ARTICLE   OPEN

Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics

Zekun Ren[1,2,7]*, Felipe Oviedo [ID][3,7]*, Maung Thway [ID][2], Siyu I. P. Tian[1,2], Yue Wang[1], Hansong Xue[2], Jose Dario Perea[3,4], Mariya Layurova[3], Thomas Heumueller [ID][4,5], Erik Birgersson[6], Armin G. Aberle [ID][2], Christoph J. Brabec[4,5], Rolf Stangl[2], Qianxiao Li[6], Shijing Sun[3], Fen Lin[2], Ian Marius Peters[3] and Tonio Buonassisi[1,3]*

Process optimization of photovoltaic devices is a time-intensive, trial-and-error endeavor, which lacks full transparency of the underlying physics and relies on user-imposed constraints that may or may not lead to a global optimum. Herein, we demonstrate that embedding physics domain knowledge into a Bayesian network enables an optimization approach for gallium arsenide (GaAs) solar cells that identifies the root cause(s) of underperformance with layer-by-layer resolution and reveals alternative optimal process windows beyond traditional black-box optimization. Our Bayesian network approach links a key GaAs process variable (growth temperature) to material descriptors (bulk and interface properties, e.g., bulk lifetime, doping, and surface recombination) and device performance parameters (e.g., cell efficiency). For this purpose, we combine a Bayesian inference framework with a neural network surrogate device-physics model that is 100× faster than numerical solvers. With the trained surrogate model and only a small number of experimental samples, our approach reduces significantly the time-consuming intervention and characterization required by the experimentalist. As a demonstration of our method, in only five metal organic chemical vapor depositions, we identify a superior growth temperature profile for the window, bulk, and back surface field layer of a GaAs solar cell, without any secondary measurements, and demonstrate a 6.5% relative AM1.5G efficiency improvement above traditional grid search methods.

INTRODUCTION

Process optimization is essential to reach maximum performance of novel materials and devices. This is especially relevant for photovoltaic devices, as numerous process variables can influence their performance. Often, process optimization is done using black-box optimization methods (e.g., Design of Experiments,[1] Grid Search,[2] Bayesian Optimization,[3,4] Particle Swarm Optimization,[5] etc.), in which selected variables are modified systematically within a range and the system's response surface is mapped to reach an optimum. These methods have shown potential for inverse design of materials and systems in a cost-effective manner, and are usually postulated as ideal methods for future self-driving laboratories.[6–13] However, traditional black-box optimization approaches have limitations: the maximum achievable performance improvement is limited by the designer's choice of variables and their ranges, artificially constraining the parameter space. Furthermore, insights into the root causes of underperformance are severely limited, often requiring secondary characterization methods or batches composed of combinatorial variations of the base samples. In contrast, recently, Bayesian inference coupled to a physics-based forward model and rapid, light-dependent and temperature-dependent, current–voltage measurements were shown to offer a statistically rigorous approach to identify the root cause(s) of underperformance in early-stage photovoltaic devices.[14] Further-more, recently, the combination of physical insights with machine-

learning models have shown good promise in development of energy materials.[15–23]

In this contribution, we consider the optimization of the synthesis temperature profile of a gallium arsenide (GaAs) solar cell using a metal organic chemical vapor deposition (MOCVD) reactor. Growth temperature is one of the most important and challenging parameters to optimize in III–V film deposition.[24,25] Previous studies showed that the growth temperature has an impact on the film's growth rate, surface morphology, dopant incorporation, and defect formation.[24,25] Other important process parameters include precursor flowrate and growth pressure. These process parameters are closely related, and the relationship can be approximated using the Ideal Gas Law in the kinetic epitaxy process.[26] Therefore, we use the growth temperature as the key optimization variable. For other secondary process variables, for example, precursor type and carrier gas flowrate, the physical relation between process variables and material properties is unclear and likely tool specific,[27] we can replace the physics-based parametrization in the first layer of the Bayesian network inference with a machine-learning model with higher capacity, such as kernel ridge regression.[28]

GaAs solar cells comprise several layers, including a back surface field (BSF), a bulk absorber, and a window layer.[29] To maximize device performance, material properties need to be optimized for each layer and interface.[20,30] An experienced researcher would
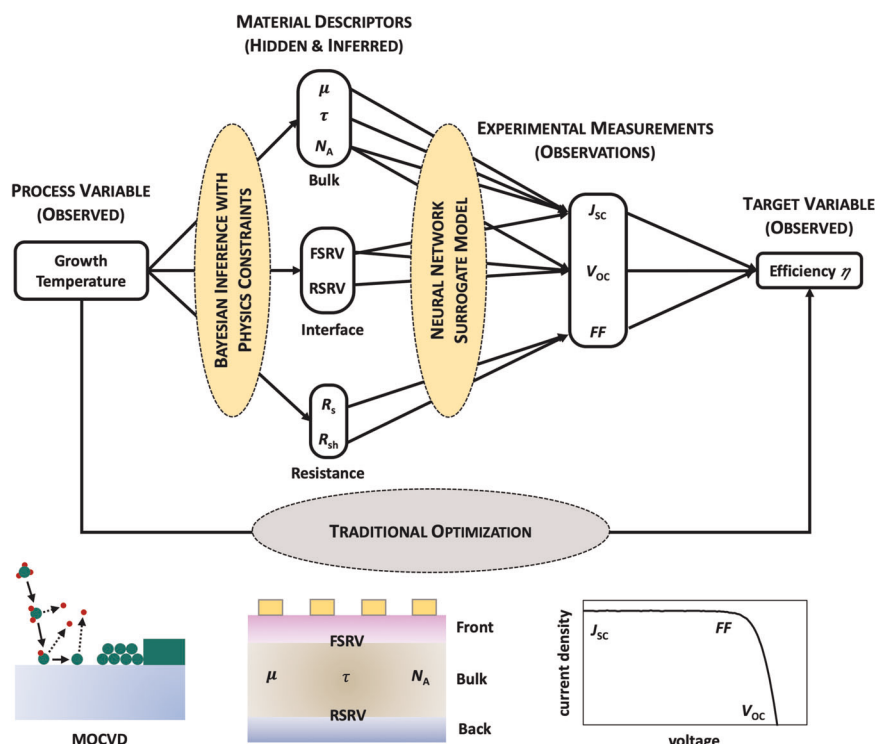
Fig. 1 Schematic of our Bayesian network-based process-optimization model, featuring a two-step Bayesian inference that first links process conditions to material descriptors, and then the latter to device performance. Our Bayesian network-based process optimization back propagates from efficiency to bulk interface properties and then to growth temperature, enabling layer-by-layer tuning of process variables.

