Physics-guided Reinforcement Learning for 3D Molecular Structures

Youngwoo Cho[*1], **Sookyung Kim**[*2], **Peggy Pk Li**[2], **Mike P Surh**[2], **T. Yong-Jin Han**[2], **Jaegul Choo**[1]
Korea University[1], Lawrence Livermore National Lab[2]
cyw314@korea.ac.kr, kim79@llnl.gov, li54@llnl.gov, surh1@llnl.gov, han5@llnl.gov, jchoo@korea.ac.kr

Abstract

Conventional methods to predict 3D molecular structure are based on iterative stochastic optimization techniques by calculating energy using physics-based electronic structure modeling such as Density Functional Theory (DFT) or Molecular Dynamics (MD) which is computationally expensive and multi-modal by in nature. As a cost-efficient and relatively deterministic alternative, we propose a novel RL-based algorithm to optimize 3D structure of single molecule based on DDPG method. Our model teaches the agent to find the best movement trajectory of each atom to reach the correct structure by guiding their movement using three reward approaches based on DFT calculation. Our experiment shows that our model successfully predicts the most energetically stable 3D structure of small aromatic-hydrocarbon molecules.

1 Introduction

In computational chemistry, a common theoretical method used to determine and optimize 3D molecular structure is by minimizing the strain between the atoms in a given molecular system. Any perturbation from the geometry will induce the system to change, so as to reduce this perturbation unless preventing by external forces. Starting from the experimental geometry of molecule, we calculate total energy of the molecule by slightly perturbing the coordinates of each atom. The calculation of total energy can be done by using simulation methods to calculate electronic structure such as DFT (Density Functional Theory) [18] or MD (Molecular Dynamics) [12]. From the variation of total energy, $\delta E(r)$, by chaining location of each atom, δr, we can estimate the derivative of the energy with respect to the position of the atom, $\delta E/\delta r$. Then, the geometry optimization algorithm use $E(r)$, $\delta E/\delta r$ and $\delta\delta E/\delta r_i \delta r_j$ to try to minimize the force. There are many geometry optimization techniques which is built on minimizing the strain and the forces on a given system between atoms such as gradient descent, conjugate gradient, or based on Newton's method (BFGS). There are two main challenges to use this conventional geometry optimization schemes to predict molecular structure.

(1) Stochasticity: The geometry optimization process seeks to find the geometry of a particular arrangement of the atom perturbed from the initial geometry. Therefore, we can find the local energy minimum nearby initial geometry, but difficult to find a global energy minimum. Therefore, the optimized geometry can be not actually the correct answer. Also, the choice of the initial coordinate system can be crucial for performing a successful optimization.
(2) Computational Cost: The process to calculate total energy is repeated until the structure is converged. Therefore, the computing cost of geometry optimization is depended by the number of iterations to calculate energy until the structure is converged. For most large systems of practical interest, it can be prohibitively expensive due to the cost to compute the second derivative of energy.

[*]These authors equally contributed.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Here, we propose to use the reinforcement learning technique as the optimization scheme to predict 3D geometry of small single molecule. Firstly, as the RL model learns the policy by repeating exploration and exploitation, we can potentially explore new structural energy surface which can be dramatically different with initial starting geometry. Therefore, the problem that output structure can be stuck on local energy minimum nearby initial geometry *[problem - 1 above]* can be resolved. Secondly, as the RL model seeks to find the best policy to achieve the goal rather than find the best geometry itself, we can potentially reduce the number of iteration of energy calculation after the agent successfully learn the policy to find the minimum energy structure. *[problem - 2 above]* As our RL model, we used deterministic policy gradient algorithm (DDPG) algorithm [16] because of following reasons: (1) We assume the **deterministic answer** (Correct molecular structure) for the problem to prevent the case of the structure converging to the local minimum. , (2) We design the action space as continuous (**policy gradient**) to allow the smooth movement of each atom in 3D space. To demonstrate the practical use of our model, we show the preliminary results to optimize the small aromatic-hydrocarbon molecules which have simple structural patterns with multiple benzene rings consisted of Carbon and Hydrogen atoms. For the rigorous and physically correct structural prediction, we elaborated physics-based DFT calculation as our reward function. By comparison study between different formats of reward function, we found the best performing reward signal.

