2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) | 979-8-3315-1635-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/SISPAD62626.2024.10733057

Global Field Heterogeneous Graph Neural Networks for Accelerating Quantum Transport Calculation

Xiaoxin Xie
School of Integrated Circuits
Peking University
Beijing, China
xiexiaoxin@pku.edu.cn

Zhijiang Wang
School of Integrated Circuits
Peking University
Beijing, China
wangzhijiang@pku.edu.cn

Yuchen Wang
School of Integrated Circuits
Peking University
Beijing, China
wangyc0909@pku.edu.cn

Fei Liu
School of Integrated Circuits
Peking University
Beijing, China
feiliu@pku.edu.cn

***Abstract*—In this work, we propose an attention-based global field heterogeneous graph neural network (GFGNN) to characterize global field and dynamic features of the open system, aiming to accelerate or even bypass the computationally demanding self-consistent iterations of NEGF and substantially improve the efficiency of quantum transport calculations. Representing the device with a heterogeneous graph largely preserves its intrinsic physical characteristics, while the global graph attention network effectively captures the propagation of nonlocal physical information and mitigates prediction accuracy issues due to device scaling. GFGNN has been verified to have strong predictive power on 2D MoS$_2$ DG-MOSFETs, with MAE(mean absolute error) as low as 2.1~3.3 meV for potential profile prediction. By incorporating GFGNN into the NEGF computing framework, an acceleration of 182.22~635.71% can be achieved while maintaining the accuracy of transport property calculations.**

***Keywords*—NEGF, Open System, Heterogeneous Graph, GAT, Acceleration**

I. Introduction (*Heading 1*)

As transistor dimensions approach the physical limits of the material, there is a growing need to develop a new generation of atomic-level simulation tools that can provide predictive insights and experimental guidance. The non-equilibrium Green's function (NEGF) method is a crucial tool for calculating the quantum transport properties of a system[1]. However, the self-consistent iteration of the transport and Poisson equations in NEGF requires significant computational resources, making the acceleration of the NEGF method a crucial issue in device simulation. In recent years, machine learning (ML) has had a significant impact on density-functional-theory (DFT) research[2-4]. The method of introducing ML in the confined/ periodic system studied by DFT has matured, while the method of introducing ML in the open system studied by NEGF is still in its early stages(as shown in Fig. 1).

In this work, we propose an attention-based global filed heterogeneous graph neural network(GFGNN) capable of characterizing global field and dynamic features for the open system problem. The flow of our acceleration method is: use the existing simulation/experimental data to train the GFGNN, when using NEGF method to calculate the transport properties of the unknown structure, we can use the trained GFGNN to predict a potential/charge distribution close enough to the real solution, which allows to reduce the number of self-consistent iterations or even skip self-consistent iterations directly. To illustrate this approach, we provide an example of its application to 2D MoS$_2$ DG-MOSFETs. For our experiments, we kept the source/drain length, doping concentration, and unit-cell type fixed. We used small channel length devices (L_{CH} = 7.1 nm~12.07 nm, step = 0.71 nm) as a training set to predict the potential distribution of larger channel devices (L_{CH} = 12.78/13.49/14.20nm). The mean absolute error (MAE) between the predicted and target values were 2.1/2.6/3.3meV, respectively. The NEGF calculation speedup has reached 182.22% to 635.71%, while maintaining accurate transport calculation results.

II. Simulation Method

A. Dataset

The transport properties of 2D MoS$_2$ DG-MOSFETs (as shown on the left side of Fig. 2) were simulated using the kp-NEGF approach. The doping concentration and length of source/drain is 2.86×10^{13} cm^{-2} and 7.1nm(L_S/L_D), respectively. The gate dielectric is HfO$_2$ of 1nm thickness and channel length(L_{CH}) varies in the range of 7.1-14.2 nm (step = 0.71nm). We set V_G = 0-0.6V (step = 0.1V) and V_{DS} = 0-0.5V (step = 0.05V). Eventually 847 device states were generated. Since each atom(unit-cell) of the device is trained independently in the GFGNN, the total amount of data is 148225.

