Bridging TCAD and AI: Its Application to Semiconductor Design

Changwook Jeong, Sanghoon Myung, In Huh, Byungseon Choi, Jinwoo Kim, Hyunjae Jang, Hojoon Lee, Daeyoung Park, Kyuhun Lee, Wonik Jang, Jisu Ryu, Moon-Hyun Cha, Jae Myung Choe, Munbo Shim, and Dae Sin Kim

(Invited Paper)

***Abstract*— There is a growing consensus that the physics-based model needs to be coupled with machine learning (ML) model relying on data or vice versa in order to fully exploit their combined strengths to address scientific or engineering problems that cannot be solved separately. We propose several methodologies of bridging technology computer-aided design (TCAD) simulation and artificial intelligence (AI) with its application to the tasks for which traditional TCAD faces challenges in terms of simulation runtime, coverage, and so on. AI-emulator that learns fine-grained information from rigorous TCAD enables simulation of process technologies and device in real-time as well as large-scale simulation such as full-pattern analysis of stress without high demand on computational resource. To accelerate atomistic molecular dynamics (MD) simulation, we have done a comparison study of descriptor-based and graph-based neural net potential, and also show their capability with large-scale and long-time simulation of silicon oxidation. Finally, we discuss the use of hybrid modeling of AI- and physics-based model for the case where physical equations are either fully or partially unknown.**

***Index Terms*— Artificial intelligence (AI), atomistic simulation, design optimization, device simulation, full-chip level modeling, machine learning, process simulation, semiconductor, technology computer-aided design (TCAD).**

I. Introduction

COMBINATION of atomistic, process, device, and circuit simulations, referred to as technology computer-aided design (TCAD), have been widely used to develop and optimize semiconductor process technologies and devices by using physical (or compact) models. Recently, with massive datasets and high computing power, data-driven models, called machine learning (ML) or artificial intelligence (AI), expand their application in many areas of semiconductor industry such as automating chip design [1], optimizing processes [2], [3], improving production yield [4]–[6], and so on.

TABLE I
COMPARISON OF PROS (◎) AND CONS (○) BETWEEN PHYSICS-BASED AND DATA-DRIVEN MODEL

Symbol		Physics-based Model	Data-driven Model
Data Cost		◎	○
Accuracy	Extrapolation	◎	○
	Interpolation	○	◎
Inverse Optimization		○	◎
Interpretability		◎	○
Prediction Time		○	◎

As shown in Table I, two approaches are complementary to each other to solve scientific or engineering problems in terms of data cost, prediction accuracy, interpretability, inverse optimization, and prediction time. As such, merging ML and physics-based model has received rapidly growing interest to completely exploit their combined potential in many areas including accelerating time-consuming scientific simulator with ML emulator [7], [8], improving generalization capability for previously unseen case [9], [10], solving partial differential equation with deep learning [11], [12], discovering underlying physics law [13]–[15], custom programming framework for a seamless integration of machine learning models into existing physical simulation environment [16], [17], and many others.

Advanced AI are also having a positive impact on TCAD simulation from many aspects such as use of ML for compact modeling [18]–[21], estimating reliability of power device [22] and defect formation energy [23], and speeding up of process and device TCAD [24]. Here we concentrate on the various applications of bridging AI and TCAD in semiconductor design to show that combining two approaches allows to meet the ever-increasing demand for large-scale and

Manuscript received March 19, 2021; revised May 18, 2021; accepted June 24, 2021. Date of publication July 14, 2021; date of current version October 22, 2021. The review of this article was arranged by Editor S.-M. Hong. *(Corresponding author: Changwook Jeong.)*

The authors are with the Computational Science and Engineering Team, the Data and Information Technology Center, Samsung Electronics Company Ltd., Hwaseong, Gyeonggi-do 18448, South Korea (e-mail: chris.jeong@samsung.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3093844.

Digital Object Identifier 10.1109/TED.2021.3093844

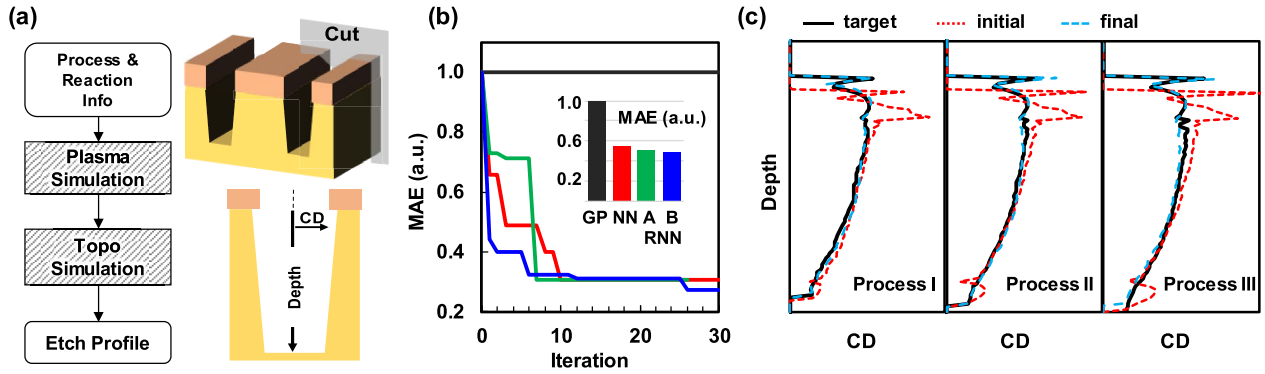

Fig. 1. (a) Etch profile simulation flow and a schematic of STI structures. (b) Convergence curve of BO with four different surrogate models: GP, baseline neural net (NN), and recurrent NN (RNN-A and RNN-B). Compared to RNN-A model, RNN-B is fine-tuned to simultaneously consider errors from three different process conditions. Inset: performance of the surrogate models after training with 100 initial samples. (c) For RNN-B model, experimentally measured profiles (target) are compared with simulation results before (initial) and after calibration (final) for three different process conditions to pattern sub-20-nm trench isolation features. After calibration, the mean absolute percentage error of simulated profiles is 5.1%. For the acquisition function which decides the next point for simulation by trading-off exploration and exploitation, conventional expected improvement (EI) is used [43]: EI is defined as $\alpha_{EI}(x) = (u\Phi(u) + \phi(u))\sigma_n(x)$ where $u = (\mu_n(x) - y_{\min})/\sigma_n(x)$ and $\Phi(\cdot), \phi(\cdot)$ are the standard Gaussian distribution and density functions respectively and $y_{\min}$ is the best current solution. Ten samples are suggested at each iteration.

