] and restrict the thickness of each layer.

$$0 \leq \sum_{j=z*n}^{z*n+z} x_j \leq 1, \quad \text{for } n \in [0\,..\,l-1] \tag{7}$$

$$c_{min} \leq r_0 \leq c_{max} \tag{8}$$

$$t_{min} \leq r_{n+1} - r_n \leq t_{max}, \quad \text{for } n \in [0\,..\,l-1] \tag{9}$$

We initialize random starting configuration within the distributions outlined for each dataset as a starting point for optimization. We perform local optimization using the trust region constrained optimization as implemented in scipy[43] with an initial trust radius of 1.0. An initial constraint penalty of 1×10^3 was applied to strongly penalize constraint violation, ensuring that concentrations stayed within the range of [0, 1.0] and total radius within specification. The criterion used for termination of local optimization is when the trust radius is less than 1×10^{-8} To search for a globally optimal particle, we repeatedly perturb the local minima and re-optimize the nanoparticle heterostructure. To achieve this, we utilize the basinhopping functionality of scipy with up to 500 steps, a step size of 0.15, and temperature value of 0.25. Following global optimization, validation of the best identified candidate structures are subjected to kMC simulations.

Supplementary Information

Additional information about data pre-processing; consideration of the effect of $NaYF_4$ structure on simulation results; details for tabular, image, and homogeneous graph representations for UCNPs; dataset biases and attempts to mitigate bias; optimal model hyperparameters; additional details regarding model validation and optimization.

Acknowledgments

This work was primarily funded by the Charter Hill Laboratory Directed Research and Development program of Lawrence Berkeley National Laboratory (LBNL), supported by the Office of Science, Office of Basic Energy Sciences (BES), of the U.S. Department of Energy (DOE) under Contract No. DE-AC02-05CH11231. This DOE-BES contract also supported work at the Molecular Foundry as well as computational resources at the National Energy Research Scientific Computing Center (NERSC, award BES-ERCAP0023292) and the Lawrencium computational cluster provided by the LBNL IT Division. We are especially grateful for the NERSC and Lawrencium low-priority queues, without which this work would not have been possible. L.A. was supported by the DOE Computational Science Graduate Fellowship under Award No. DE-SC0022158. E.W.C.S.-S. was supported by the Carnegie Bosch Institute Postdoctoral Fellowship. We thank Karen Chiao for helpful discussions related to optimizing C++ code.

Data Availability

The SUNSET dataset is freely available for download on Figshare (https://figshare.com/s/49222bae78f228363897). Each subset (e.g. SUNSET-1) is presented in Javascript Object Notation (JSON) format; separate JSON files are also provided for ID and OOD collections for each subset. The hetero-GNN models used for optimization and UCNP design are freely downloadable as checkpoint files on Figshare (https://doi.org/10.6084/m9.figshare.27941694.v1). The nanoparticle structures discovered by optimizing the hetero-GNN model are available on Figshare (https://doi.org/10.6084/m9.figshare.27973206).

Code Availability

The `RNMC` program, which contains the `NPMC` kMC tool, can be found at https://github.com/BlauGroup/RNMC. Code defining the machine learning representations, data featurization, and model training can be found athttps://github.com/BlauGroup/NanoParticleTools.

Conflicts of Interest

There are no conflicts of interest to declare.

Author Contributions

E.S. and D.B. implemented nanoparticle kinetic Monte Carlo in C++, and X.X. provided validation against a previous implementation. E.S. and X.X. built simulation analysis capabilities in Python. E.S. implemented the high-throughput workflow infrastructure. E.S. and X.X. ran kMC simulations to construct SUNSET. E.S. implemented RFR, FCNN, CNN, and homogeneous GNN models and trained on SUNSET data. E.S. and L.A. implemented the heterogeneous GNN model and trained on SUNSET data. L. A. and E.S. implemented the on-the-fly data augmentation scheme. E.S. implemented and carried out global optimization, and then ran kMC on the optimized particles to validate predicted intensities. B.S.-L. provided important input on components of the GNN architecture and global optimization. E.S., L.A., E.S.-S., E.M.C., and S.M.B. wrote the manuscript. E.M.C. and S.M.B. conceived of and supervised the project.