grow and characterize each layer (emitter, base, window, and BSF) separately to map the process variable to material properties, in an attempt to gain physical insights to optimize the final solar cells efficiency.[25,31–33] In this context, optimizing growth temperature of GaAs solar cells becomes an optimization scenario in which one process variable (temperature) affects several material descriptors in various device layers. With the assistance of a solar cell physical simulator and additional characterization techniques, the optimal growth temperature for each layer could be pinpointed and the whole device stack could be grown using the optimized growth profile. However, this expert approach requires fabricating many auxiliary samples at varying conditions with multiple layer variations, and use secondary characterization measurements, such as secondary ion mass spectroscopy (SIMS) and time-resolved photoluminescence (TR-PL), to confirm root causes of underperformance. These characterization techniques are significantly more costly than current–voltage (JV) measurements, the primary proxy of solar cell performance. This problem mirrors the challenges in optimizing other multi-layer energy systems and semiconductors, including light-emitting diodes, power electronics, thermoelectrics, batteries, and transistors.

To address this challenge, we combine several machine-learning techniques to infer the effects of a given process variable on different device layers. To avoid performing expensive characterization, such as SIMS or TR-PL, we perform automated JV measurement at multiple illumination intensities (JVi) as the input for the algorithm. To speed up our calculations, we employ a physics-based "surrogate" model that mimics a complex physical model, in this case solar cell growth. Our surrogate model consists of a two-step Bayesian inference method, typically referred as Bayesian network or hierarchical Bayes,[34–36] with relations between layers constrained by physical laws. Embedded therein is a surrogate device-physics model, which operates >100× faster

(shown in Supplementary Fig. 1) than a numerical device-physics solver.

Figure 1 shows the schematic of our Bayesian network. We propose three methodological innovations in this approach. First, we create a parameterized process model by imposing physics-based constraints to couple the process-optimization variable (e.g., growth temperature) to the resulting material's bulk and interface properties (e.g., lifetime). This parametrization limits the number of fitting variables in the first layer of our Bayesian inference model, reducing the risk of overfitting, and provides a degree of interpretability. Second, we add another inference layer inside a numerical device-physics model, linking the inferred bulk and interface properties to the solar cell performance measures (e.g., JVi characteristics, quantum efficiency, and energy conversion efficiency). Extraction of underlying materials descriptors from JVi curves, previously demonstrated in ref.,[14] enables us to trace the root cause(s) of device underperformance to a specific material or interface property. Third, we achieve a >100× acceleration in inference by replacing the solar cell model, a traditional PDE (partial differential equation) numerical model, with a highly accurate neural network surrogate model. This enables us to update the posterior probability distribution for our Bayesian network inference model over a vast parameter space.

In Fig. 1, we also show the difference between our Bayesian network-based optimization and the traditional black-box optimization. As only low-cost evaluations (JVi measurement) are performed for solar cell characterization, accurate extraction of underlying material properties requires performing Bayesian inference using a device-physics model.[14] Traditional optimization approaches often make use of a purely black-box surrogate model[3] to map the relation between process variables and device performance directly, without any insights about material properties in the device. In contrast, our Bayesian network inference

connects target variables to material descriptors, then to process conditions. It provides rich, layer-by-layer information about critical material properties that affects device's electrical performance. In this study, we chose to map doping concentration in emitter and bulk, bulk lifetime (τ), front and rear (indium gallium phosphide) InGaP/GaAs surface recombination velocities (SRVs) to growth temperature using, and customize the growth temperature that maximize those desired material properties. Replacing the traditional optimization (process variable-device performance) with our Bayesian network-based optimization (process variable-material properties-device performance) feasibly enables us to expand the variable space, and identify design process windows that selectively improve specific materials, layers, and interfaces inside a solar cell. This results in vastly improved device performance and process interpretability in few MOCVD fabrication rounds with a single temperature sweep.

To demonstrate the potential of our approach, we use our Bayesian network to characterize and optimize, in a single temperature sweep consisting of five MOCVD fabrication rounds, the process temperature of a GaAs solar cell. Our devices have a baseline efficiency of ~16% without an anti-reflection coating (ARC). Our Bayesian network approach identifies the optimal set of process conditions that translate into maximum performance under the physical model and real process constraints. The physical insights from the Bayesian network inference suggest an optimal growth temperature profile, allowing a significant 6.5% relative increase in average AM1.5G efficiency above baseline in a single temperature sweep (sixth MOCVD run). This result verifies the capacity of our approach to find optimal process windows with little intervention from the experimentalist, no secondary characterization techniques or auxiliary samples, and with performance beyond experimentalist-constrained optimization.

RESULTS

As illustrated in Fig. 1, we construct a Bayesian network to link the process variables with each material and device property in the GaAs solar cells. Hereafter, we optimize each material property separately to maximize the final device performance. The Bayesian network consists of four parts.

Parameterization of process variables by embedding physics knowledge

This section describes how we define physics-based relations between process variables and materials descriptors, embedding physics domain knowledge, and ensuring faster and better convergence of our Bayesian optimization algorithm. This corresponds to the progression from "Process Conditions" to "Materials Descriptors" in Fig. 1. Device fabrication of solar cells is expensive, thus it is essential to explore the process variable space efficiently.[37] From a machine-learning point of view, we leverage the existing knowledge from literature and embed such domain knowledge as prior parameterization to constrain the variable space, for example, Eq. [2]. The parameterization connects process variables with underlying material and interface properties. In this study, we chose to infer emitter and bulk doping concentration, bulk lifetime (τ), front and back SRV as the intermediate material properties because they play a critical role in determining the device electrical performance,[29] and each property is layer or interface specific. In the case of chemical vapor deposition (CVD), recognizing that growth temperature affects several thermally and kinetic activated processes,[38] we embed such knowledge and enforce an exponential dependence of underlying material properties based on the modified Arrhenius equation[39–41] (Eq. [2]).

