Embedding Physics Domain Knowledge into a Bayesian Network Enables Layer-by-Layer Process Innovation for Photovoltaics

Zekun Ren1*, Felipe Oviedo2*, Muang Thway3, Siyu I.P. Tian1, Yue Wang1, Hansong Xue2, Jose Dario Perea4,2, Mariya Layurova2, Thomas Heumueller4,5, Erik Birgersson6, Armin Aberle3, Christoph J. Brabec4,5, Rolf Stangl3, Shijing Sun2, Fen Lin3, Ian Marius Peters2, Tonio Buonassisi1,2*

1Singapore-MIT Alliance for Research and Technology SMART, 138602 Singapore
2Massachusetts Institute of Technology, Cambridge MA 02139, U.S.A.
3Solar Energy Research Institute of Singapore (SERIS), 117574 Singapore
4Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander University Erlangen-Nürnberg, 91058 Erlangen, Germany
5Helmholtz Institute HI-ErN, Forschungszentrum Jülich, Immerwahrstrasse 2, D-91058 Erlangen, Germany
6National University of Singapore, 119077, Singapore
*These authors contributed equally.

& Corresponding authors:
Zekun Ren
Singapore-MIT Alliance for Research and Technology, 138602 Singapore
+65-6516-8603
zekun@smart.mit.edu

Felipe Oviedo
77 Massachusetts Ave., Bldg. 35-135, Cambridge MA 02139
+1-617-642-1529
foviedo@mit.edu

Tonio Buonassisi
77 Massachusetts Ave., Bldg. 35-213, Cambridge MA 02139
+1-617-324-5130
buonassisi@mit.edu

Abstract: Process optimization of photovoltaic devices is a time-intensive, trial-and-error endeavor, without full transparency of the underlying physics, and with 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 that identifies the root cause(s) of underperformance with layer-by-layer resolution and reveals alternative optimal process windows beyond global black-box
optimization. Our Bayesian-network approach links process conditions to materials descriptors (bulk and interface properties, e.g., bulk lifetime, doping, and surface recombination) and device-performance parameters (e.g., cell efficiency), using a Bayesian inference framework with an autoencoder-based surrogate device-physics model that is 100x faster than numerical solvers. With the trained surrogate model, our approach is robust and reduces significantly the time-consuming experimentalist intervention, even with small numbers of fabricated samples. To demonstrate our method, we perform layer-by-layer optimization of GaAs solar cells. In a single cycle of learning, we find an improved growth temperature for the GaAs solar cells without any secondary measurements, and demonstrate a 6.5% relative AM1.5G efficiency improvement above baseline and traditional black-box optimization methods.

**Keywords:** solar cell, process optimization, solar device, machine learning, Bayesian network, Autoencoder, neural network, GaAs, process innovation, deep learning
I. Introduction:

Process optimization is essential to extract maximum performance from novel materials and devices. This is especially relevant for photovoltaic devices, as numerous process variables can influence the performance of these devices. Often, process optimization is performed using black-box optimization methods, (e.g., Design of Experiments\textsuperscript{1}, Bayesian Optimization\textsuperscript{2,3}, Particle Swarm Optimization\textsuperscript{4}), 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\textsuperscript{5-12}. 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, the parameter space is artificially constrained, and 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 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\textsuperscript{13}. Furthermore, recently, the combination of physical insights with machine learning models have shown good promise in development of energy materials\textsuperscript{14-21}.

In this contribution, we consider the use case of optimizing the synthesis time-temperature profile of a gallium arsenide (GaAs) solar cell. GaAs solar cells comprise several layers,
including a back contact, a bulk absorber, and a front contact\textsuperscript{22}; to maximize device performance, one must optimize the bulk and interface properties of each layer and interface\textsuperscript{19,23}. In essence, this is an optimization scenario in which one process variable (temperature) affects several materials descriptors in different device layers, and decoupling their various influences is challenging with only one macroscopic electrical response variable (efficiency). The usual approach requires fabricating dozens or hundreds of samples at varying conditions with various layer variations, and use secondary characterization measurements to confirm root causes of underperformance. The process is very time consuming, and requires a significant amount of supervision and expertise by the experimentalist. This mirrors the challenges in optimizing other multi-layer energy systems, including light-emitting diodes, power electronics, thermoelectrics and batteries, to name a few.