References

[1] Voznyy, O., Sutherland, B.R., Ip, A.H., Zhitomirsky, D., Sargent, E.H.: Engineering charge transport by heterostructuring solution-processed semiconductors. Nature Reviews Materials **2**(6), 17026 (2017) https://doi.org/10.1038/natrevmats.2017.26

[2] Sun, W., Zheng, Y., Yang, K., Zhang, Q., Shah, A.A., Wu, Z., Sun, Y., Feng, L., Chen, D., Xiao, Z., *et al.*: Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science advances **5**(11), 4275 (2019)

[3] Sanchez-Lengeling, B., Aspuru-Guzik, A.: Inverse molecular design using machine learning: Generative models for matter engineering. Science **361**(6400), 360–365 (2018)

[4] Wu, S., Kondo, Y., Kakimoto, M.-a., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., *et al.*: Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. Npj Computational Materials **5**(1), 66 (2019)

[5] Skripka, A., Lee, M., Qi, X., Pan, J.-A., Yang, H., Lee, C., Schuck, P.J., Cohen, B.E., Jaque, D., Chan, E.M.: A generalized approach to photon avalanche upconversion in luminescent nanocrystals. Nano Letters **23**(15), 7100–7106 (2023) https://doi.org/10.1021/acs.nanolett.3c01955 https://doi.org/10.1021/acs.nanolett.3c01955. PMID: 37471584

[6] Zhong, Y., Tian, G., Gu, Z., Yang, Y., Gu, L., Zhao, Y., Ma, Y., Yao, J.: Elimination of photon quenching by a transition layer to fabricate a quenching-shield sandwich structure for 800 nm excited upconversion luminescence of nd3+-sensitized nanoparticles. Advanced Materials **26**(18), 2831–2837 (2014) https://doi.org/10.1002/adma.201304903 https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201304903

[7] So, S., Mun, J., Rho, J.: Simultaneous inverse design of materials and structures via deep learning: Demonstration of dipole resonance engineering using core–shell nanoparticles. ACS Applied Materials & Interfaces **11**(27), 24264–24268 (2019) https://doi.org/10.1021/acsami.9b05857 https://doi.org/10.1021/acsami.9b05857. PMID: 31199610

[8] Peurifoy, J., Shen, Y., Jing, L., Yang, Y., Cano-Renteria, F., DeLacy, B.G., Joannopoulos, J.D., Tegmark, M., Soljačić, M.: Nanophotonic particle simulation and inverse design using artificial neural networks. Science Advances **4**(6), 4206 (2018) https://doi.org/10.1126/sciadv.aar4206 https://www.science.org/doi/pdf/10.1126/sciadv.aar4206

[9] Xia, X., Sivonxay, E., Helms, B.A., Blau, S.M., Chan, E.M.: Accelerating the design of multishell upconverting nanoparticles through bayesian optimization. Nano Letters **23**(23), 11129–11136 (2023) https://doi.org/10.1021/acs.nanolett.3c03568

[10] Liu, G.-X., Liu, J.-F., Zhou, W.-J., Li, L.-Y., You, C.-L., Qiu, C.-W., Wu, L.: Inverse design in quantum nanophotonics: combining local-density-of-states and deep learning. Nanophotonics **12**(11), 1943–1955 (2023) https://doi.org/10.1515/nanoph-2022-0746

[11] Hamachi, L.S., Yang, H., Jen-La Plante, I., Saenz, N., Qian, K., Campos, M.P., Cleveland, G.T., Rreza, I., Oza, A., Walravens, W., Chan, E.M., Hens, Z., Crowther, A.C., Owen, J.S.: Precursor reaction kinetics control compositional grading and size of cdse1-xsx nanocrystal heterostructures. Chem. Sci. **10**, 6539–6552 (2019) https://doi.org/10.1039/C9SC00989B

[12] Albrecht, M., Hu, G., Guhr, I.L., Ulbrich, T.C., Boneberg, J., Leiderer, P., Schatz, G.: Magnetic multilayers on nanospheres. Nature Materials **4**(3), 203–206 (2005)

[13] Xu, M., Liang, T., Shi, M., Chen, H.: Graphene-like two-dimensional materials. Chemical reviews **113**(5), 3766–3798 (2013)

[14] Chen, Y., Fan, Z., Zhang, Z., Niu, W., Li, C., Yang, N., Chen, B., Zhang, H.: Two-dimensional metal nanomaterials: synthesis, properties, and applications. Chemical reviews **118**(13), 6409–6455 (2018)

[15] Xie, C., Niu, Z., Kim, D., Li, M., Yang, P.: Surface and interface control in nanoparticle catalysis. Chemical reviews **120**(2), 1184–1249 (2019)

[16] Haghighatlari, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X., Head-Gordon, T.: Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. Chem **6**(7), 1527–1542 (2020)

[17] Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (1988) https://doi.org/10.1021/ci00057a005 . Accessed 2023-06-20