The detailed schematic of the Bayesian network inference is shown in Fig. 2. To illustrate the flow of our approach, we use the optimization of acceptor (Zn) doping concentration in a GaAs solar cell as a showcase. Our approach can be represented as a two-step Bayesian inference procedure using conditional probability (Eq. [1]):

$$P(J|T) = \int P(J, N_A|T) \, dN_A = \int P(N_A|T) * P(J|N_A, T) \, dN_A, \quad (1)$$

where $P(N_A|T)$ is the conditional probability of Zn acceptor doping levels given the process temperature. We parameterize the prior ($P(N_A|T)$) based on existing literature and our physical knowledge. Recognizing that MOCVD growth is a kinetic process,[38] we enforce an Arrhenius equation-type of parameterization to link the underlying material properties with growth temperatures. Zn doping level can be represented in the modified Arrhenius equation [2]:

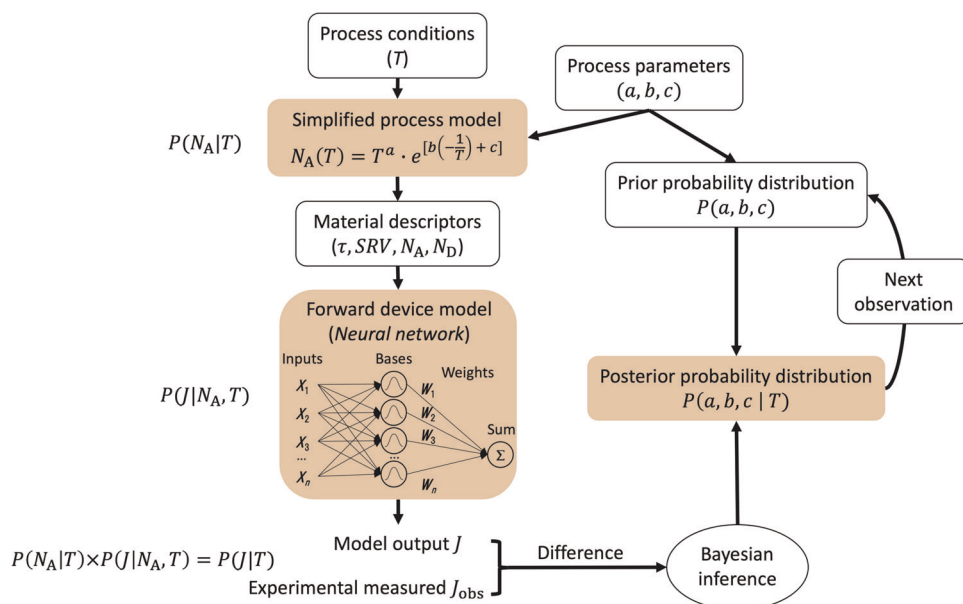$$N_A(T) = T^a \exp\left[b\left(-\frac{1}{T}\right) + c\right], \quad (2)$$

Fig. 2 Architecture of our Bayesian inference network to identify new windows for process optimization.

where (a, b, c) are latent process parameters that are inferred from the Bayesian framework. b and c correspond to the activation energy and pre-exponential factor in the traditional Arrhenius equation. a is the temperature dependence of the pre-exponential factor ($\ln(c)$). Aside from Zn doping concentration, Si doping concentration, bulk minority carrier lifetime (τ), and front and back SRVs are also parameterized in the same fashion.

The modified Arrhenius equation form for the doping concentration agrees well with trends reported in the literature.[42–44] There is insufficient literature and domain expertise to directly relate bulk and interface properties with the growth temperature. However, a previous study has shown that τ and SRVs are correlated with doping concentration.[45,46] Note that performing the fitting of Eq. [2] can be an implicit hypothesis test. A small a value suggests that the pre-exponential factor temperature dependence is suppressed, and that the Arrhenius relationship governs the temperature dependence of the particular bulk, interface, or resistance property. On the other hand, a big a value suggests a larger contribution of the pre-exponential factor to temperature dependence, indicating a deviation from a pure Arrhenius-like regime at a given temperature. Additional domain knowledge is embedded in the prior by setting hard constraints for the material properties. The ranges of the five inferred material properties are shown in Supplementary Table 2.

Inference of material and device properties from device measurements

This section describes the progression from "Materials Descriptors" to "Target Variable: Performance" in Fig. 1. Inference of underlying material properties from JVi measurements is used to trace the root cause(s) of device underperformance to specific material or interface properties. We further extend the connection between process variables and material properties to device measurements by adding an additional inference layer. The forward model of this inference layer is a numerical device-physics model, linking the inferred bulk and interface properties to solar cell device parameters (e.g., JV characteristics, quantum efficiency, and conversion efficiency).

Following the above example, $P(J|N_A,T)$ is the conditional probability of a set of JVi observations at a series of fixed illumination intensities given the underlying material parameters (Zn doping concentration). The material property–JVi relation is extensively investigated and can be solved numerically using a well-developed device model from literature.[29,47–49] A well-calibrated PC1D model[47] is used in this work. However, numerical simulation is computationally expensive in the Bayesian framework (which requires hundreds of thousands of simulated runs to provide adequate posterior probability estimation) and makes it difficult to integrate new features into the model. Furthermore, experimental JVi observations contain experimental noise that causes deviations from simulated JVi curves.