B. Confiend/periodic system VS open system

In this section, we will examine the distinctions and interconnections between confined/periodic and open systems, with a particular focus on their application in the context of neural network algorithm implementation. In the case of a confined/periodic system, by approximation, for each atom, we can consider only the influence of atoms within the surrounding radius R0 on it Fig. 1(a). Regardless of the size of the system, there are a finite number of types of central atoms and local environments, so as long as we have iterated through all the types of central atoms and local environments to be studied in our training, we can decompose the arbitrarily large system into many subproblems with central atoms and their local environments. After summarizing the contributions of each subproblem we can predict some properties of the system to be studied as a whole.

Unlike DFT method, NEGF method address the issue of an open system where electrodes are introduced. This is illustrated in the Fig. 1(b). The electrodes are capable of exchanging electrons with the atoms of the system and/or keeping these atoms under a global electric field. At this point, if we still follow the same approach as in the confined/periodic boundary case, problems arise: (1) Lack of means to represent the different boundary conditions imposed by the contact electrodes. (2) Limitations of the local approximation description.

In an open system, it is necessary to devise novel methodologies for addressing challenges.

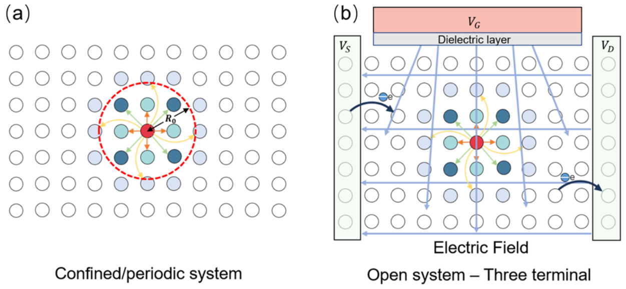

Fig. 1. (a) Confined/periodic system. Consider the influence of atoms within R0 around the center atom. (b) Open system with three terminal: source electrode, drain electrode and gate electrode. The gate brings an additional electric field, but no electrons are exchanged.

C. Heterogeneous Graph of MOSFET

In the study of neural networks, it is very important to adequately characterize the input data. The characterization serves as a bridge between the input data and the neural network, which is related to both the extent to which the information in the input data can be retained and represented, and the ability of the neural network to efficiently learn and capture the knowledge.

To be specific, in the MOSFET structure, we have to deal not only with the introduction of the source/drain electrodes, but also with the introduction of the gate electrodes. The key for us to deal with these problems is to abstract the device structure using a heterogeneous graph[5], as shown in Fig. 2. We abstract the MOSFET into three types of nodes: device nodes, electrode nodes and gate nodes. Device nodes contain three features: atom(unit-cell) type(Z), doping concentration(D), and potential(P^D). In the dataset, since Neumann boundary condition is used at the source/drain electrode, electrode nodes and device nodes are not differentiated in the heterogeneous graph according to our method, so the device nodes mentioned later actually include electrode nodes. Device nodes are connected to each other by bi-directional edges with a distance feature(R^D). Gate nodes contain only the potential feature(P^G). Gate nodes and device nodes are connected by unidirectional edges with a distance feature(R^G).

In the device graph contains several nodes and edges, all of them have their own features, from the basic physics, we can categorize these features, for example: atom(unit-cell) type, doping, inter-atomic distance, and gate node potentials are fixed features, while the potentials of device nodes are indeterminate variable features, and the two types of features play different roles in the network: only the variable features will be updated in the propagation of network, and the process is affected by the joint influence of the fixed features and the variable features. These are meticulously processed in the GF-GAT layer(Fig. 3).

Note that, all features must pass through the embedding layer (Z, D, P^D, P^G) or be expanded using the Gaussian basis (R^D, R^G) before being input to the GF-GAT layer(as shown on the right side of Fig. 2). The construction of this graph includes our full consideration of the physical mechanisms of the MOSFET.