To address this challenge, we combine several machine learning techniques to infer the effects of a given process variable on different device layers, thus enabling accelerated and highly automated layer-by-layer optimization. To speed up our calculations, we employ a physics-based “surrogate” model that mimics the behavior of 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\textsuperscript{24-26}, with relations between layers constrained by physical laws. Embedded therein is a surrogate device-physics model, which operates $>100x$ faster (shown in Figure S3 in Supplementary Information) than a numerical device-physics solver.
Figure 1 shows the schematic of our Bayesian Network. We perform three methodology innovations in the design of this two-step Bayesian inference. First, we create a parameterized process model by imposing physics-based constraints to couple process optimization variables \((e.g.,\) growth temperature) to the resulting material’s bulk and interface properties \((e.g.,\) lifetime). This enables us to limit the number of fitting variables in the first layer of our Bayesian inference model, reducing the risk of overfitting. Second, we add another inference layer inside a numerical device-physics model, linking the inferred bulk and interface properties with solar cell device parameters \((e.g.,\) current-voltage characteristics, quantum efficiency, and conversion efficiency). Extraction of underlying materials descriptors from illumination current-voltage curves, previously demonstrated in Ref\(^{13}\), enables us to trace the root cause(s) of device underperformance to a specific material or interface property. Third, we achieve a >100x acceleration in inference by replacing the traditional numerical device-physics model’s partial differential equations with a highly accurate neural network surrogate model. This enables us to update the posteriors for our Bayesian network model over a vast parameter space.

In Figure 1, we also show the difference between our Bayesian network and the traditional black-box optimization approach. Traditional optimization approaches make use of purely black-box surrogate models\(^2\) such as Response Surface Models, Gaussian Process or Random Forest Regression, to map the relation between process variables and device performance directly, obfuscating any insights about material properties in the device. In contrast, our Bayesian network backpropagates from target variables to material descriptors, then to process conditions. It provides rich, layer-by-layer information about material properties such as doping concentration, bulk lifetime \((\tau)\), surface recombination
velocity (SRV) and parasitic resistances. Replacing the black-box surrogate model with our Bayesian Network feasibly enables us to expand the variable space and 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 cycles of learning.
Figure 1. Schematic of our Bayesian-network-based process-optimization model, featuring a two-step Bayesian inference that first links process conditions to materials descriptors, then the latter to device performance. Our Bayesian-network-based process-optimization back propagates from efficiency to bulk interface properties then to growth temperature, enabling layer-by-layer tuning of process variables.

To demonstrate the potential of our approach, we use our Bayesian network to characterize and optimize, in a single learning step, the process parameters of a solar cell. As test case, we use gallium arsenide (GaAs)-based devices fabricated by Metal Organic Chemical Vapor Deposition (MOCVD) with baseline efficiency of ~16% without Anti-Reflection Coating (ARC). Our Bayesian network approach allows us to identify an optimal set of
process conditions that translate into a maximum performance under the physical model and real process constraints. In the GaAs case, the physical insights of the Bayesian network suggest an optimal growth temperature profile, allowing a significant 6.5% relative (1.4% absolute) increase in average AM1.5G efficiency above baseline in a single cycle of learning. This result verifies the capacity of our approach to find optimal process windows with little experimentalist intervention, and surpass experimentalist-constrained black-box optimization.

II. Results and Discussion:

Bayesian network development

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

1. *Parameterization of process variables by embedding physics knowledge:*

This section describes how we define physics-based relationships between process variables and materials descriptors, embedding physics domain knowledge, and ensuring faster convergence of our Bayesian optimization algorithm. This corresponds to the progression from “Process Conditions” to “Materials Descriptors” in Figure 1. Device fabrication of solar cells is expensive, thus it is essential to explore the process variable space efficiently\(^{27}\). 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 parameter space e.g. Equation [2]. The parameterization connects process variables with underlying material and interface properties (e.g., doping levels, front SRV, back SRV, etc.). In the case of chemical vapor deposition (CVD), recognizing that growth temperature affects several thermally and kinetic activated processes, we embed such knowledge and enforce an exponential dependence of underlying material properties based on the Arrhenius equation, including high-order effects (Equation [2]).

The detailed schematic of the Bayesian network is shown in Figure 2. To illustrate the flow of Bayesian network, we will use optimization of acceptor (zinc) doping concentration in a GaAs solar cell as a showcase. Bayesian network can be represented as two-step Bayesian inference using the representation of conditional probability (Equation [1]).

\[
P(J|T) = P(N_A|T) \times P(J|N_A,T) \tag{1}
\]

\(P(N_A|T)\) is the conditional probability of zinc 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, we enforce an Arrhenius equation-type of parameterization to link the underlying material properties with growth temperatures. Zinc doping level can be represented in Equation [2].

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

(a, b, c) are latent process parameters that are inferred from the Bayesian framework. Aside from zinc doping concentration, bulk minority carrier lifetime (\(\tau\)), front and back interface
recombination velocities (SRV’s), and series and shunt resistances are also parameterized in the same fashion.