Replacement of numerical solver with a robust neural network surrogate model

To circumvent the computational complexity of the numerical device-physics model and the discrepancies between experimental and simulated JVi curves, we replace the numerical simulator with a surrogate deep neural network. Figure 3 shows a schematic of the model, consisting of two parts: (1) a denoising Autoencoder (AE)[50] that takes noisy JVi curves as input and reconstructs noise-free JVi curves. In our case, the training data are 20,000 simulated JVi curves, computed with a device-physics model, and augmented with Gaussian noise that mimics experimental noise. The Gaussian noise has a 0 mean and 0.2% variance, which are estimated from the repeated JV measurements. The structure of the encoder network is shown in Supplementary Fig. 2, and consists of three convolutional and two dense layers in the encoder and three convolution transpose and two dense layers in the decoder. The decoder is a mirror of the encoder network, with transposed convolution layers replacing the convolutional layers. The denoising training of the architecture provides robustness to experimental noise. (2) A regression model that predicts the JVi
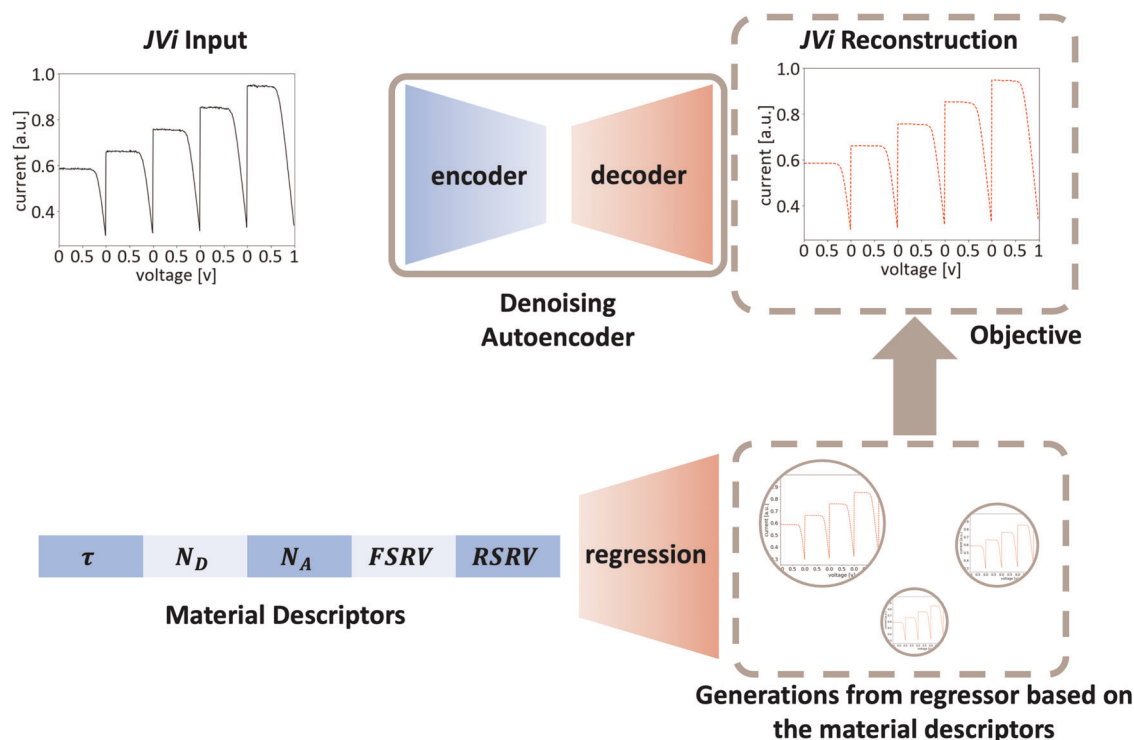

JVi Input

JVi Reconstruction

encoder **decoder**

Denoising Autoencoder

Objective

τ | N_D | N_A | FSRV | RSRV | **regression**

Material Descriptors

Generations from regressor based on the material descriptors

Fig. 3 Schematic of neural network surrogate model to infer material descriptors from *JVi* curves. (In this figure, five sequential JV curves are shown as inputs and outputs, with increasing illumination intensity).

curves based on underlying material properties. The regression model has the same structure as the decoder used in the denoising AE. $P(J|N_A,T)$ thus can be computed using this surrogate neural network model.

To create the training dataset, we first randomly sample a set of 20,000 random material properties (τ, FSRV, RSRV, Zn, and Si doping concentration) from uniform probability distributions. The threshold of the uniform distribution is shown in Supplementary Table 2. Then, we use scripted PC1D[51] to numerically simulate a set of 20,000 JVi curves from the chosen material descriptors. Although domain expertise is required in setting up the numerical PC1D model, this exercise is a one-time implementation for each material system. Subsequently, we augment the simulated JVi curves with Gaussian noise to mimic the experimental measurements and feed the noisy JVi to train the denoising AE. Figure 3 shows the noise-free JVi curves after we feed the experimental data to the AE. Hereafter, we train the neural network regression model to predict JVi curves from material descriptors in the latent space of the AE. The loss for both AE and regression model is chosen to constraint the latent parameter space to the five variables of interest, and is minimized using the ADAM gradient descent algorithm with a batch size of 128 and an initial learning rate of 0.0001. We split the JVi curves into 80 and 20% for training and testing purposes. The numerical solver in the Bayesian network is then replaced by the regression model. The surrogate model is significantly more computationally efficient than the numerical solver. Supplementary Fig. 1 shows the acceleration by adapting the neural network surrogate for calculation of a set of JVi curves. The surrogate model, running on a GPU, is 130 times faster than the PC1D numerical solver and 700 times faster if the numerical solver is called externally.

Once the device model is trained, we connect these previous two Bayesian inference steps into a hierarchical structure using Eq. [1]. A posterior probability to every combination of latent fitting parameters (a, b, c) is assigned. This probability is represented by a multivariate probability distribution over all possible combinations of model fit parameters. This probability is modified every time new data (JVi measurement) is observed. We apply a Markov Chain Monte Carlo (MCMC) method for sampling the posterior distribution of latent parameters (a, b, c); this achieves an acceleration of Bayesian inference computation time comparable or superior to the successive grid subdivision method.[2] Specifically, we implement the affine-invariant ensemble sampler of MCMC proposed by Goodman and Weare[52] using an external library.[53] With each newly observed JVi measurement, the posterior distributions of the latent process parameters (a, b, c) are updated. Hereafter, the material descriptor (Zn doping concentration as a function of growth temperature ($N_A(T)$) can be obtained from Eq. [2].

In an analogous way, other descriptors, such as the doping levels of other species and bulk and interface recombination properties, can be obtained as a function of the process variables and adequate prior parametrizations. We use this result to optimize the MOCVD growth temperature of a set of GaAs solar cells.

Optimizing solar cells using our Bayesian network inferred results

After we map the growth temperature to the material properties, we apply grid search method with 10 °C resolution to find the growth temperature that maximizes the desired material properties and minimizes the undesired properties (maximize τ and minimize SRVs) for the solar cell. Mathematically, we can define the optimization procedure enabled by our Bayesian network model as:

$$x^* = \arg\max(h(g_i(x))). \tag{3}$$

The variable x^* is the set of process variable, specifically the

MOCVD growth temperature, that produce the largest solar cell efficiency. We first estimate the function $g_i(x)$, which models how the set of underlying material properties changes with the process variable. Hereafter, the cell efficiency can be maximized by plugging material properties $y_i = g_i(x)$ to $h(y_i)$, which models the relation between material properties and the final solar cell performance (JVi curves). $h(y_i)$ can be solved numerically and, in our case, is estimated using a neural network. The material properties extracted can be exploited to find x^* that maximizes the cell efficiency. As $g_i(x)$ determines the functional relation of material descriptors and the process variable, we can tailor our process variable to maximize the desired materials properties, such as lifetime, and minimize the undesired properties, such as SRVs, in selected locations across the devices.