high-throughput simulations, which has been traditionally addressed by the increase in computing power and advanced numerical methods. Section II presents use cases of AI for TCAD simulations. In Section III, we discuss how AI could be of help when analytic description of physical model is not available. Finally, our conclusions are summarized in Section IV.

II. Machine Learning for TCAD

This section describes the uses of ML/AI for TCAD simulation: ML-based calibration of complex physical model, AI-based emulator for process and device TCAD simulation, and coarse-grained AI model for full-chip level TCAD as well as for atomistic simulation.

A. Calibration of Parameterized Physical Models

Industrial TCAD often use simplified physical approximations to account for complex physics as well as to interpret experimental results. Parameter values of the simplified physical model are estimated from either rigorous simulation results or experimental observations through a procedure referred to as parameter calibration. To correctly identify parameter values makes the simplified physical model robust to the modeling of important physical processes, e.g., generalization of model to novel instances that were not seen before. Recent advances of ML accelerate to automate the calibration procedures. The main components of parameter auto-calibration are to define search space, to choose surrogate model, and to optimize parameter values.

Search space describes the set of possible parameters to consider and depends on the specific application. The design of these search spaces currently relies on TCAD expert-designed set of models as starting points. A surrogate model is to approximate its predictions as accurately as possible to "expensive-to-evaluate" physics-based simulator, and as a rule, it is much less demanding than the latter in terms of runtime and computational resources. To this end, AI-based surrogate model is a good candidate, but the model should be carefully chosen and constructed in order to consider types of simulation as well as data sampling and generation strategy. Optimization method determines how to explore the search space in order to find a good set of parameters. Depending on how they determine which parameters to evaluate, various adaptive methods have also been introduced—e.g., evolutionary search [25], gradient-based optimization [26], and Bayesian optimization (BO) [27], [28].

Full physics-based simulation for an etching process, one of the most important processes in semiconductor manufacturing, is a good illustrative example for which the use of machine learning methods is attractive for automatic calibration. As shown in Fig. 1(a), the simulation is quite challenging because advanced physical models require more parameters to be calibrated in order to correctly describe various reactions among gases, by-products, and plasma species as well as surface reactions at silicon surface. For illustrative purposes, automatic calibration of etching model with parameters being $\sim$100 is done with BO to match observed etch profiles for shallow trench isolation (STI) processes. The search space of calibration parameters is reduced to $\sim$100 parameters based on an expert's knowledge. To deal with such high dimensions, a careful choice of surrogate model is necessary.

In conventional BO, Gaussian process (GP) is widely adopted as a surrogate model to make a prediction with uncertainty estimates which are used to sequentially decide where to sample next in order to find better solutions. We instead use ensemble of neural network (NN) to quantify prediction uncertainty [29] as well as specifically designed to represent etching process, because standard GP model performs poorly on high dimensional problems while many variations of GP developed to address the issues [30].

For different surrogate models, Fig. 1(b) shows comparisons of performance on the parameter calibration for the etch simulator. Inset of Fig. 1(b) compares the mean absolute error (MAE) between the measured and simulated etch profiles

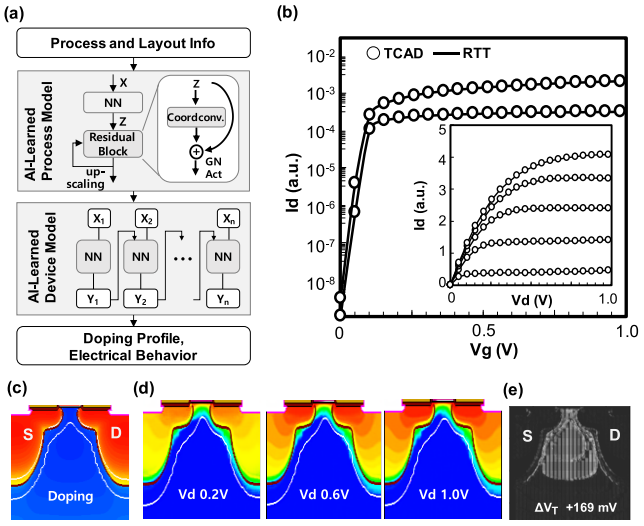

Fig. 2. (a) Schematic of AI-based emulator (RTT) to mimic the TCAD process and device simulation. A thousand of training examples per simulation is used for training, and up-sampling methods are conducted to double the image resolution before passing each residual block. (b) Comparison of *I*–*V* curves between TCAD simulation and RTT. (c) and (d) RTT-generated junction and current density contours. (e) For a case where threshold voltage (VT) is increased by ∼170 mV with respect to reference device, RTT with xAI techniques can identify important doping regions that mostly explains the increase of VT.

with different surrogate models after training with 100 initial samples: GP, baseline NN, and recurrent NN (RNN). It can be seen that RNN models (denoted as A and B) are more suitable to learn sequential relationship such as etching profile shape, which results in rapid convergence of MAE between observed and simulated etch profiles in a small number of iterations. In comparison, BO with standard GP does not converge within 30 iterations. As BO with RNN-B model progressed, Fig. 1(c) shows how simulated etch profiles reach observed profiles for three different etching conditions, and final profiles well match the observed profiles with errors being about 1%.