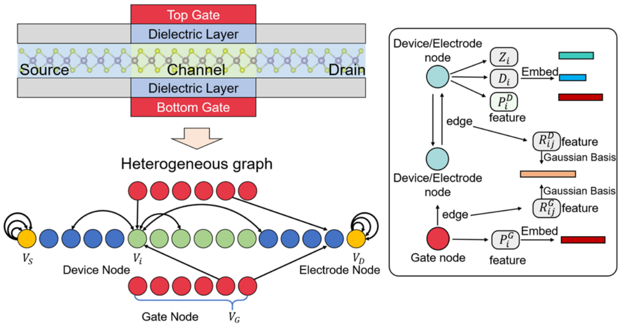

Fig. 2. Heterogeneous graph abstraction of MOSFET. Yellow represents electrode nodes, blue and green represent device nodes, and red represents gate nodes. Electrode nodes and device nodes have features including: atomic(unit-cell) type, doping, and potential, and these nodes are connected to each other by bidirectional edges. Gate nodes have only potential feature, which are connected to electrode/device nodes by unidirectional edges. Both bidirectional and unidirectional edges have a distance feature. Notably, electrode nodes have multiple edges pointing towards themselves, characterizing the semi-infinite electrodes.

D. GF-GAT

The most central part of GFGNN is the GF-GAT layer. Inspired by Transformer's global attention mechanism[6], and taking advantage of the simplicity of the GAT network in attention computation[4,7], we propose the so-called Global Field Graph Attention(GF-GAT) Layer. Unlike the general GAT, instead of setting a truncation radius, we let each device node in the graph pay attention to all other nodes, and the degree of influence of different nodes on the central node is jointly determined by their respective node features and distance features. After propagating the field information through the GF-GAT layers, it enters a fully connected layer and outputs the predicted potential.

As shown in Fig. 3, before the NEGF transport calculation enters the self-consistent iteration, we first use the trained GFGNN to predict the final potential distribution based on the device structure and the applied voltages. Then we can use the predicted potential distribution as the initial value of the self-consistent iteration to reduce the number of iteration steps. Furthermore, within the acceptable accuracy, we can skip the self-consistent iteration and calculate the transport property directly.

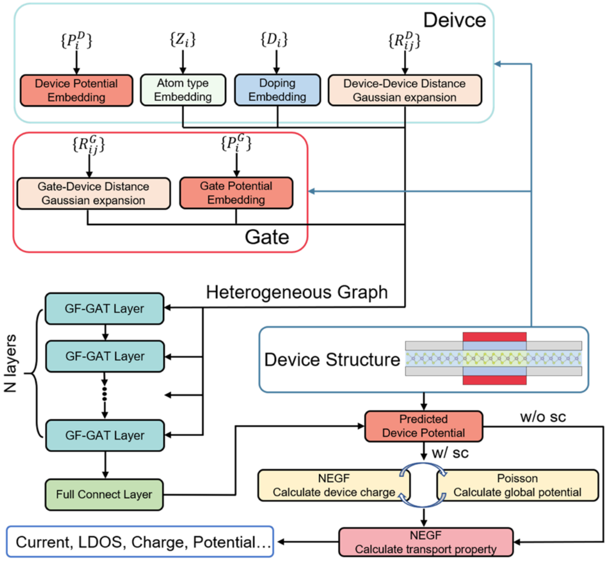
Deivce
$\{P_i^D\}$
$\{Z_i\}$
$\{D_i\}$
$\{R_{ij}^D\}$
Device Potential Embedding
Atom type Embedding
Doping Embedding
Device-Device Distance Gaussian expansion
$\{R_{ij}^G\}$
$\{P_i^G\}$
Gate-Device Distance Gaussian expansion
Gate Potential Embedding
Gate
Heterogeneous Graph
N layers
GF-GAT Layer
GF-GAT Layer
GF-GAT Layer
Full Connect Layer
Device Structure
Predicted Device Potential
w/o sc
w/ sc
NEGF Calculate device charge
Poisson Calculate global potential
NEGF Calculate transport property
Current, LDOS, Charge, Potential…

Fig. 3. Architecture and workflow of the neural network. Firstly, the features are embedded or expanded, then the device graph is fed into a network consisting of GF-GAT Layers and a full connect layer for training, and finally, the trained GFGNN can be embedded into the simulation framework of NEGF for prediction and acceleration.