As a baseline for testing our methodology, we fabricate five batch of GaAs solar cells (four cells per batch), sweeping a range of constant growth temperatures. The GaAs solar cell structure consists of multiple InGaP and GaAs layers (Supplementary Fig. 3), and all solar cell layers are grown at the same temperature in one MOCVD experiment. In five experiments, a temperature range of 530–680 °C, with 20–50 °C temperature intervals, is explored. The films are fabricated into 1 cm^2 solar cells, without ARCs. Detailed growth and fabrication procedures can be found in the Experimental procedures section. JVi measurements under multiple illumination intensities (0.1–1.1 suns) are performed. Figure 4 shows the inferred material properties as a function of MOCVD growth temperatures. We can see that the logarithm value of p-type (Zn) doping level, n-type (Si) doping level, and FSRV have an almost linear correlation with $1/T$, suggesting a good agreement with the standard Arrhenius equation, while the bulk lifetime and
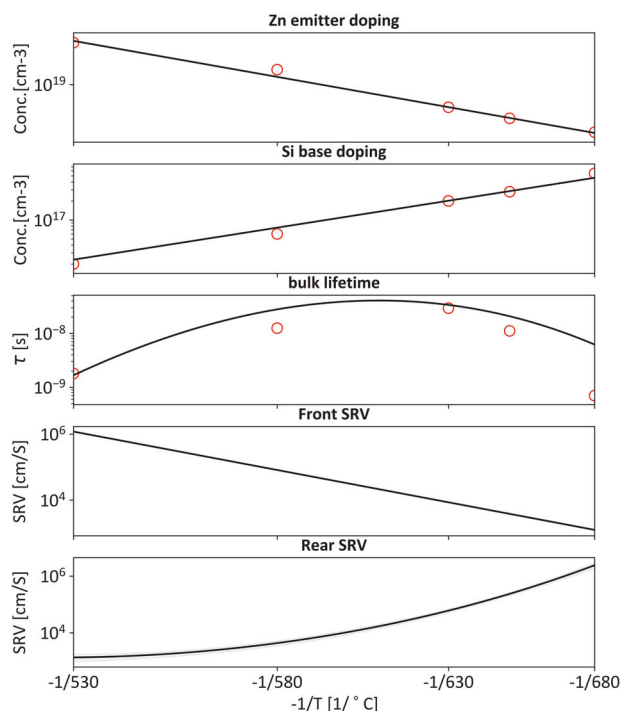

Fig. 4 Bayesian network reveals how each material descriptor (bulk and interface property) varies with processing conditions. Black lines show inferred values of material descriptors as a function of growth temperature; red circles show experimental validation of material descriptors using SIMS and TR-PL. Doping concentrations of different species (Zn and Si), bulk lifetime, and InGaP/GaAs interface SRV can be inferred from finished device measurements using this approach. The x-axis is $-1/T$ and the y-axis is in logarithm scale to illustrate whether the material property follows the Arrhenius equation (linear trend).

Table 1. Fitted mean value of latent parameter (*a*, *b*, *c*) for different material properties.

	a	*b*	*c*
Zn doping	0.0018	−0.1494	−0.1948
Si doping	0.0016	0.1551	0.2970
Bulk lifetime	−4.5973	2.7984	2.3687
Front SRV	0.0015	−0.1440	−0.1892
Rear SRV	2.1194	−1.1119	−0.7300

Table 2. Temperature profile of GaAs solar cells.

Layer	Optimal growth temperature suggested by Bayesian network (°C)
Buffer	580
InGaP BSF	580
GaAs bulk	620
InGaP window	650
GaAs contact	650

RSRV exhibits nonlinear relationships. To trace the root causes, mean of (*a*, *b*, *c*) values for each material properties after the MCMC run are attached in Table 1. The full distribution of the (*a*, *b*, *c*) values extracted from the Bayesian network is shown in Supplementary Fig. 4. The *a* values for both the Zn and Si doping concentration are close to zero (<0.002), indicating a negligible temperature dependence in the pre-exponential factor and the Arrhenius regime is dominant. This agrees well with the trend reported in the literature for various dopant species.[42–44] The *a* value for FSRV is also insignificant, and FSRV has similar trend as the Zn doping concentration. We postulate that this is due to SRVs being affected by doping concentrations,[45,46] and the dominant recombination mechanism in the front interface being related to Zn doping level. The *a* value for effective bulk lifetime is significant (−4.59), indicating a strong temperature dependence on the pre-exponential factor and thus non-Arrhenius regime. We postulate that this could result from the existence of both Zn and Si dopant in the GaAs bulk layer, as there is a competing contribution from the two dopants. The *a* value for RSRV lies between the value of bulk lifetime and Si doping levels. The RSRV slightly follows the linear trend of Si doping levels; however, we postulate that the subsequent bulk, front, and contact layers' growth impact on the rear interface's quality,[45] and contributes to the non-Arrhenius behavior.

To validate the inferred doping concentrations and lifetime from our Bayesian network approach, we grow auxiliary structures (e.g., single-layer structure to conduct SIMS measurement and heterostructure for TR-PL measurements). The red circles in the first three subplots of Fig. 4 represent the results from those independent auxiliary measurements. The experimental details are shown in Supplementary Fig. 5. It is evident that the independent measurements agree well with the inferred material properties.

It is interesting to note that each parasitic recombination parameter (bulk lifetime, FSRV, and RSRV) has its minimum/maximum at a different growth temperature. The bulk lifetime peaks around 620 °C, which is close to our baseline process temperature (630 °C). The front and rear SRVs exhibit opposite trends when growth temperature increases. Instead of growing the whole GaAs stack at the same temperature, Fig. 4 indicates that the back contact, bulk, and front contact should each be grown at a different temperature to optimize performance.

This knowledge gained by the Bayesian network enables us to formulate a new time temperature profile (Table 2) for our GaAs devices (labeled "Bayes Net" in Fig. 5). We performed an additional MOCVD experiment by selecting the growth temperature show in Table 2 that minimizes recombination at each layer or interface in a 10 °C resolution (hardware tolerance) grid. We avoid extreme conditions (e.g., 680 °C), which show deterioration of overall device performance (Fig. 5) despite inferred layer-specific improvements (Fig. 4).