B. Real-Time TCAD (RTT)

Calibrated TCAD models were used to develop and optimize semiconductor process for the past 25 years. Although many numerical methods [31], [32] have been developed to make TCAD simulation fast, typical process and device simulation takes about from a few CPU-hours to a 100 CPU-hours to run and the running time significantly increases with complexity of models. Furthermore, optimization of semiconductor often requires to run about 1000 simulation experiments to obtain the optimal solution depending on the number of possible input variables. Simulation-based optimization which integrates optimization techniques into simulation relaxes this difficulty. However, to obtain the optimal solution with minimum computation and time as well as to interpret experiments on the fly, there has been recently attempts to replace physics-based simulations with approximate detailed simulations based on machine learning model, i.e., AI emulator [7], [24].

Fig. 2(a) shows a schematic of AI-based emulator (RTT) to mimic the TCAD process and device simulation. First, to generate doping profiles as shown in Fig. 2(c), AI-learned process emulator use a tailored convolutional neural network (CNN) with techniques such as residual blocks [33], coordinate convolution [34], and group normalizations [35] which are designed to prevent gradient vanishing, naturally embed position information to RTT model, and ensure the efficient training of the model, respectively. Second, for a device emulator we adopted a RNN structure which is specially designed to learn sequential data, because current versus voltage (I–V) curves can be regarded as sequential data and current values of neighboring voltage points are highly correlated. Fig. 2(b) shows that the emulator reproduces well the relationship between drain current and gate voltage and drain voltage, and cross section of device with current density contours at different drain bias is shown in Fig. 2(d). To improve a prediction accuracy for I–V, trans-conductance characteristics is considered simultaneously while training the AI model.

Once RTT models are accurately trained, they can be directly used to find the optimal process condition to achieve target performances. By using the RTT models, optimization of the semiconductor process can be done in the order of minutes within 1% error [24]. This level of accuracy is achieved with a 1000 training examples per simulation. On top of reducing runtime of simulation-based optimization by several order of magnitude, explainable AI (xAI) techniques [36] help humans to understand the results of AI-emulator. Example shown in Fig. 2(e) demonstrates that xAI can identify important doping regions that are attributed to threshold voltage increase of ∼170 mV with respect to reference device. This is helpful in assisting human comprehension.

C. Full-Chip Level Simulation

Along with predicting device-level behavior, TCAD is also expected to account for layout-dependent issues such as well-proximity effects and stress effects during the earliest stages of design, because such effects introduce variability to circuit design and significantly impact device performance as well as mechanical issues. For example, controlling stress in the film stack and pattern-dependent stress variations has become a major challenge. Simulations of such mechanical issue, however, require large range of simulation dimension from nanometer-, micrometer-, and to millimeter-size.

To model the huge dimensional range with good accuracy, large number of meshes are required. These are main reasons that make the simulation of full-chip analysis computationally very expensive, sometimes prohibitively so on. Thus, the standard commercial finite element analysis (FEA) tools can only simulate very small cross sections at a time. The physical insights from the small-domain simulation is used to generate guidelines either to alleviate stress-related issues or to improve performance, and then the guidelines is applied to standard cell layouts [37]–[39]. The hand-crafted guidelines (or a simplified analytical model), however, does not cover all possible cases of pattern dependencies. To mitigate this issue and to enable a full-chip stress analysis, several methods are proposed for computational efficiency by employing structural [40] and model [41] simplification. Such methods often compromise the accuracy and run time of simulation.

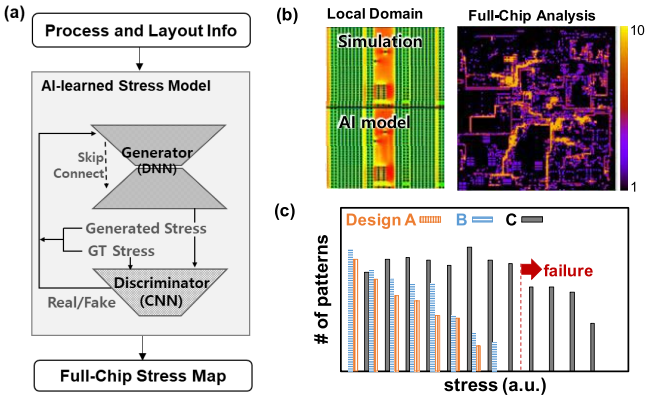

Fig. 3. (a) Schematic of AI-based emulator for full-chip stress simulation. Simulation data for local area of layout patch is used to train AI model, then AI-learned model predicts stress fields over an entire region of full-chip. (b) Left: examples of the AI-emulators output compared to the simulations output taken from the test dataset. R-square value of $>$0.95 is achieved between TCAD and AI model. Right: examples of full-chip stress analysis with AI-learned stress model highlighting stress hot-spot. (c) Examples that AI-learned model systematically characterizes the space of all geometric configurations appearing in a design layout A, B, and C, and assess the risk of stress-induced failure.

Fig. 3 demonstrates that AI-learned model could provide a way to assess a fast and accurate full-chip mechanical issues and takes less than a day for a full-chip analysis when using four GPUs. The proposed model is executed for every $\sim$10 μm $\times$ 10 μm square covering the entire chip at $\sim$10-nm resolution. However, the actual physical coverage of the input data corresponding to each of these output regions is much larger, since it must take into account the possible long-range stress fields. To achieve this level of spatial context, information from $\sim$10$\times$ larger area than output layout patch is required for predictions being made on the center of layout patch. Processing such large area at full resolution requires a significant amount of memory, so spatial downscaling techniques can be used to reduce the spatial dimension of such a large input patch while keeping the relevant information.