The Arrhenius equation form for the doping concentration agrees well with literature reported trend\textsuperscript{30-32}. There is insufficient literature and domain expertise to directly relate bulk and interface properties with the growth temperature. However, previous study has shown that \( \tau \) and SRV’s are correlated with doping concentration\textsuperscript{33,34}. We introduce the quadratic term \((a \left(-\frac{1}{T}\right)^2)\) to account for additional complexity introduced by bulk and interface properties. Note that Equation [2] can be an implicit hypothesis test. A large \( b/a \) ratio suggests that higher-order terms are suppressed, and that the Arrhenius relationship governs the temperature dependence of the particular bulk, interface, or resistance property. On the other hand, a smaller \( b/a \) ratio suggests stronger participation of higher-order terms, indicating a deviation from pure Arrhenius behavior at some temperature.

Additional domain knowledge is embedded in the prior by setting hard constraints for the material properties. Range of the 5 inferred material properties are shown in Table S2.

2. *Inference of material and device properties from device measurements*

This section describes the progression from “Materials Descriptors” to “Target Variable: Performance” in Figure 1. Inference of underlying material properties from voltage-dependent current (\(JV\)) measurements of solar cells under different illumination intensities has been previously demonstrated in Ref.\textsuperscript{13}. This innovation enables us to trace the root cause(s) of device underperformance to specific material or interface properties. We further extend the connection between process variables and materials properties to
device measurements by adding this 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., current-voltage characteristics, quantum efficiency, and conversion efficiency).

Following the above example, $P(J|N_A,T)$ is the conditional probability of $JV$ observations at different illumination intensity given the underlying material parameters (zinc doping concentration). The material property-$JV$ relationship is extensively investigated and can be solved numerically using a well-developed device model from literature$^{22,35-37}$. A well calibrated PC1D model$^{19}$ 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 conditional probability estimation) and makes it difficult to integrate new features into the model.

3. Replacement of numerical solver with neural-network surrogate model

To circumvent the computational complexity of the forward device model (numerical solver), we apply a transfer learning scheme$^{38}$ to replace the numerical simulator with a surrogate deep neural network. The schematic of the surrogate model is shown in Figure 6. We first generate 50,000 illumination current-voltage ($JV_i$) curves for our GaAs architecture with different material properties in each layer ($\tau$, $SRVs$, doping and resistances) using the forward device model (PC1d solver)$^{39}$. We train an autoencoder (AE)$^{40}$ neural network to predict the underlying material properties and reconstruct $JV_i$ curves from the latent space. The AE consists of 3 convolution and 2 dense layers in the
encoder and 3 convolution transpose and 2 dense layers in the decoder. Detailed structure of the AE is shown in the method section. \( P(J|N_A, T) \) thus can be computed using the surrogate neural network model. The computational speed for our GaAs example herein is improved by >100x by replacing the PDE numerical solver with this carefully trained neural network\(^{41}\).

Once the device model is completed, we connect these previous two Bayesian inference steps into a hierarchical structure using Equation [1]. A posterior probability to every combination of latent 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 when one observes new data (\( JV_i \) measurement).

We apply 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\(^{42}\). Specifically, we implement the affine-invariant ensemble sampler of MCMC proposed by Goodman & Weare\(^{43}\). With each newly observed \( JV_i \) measurement, the posterior distribution of latent process parameters \((a, b, c)\) are updated. Hereafter, material descriptor (zinc doping concentration as a function of growth temperature \((N_A(T))\) can be obtained from Equation [2].

In an analogous way, 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.
4. Optimizing solar cells using a Bayesian Network:

After fitting the Bayesian network with current-voltage data for various processing conditions, the model links the process variables to the material properties. This surrogate model better captures the latent relations between physical variables and process variables than common black-box models. We can define the optimization procedure enabled by our Bayesian network as:

$$x^* = \arg \max (h(g_i(x)))$$

[3]

The variable $x^*$ is the set of process variables, 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 process variables. 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 parameters and final solar cell performance (JV curves). $h(y_i)$ can be solved numerically and, in our case, is estimated using a neural network. To find $x^*$ that maximizes the cell efficiency, more insights in process optimization can be obtained. As $g_i(x)$ determines the functional relation of materials descriptors and process variables and estimates its uncertainty, we can tailor our process variables to maximize the desired materials properties, such as lifetime or shunt resistance, and minimize the undesired properties in selected locations across the devices.
GaAs solar cell optimization

As baseline for testing our methodology, we fabricate a set of GaAs solar cells, sweeping a range of MOCVD growth temperatures. Subsequently, our optimization procedure uses the knowledge of the temperature dependence of materials descriptors extracted from the Bayesian network to improve the device in single cycle of learning. Material and device properties are optimized independently, and the process temperature is tailored for each layer, guided by the Bayesian network. As each layer and interface can be optimized in a single step, this reduces the time and cost of process optimization and characterization significantly, and increases the accuracy of root-cause analysis / troubleshooting.
Figure 3. Bayesian network reveals how each material descriptor (bulk and interface property) varies with processing conditions. Black lines show inferred values of materials descriptors as a function of growth temperature; red circles show experimental validation of materials 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.