Figure 5 shows the spread of GaAs cells' efficiency for the five MOCVD experiments and the additional MOCVD run with the customized growth profile. Without additional insights on material properties from our Bayesian network, cell efficiency become the sole optimization target. The grid search on growth temperature
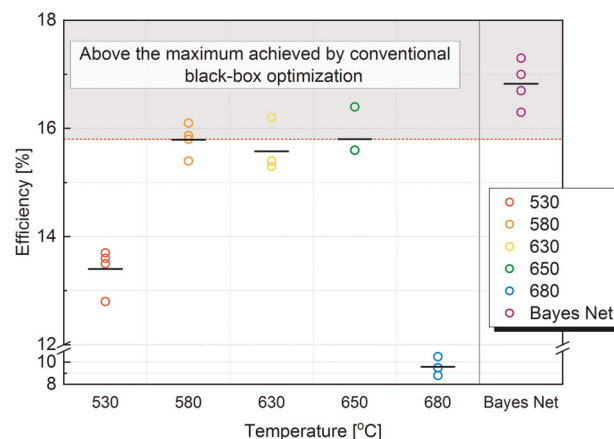

Fig. 5 Comparison of "grid search optimization" versus our approach using a Bayesian network (Bayes Net). GaAs cell efficiency varies with growth temperature, reaching an average maximum between 580 °C and 650 °C. Our Bayesian network-informed process (labeled "Bayes Net") tunes the growth temperature of each layer to minimize recombination (Fig. 4), increasing efficiency by 6.5% relative. The gray area represents the additional efficiency gain that cannot be achieved using conventional grid search optimization. Please note that black-box optimization methods searching in the constant temperature space would have underperformed compared to the Bayesian network results.

suggests growing the whole solar cell stack at 580 °C or 650 °C are the optimal growth scheme. This temperature sweep (i.e., a single cycle of learning) gives us an efficiency improvement of 1.4% relative above our baseline efficiency (630 °C) after five MOCVD runs. The sixth MOCVD run that tunes growth temperatures of each layer (Table 2), thereby minimizing layer-specific recombination, achieve a 6.5% relative improvement over the baseline, well exceeding the conventional approach. Auxiliary one-Sun *JV* and external quantum efficiency (EQE) measurements are performed to trace the root causes of efficiency improvement using the customized temperature profile (Fig. 6). It shows that both J_{SC} and V_{OC} are responsible for the efficiency improvements in our "Bayes Net" growth temperature profile. EQE shows that photo-response at wavelengths <820 nm (corresponding to an optical penetration depth comparable to our 2-μm-thick absorber) is improved, indicating significant reduction in recombination for the front and bulk layers. We perform Bayesian inference (second layer in the Bayesian network) to extract the material properties of measured *JVi* curve of this cell and our baseline. The mean values of FSRV, RSRV, and τ of the cells grown using "Bayes Net" profile (Table 2) are 1.2×10^3 cm/s, 5.4×10^4 cm/s, and 29 ns, while the best baseline values are 4.1×10^3 cm/s, 6.1×10^4 cm/s, and 26 ns. These values agree qualitatively well with the EQE observations from auxiliary measurement, which show the front surface and the bulk benefiting the most from the "Bayes Net" temperature profile, possibly because these were the highest-temperature steps, and
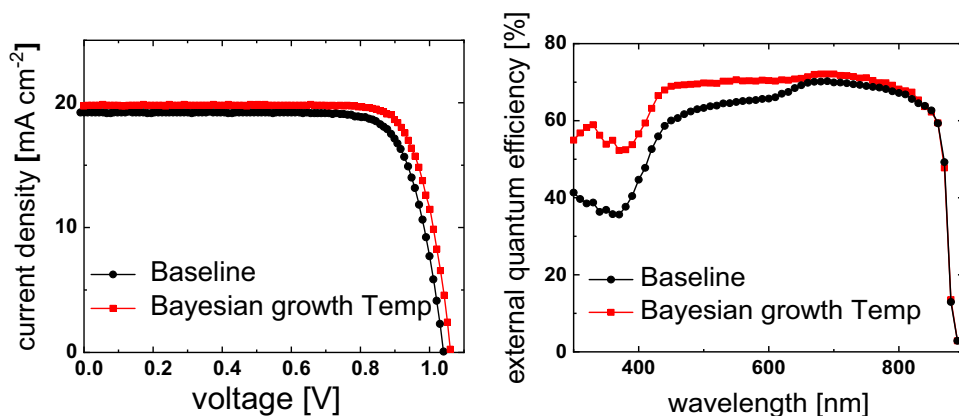
Fig. 6 *JV* and EQE measurement of GaAs solar cells with the custom growth temperature profile informed by our Bayesian network (red) and baseline 630 °C (black).

that may have partially erased the thermal history of the underlying rear-surface layer.

All cells reported herein do not have ARCs; the best cells shown in the figure are estimated to have efficiencies in the 24–25% range with optimal double-layer ARCs. The efficiency value is near state of the art for a single-junction GaAs with substrate[54] grown in an academic setting. Other process parameters, for example, epitaxial lift-off and contact grid optimization, are required to reach record efficiencies.[30] Nevertheless, the recombination gains enabled by the variable-temperature profile by our Bayesian network should translate to these advanced cell architectures. It is important to note that, given the shape of the function to optimize, any other black-box optimization methods in the constant temperature space would have underperformed in comparison to the Bayesian network. Growing the device stack at the constant temperature will never achieve the same level of improvement as what is demonstrated using the Bayesian network. This is the case because tuning layer-by-layer growth temperature only becomes evident when we perform Bayesian inference to map the *JVi* measurements to underlying material properties. This demonstrates how additional insights gained via Bayesian network-based optimization can be translated into device performance that exceeds black-box optimization. One could argue that similar performance can be achieved by following the "expert" approach to perform single-layer optimization before incorporating them into a device stack. However, many auxiliary structures' growth and secondary measurements will be required in this case. Fifteen SIMS and TR-PL samples were grown in this study for model validations. The fact that the optimal variable-temperature profile is found after a single temperature sweep of five MOCVD runs at constant temperatures, verifies the potential of our method to accelerate time-consuming device optimization significantly, limiting the need of synthesizing auxiliary samples and performing secondary measurements. Lastly, we can modify the Bayesian network approach by replacing the physics parametrization (Arrhenius equation) with regularized black-box regression, in those cases when the physics between the process variable and material properties are unclear or complex to model. In Supplementary Fig. 6, we demonstrate that the temperature can be mapped to the material properties manifold with a similar accuracy by replacing the Arrhenius equation parametrization with a black-box regression model (kernel ridge regression using radial basis function[28]). Incorporating black-box regression in the first layer of the Bayesian network enables one to describe complex process variables. However, the performance of the black-box regression will be affected by the hyperparameter values shown in Supplementary Fig. 6. Because the experiments are expensive in our case and data scarcity, the

initial selection of hyperparameters in the black-box regression can be critical. Furthermore, the interpretability of the Bayesian network will surfer as latent process parameters (*a*, *b*, *c*) cannot be inferred in the black-box regression case.