2 Related Work

Recently, with the popularity of deep learning, there has been many works in applying deep learning to predicting molecular attributes [21, 4, 25, 14, 8, 6]. Most of advances in material AI built upon graph network representing 2D molecular structure as graph which represent node as atom and edge as the bond between atom. [7, 14, 8] However, the idea to optimize various desired physical property objectives using graphical network can be challenging. The main difficulty arises because these property objectives are difficult to be featurized [4] and non-differentiable. Furthermore, the labeled molecular database is significantly limited. As the distribution of the molecules is vast, it is challenging that the supervised neural net based model learns the entire distribution of chemical space to predict meaningful desired properties of specific target material in limited data. As the alternative method, there have been several advances in applying reinforcement learning to learn physical properties of molecules. Reinforcement learning based approaches specifically has unique advantages to be applied to molecular prediction. First, desired molecular properties such as drug-likeness [1, 25, 17] and structural attributes such as space group or density [22] are complex and non-differentiable. Therefore, It is difficult to be featurized and directly formulated into the objective function of graph generative models. In contrast, reinforcement learning is capable of directly representing hard constraints and desired properties through the design of state, action and reward function. Second, reinforcement learning allows active exploration of the molecule space beyond samples in a dataset. Therefore, reinforcement learning based approach can be the alternative of supervised deep learning [26, 11, 8, 5, 14, 13]. With above strengths, several goal-directed molecule design models based on reinforcement learning have been proposed recently. You et al. [25] proposed GCPN (Graph Convolutional Policy Network) for goal-directed graph generation through reinforcement learning. Gabriel et al. [10] proposed ORGAN model (Objective Reinforced Generative Adversarial Networks) for generating drug-like SMILES strings. Mariya et al. [19] proposed reinforcement learning based SMILE generator also using the combined approach with GAN. However, existing molecule design models limits the structural dimension as 2D and trained their model with data-driven reward function which is often neglecting correct physics. In this work, we designed DFT-based reward function which can learn the physically correct policy and extend output dimension of molecular structure to 3D.

3 Methods

In this work, we tackled the problem to learn 3D structure of small single molecules in aromatic-hydrocarbon family. Figure 1 shows the overview of reinforcement learning setting of our proposed model. We start from the initial structure which is the random configuration of the target molecule. According to the initial structure, initial state tensor (S) is defined. First, the agent takes action (a) to change location of every atoms in 3D space from current state molecular geometry. Second, according to the action, the 3D molecular geometry is updated. Lastly, the DFT environment evaluate total energy (E_{tot}) from updated molecular geometry. Based on the change of energy, the agent obtain the reward (r) by designed reward function (f). By iterating this repeating exploration and exploitation, the agent learns the policy to move every atoms to reach to the correct 3D molecular

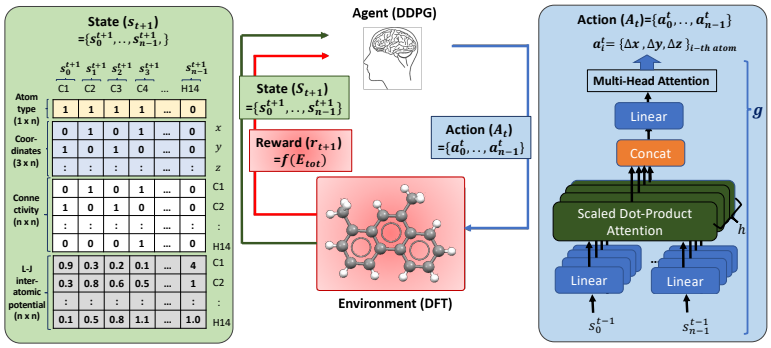

Figure 1: Our reinforcement learning setting to optimize 3D structure of single molecule

geometry (answer). We used DFT calculation as the environment to update molecular geometry and calculate total energy from the given geometry. In this section, we explain details of reinforcement learning setting, and construction of environment.

3.1 Reinforcement Learning Setting

(1) State :
Initial Structure: For every episode, we uses different randomized initial structure of target molecule. To generate initial structure without significantly distort or break chemical bonds of the target molecule, we defines rule-based algorithm as following.

1. Given SMILES [20] string of the target molecule, the 2D chemical configuration with approximated bond-lengths is obtained using RDkit [15].
2. Find two principle components of 2D structure using PCA and re-align structure accordingly.
3. Replace CH_3 (C with 3 Hs) units as theoretical tetrahedron geometry, and perturb z-direction of atoms outside of CH_3 units between $\pm\alpha$. The value of α is the maximum of z-distortion without breaking the chemical bonds of the molecule. We found the α by sensitivity experiment using DFT.