[18] Krenn, M., Häse, F., Nigam, A., Friederich, P., Aspuru-Guzik, A.: Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Machine Learning: Science and Technology **1**(4), 045024 (2020) https://doi.org/10.1088/2632-2153/aba947 . Accessed 2023-05-01

[19] O'Brien, M.N., Jones, M.R., Mirkin, C.A.: The nature and implications of uniformity in the hierarchical organization of nanomaterials. Proceedings of the National Academy of Sciences **113**(42), 11717–11725 (2016)

[20] Liang, R., Xue, Y., Fu, X., Le, A.N., Song, Q., Qiang, Y., Xie, Q., Dong, R., Sun, Z., Osuji, C.O., *et al.*: Hierarchically engineered nanostructures from compositionally anisotropic molecular building blocks. Nature Materials **21**(12), 1434–1440 (2022)

[21] Brown, K.A., Brittman, S., Maccaferri, N., Jariwala, D., Celano, U.: Machine learning in nanoscience: big data at small scales. Nano Letters **20**(1), 2–10 (2019)

[22] Zhou, J., Shen, L., Costa, M.D., Persson, K.A., Ong, S.P., Huck, P., Lu, Y., Ma, X., Chen, Y., Tang, H., *et al.*: 2dmatpedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. Scientific data **6**(1), 86 (2019)

[23] Reker, D., Rybakova, Y., Kirtane, A.R., Cao, R., Yang, J.W., Navamajiti, N., Gardner, A., Zhang, R.M., Esfandiary, T., L'Heureux, J., *et al.*: Computationally guided high-throughput design of self-assembling drug nanoparticles. Nature nanotechnology **16**(6), 725–733 (2021)

[24] McCandler, C.A., Dahl, J.C., Persson, K.A.: Phosphine-stabilized hidden ground states in gold clusters investigated via a au n (ph3) m database. ACS nano **17**(2), 1012–1021 (2022)

[25] Barnard, A., Motevalli, B., Parker, A., Fischer, J., Feigl, C., Opletal, G.: Nanoinformatics, and the big challenges for the science of small things. Nanoscale **11**(41), 19190–19201 (2019)

[26] Hlaváček, A., Farka, Z., Mickert, M.J., Kostiv, U., Brandmeier, J.C., Horák, D., Skládal, P., Foret, F., Gorris, H.H.: Bioconjugates of photon-upconversion nanoparticles for cancer biomarker detection and imaging. Nature Protocols

17(4), 1028–1072 (2022)

[27] Lu, Y., Zhao, J., Zhang, R., Liu, Y., Liu, D., Goldys, E.M., Yang, X., Xi, P., Sunna, A., Lu, J., *et al.*: Tunable lifetime multiplexing using luminescent nanocrystals. Nature Photonics **8**(1), 32–36 (2014)

[28] Sanders, S.N., Schloemer, T.H., Gangishetty, M.K., Anderson, D., Seitz, M., Gallegos, A.O., Stokes, R.C., Congreve, D.N.: Triplet fusion upconversion nanocapsules for volumetric 3d printing. Nature **604**(7906), 474–478 (2022)

[29] He, M., Pang, X., Liu, X., Jiang, B., He, Y., Snaith, H., Lin, Z.: Monodisperse dual-functional upconversion nanoparticles enabled near-infrared organolead halide perovskite solar cells. Angewandte Chemie International Edition **55**(13), 4280–4284 (2016)

[30] Teitelboim, A., Tian, B., Garfield, D.J., Fernandez-Bravo, A., Gotlin, A.C., Schuck, P.J., Cohen, B.E., Chan, E.M.: Energy transfer networks within upconverting nanoparticles are complex systems with collective, robust, and history-dependent dynamics. The Journal of Physical Chemistry C **123**(4), 2678–2689 (2019) https://doi.org/10.1021/acs.jpcc.9b00161

[31] Zichi, L., Barter, D., Sivonxay, E., Spotte-Smith, E.W.C., Srinivaas, M.R., Chan, E., Persson, K.A., Blau, S.M.: RNMC: kinetic Monte Carlo implementations for complex reaction networks. The Journal of Open Source Software **9**(104), 7244 (2024)

[32] Wu, X., Zhang, Y., Takle, K., Bilsel, O., Li, Z., Lee, H., Zhang, Z., Li, D., Fan, W., Duan, C., Chan, E.M., Lois, C., Xiang, Y., Han, G.: Dye-sensitized core/active shell upconversion nanoparticles for optogenetics and bioimaging applications. ACS Nano **10**(1), 1060–1066 (2016) https://doi.org/10.1021/acsnano.5b06383 https://doi.org/10.1021/acsnano.5b06383. PMID: 26736013