DISCUSSION

We developed and applied a Bayesian network approach to GaAs solar cell growth optimization. This approach enables us to exceed our baseline efficiency by 6.5% relative, by tuning process variables layer by layer, in just six MOCVD experiments. Our approach is enabled by implementing physics-informed relations between process variables and materials descriptors, and embedding these into a Bayesian network. We link these material descriptors to device performance using a neural network surrogate model, which is 100× faster than numerical device simulation. The small number of growth (MOCVD) runs necessary to implement this layer-by-layer process-optimization scheme translate into significant cost and time reductions compared to common black-box optimization methods. We believe this approach can generalize to other solar cell materials,[55,56] as well as other systems with physics-based or black-box relations between process variables and materials descriptors, and physics-based device performance models. Our surrogate model can replace common models in closed-loop black-box optimization, such as a Gaussian process regression in Bayesian optimization, providing good functional fitting and physical insights.

METHODS

Experimental procedures

The top GaAs cell was fabricated on epi-ready <100>-oriented GaAs on-axis wafers using an AIXTRON Crius MOCVD reactor. The growth is performed under a reactor pressor of 100 mbar using TMGa, TMIn, AsH$_3$, and PH$_3$ as precursors and 32 standard liters per minute H$_2$ as carrier gas. It has a 3 μm *n*-doped GaAs base (Si dopant) and 100 nm *p*$^+$-doped GaAs emitter (Zn dopant). Highly doped InGaP is used as the window (Zn dopant) and BSF layer (Si dopant). *p*$^+$-doped GaAs layer (carbon dopant) is added at the front surface to achieve an ohmic contact to the metal fingers. The solar cells are metalized using an E-beam evaporator and a shadow mask to fabricate a grid pattern with ~8% shading. A SIMS measurement is conducted for the GaAs films that are grown in the same batch before metallization. We additionally grow *n*-doped InGaP/GaAs/InGaP heterostructure with two different base thicknesses (500 and 1000 nm) to measure the bulk lifetime of the *n*-doped GaAs bulk.[57] The growth conditions for the heterostructure are identical to the conditions for GaAs solar cells.

DATA AVAILABILITY

Part of the experimental and simulated dataset for training the neural network surrogate model and predicting the optimal growth conditions from Bayesian network is available in the following GitHub repository: https://github.com/PV-Lab/BayesProcess. Additional data supporting the findings of this study is available from the authors upon reasonable request.

CODE AVAILABILITY

The code used for Bayesian network and neural network surrogate to predict material properties is also available at https://github.com/PV-Lab/BayesProcess. Additional code supporting the findings of this study is available from the authors upon reasonable request.