State Matrix: We denote $S_t = \{s_0^t, \ldots, s_{n-1}^t\}$ as a state matrix at step t, where $\mathbf{s}_i^t \in \mathbb{R}^{(4+2n)\times 1}$ for $i = 0, \ldots, n-1$, is a 1-D state vector of $i-th$ atom where n is the total number of atoms in target molecule. The order to traverse atoms at target molecule in State matrix, S_t,(ex: the order of atoms at row direction in State matrix) is defined by Breadth-first search of 2D molecular graph obtained from the 2nd step to generate initial structure above. After projecting structure with two principle components, we always start to traverse atoms from the one located on lowest value of two axes. The 1-D state vector of $i-th$ atom, $\mathbf{s}_i^t$, is consisted with $4+2n$ elements, such as $[\{t_i\}, \{x_i, y_i, z_i\}, \{c_0, \ldots, c_{n-1}\}^i, \{p_0, \ldots, p_{n-1}\}^i]$, where $\{t_i\}$ is the type of atom (C=1, H=0), $\{x_i, y_i, z_i\}$ is coordinate value of $i-th$ atom in 3D space, $\{c_0, \ldots, c_{n-1}\}^i$ is connectivity information between $i-th$ atom and all other atoms with traversing order, and $\{p_0, \ldots, p_{n-1}\}^i$ is Lennard-Jones inter-atomic potential [24] between $i-th$ atom and all other atoms.
Stopping Criteria: There are three stopping criteria of the episode. First, when the action makes connected two atoms way too far from each other and break the chemical bond of target molecule, the episode is terminated with negative reward (-1). Second, when the action makes connected two atoms way too close from each other and makes the DFT-error, the episode is terminated with negative reward (-1). Third, when the calculated energy from DFT are in the range $E_{ans} \pm 3.0eV$, where E_{ans} is the total energy of the correct 3D structure of target molecule, the episode is terminated with positive reward (+1).

(2) Action : We denote $A_t = \{a_0^t, \ldots, a_{n-1}^t\}$ as a action matrix at episode step t where n is the total number of atoms and $\mathbf{a}_i^t \in \mathbb{R}^{3\times 1}$ for $i = 0, \ldots, n-1$, is a 1-D action vector of $i-th$ atom. Each $\mathbf{a}_i^t$ is representing the movement of $i-th$ atom in xyz coordinate. As shown in right blue box at Figure 1, action vector, a_i^t, is modeled by g given previous states. Formally, $a_i^t = g(s_0^{t-1}, \ldots, s_{n-1}^{t-1})$. We use self-attention networks for g to consider the interaction between state of each atom properly to predict the movement of each atom.

(3) Reward : Designing reward is critical to teach agent how to perceive the environment and figure out the best policy. For comparison study, we designed three reward functions based on the total energy calculated from DFT. We denote $r_t = f(E_{tot})$ as a reward at episode step, t.

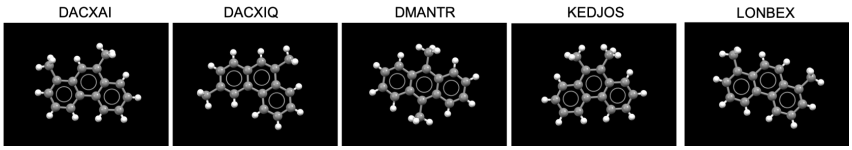

Figure 2: 5 aromatic-hydrocarbon molecules according to crystal structure ID. (C: grey, H: white)

(a) Continuous Reward around E_{ans}: Reward is defined as the continuous value increasing as it approaches to E_{ans}, the total energy with the optimal 3D structure. Formally, $r_t = \frac{a_0}{a_1 + abs(E_{tot} - E_{ans})}$. We set $a_0 = 4.0$ and $a_1 = 1.0$ to make the maximum reward before terminating episode by finding answer (energy in a range of $E_{ans} \pm 3.0eV$) as +1.
(b) Discrete Reward around E_{ans}: The agent only gets reward, +1, when it terminates episode by finding answer($E_{ans} \pm 3.0eV$). Otherwise, it gets 0 or -1 when it terminates with errors.
(c) Continuous Reward in Discrete Range: Inspired by [3], we design reward as the continuous value which increases as decreasing total energy. First, from the experiments with five target aromatic hydrocarbon molecules(Figure 2), we obtained the energy range that the optimal structure of the target molecules can be in. We set E_{min} as -5.0 eV from the minimum total energy among five molecules, and E_{max} as E_{min}+ 100.0 eV. We found E_{min} as -16810.0 eV and E_{max} as -16710.0 eV. We define reward as, $r_t = \frac{b_0 * (E_{max} - E_{tot})}{abs(E_{max} - E_{min})}$. We set $b_0 = 1.0$ to make the maximum reward before terminating as closer to +1. Unlike 1 and 2, the total energy of the answer, E_{ans}, is not required before starting the episode. Therefore, this reward can be applied to any unknown aromatic hydrocarbon molecules with same size as our test cases without pre-evaluation of DFT energy.