III. Result and Discussion

In the experiment with MOSFETs we performed two tests:

(1) To test the extrapolation capability of GFGNN in a three terminal device: we fix the length of the source-drain region unchanged ($L_S = L_D = 7.1$nm), vary the length of the channel region $L_{CH} = 7.1$nm~12.07nm(step = 0.71nm), and the data of a total of 8 MOSFETs is used as the training and validation sets. The length of the channel region L_{CH} is then extended to [12.78nm, 13.49nm, 14.20nm] as the test set. In particular, V_D varies in the range of 0V~0.5V(step = 0.05V) and V_G varies in the range of 0V~0.6V(step = 0.1V). The training data contains a total of 103,180 data points.

(2) GFGNN-predicted potential is directly computed for the transport properties in a non-self-consistent step: in previous tests, we have added the GFGNN-predicted results as initial values to the self-consistent iteration, and we have been able to achieve good acceleration as long as the GFGNN predicted results are sufficiently close to the end value of the iteration. Here, we will input the predicted results from test(1) directly into the non-self-consistent step as the iteration end value to compute the transport properties. Since the self-consistent iterative process is completely skipped, this method can achieve the best acceleration results, but accordingly, the loss of accuracy needs to be quantitatively considered.

A. Result of test (1)

In Fig. 4(a-c), we show the comparison between the predicted and target values of the potential distribution at $V_D = 0.25$V for $L_{CH} = $ [12.78nm, 13.49nm, 14.20nm], and the curves are in high agreement (with little difference in the error when V_D is taken at other values). In Fig. 4(d-f), we plotted the distribution of MAE between the predicted and target values in the three devices and calculated the average MAE, which are 0.0023eV,0.0046eV,0.0067eV, respectively, all in the meV order of magnitude.

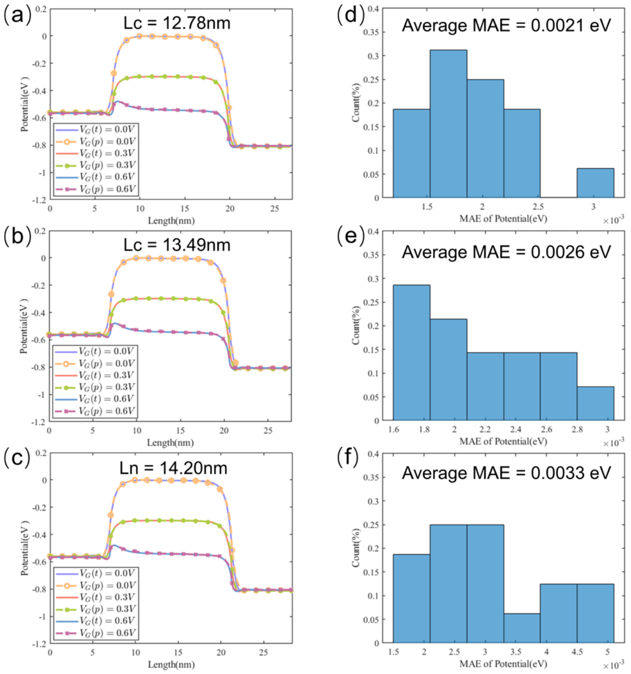

Fig. 4. Extrapolation capability test of GFGNN in MOSFETs. Comparison between the predicted and target potentials of the MOSFETs for fixed source/drain length(L_S/L_D) = 7.1nm and channel length(L_{CH}) = (a) 12.78 nm (b) 13.94 nm (c) 14.20 nm, respectively, which contains the case of $V_D = 0.25$V and $V_G = $ 0V/0.3V/0.6V. (d-f) MAE distribution predicted for all voltage points for cases in (a-c), where the average MAE is 0.0021eV/0.0026eV/0.0033eV respectively.