[33] Corso, G., Stark, H., Jegelka, S., Jaakkola, T., Barzilay, R.: Graph neural networks. Nature Reviews Methods Primers **4**(1), 17 (2024)

[34] Shi, R., Mudring, A.-V.: Phonon-mediated nonradiative relaxation in ln3+-doped luminescent nanocrystals. ACS Materials Letters **4**(10), 1882–1903 (2022)

[35] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[36] Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)

[37] Chan, E.M., Levy, E.S., Cohen, B.E.: Rationally designed energy transfer in upconverting nanoparticles. Advanced Materials **27**(38), 5753–5761 (2015) https:

//doi.org/10.1002/adma.201500248

[38] Szalkowski, M., Kotulska, A., Dudek, M., Korczak, Z., Majak, M., Marciniak, L., Misiak, M., Prorok, K., Skripka, A., Schuck, P.J., Chan, E.M., Bednarkiewicz, A.: Advances in the photon avalanche luminescence of inorganic lanthanide-doped nanomaterials. Chem. Soc. Rev., 10–1039400177 (2025) https://doi.org/10.1039/D4CS00177J

[39] Chan, E.M.: Combinatorial approaches for developing upconverting nanomaterials: high-throughput screening, modeling, and applications. Chemical Society Reviews **44**(6), 1653–1679 (2015)

[40] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)

[41] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.: E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature communications **13**(1), 2453 (2022)

[42] Xu, X., Zhou, Z., Liu, Y., Wen, S., Guo, Z., Gao, L., Wang, F.: Optimising passivation shell thickness of single upconversion nanoparticles using a time-resolved spectrometer. APL photonics **4**(2) (2019)

[43] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2

[44] Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, ??? (2000)

[45] Gargas, D.J., Chan, E.M., Ostrowski, A.D., Aloni, S., Altoe, M.V.P., Barnard, E.S., Sanii, B., Urban, J.J., Milliron, D.J., Cohen, B.E., al.: Engineering bright sub-10-nm upconverting nanocrystals for single-molecule imaging. Nature Nanotechnology **9**(4), 300–305 (2014)

[46] Tian, B., Fernandez-Bravo, A., Najafiaghdam, H., Torquato, N.A., Altoe, M.V.P., Teitelboim, A., Tajon, C.A., Tian, Y., Borys, N.J., Barnard, E.S., al.: Low irradiance multiphoton imaging with alloyed lanthanide nanocrystals. Nature Communications **9**(1), 3082 (2018)

[47] Xie, X., Gao, N., Deng, R., Sun, Q., Xu, Q.-H., Liu, X.: Mechanistic investigation

of photon upconversion in nd3+-sensitized core–shell nanoparticles. Journal of the American Chemical Society **135**(34), 12608–12611 (2013) https://doi.org/10.1021/ja4075002 https://doi.org/10.1021/ja4075002. PMID: 23947580

[48] Wahl, C.B., Aykol, M., Swisher, J.H., Montoya, J.H., Suram, S.K., Mirkin, C.A.: Machine learning–accelerated design and synthesis of polyelemental heterostructures. Science Advances **7**(52), 5505 (2021) https://doi.org/10.1126/sciadv.abj5505 https://www.science.org/doi/pdf/10.1126/sciadv.abj5505

[49] Yan, Y., Such, G.K., Johnston, A.P., Lomas, H., Caruso, F.: Toward therapeutic delivery with layer-by-layer engineered particles. ACS nano **5**(6), 4252–4257 (2011)

[50] Lee, J.-H., Loya, P.E., Lou, J., Thomas, E.L.: Dynamic mechanical behavior of multilayer graphene via supersonic projectile penetration. Science **346**(6213), 1092–1096 (2014)

[51] Rosen, A.S., Gallant, M., George, J., Riebesell, J., Sahasrabuddhe, H., Shen, J.-X., Wen, M., Evans, M.L., Petretto, G., Waroquiers, D., *et al.*: Jobflow: Computational workflows made simple. Journal of Open Source Software **9**(93), 5995 (2024)

[52] Jain, A., Ong, S.P., Chen, W., Medasani, B., Qu, X., Kocher, M., Brafman, M., Petretto, G., Rignanese, G.-M., Hautier, G., *et al.*: Fireworks: a dynamic workflow system designed for high-throughput applications. Concurrency and Computation: Practice and Experience **27**(17), 5037–5059 (2015)