4 Experiment and Results

To simplify the problem, we hand-picked 5 simple aromatic-hydrocarbon molecules from Cambridge Structure Database [9] (Figure 2) which have three benzene rings consisted with 16 Carbon(C) and 14 Hydrogen(H) atoms. We constructed general simulated physical environment setting based on DFT which can represent each of test molecule. We used DFT calculation as the environment to update molecular geometry and calculate total energy from the given geometry. We use FHI-aims DFT software-package [2]. We compute energy of the target molecule in the vacuum condition (No k-point mesh) and used the local-density approximation (LDA) [23].

For the comparison, we trained the policy for each target molecule using three reward approaches above and optimized the 3D structure of target molecule using the learned policy. Figure 3-(a,b) shows the learning curves to optimize DACXAI molecule over 1850 episodes, where (a) shows the reward that agent obtains over episodes and (b) shows the critic loss in DDPG defined as the difference between time-dependent target with respect to the output of the critic network sampled for each step. As Figure 3-(a,b) shown, reward-(b) (discrete reward: orange) outperforms than continuous rewards ((a) blue, (c) green) with highest rewards and lowest losses. Reward-(c) shows slightly better than reward-(a) in its reward and loss curve. We hypothesize that a primary reason that reward-(c) is better than reward-(a) is that it is easier task for the agent to optimize the structure with lowest energy (reward-(c)) than locate the structure in the specific energy windows (reward-(a)). Figure 3-(c) shows the inference results to find optimal DACXAI structure with the policy trained with the reward-(b) which is the best performing reward among three. E_{ans} is -16810 eV and yellow region on graph denotes the region considered as the answer ($E_{ans} \pm 3eV$). Two episodes found answer in 3 steps and one episode found answer in 11 steps, which is significantly less than steps required for conventional method by energy gradient, which usually is more than several hundreds.

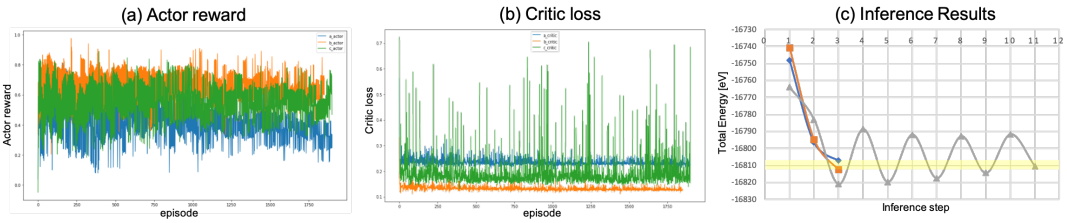

Figure 3: Learning curves (a,b) and inference results(c) to optimize DACXAI molecule.

Acknowledgement

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

References

[1] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90, 2012.

[2] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. The fritz haber institute ab initio molecular simulations package (fhi-aims), 2009.

[3] F. Curtis, X. Li, T. Rose, A. Vazquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli, and N. Marom. Gator: a first-principles genetic algorithm for molecular crystal structure prediction. *Journal of chemical theory and computation*, 14(4):2246–2264, 2018.

[4] M. D Segall. Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Current pharmaceutical design*, 18(9):1292–1310, 2012.

[5] H. Dai, Y. Tian, B. Dai, S. Skiena, and L. Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[6] P. Ertl, R. Lewis, E. Martin, and V. Polyakov. In silico generation of novel, drug-like chemical matter using the lstm neural network. *arXiv preprint arXiv:1712.07449*, 2017.

[7] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[8] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[9] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

[10] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

[11] T. Hester and P. Stone. Texplore: real-time sample-efficient reinforcement learning for robots. *Machine learning*, 90(3):385–429, 2013.

[12] W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[13] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.

[14] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1945–1954. JMLR. org, 2017.

[15] G. Landrum. Rdkit documentation. *Release*, 1:1–79, 2013.

[16] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[17] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[18] R. G. Parr. Density functional theory of atoms and molecules. In *Horizons of Quantum Chemistry*, pages 5–15. Springer, 1980.

[19] M. Popova, O. Isayev, and A. Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[20] J. Schwartz, M. Awale, and J.-L. Reymond. Smifp (smiles fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules. *Journal of chemical information and modeling*, 53(8):1979–1989, 2013.

[21] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2017.

[22] A. L. Spek. Structure validation in chemical crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 65(2):148–155, 2009.

[23] C. Stampfl, W. Mannstadt, R. Asahi, and A. J. Freeman. Electronic structure and physical properties of early transition metal mononitrides: Density-functional theory lda, gga, and screened-exchange lda flapw calculations. *Physical Review B*, 63(15):155106, 2001.

[24] L. Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.

[25] J. You, B. Liu, Z. Ying, V. Pande, and J. Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, pages 6410–6421, 2018.

[26] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.