It is also worth noting that the potential distribution predicted by GFGNN shows a bump in potential between the source and the channel, which is not present between the drain and the channel. This is a reflection of the DIBL effect in device physics, which again demonstrates GFGNN's ability to capture and learn from the intrinsic physics. Next, we embedded the trained GFGNN into the NEGF computational framework to verify the acceleration effect. The good acceleration effect of the GFGNN at each voltage point is demonstrated in Fig. 5(a-c), and the acceleration is 184.85%, 189.74% and 182.22%, respectively. In the $I_D - V_G$ curves (Fig. 5(d-f)), NEGF and GFGNN+NEGF are also in high agreement with R2 = 0.9992/1.0000/0.9999, respectively.

B. Result of test (2)

Unlike previous test, we take the potential predicted by the GFGNN directly as an input to the non-self-consistent step. Since the accuracy of the GFGNN predictions is sufficiently high, the computed transport properties maintain an accuracy close to that of the conventional computational framework, but the computational speed is dramatically improved. Fig. 6(a-c) illustrate the comparison between the NEGF and the GFGNN+NEGF $I_D - V_G$ curves without self-consistent, both of which have R2 = 0.9997/0.9961/0.9996, respectively. Fig. 6(d-f) demonstrates the LDOS as well as the error distribution for both.

To quantify the speedup, we measured the computational time in terms of the number of transport equation solutions that need to be performed in the computation (since the transport equations are the main time consuming part of the NEGF computation). In the calculations in Fig. 6(a-c), NEGF requires 68/81/89 steps, whereas in the non-self-consistent NEGF+GFGNN it requires 14 steps, with a speedup of 485.71%, 578.57% and 635.71%.

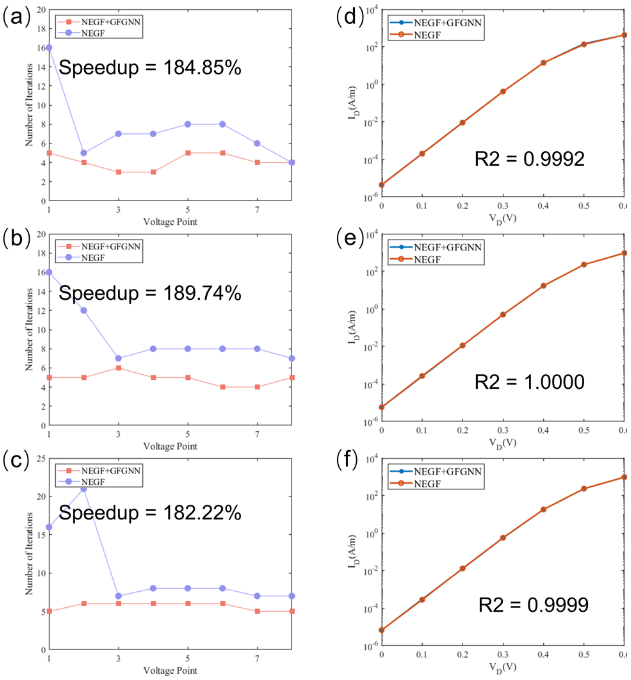

Fig. 5. Acceleration capability and computational accuracy of GFGNN-embedded NEGF in MOSFETs. (a-c) Comparison of the number of iterations per voltage point(V_G) between conventional NEGF and GFGNN-embedded NEGF for the cases of L_{CH} = 12.78 nm/13.94 nm/14.20 nm, with the acceleration = 184.85%, 189.74% and 182.22% respectively. (d-f) are the comparison of I_D-V_G curves calculated by conventional NEGF and GFGNN-embedded NEGF in the above cases, with R2 = 0.9992/1.0000/0.9999, respectively.