We grow the GaAs films at temperatures ranging from 530°C to 680°C, with 20–50°C temperature intervals. The films are fabricated into 1 cm² solar cells, without ARC’s. Detailed growth and fabrication procedures can be found in the Experimental Procedures section. JV measurements under different illumination intensities (0.1–1.1 suns) are performed. We apply our Bayesian network to assess the effects of growth temperature on
materials descriptors, using the measured $JVi$ curves. Emitter doping concentration, base doping concentration, bulk lifetime, front surface recombination velocity (FSRV) and rear surface recombination velocity (RSRV) are selected as underlying material and device properties that are to be inferred and optimized in the Bayesian network. Figure 3 shows the inferred material properties for our GaAs cells under various growth temperatures. We can see that the $p$-type dopant (zinc) and $n$-type dopant (silicon) respond differently to temperature changes. Zinc doping level decreases as temperature increases, while silicon doping level increases. We conduct secondary ion mass spectroscopy (SIMS) and time-resolved photoluminescence (TR-PL) measurements to validate our inferred estimates, shown in the first three plots of Figure 3. The experimental details of those measurements are shown in the supplementary information. The red dots in Figure 3 are the extracted doping concentrations and bulk lifetime for the GaAs cells grown at different temperature. It is explicit that the independent experimental measurements agree quantitatively well with the estimates inferred by the Bayesian network.

It is interesting to note, that each parasitic recombination parameter (bulk lifetime, front SRV and rear SRV) 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, the Bayesian network 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 for our GaAs devices (labeled “Bayes Net” in Figure 4), achieving
significant efficiency gains. We grow the different device layers at different temperatures, seeking to minimize recombination at each layer or interface. The rear InGaP back-surface-field layer is grown at 580°C, the GaAs base and emitter layers at 620°C, and the front-window InGaP layer at 650°C, in a GaAs solar cell stack. We avoid extreme conditions (e.g., 680°C), which show deterioration of overall device performance (Figure 4) despite inferred layer-specific improvements (Figure 3).

![Graph showing efficiency vs. temperature with Bayesian network approach]

**Figure 4.** Comparison of “black-box optimization” versus our approach using a Bayesian network (Bayes Net). GaAs cell efficiency varies as a function of growth temperature, reaching an average maximum between 580 and 650°C. Our Bayesian-network-informed process (labeled “Bayes Net”) tunes the growth temperature of each layer to minimize recombination (Figure 4), increasing efficiency by 6.5% relative. The grey area represents the additional efficiency gain that cannot be achieved using conventional black-box optimization.

Figure 4 shows the spread of GaAs cells efficiency as a function of growth temperature. If efficiency is used as the performance indicator, the traditional process optimization will
stop between 580°C and 650°C, as the efficiency decreases on either side. Under this scenario, the traditional approach will give us an efficiency improvement of 1.4% relative and 0.28% absolute above our baseline efficiency (630°C). Using the Bayesian network approach to tune growth temperatures of each layer, thereby minimizing layer-specific recombination, we achieve a 6.5% relative (1.4% absolute) improvement over baseline, well exceeding the conventional approach. This demonstrates how additional insights gained via Bayesian-network-based optimization can be translated into device performance that exceeds black-box optimization. Additionally, the fact that the optimal temperature profile is found in a single cycle of learning, verifies the potential of our method to accelerate time-consuming device optimization significantly, limiting the need of synthesizing auxiliary samples and performing secondary measurements. Further efficiency gains may be achieved, by wrapping our entire framework within an optimization algorithm (e.g., Bayesian optimization), further optimizing individual growth temperatures for each layer and capturing non-linear effects.