Generative model like a conditional generative adversarial network (GAN) [42] is used to account for process and material information. Architecture of generator and discriminator are fine-tuned with deeper CNN and coordinate convolution based on U-Net [43], and the model is trained with $\sim$100 pair image of TCAD versus layout. Image data augmentation techniques are utilized to reduce the amount of training samples and to make network architectures dealing with rotation and translation-invariances for improving performance and robustness of ML models. Fig. 3(b) and (c) shows examples of full-chip stress analysis for all possible layout patterns in three different designs. By applying AI-learned model, we can analyze and quantify variations of layout design, which are one of the sources of the yield-limiting layout configurations. Therefore, the risk of stress-induced failures can be quantitatively compared between designs, for example, design A versus design B versus design C.

D. Neural Net Potential

Atomistic simulation, a method to calculate physical properties directly from basic principle of quantum mechanics, is a preferable method for understanding and predicting material properties without relying on analytic form of physical model. For example, molecular dynamics (MD) simulations are widely used to model the motion of atom and molecules and then to predict properties of materials and identify the reaction mechanisms. The *ab initio* MD (AIMD) can apply to a various task, but it has limited time and length scale of simulation due to expensive simulation cost. Classical MD, in which the interaction between atoms is described by an empirical interatomic potential, has been used to mitigate such computational cost issue, but has its own limitations: the accuracy of the underlying potential determines the reliability of the simulation and the construction of such empirical potential is time-consuming and complex task. Bridging computational efficiency and accuracy with machine learning is also an active research topic for atomistic simulation. Recently, ML-based interatomic potentials are gaining attention as they can reproduce potential energy surfaces of *ab initio* calculations, with a much lower computational cost [44].

The atomic-scale understanding of material behavior at semiconductor process becomes more important as device have scaled down to sub-nano scale: one of such examples is the thermal oxidation process of Si. Due to the limitation of the time and length scale of density-functional theory (DFT) calculation, the atomic scale simulation of the oxide growth process cannot be modeled. Thus, the detailed mechanism of defect formation during the oxidation process is still not clearly understood.

For Si oxidation process, neural network potential (NNP) is generated based on symmetry function based methods [45]: 70 symmetry functions (16 radial, 54 angular) for both atom types (Si, O) are used and NN consists of two hidden layers with 30 nodes to minimize computational cost. In Fig. 4(a), oxidation for several atomic layers of Si is observed with NNP-MD simulation at 800 K. In this simulation, the diffusion of O_2 molecules in the SiO_2 layers is bypassed by depositing $\sim$200 O_2 molecules directly at the interface region. The simulation is performed with more than $>$10 k atoms ($\sim$8 nm $\times$ 8 nm $\times$ 6 nm cell) during $>$10 ns within a few days without relying on large computing resources. To confirm that NNP faithfully represents the oxidation process at the interface, the bond length of the Si-O bond in the newly generated SiO_2 is compared with one in the initial pre-deposited SiO_2 which is generated with DFT. As shown in Fig. 4(b), the NNP shows good agreement with DFT results. Fig. 4(c) shows the number of suboxide ions (Si^{1+}, Si^{2+}, Si^{3+}) remains nearly constant and the number of Si^{4+}(Si^0) is increased (decreased), indicating that new SiO_2 layer is formed during simulation. The behavior of suboxide ions is consistent with previous research [46].

While the hand-crafted symmetry function based approach [47] has been widely adopted to approximate DFT, graph neural network (GNN) has recently shown its superiority with lots of variants [48] by comprehensively and automatically aggregating neighbor atomic information and embedding atom representation. When predicting total energy of SiO_2 systems with atom impurity (Cl or H), Fig. 4(d) shows the results of GNN models performance compared to

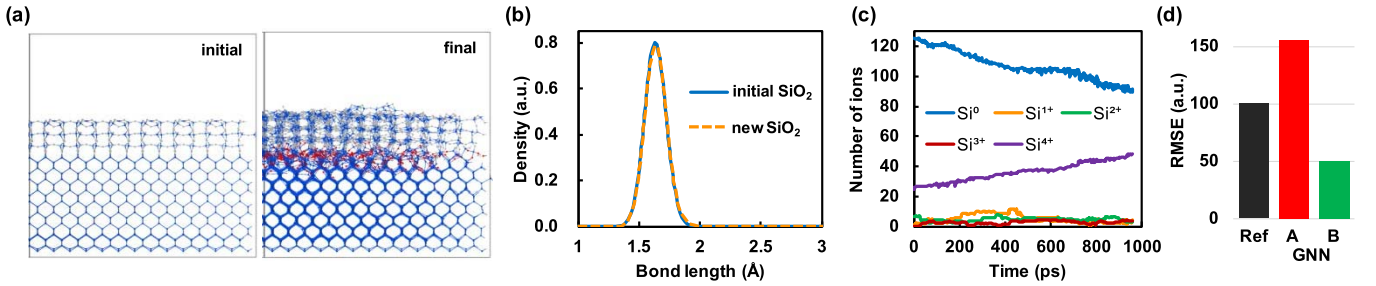

Fig. 4. Data and figures for (a)–(c) are from [45]. (a) NNP-based MD simulation of Si oxidation process at the interface. Initial and final configuration of large-scale oxidation simulation with >10 000 atoms during a few nanoseconds. Si atom is shown in blue, O atom in gray, and newly generated O atoms in red. (b) Comparison of Si-O bond length between initial oxide with DFT and generated new oxide with NNP. (c) Number of Si ions during the oxidation process which is in good agreement with DFT results. (d) Comparison of NNP model performance in term of the root mean squared error (RMSE) on energy: reference is descriptor-based NNP. GNN-B model simultaneously considers distance and angles between atoms, while GNN-A does not consider angular information from atom to another by assuming that the angular information can be considered in the series of node-edge interactions.

conventional NNP which uses radial and angular symmetry function based descriptor. It is shown that performance of GNN-based NNP depends on capability of GNN model that reflects the position of atoms into distance and angles between neighbors respectively [49]. Compared to conventional NNP, proposed GNN-based NNP reduces RMSE by ∼49% on average for two impurity dataset.