IV. Conclusion

We propose an attention-based global filed heterogeneous graph neural network(GFGNN) capable of describing global field and dynamic feature of the open system. Using 2D MoS_2 DG-MOSFETs as an example, we demonstrate the powerful capability of the GFGNN and achieve 182.22~635.71% speedup for the NEGF computation. Although the examples are relatively simple, our method is applicable to more complex device structures and more feature types. We believe that the GFGNN-based acceleration method for transport computation extends the research paradigm of data-driven scientific discovery from the field of DFT to NEGF. And similarly, as an atomic (unit-cell) level neural network, which is capable of learning transport properties from the most intrinsic point of view. Combined with networks used in DFT, it is capable of achieving quasi-DFT+NEGF level transport property prediction, which will have a great facilitating effect on the exploration and discovery of new materials and devices in the future.

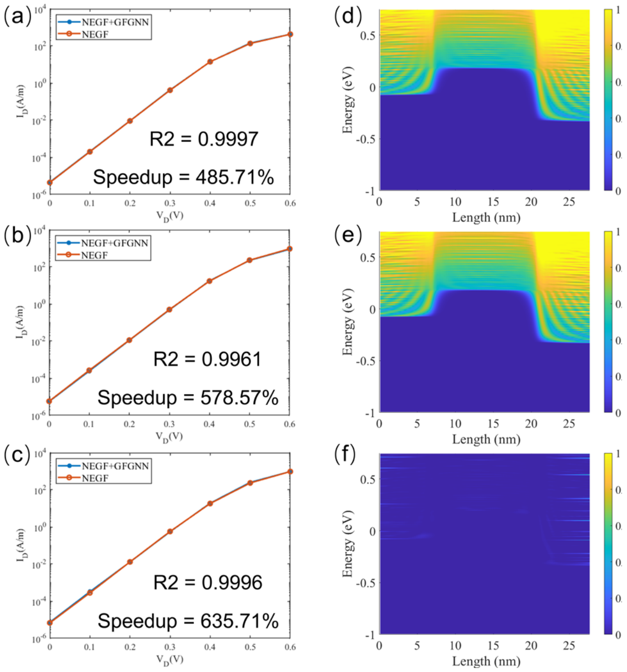

Fig. 6. Acceleration capability and computational accuracy of GFGNN-embedded NEGF(Skip the self-consistent iteration) in MOSFETs. (a-c) Comparison of I_D-V_G curves between conventional NEGF and GFGNN-embedded NEGF for the cases of L_{CH} = 12.78 nm/13.94 nm/14.20 nm, with R2 = 0.9997/0.9961/0.9996 and the acceleration = 485.71%, 578.57% and 635.71% respectively. (d,e) LDOS caculated by conventional NEGF and GFGNN-embedded NEGF. (f) The error distribution.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grants No. T2293703, Grants No. T2293702, and No. T2293700) and the 111 Project (B18001). We thank the high performance computing platform of Peking University for computation facilities.

References

[1] S. Datta, "Quantum Transport: Atom to Transistor", Cambridge University Press, Cambridge, 2005.

[2] T. Xie, and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," Phys. Rev. Lett., vol. 120, pp. 145301, 2018.

[3] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, Y. Xu. "Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation," Nat. Comput. Sci., vol. 2, pp. 367–377, 2022.

[4] Z. Wang, S. Ye, H. Wang, Q. Huang, J. He, S. Chang. "Graph representation-based machine learning framework for predicting electronic band structures of quantum-confined nanostructures." Sci. China Mater., vol. 65, pp. 3157–3170, 20.

[5] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous Graph Neural Network," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 793–803.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. "Attention is All you Need," in Advances in neural information processing systems, vol. 30, 2017.

[7] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio. "Graph Attention Networks," in stat, vol. 1050, no. 20, pp. 10–48550, 2017.