**Figure 5.** 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)
Figure 5 shows that both $J_{SC}$ and $V_{OC}$ are responsible for the efficiency improvements in our “Bayes Net” growth temperature profile and the external quantum efficiency ($EQE$), indicating photoresponse at wavelengths less than 820 nm (corresponding to an optical penetration depth comparable to our 2-μm thick absorber) is clearly improved. Subsequently, we feed the measured $JV_i$ curve of this cell to the autoencoder model to extract the bulk and interface properties. The extracted FSRV, RSRV and $\tau$ is $1.2 \times 10^3$ cm/S, $5.4 \times 10^3$ cm/S and 39μs. This suggests the front surface and the bulk benefit the most from the “Bayes Net” optimized thermal profile, possibly because these were the highest-temperature steps, which may have partially erased the thermal history of the underlying rear-surface layer. All cells reported herein do not have anti-reflection coatings; the best cells shown in the figure are estimated to have efficiencies in the 24% to 25% range with optimal double-layer ARC’s. The efficiency value is near state-of-the-art for a single-junction GaAs with substrate grown in an academic setting. Other process parameters, e.g., epitaxial lift-off and grid optimization, are required to reach record efficiencies. Nevertheless, the recombination gains enabled by the variable-temperature process informed by our Bayesian network should translate also to these advanced cell architectures. To our knowledge, reports of variable-temperature processing to minimize layer-specific recombination in GaAs devices are not commonplace in the literature, suggesting a non-intuitive finding by our Bayesian network.

III. Conclusions:
We developed and applied a Bayesian-network approach to GaAs solar cell growth optimization. This approach enables us to exceed conventional “black-box” efficiency optimization by 6.5% relative, by tuning process variables layer-by-layer, in a single shot. Our approach is enabled by implementing physics-informed relations between process variables and materials descriptors, and embedding these into a Bayesian network. We connect these materials descriptors to device performance using a surrogate model consisting of a trained autoencoder, which is 100x faster than numerical device simulation. The number of growth (MOCVD) runs necessary to implement this layer-by-layer process-optimization scheme translates into significant cost and time reductions. We believe this approach can generalize to other solar cell materials\cite{45,46}, as well as other systems with physics-based relationships between process variables and materials descriptors, and physics-based device-performance models. The surrogate model can replace common models in closed-loop black-box optimization, such as a Gaussian Process model in Bayesian optimization, providing good functional fitting, physical insights and uncertainty prediction.

IV. Method

Neural network surrogate model to infer material descriptors: To improve the computation speed for Bayesian network, we replace the numerical PC1d model with a conditional autoencoder neural network model. The schematic of our model is shown in Figure 6. It consists of two parts: (1) the encoder that maps the $JV_i$ curves and into a latent space with 5 dimensions. The latent parameters are conditioned on the material descriptors that are used to generate $JV_i$ curves using the numerical solver and the latent in each dimension
corresponds to one of the underlying material descriptors. The geometry of the encoder neural network is shown in Figure 6. (2) the decoder that takes the latent parameters and generates the $JVi$ curves. The decoder is a mirror of the encoder network, with transposed convolutions layers replacing the convolutional layers.

Figure 6. a) Schematic of conditional autoencoder network to infer material descriptors from $JVi$ curves.

To create the training dataset, we first randomly sample 50,000 set of material descriptors ($\tau$, $FSRV$, $RSRV$, series and shunt resistances) within physically meaningful range. We use scripted-PC1d\textsuperscript{39} to numerically simulate a set of $JVi$ curves that can be experimentally measured. Although domain expertise is required in setting up the numerical PC1d model, this exercise is a one-time implementation for each material system. Hereafter, we feed those simulated $JVi$ curves and material descriptor values to train the neural network. The losses in the conditional autoencoder is defined as the mean squared differences between the reconstructed $JVi$ and input $JVi$ curves plus the mean squared differences between the latent parameters and input material descriptors. The loss 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 $JV_i$ curves into 80% and 20% for training and testing purposes. After the autoencoder is trained, we plug it back into the Bayesian network to generate the simulated $JV_i$ curves as a replacement of the numerical solver. The autoencoder is much more computationally efficient than the numerical solver. Figure S2 shows the acceleration by adapting the autoencoder in calculating a set of $JV_i$ curves in the MCMC run. The autoencoder runs on a GPU is 130 times faster than the PC1d numerical solver and 700 times faster if the solver is called externally.

**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 100mbar using TMGa, TMIn, AsH$_3$ and PH$_3$ as precursors and 32 standard liters per minute (slm) H$_2$ as carrier gas. It has a 3 µm $n$-doped GaAs base (Si dopant) and 100 nm $p^+$-doped GaAs emitter (zinc dopant). Highly doped InGaP is used as the window (zinc dopant) and back surface field (BSF) layer (silicon dopant). $p^+$-doped GaAs layer (carbon dopant) are added at the front surface to achieve ohmic contact to the metal fingers. The solar cells are metalized using E-beam evaporator and a shadow mask to fabricate a grid pattern with ~8% shading. SIMS measurement is conducted for the GaAs films that are grown at in the same batch before metallization. We additionally grow $n$-doped InGaP/GaAs/InGaP heterostructures with two different base thickness (500nm and 1000nm) to measure the bulk lifetime of the $n$-doped GaAs bulk$^{47}$. The growth conditions for the heterostructure are identical to the conditions for GaAs solar cells.
Data availability: The experimental and simulated dataset for training the autoencoder and predicting the optimal growth conditions from Bayesian network is available in the following GitHub repository: https://github.com/PV-Lab/BayesProcess. Additional data that support the findings of this study are available from the authors upon reasonable request.