III. Discussion

TCAD is based on the physical and/or mathematical model which accurately describes the processes governing the observed data. Research communities in many disciplines (e.g., material science, electrical engineering, etc.) have devoted to perfecting them, which form the backbone of process and device modeling for silicon-based devices. Often in some cases for which experimental data is available, however, the formal analytic descriptions of governing equations remain unattainable. Bridging ML and TCAD could help to resolve such issues, which will be discussed in this section.

A. Data-Driven Construction of Key Physics

Recently, data-driven discovery of unknown governing equations has gained attention [13], [14]. To illustrate finding of underlying physical function for TCAD simulation, a tunnel field-effect transistor (TFET) is used as an example. The device is aimed at low-power applications that employ band-to-band tunneling (BTBT) to obtain a sub-thermionic subthreshold swing, in which the BTBT rate can be described by the energy-space integral of the single dominant imaginary band [50]. The BTBT of direct band materials with a single isolated imaginary band can be easily computed by using the Kane model. For materials which have indirect band gap such as Si, however, the single dominant imaginary band term cannot be easily defined due to complex interaction between different imaginary bands. So, the accurate BTBT of Si can be calculated by using the full-band quantum transport such as non-equilibrium Green's function (NEGF) with large computational cost. To simulate real-world devices in a short time, generalized Kane model is commonly used for computational advantages.

For InAs-based TFET shown in Fig. 5(a), virtual experiment is done with the NEGF method. Two types of AI model are established to predict I–V curve of the TFET for different gate lengths: a baseline AI model directly learns and predicts I–V curve, whereas Wentzel-Kramers-Brillouin (WKB)-Net learns to match predicted I–V against virtual experiment by reconstructing "effective" imaginary band. Fig. 5(b) shows the flowchart describing how the WKB-Net computes tunneling current (I) based on the Landauer formular for current

$$I = \frac{2q}{h}\int T(E)\cdot\left(f^1 - f^2\right)dE \tag{1}$$

where $f^{1,2}$ is the Fermi-Dirac distribution for source and drain and the transmission T is approximated to T_{WKB}, WKB approximation to the exact transmission coefficient from tunneling point a to b

$$T(E) \approx T_{\mathrm{WKB}}(E) = \exp\left(-2\int_a^b k(x,E)dx\right) \tag{2}$$

where $k(E)$ represent the imaginary tunneling path at a given energy E. It is noteworthy that WKB-Net outperforms the baseline AI model in both the extrapolability and interpretability, thanks to its physics-inspired architecture. Fig. 5(c) shows that WKB-Net predicts well even when the point of estimation lies outside of training data with respect to gate length and gate–drain voltage, whereas the baseline model diverges, giving wildly different predictions outside of the training domain (Though not shown, the baseline model makes good prediction in the training domain, i.e., for the interpolation.) Furthermore, it can be seen in Fig. 5(d) that WKB-Net can correctly infer "effective" energy-wavevector relationship (imaginary band), a familiar concept to most semiconductor physicists. One can interpret the learned imaginary band as its physics-based counterpart. This demonstrates a possibility of replacing unknown or complex imaginary band with "effective" imaginary band given by AI model.

The main limitation of WKB-Net and similar approaches is that their success highly depends on the pre-defined and domain-specific inductive bias (e.g., generalized Kane model for WKB-Net). To overcome this issue, exploiting more general physics, e.g., continuity [51] or symmetry [10], is recently

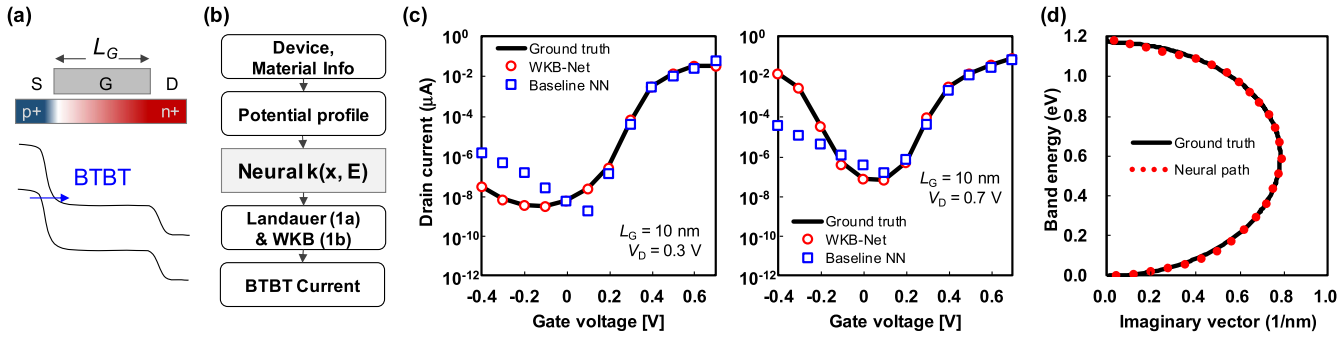

Fig. 5. (a) Schematic tunnel FET (TFET) structure for simulation. For InAs nanowireTFET, device/material parameters in [42] are used. (b) Flowchart for computing BTBT current with WKB-Net consisting of three hidden layers (48-64-48 neurons with ReLU). WKB-Net first learns "effective" imaginary band to compute BTBT current. For training baseline AI models, total 270 samples are generated for LG = [20, . . . , 25 nm], VG = [−0.1, −0.0, . . . , 0.7 V], and VD = [0.4, 0.45, . . . , 0.6 V] and test are done for LG = [10, 30 nm], VG = [−0.4, −0.3, . . . , 0.7 V], and VD = [0.3, 0.7 V]. All models are trained with Adam optimizer [47] whose initial learning rate is 10–4, during 5000 epochs under full-batch scheme. (c) Performance comparison of baseline neural net (blue) and WKB-Net (red) for test cases which are previously unseen. Test RMSE of WKB-Net (baseline NN) is 0.087 (2.23). (d) Inferred imaginary band by WKB-Net, which is in a good agreement with ground truth.