REFERENCES

1. Hicks, C. R. & Turner, K. V. *Fundamental Concepts in the Design of Experiments* Vol. 40 (Oxford University Press, New York, 1999).
2. Kurchin, R., Romano, G. & Buonassisi, T. Bayesim: a tool for adaptive grid model fitting with Bayesian inference. *Comput. Phys. Commun.* **239**, 161–165 (2019).
3. Frazier, P. I. A tutorial on Bayesian optimization. (2018). Preprint at https://arxiv.org/abs/1807.02811.
4. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. N. Phoenics: A Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
5. Kennedy, J. In *Encyclopedia Machine Learning*. Particle Swarm Optimization (ed). 760–766 (Springer, 2010).
6. Yuan, R. et al. Accelerated discovery of large electrostrains in $BaTiO_3$-based piezoelectrics using active learning. *Adv. Mater.* **30**, 1702884 (2018).
7. Xue, D. et al. Accelerated search for $BaTiO_3$-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proc. Natl Acad. Sci. USA* **113**, 13301–13306 (2016).
8. Yamawaki, M., Ohnishi, M., Ju, S. & Shiomi, J. Multifunctional structural design of graphene thermoelectrics by Bayesian optimization. *Sci. Adv.* **4**, 4192 (2018).
9. Kumar, J. N. et al. Machine learning enables polymer cloud-point engineering via inverse design. *npj Comput. Mater.* **5**, 73 (2019). https://doi.org/10.1038/s41524-019-0209-9.
10. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
11. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
12. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the harvard clean energy project: the use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
13. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
14. Brandt, R. E. et al. Rapid photovoltaic device characterization through Bayesian parameter estimation. *Joule* **1**, 843–856 (2017).
15. Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **5**, 60 (2019).
16. Sun, S. et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **3**, 1437–1451 (2019).
17. Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018).
18. Ong, S. P. Accelerating materials science with high-throughput computations and machine learning. *Comput. Mater. Sci.* **161**, 143–150 (2019).
19. Bai, Y. et al. Accelerated discovery of organic polymer photocatalysts for hydrogen evolution from water through the integration of experiment and theory. *J. Am. Chem. Soc.* **141**, 9063–9071 (2019).
20. Baloch, A. A. et al. Full space device optimization for solar cells. *Sci. Rep.* **7**, 11984 (2017).
21. Cao, B. et al. How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics. *ACS Nano* **12**, 7434–7444 (2018).
22. Huang, Q. Physics-driven Bayesian hierarchical modeling of the nanowire growth process at each scale. *IIE Trans.* **43**, 1–11 (2010).
23. Chen, C. H. et al. Physics-prior Bayesian neural networks in semiconductor processing. *IEEE Access* **7**, 130168–130179 (2019).
24. Bass, S. et al. Effect of growth temperature on the optical, electrical and crystallographic properties of epitaxial indium gallium arsenide grown by MOCVD in an atmospheric pressure reactor. *J. Cryst. Growth* **79**, 378–385 (1986).
25. Schmieder, K. J. et al. Effect of growth temperature on GaAs solar cells at high MOCVD growth rates. *IEEE J. Photovolt.* **7**, 340–346 (2016).
26. Xenidou, T. C. et al. An experimental and computational analysis of a MOCVD process for the growth of Al films using DMEAA. *Surf. Coat. Technol.* **201**, 8868–8872 (2007).
27. Ren, Z. et al. Ultra-thin GaAs double-junction solar cell with carbon-doped emitter. *IEEE J. Photovolt.* **8**, 1627–1634 (2018).
28. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
29. Ren, Z. et al. Numerical analysis of radiative recombination and reabsorption in GaAs/Si tandem. *IEEE J. Photovolt.* **5**, 1079–1086 (2015).
30. Kayes, B. M. et al. 27.6% conversion efficiency, a new record for single-junction solar cells under 1 sun illumination. In *Photovoltaic Specialists Conference (PVSC), 2011 37th IEEE*. 000004–000008 (IEEE, 2011).
31. Ikeda, E., Takamoto, T., Kurita, H. & Ohmori, M. 17% efficient InGaP/GaAs solar cells by improved minority carrier lifetime. In *Proceedings of 1994 IEEE Sixth International Conference on Indium Phosphide and Related Materials (IPRM)* 500–503 (IEEE, 1994).
32. Augustine, G., Rohatgi, A. & Jokerst, N. M. Base doping optimization for radiation-hard Si, GaAs, and InP solar cells. *IEEE Trans. Electron. Dev.* **39**, 2395–2400 (1992).
33. Sodabanlu, H. et al. Extremely high-speed GaAs growth by MOVPE for low-cost PV application. *IEEE J. Photovolt.* **8**, 887–894 (2018).
34. Lenk, P. J., DeSarbo, W. S., Green, P. E. & Young, M. R. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Mark. Sci.* **15**, 173–191 (1996).
35. Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J. & Jarvis, E. D. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**, 3594–3603 (2004).
36. Pearl, J. *Causality: Models, Reasoning and Inference* Vol. 29 (Springer, 2000).
37. Correa-Baena, J.-P. et al. Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2**, 1410–1420 (2018).
38. Coltrin, M. E. & Mitchell, C. C. Mass transport and kinetic limitations in MOCVD selective-area growth. *J. Cryst. Growth* **254**, 35–45 (2003).
39. Laidler, K. J. The development of the Arrhenius equation. *J. Chem. Educ.* **61**, 494 (1984).
40. Petty, H., Arakawa, E. & Baird, J. Methods to determine the temperature dependence of the pre-exponential factor of the Arrhenius equation from thermogravimetric data. *J. Therm. Anal.* **11**, 417–422 (1977).
41. Laidler, K. J. A glossary of terms used in chemical kinetics, including reaction dynamics (IUPAC Recommendations 1996). *Pure Appl. Chem.* **68**, 149–192 (1996).
42. Kuech, T., Veuhoff, E. & Meyerson, B. Silicon doping of GaAs and $AlGa_{1-x}As$ using disilane in metalorganic chemical vapor deposition. *J. Cryst. Growth* **68**, 48–53 (1984).
43. Hudait, M. K., Modak, P., Hardikar, S. & Krupanidhi, S. Zn incorporation and band gap shrinkage in p-type GaAs. *J. Appl. Phys.* **82**, 4931–4937 (1997).
44. Sun, S., Armour, E., Zheng, K. & Schaus, C. Zinc and tellurium doping in GaAs and $AlGa_{1-x}As$ grown by MOCVD. *J. Cryst. Growth* **113**, 103–112 (1991).
45. Aspnes, D. Recombination at semiconductor surfaces and interfaces. *Surf. Sci.* **132**, 406–421 (1983).
46. Lush, G. B. et al. A study of minority carrier lifetime versus doping concentration in n-type GaAs grown by metalorganic chemical vapor deposition. *J. Appl. Phys.* **72**, 1436–1442 (1992).
47. Clugston, D. A. & Basore, P. A. PC1D version 5: 32-bit solar cell modeling on personal computers. In *Photovoltaic Specialists Conference, 1997, Conference Record of the Twenty-Sixth IEEE*. 207–210 (IEEE, 1997).
48. Niemegeers, A., Gillis, S. & Burgelman, M. A user program for realistic simulation of polycrystalline heterojunction solar cells: SCAPS-1D. In *Proceedings of the Second World Conference on Photovoltaic Energy Conversion, JRC, European Commission, juli.* 672–675 (1998).
49. Guide, S. D. U. & Version, G. Synopsis Inc. (2012). https://www.synopsys.com/silicon/tcad/device-simulation/sentaurus-device.html.
50. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*. 1096–1103 (ACM, 2008).
51. Haug, H., Olaisen, B. R., Nordseth, Ø. & Marstein, E. S. A graphical user interface for multivariable analysis of silicon solar cells using scripted PC1D simulations. *Energy Procedia* **38**, 72–79 (2013).
52. Goodman, J. & Weare, J. Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.* **5**, 65–80 (2010).

53. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Publ. Astron. Soc. Pac.* **125**, 306 (2013).

54. Miller, O. D., Yablonovitch, E. & Kurtz, S. R. Strong internal and external luminescence as solar cells approach the Shockley–Queisser limit. *IEEE J. Photovolt.* **2**, 303–311 (2012).

55. Mangan, N. M. et al. A path to 10% efficiency for tin sulfide devices. In *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)* 2373–2378 (IEEE, 2014).

56. Cui, P. et al. Planar p–n homojunction perovskite solar cells with efficiency exceeding 21.3%. *Nat. Energy* **4**, 150 (2019).

57. Ahrenkiel, R., Keyes, B. & Dunlavy, D. Intensity-dependent minority-carrier lifetime in III–V semiconductors due to saturation of recombination centers. *J. Appl. Phys.* **70**, 225–231 (1991).

AUTHOR CONTRIBUTIONS

Z.R. and F.O. contributed equally for this work. Z.R., F.O., and T.B. conceived of the research. Z.R. and F.O. developed and tested the Bayesian network, with key intellectual contributions from S.I.P.T., H.X., J.D.P., M.L., T.H., E.B., C.J.B., R.S., S.S., Q.L., I.M.P., and T.B. Z.R. and F.O. developed the neural network surrogate model for fast inference. Z.R. fabricated the GaAs solar cells with the help from Y.W., M.T., A.A., and F.L. Z.R, F.O, and T.B. wrote the manuscript, with input from all co-authors.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-0277-x.

Correspondence and requests for materials should be addressed to Z.R., F.O. or T.B.

Reprints and permission information is available at http://www.nature.com/reprints

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