Code availability: The code used for Bayesian network and surrogate autoencoder to predict material properties is also available at https://github.com/PV-Lab/BayesProcess.

Acknowledgements: We acknowledge Kaicheng Zhang and Ning Li from FAU and Qianxiao Li from NUS for helpful discussions. We acknowledge Chuanseng Tan from NTU for providing fabrication facilities for the cell development at Nanyang NanoFabrication Centre. This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and its Energy Innovation Research programme EIRP-13 (Award No. NRF2015EWT-EIRP003-004) (supporting GaAs device fabrication), by the National Research Foundation (NRF) Singapore through the Singapore Massachusetts Institute of Technology (MIT) Alliance for Research and Technology’s Low Energy Electronic Systems research program (supporting autoencoder and physics-constrained Bayesian inference algorithm development), by the U.S. Department of Energy Photovoltaic Research and Development Program under Award DE-EE0007535 (supporting Bayesian optimization algorithm development), and by a TOTAL SA research grant funded through MITei (supporting ML algorithm framing and application).
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 H.X., J.D.P., M.L., T.H., C.J.B., R.S., S.S., 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.

Declaration of Interest: The authors declare no competing interests.

References:

1. Hicks, C. R. & Turner, K. V. Fundamental concepts in the design of experiments. Vol. 40 (Oxford University Press New York, 1999).
2. Frazier, P. I. A tutorial on Bayesian optimization. arXiv preprint arXiv:.02811 (2018).
3. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: A Bayesian optimizer for chemistry. ACS central science 4, 1134-1145 (2018).
4. Kennedy, J. Particle swarm optimization. Encyclopedia of machine learning, 760-766 (2010).
5. Yuan, R. et al. Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning. Advanced Materials 30, 1702884 (2018).
6. Xue, D. et al. Accelerated search for BaTiO3-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. Proceedings of the National Academy of Sciences 113, 13301-13306 (2016).
7. Yamawaki, M., Ohnishi, M., Ju, S. & Shiomia, J. Multifunctional structural design of graphene thermoelectrics by Bayesian optimization. Science Advances 4, 4192 (2018).
8. Kumar, J. N. et al. Machine learning enables polymer cloud-point engineering via inverse design. arXiv preprint arXiv:1812.11212 (2018).
9. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS central science 4, 268-276 (2018).
10. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. Science 361, 360-365 (2018).
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. *Advanced Functional Materials* **25**, 6495-6502 (2015).

12 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).

13 Brandt, R. E. *et al.* Rapid photovoltaic device characterization through Bayesian parameter estimation. *Joule* **1**, 843-856 (2017).

14 Oviedo, F. *et al.* Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* **5**, 60 (2019).

15 Sun, S. *et al.* Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis. *Joule* (2019).

16 Ren, F. *et al.* Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science advances* **4**, eaaq1566 (2018).

17 Ong, S. P. Accelerating materials science with high-throughput computations and machine learning. *Computational Materials Science* **161**, 143-150 (2019).

18 Bai, Y. *et al.* Accelerated discovery of organic polymer photocatalysts for hydrogen evolution from water through the integration of experiment and theory. *Journal of the American Chemical Society* (2019).

19 Baloch, A. A. *et al.* Full space device optimization for solar cells. *Scientific reports* **7**, 11984 (2017).

20 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).

21 Huang, Q. Physics-driven Bayesian hierarchical modeling of the nanowire growth process at each scale. *IIE transactions* **43**, 1-11 (2010).

22 Ren, Z. *et al.* Numerical analysis of radiative recombination and reabsorption in GaAs/Si tandem. *IEEE Journal of Photovoltaics* **5**, 1079-1086 (2015).

23 Kayes, B. M. *et al.* in *Photovoltaic Specialists Conference (PVSC), 2011 37th IEEE*. 000004-000008 (IEEE).

24 Lenk, P. J., DeSarbo, W. S., Green, P. E. & Young, M. R. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science* **15**, 173-191 (1996).

25 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).

26 Pearl, J. *Causality: models, reasoning and inference*. Vol. 29 (Springer, 2000).

27 Correa-Baena, J.-P. *et al.* Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* (2018).

28 Coltrin, M. E. & Mitchell, C. C. Mass transport and kinetic limitations in MOCVD selective-area growth. *Journal of crystal growth* **254**, 35-45 (2003).