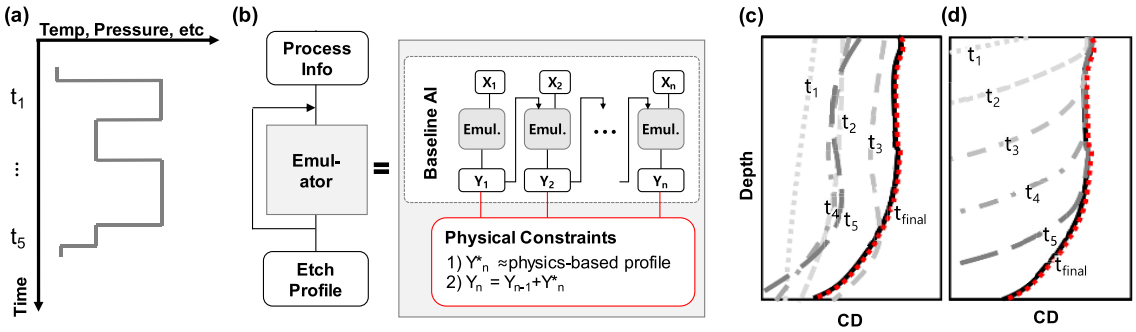

Fig. 6. (a) Schematic of the input recipes for etch processes, in which input parameters like temperature, flow rate, pressures and so on are timely correlated. (b) Illustration of AI model to predict etch profiles. For the model to learn the etching process, we apply two constraints to the model (see red box): 1) the first forces the etched profile predicted by the model to follow a physical behavior and 2) silicon is always etched as time steps progress. The cumulative etch profile until nth step is denoted as $\boldsymbol{Y_n}$, i.e., $\boldsymbol{Y_n} = \boldsymbol{Y_{n-1}} + \boldsymbol{Y_n^*}$ where $\boldsymbol{Y_n^*}$ refers to the etch profile at *n*th step. (c) and (d) Prediction results of etch profiles at initial (t_1), intermediate ($t_2 \sim t_5$), and final time steps (tfinal). Observed profiles are shown in red dashed lines. (c) Result of the baseline model cannot predict etch profiles as time progresses, and (d) while the result with physics-based constraints shows physically sound etch profiles with time.

getting attention in both physics and machine learning communities. Finding general inductive bias for semiconductor design would be an interesting future work.

B. Physics-Based Model for Purely Data-Driven AI Model

In Section II-A, it is shown that parameters for physical TCAD model for an etching process are auto-calibrated with advances of ML technique to match observed data. There are, however, cases where it is difficult to accurately describe the etch processes with TCAD because the physical models in the simulator cannot capture all the underlying governing equations as well as relevant features in experiments such as type of process scheme, equipment information, and so on. In this case, purely data-driven approach instead of TCAD is an attractive option to explore with abundant experimental data.

Sequence-to-sequence model can be a baseline model for etching process because both inputs and outputs are sequential data in time as shown in Fig. 6(a) [3]. Baseline model learns to generate the etched STI profile with respect to input recipes—type of gas, temperature, pressure, etc., and then adapts its shape from categorical inputs such as equipment information. Fig. 6(c) shows how the baseline model works and that the predicted profile at a final time step (denoted as t_{final}) well matches actual process results (red dashed line). It turns out, however, that the baseline AI model cannot properly predict profile evolution as the time progresses ($t_1 \sim t_5$). For example, the model learns that the STI is vertically etched down to bottom layer in the beginning of processes, which is unphysical. This is because etching profiles at intermediate steps are usually not available or enough for the model to learn the physically sound behavior between each time steps.

Physics-based constraints have a potential to alleviate the challenges that the model predict beyond what it has seen before, i.e., improving the extrapolation capability which is one of main strength of physics-based simulation compared to AI model. A physical compact model for etching process or AI-emulator for TCAD simulation as discussed in Section II are used to impose a physics-based topology constraint on the baseline model to learn causality between each steps as depicted in Fig. 6(b). Fig. 6(d) demonstrates that incorporating such physical insights from TCAD helps AI model predict correctly profile evolution at intermediate steps. This can

help extract more physically meaningful insight that can be interpreted by domain experts, when designing new process scheme.

IV. Conclusion

In this study, we primarily focused on challenges of current TCAD approaches and how the ML can alleviate the limitations of TCAD simulations. For calibration of parameterized model, the recent advance of Bayesian optimization helps to deal with high dimensionality for complex physical models and make the physical model robust. We have also demonstrated that a time-consuming physics-based TCAD is replaced with real-time TCAD, AI-emulator for process and device TCAD, which enables tasks such as on-the-fly prediction of device characteristics, extensive searches for optimal designs, and an automated reasoning for the prediction and optimal solution. The full-pattern quantitative analysis of mechanical stress demonstrates that TCAD simulation coverage can be expanded to full-chip level with prudently designed AI-emulator considering short- and long-range stress impact. In the atomistic level, simulation of Si oxidation with NNP shows that machine learning pushes the limit of MD simulation with *ab initio* accuracy to more realistic system size and time scales which was previously unfeasible with AIMD.