29 Laidler, K. J. The development of the Arrhenius equation. *Journal of Chemical Education* **61**, 494 (1984).

30 Kuech, T., Veuhoff, E. & Meyerson, B. Silicon doping of GaAs and AlxGai–xAs using disilane in metalorganic chemical vapor deposition. *Journal of crystal Growth* **68**, 48-53 (1984).
Hudait, M. K., Modak, P., Hardikar, S. & Krupanidhi, S. Zn incorporation and band gap shrinkage in p-type GaAs. *Journal of applied physics* **82**, 4931-4937 (1997).

Sun, S., Armour, E., Zheng, K. & Schaus, C. Zinc and tellurium doping in GaAs and AlxGa1−xAs grown by MOCVD. *Journal of crystal growth* **113**, 103-112 (1991).

Aspnes, D. Recombination at semiconductor surfaces and interfaces. *Surface Science* **132**, 406-421 (1983).

Lush, G. B. *et al.* A study of minority carrier lifetime versus doping concentration in n-type GaAs grown by metalorganic chemical vapor deposition. *Journal of applied physics* **72**, 1436-1442 (1992).

Clugston, D. A. & Basore, P. A. in *Photovoltaic Specialists Conference, 1997., Conference Record of the Twenty-Sixth IEEE.* 207-210 (IEEE).

Niemegeers, A., Gillis, S. & Burgelman, M. in *Proceedings of the 2nd World Conference on Photovoltaic Energy Conversion, JRC, European Commission, juli.* 672-675.

Guide, S. D. U. & Version, G. Synopsis. *Inc., Jun* (2012).

Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345-1359 (2010).

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).

Hinton, G. E. & Zemel, R. S. in *Advances in neural information processing systems.* 3-10.

Ren, Z. *et al.* Physics-guided characterization and optimization of solar cells using surrogate machine learning model. *IEEE PVSC 46* (2019).

Kurchin, R., Romano, G. & Buonassisi, T. Bayesim: a tool for adaptive grid model fitting with Bayesian inference. *Computer Physics Communications* (2019).

Goodman, J. & Weare, J. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science* **5**, 65-80 (2010).

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

Mangan, N. M. *et al.* in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC).* 2373-2378 (IEEE).

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

Ahrenkiel, R., Keyes, B. & Dunlavy, D. Intensity-dependent minority-carrier lifetime in III-V semiconductors due to saturation of recombination centers. *Journal of applied physics* **70**, 225-231 (1991).
Supplementary Information

Embedding Physics Domain Knowledge into a Bayesian Network Enables Layer-by-Layer Process Innovation for Photovoltaics

Zekun Ren¹*, Felipe Oviedo²*, Muang Thway³, Siyu I.P. Tian¹, Yue Wang¹, Hansong Xue³, Jose Dario Perea⁴, Mariya Layurova², Thomas Heumueller⁴,⁵, Erik Birgersson⁶, Armin Aberle³, Christoph J. Brabec⁴,⁵, Rolf Stangl³, Shijing Sun², Fen Lin³, Ian Marius Peters², Tonio Buonassisi¹,²*

¹Singapore-MIT Alliance for Research and Technology SMART, 138602 Singapore
²Massachusetts Institute of Technology, Cambridge MA 02139, U.S.A.
³Solar Energy Research Institute of Singapore (SERIS), 117574 Singapore
⁴Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander University Erlangen-Nürnberg, 91058 Erlangen, Germany
⁵Helmholtz Institute HI-ErN, Forschungszentrum Jülich, Immerwahrstrasse 2, D-91058 Erlangen, Germany
⁶National University of Singapore, 119077, Singapore
*These authors contributed equally.
I. Schematic of encoder used in the autoencoder surrogate model:

Figure S1. a) geometry of the encoder, the decoder is a mirror of the encoder with convolutional transpose layers.

The geometry of the encoder neural network is shown in Figure S1. (2) the decoder that takes the latent parameters and generates the \( JVi \) curves. The decoder is a mirror of the encoder network, with transposed convolutions layers replacing the convolutional layers.
II. Computation time comparison between the neural network and numerical solver:

The simulated $JV_i$ curves consists of 30 $JV$ curves (5 different process conditions with 6 illumination intensities). The Bayesian network has MCMC chain with 20,000 samples indicating that those 30 $JV$ curves are computed 20,000 times to update the posterior. Using the numerical solver, in this case, is extremely time consuming. Herein, we compare the runtime of simulating 30 $JV$ curves using the numerical solver and the autoencoder surrogate model as shown in Figure S2. PC1d is chosen in our work as it is one of the fastest and well-studied numerical solver\(^1\). Although other Python-based numerical solver has been developed\(^2\), its computation speed is not faster. In Figure S3, runtime of the full .exe run accounts for calling the scripted PC1d\(^3\) program in Python. To eliminate the external file reading time from the numerical solver to Python, we also generate 30 $JV$ curves using the batch mode within the PC1d program. Figure S3 shows that generating 30 $JV$ curves using the autoencoder with a GPU is more than 130 times faster than generating it within the PC1d programs and 700 times faster if calling PC1d externally in python. The GPU used in this work is a Nvidia GTX 1070 and the CPU used is an Intel i7 3770.
Figure S2. Computation time of 30 JV curves that are updated every MCMC run using autoencoder (AE) and Numerical solver (PC1d). Computation time of the full .exe run accounts for calling the scripted PC1d program in python. Computation time of PC1d only accounts for numerically generating JVs within the program.
III. GaAs solar cell architecture:

The following schematic (not to scale) shows the GaAs solar cell fabricated according to the Experimental Procedures section.

Figure S3. Schematic (not to scale) of the GaAs solar cell architecture used in this study.
IV. SIMS and TR-PL results of GaAs solar cell:

This section describes the experimental details of SIMS and TR-PL measurements of GaAs solar cells grown at different temperature. The extracted doping concentration and bulk lifetimes serves as independent validation to the Bayesian network inference results shown in Figure 3 in the main body of the manuscript.

Figure S4 shows the zinc and silicon distributions in the emitter and bulk, respectively, as a function of growth temperature.
Figure S4. Si and Zn dopant distribution in the GaAs solar cell under different growth temperatures.

Table S1 shows the extrapolated bulk lifetime ($\tau$) from the $n$-doped InGaP/GaAs/InGaP heterostructure with two thicknesses (500nm and 1000nm) from the TRPL measurement. The doping level and layer thickness for InGaP layers are identical to the back-surface field layer used in the GaAs solar cell structure. To extract the bulk lifetime of the n-doped GaAs, we numerically solve this differential equation for the excess carrier concentration as a function of distance to the front surface and time ($\Delta n(x,t)$).

$$\frac{\partial \Delta n}{\partial t} = G + D \frac{\partial^2 \Delta n}{\partial x^2} - R \quad [S1]$$

$G$ is the generation rate, $D$ is the diffusion coefficient, and $R$ is the recombination rate in the above equations. We set the initial condition: $\Delta n(x, 0) = 0$ with boundary conditions where $S$ is the surface recombination velocities and $d$ is the sample thickness:
\[ \frac{\partial \Delta n}{\partial x} (x = 0, t) = \frac{S}{D} \cdot \Delta n(x = 0, t); \quad \frac{\partial \Delta n}{\partial x} (x = d, t) = -\frac{S}{D} \cdot \Delta n(x = d, t) \quad [S2] \]

Hereafter, effective bulk lifetime is extracted by fitting the following equation.

\[ \frac{1}{\tau_{\text{eff}}} = \frac{1}{\tau_{\text{bulk}}} + \frac{2S}{d} \quad [S3] \]

Table S1. Effective bulk lifetime of n-doped GaAs grown at different temperatures.

| Temperature [°C] | \(\tau\) [ns] |
|------------------|---------------|
| 530              | 1.8           |
| 580              | 12.5          |
| 630              | 29.3          |
| 650              | 11.1          |
| 680              | 0.6           |
V. Constraints for the Bayesian network prior:

Table S2 describes the constraints for Bayesian network prior. Instead of directly setting constraints on \((a,b,c)\), we set hard constraints for the material properties that are computed from them in Equation [2] from the manuscript. The doping concentration, \(\tau\) and \(SRV\) should have physically meaningful values. If they are not within those ranges shown in Table S2, the posterior probability will be set to negative infinity in the Bayesian network.

_**Table S1. effective bulk lifetime of n-doped GaAs grown at different temperatures.***_

| Parameter                  | Range          |
|----------------------------|----------------|
| Zn doping concentration [cm\(^{-3}\)] | \(1\times10^{16}\)~\(1\times10^{20}\) |
| Si doping concentration [cm\(^{-3}\)] | \(1\times10^{16}\)~\(1\times10^{20}\) |
| Bulk lifetime [s]          | \(1\times10^{-9}\)~\(1\times10^{-5}\) |
| Front SRV [cm/s]           | \(1\times10^{2}\)~\(1\times10^{8}\) |
| Back SRV [cm/s]            | \(1\times10^{2}\)~\(1\times10^{8}\) |
References:

1 Clugston, D. A. & Basore, P. A. in *Photovoltaic Specialists Conference*. 207-210 (IEEE, 1997).
2 Alonso-Álvarez, D. et al. Solcore: a multi-scale, Python-based library for modelling solar cells and semiconductor materials. *Journal of Computational Electronics* **17**, 1099-1123 (2018).
3 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).