We have also illustrated how ML and TCAD are coupled with their combined strengths in the case when physically parameterized model is not available. With virtual measurement of TFET, ML model is shown to discover unknown physical function which can be used as a simplified model instead of rigorous but expensive simulation. Finally, prediction capability of pure AI model can be enhanced with physics-based AI emulator, one way to harness physics-based model, which guides machine learning model to follow a physically meaningful behavior when extrapolation needs.

The results presented here shed new light on use of ML in TCAD and also show that the bridging ML and TCAD with their combined strengths provides a useful tool for a predictive modeling and optimization for early design of semiconductor process and device.

Acknowledgment

Changwook Jeong would like to thank Dr. Seong Hoon Jin and Dr. Hong-Hyun Park for the helpful discussion.

References

[1] A. Mirhoseini *et al.*, "Chip placement with deep reinforcement learning," Apr. 2020, *arXiv:2004.10746*. Accessed: Mar. 13, 2021. [Online]. Available: http://arxiv.org/abs/2004.10746

[2] S. Shim, S. Choi, and Y. Shin, "Machine learning (ML)-based lithography optimizations," in *Proc. IEEE Asia Pacific Conf. Circuits Syst. (APCCAS)*, Oct. 2016, pp. 530–533, doi: 10.1109/APCCAS.2016.7804021.

[3] S. Myung *et al.*, "A novel approach for semiconductor etching process with inductive biases," presented at the NeurIPS Workshop Interpretable Inductive Biases Physically Structured Learn., 2020.

[4] H. Lee, Y. Kim, and C. O. Kim, "A deep learning model for robust wafer fault monitoring with sensor measurement noise," *IEEE Trans. Semiconductor Manuf.*, vol. 30, no. 1, pp. 23–31, Feb. 2017, doi: 10.1109/TSM.2016.2628865.

[5] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Trans. Semiconductor Manuf.*, vol. 30, no. 2, pp. 135–142, May 2017, doi: 10.1109/TSM.2017.2676245.

[6] E. Kim, S. Cho, B. Lee, and M. Cho, "Fault detection and diagnosis using self-attentive convolutional neural networks for variable-length sensor data in semiconductor manufacturing," *IEEE Trans. Semiconductor Manuf.*, vol. 32, no. 3, pp. 302–309, Aug. 2019, doi: 10.1109/TSM.2019.2917521.

[7] M. F. Kasim *et al.*, "Building high accuracy emulators for scientific simulations with deep neural architecture search," Oct. 2020, *arXiv:2001.08055*. Accessed: Mar. 8, 2021. [Online]. Available: http://arxiv.org/abs/2001.08055

[8] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annu. Rev. Phys. Chem.*, vol. 71, no. 1, pp. 361–390, Apr. 2020, doi: 10.1146/annurev-physchem-042018-052331.

[9] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," 2017, *arXiv:1710.11431*. [Online]. Available: http://arxiv.org/abs/1710.11431

[10] I. Huh, E. Yang, S. J. Hwang, and J. Shin, "Time-reversal symmetric ODE network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19016–19027. Accessed: Mar. 8, 2021. [Online]. Available: https://papers.nips.cc/paper/2020/hash/db8419f41d890df802dca330e6284952-Abstract.html

[11] J. Han, A. Jentzen, and W. E, "Solving high-dimensional partial differential equations using deep learning," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 34, pp. 8505–8510, Aug. 2018, doi: 10.1073/pnas.1718942115.

[12] H. Pham, X. Warin, and M. Germain, "Neural networks-based backward scheme for fully nonlinear PDEs," Jul. 2019, *arXiv:1908.00412*. [Online]. Available: http://arxiv.org/abs/1908.00412

[13] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 24, pp. 9943–9948, Jun. 2007, doi: 10.1073/pnas.0609476104.

[14] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, Apr. 2009, doi: 10.1126/science.1165893.

[15] S. H. Rudy, J. Nathan Kutz, and S. L. Brunton, "Deep learning of dynamics and signal-noise decomposition with time-stepping constraints," *J. Comput. Phys.*, vol. 396, pp. 483–506, Nov. 2019, doi: 10.1016/j.jcp.2019.06.056.

[16] *NeurIPS 2020: JAX MD: A Framework for Differentiable Physics*. Accessed Mar. 13, 2021. [Online]. Available: https://neurips.cc/virtual/2020/public/poster_83d3d4b6c9579515e1679aca8cbc8033.html

[17] Y. Hu *et al.*, "DiffTaichi: Differentiable programming for physical simulation," Oct. 2019, *arXiv:1910.00935*. Accessed: Mar. 13, 2021. [Online]. Available: http://arxiv.org/abs/1910.00935

[18] Y. Kim, S. Myung, J. Ryu, C. Jeong, and D. S. Kim, "Physics-augmented neural compact model for emerging device technologies," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 257–260, doi: 10.23919/SISPAD49475.2020.9241638.

[19] X. Gao, A. Huang, N. Trask, and S. Reza, "Physics-informed graph neural network for circuit compact model development," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 359–362, doi: 10.23919/SISPAD49475.2020.9241634.

[20] C.-C. Liu, Y. Li, Y.-S. Yang, C.-Y. Chen, and M.-H. Chuang, "Automatic device model parameter extractions via hybrid intelligent methodology," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 355–358, doi: 10.23919/SISPAD49475.2020.9241613.

[21] J. Wang, Y.-H. Kim, J. Ryu, C. Jeong, W. Choi, and D. Kim, "Artificial neural network-based compact modeling methodology for advanced transistors," *IEEE Trans. Electron Devices*, vol. 68, no. 3, pp. 1318–1325, Mar. 2021, doi: 10.1109/TED.2020.3048918.

[22] H. Yamasaki *et al.*, "Power device degradation estimation by machine learning of gate waveforms," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 335–338, doi: 10.23919/SISPAD49475.2020.9241607.

[23] D. Milardovich, M. Jech, D. Waldhoer, M. Waltl, and T. Grasser, "Machine learning prediction of defect formation energies in a-SiO_2," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 339–342, doi: 10.23919/SISPAD49475.2020.9241609.

[24] S. Myung *et al.*, "Real-time TCAD: A new paradigm for TCAD in the artificial intelligence era," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 347–350, doi: 10.23919/SISPAD49475.2020.9241622.

[25] D. Corne and M. A. Lones, "Evolutionary algorithms," in *Handbook Heuristics*, R. Martí, P. Panos, M. G. C. Resende, Eds. Cham, Switzerland: Springer, 2018, pp. 1–22, doi: 10.1007/978-3-319-07153-4_27-1.

[26] S. Ruder, "An overview of gradient descent optimization algorithms," Jun. 2016, *arXiv:1609.04747*. Accessed: May 11, 2021. [Online]. Available: http://arxiv.org/abs/1609.04747

[27] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," Dec. 2010, Accessed: May 11, 2021. *arXiv:1012.2599*. [Online]. Available: http://arxiv.org/abs/1012.2599

[28] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2012, pp. 2951–2959. Accessed: May 11, 2021. [Online]. Available: https://papers.nips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

[29] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12. Accessed: May 11, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[30] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gaussian process meets big data: A review of scalable GPs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4405–4423, Nov. 2020, doi: 10.1109/TNNLS.2019.2957109.

[31] O. Schenk, S. Rollin, and A. Gupta, "The effects of unsymmetric matrix permutations and scalings in semiconductor device and circuit simulation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 23, no. 3, pp. 400–411, Mar. 2004, doi: 10.1109/TCAD.2004.823345.

[32] S. Sho and S. Odanaka, "A hybrid MPI/OpenMP parallelization method for a quantum drift-diffusion model," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Kamakura, Japan, Sep. 2017, pp. 33–36, doi: 10.23919/SISPAD.2017.8085257.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[34] R. Liu *et al.*, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12, Accessed: Mar. 17, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/60106888f8977b71e1f15db7bc9a88d1-Abstract.html

[35] Y. Wu and K. He, "Group normalization," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01261-8_1.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[37] C.-C. Wang, W. Zhao, F. Liu, M. Chen, and Y. Cao, "Modeling of layout-dependent stress effect in CMOS design," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, New York, NY, USA, Nov. 2009, pp. 513–520, doi: 10.1145/1687399.1687496.

[38] V. Joshi, V. Sukharev, A. Torres, K. Agarwal, D. Sylvester, and D. Blaauw, "Closed-form modeling of layout-dependent mechanical stress," in *Proc. 47th Design Autom. Conf. (DAC)*, New York, NY, USA, Jun. 2010, pp. 673–678, doi: 10.1145/1837274.1837445.

[39] J. Xue, Y. Deng, Z. Ye, H. Wang, L. Yang, and Z. Yu, "A framework for layout-dependent STI stress analysis and stress-aware circuit optimization," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 3, pp. 498–511, Mar. 2012, doi: 10.1109/TVLSI.2010.2102374.

[40] K.-B. Chang *et al.*, "The novel stress simulation method for contemporary DRAM capacitor arrays," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2013, pp. 424–427, doi: 10.1109/SISPAD.2013.6650665.

[41] P. Nicole An and P. A. Kohl, "Modeling simplification for thermal mechanical analysis of high density chip-to-substrate connections," *J. Electron. Packag.*, vol. 133, no. 4, Dec. 2011, Art. no. 041004, doi: 10.1115/1.4005289.

[42] M. Mirza and S. Osindero, "Conditional generative adversarial nets," Nov. 2014, *arXiv:1411.1784*. Accessed: Mar. 17, 2021. [Online]. Available: http://arxiv.org/abs/1411.1784

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[44] J. Wu, Y. Zhang, L. Zhang, and S. Liu, "Deep learning of accurate force field of ferroelectric HfO_2," *Phys. Rev. B, Condens. Matter*, vol. 103, no. 2, Jan. 2021, Art. no. 024108, doi: 10.1103/PhysRevB.103.024108.

[45] K. Lee, "Neural network potential: From code development to application," M.S. thesis, Seoul Nat. Univ., Seoul, South Korea, 2019. Accessed: Mar. 18, 2021. [Online]. Available: https://s-space.snu.ac.kr/handle/10371/161956

[46] S. Ogawa, A. Yoshigoe, S. Ishidzuka, Y. Teraoka, and Y. Takakuwa, "Si(001) surface layer-by-layer oxidation studied by real-time photoelectron spectroscopy using synchrotron radiation," *Jpn. J. Appl. Phys.*, vol. 46, no. 5S, p. 3244, May 2007, doi: 10.1143/JJAP.46.3244.

[47] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, no. 14, Apr. 2007, Art. no. 146401, doi: 10.1103/PhysRevLett.98.146401.

[48] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 991–1001. Accessed: Mar. 13, 2021. [Online]. Available: https://papers.nips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html

[49] J. Klicpera, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," presented at the 8th Int. Conf. Learn. Represent., Apr. 2020. Accessed: Mar. 8, 2021. [Online]. Available: https://iclr.cc/virtual/poster_B1eWbxStPH.html

[50] M. Luisier and G. Klimeck, "Simulation of nanowire tunneling transistors: From the Wentzel-Kramers-Brillouin approximation to full-band phonon-assisted tunneling," *J. Appl. Phys.*, vol. 107, no. 8, Apr. 2010, Art. no. 084507, doi: 10.1063/1.3386521.

[51] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Dec. 2018, pp. 6572–6583.