3D Molecular Representation and Modeling using Deep Learning

Tomohide Masuda, Matthew Ragoza, David Ryan Koes

Department of Computational & Systems Biology, School of Medicine, University of Pittsburgh

Abstract

We previously developed GNINA[1-3], a deep learning framework for molecular docking that uses a three-dimensional grid-based atom density representation of protein-ligand complexes as input (3DCNN). Here we extend this work to *de novo* drug design by training models that output this atomic density representation. We utilize deep generative models, such as variational autoencoders (VAE) and generative adversarial networks (GANs) to learn a latent space of 3D molecular conformations. We can sample novel molecular geometries from this latent space.

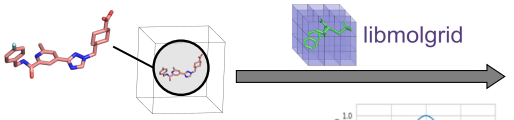
Purpose

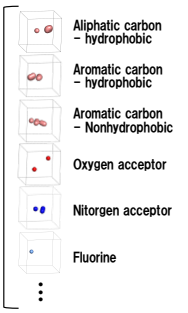
Generation of 3D molecules using 3DCNN and confirmation of "proof of concept"

Methods

➢ 3D grid molecular representation

https://gnina.github.io/libmolgrid/

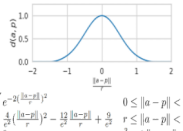




Atomic Density

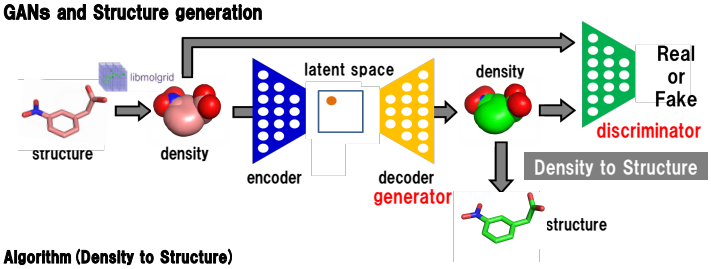
Represent atoms as continuous Gaussians

- 24x24x24 Å grid with 0.5 Å resolution
- 16 receptor atom type channels
- 19 ligand atom type channels
- Randomly rotate/translate



$$d(a,p) = \begin{cases} e^{-2\left(\frac{\|a-p\|}{r}\right)^2} & 0 \le \|a-p\| < r \\ \frac{4}{e^2}\left(\frac{\|a-p\|}{r}\right)^2 - \frac{12}{e^2}\frac{\|a-p\|}{r} + \frac{9}{e^2} & r \le \|a-p\| < \frac{3}{2}r \\ 0 & \frac{3}{2}r \le \|a-p\| \end{cases}$$

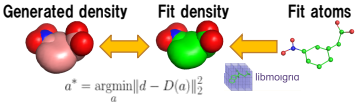
➢ GANs and Structure generation



Algorithm (Density to Structure)

1) Atom fitting

Convert generated continuous density to discrete atoms

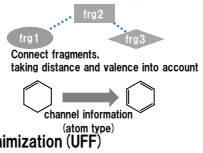


$$a^* = \underset{a}{\operatorname{argmin}} \|d - D(a)\|_2^2$$

- Find a set of atoms a^* that minimizes the squared error between a target density d and the density produced by the set of atoms $D(a)$
- Some conditions such as atom types and the number of them can be imposed

2) Bond adding

1. Recognized by RDKit/Open Babel
2. Modify bond information based on distance between fragments, valence and aromaticity



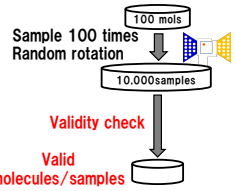
3. Minimization (UFF)

➢ Test scheme

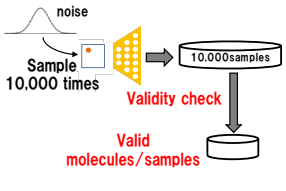
Validity check

1. Recognized by RDKit/Open Babel
   - -multiple fragment
   - -valence
   - -force field
2. Minimization (UFF)

1) Around a target molecule



2) Random (only for VAE)



Data set

PDBbind refined set

- 3,765 protein-ligand complexes

PDBbind

Redocking with Autodock Vina

- Only use poses with RMSD < 2 Å from crystal structure
- 8,648 poses in total (~2.3 per target)

Clustered cross-validation

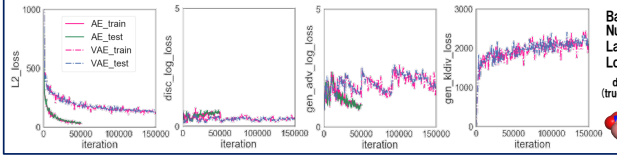
- Cluster targets by protein sequence and ligand fingerprint similarity
- Hold out ⅓ of clusters as test set

Independent test set

- 100 compounds were randomly selected from commercial database (Molport)
- Conformers were generated using RDKit

MolPort

Training



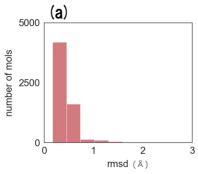
Batch size : 25
Number of G/D iterations : 2
Latent space : 1024
Loss : L2 + GAN + KLDiv

Results

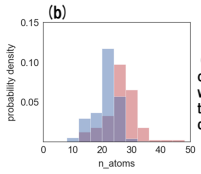
➢Autoencoder

Fitting constraint	# of valid molecules	Valid samples	Reproduced molecules
w*	100/100	6,684/10,000	52/100
w/o	100/100	5,135/10,000	5/100

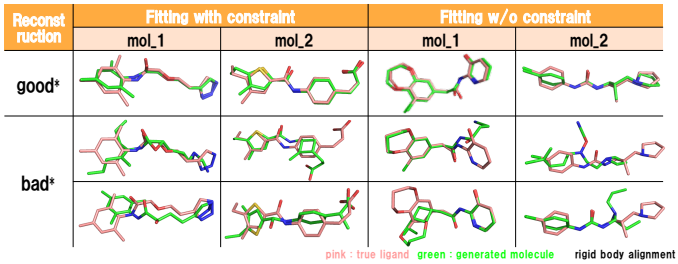
* Atom types information of true ligand is given



(a) Hitstogram about RMSD between fit atoms for a true density and a generated one in the case of atom fitting with atom types constraint



(b) Distribution of the number of atoms in true ligands and valid generated molecules in the case of atom fitting without constraint

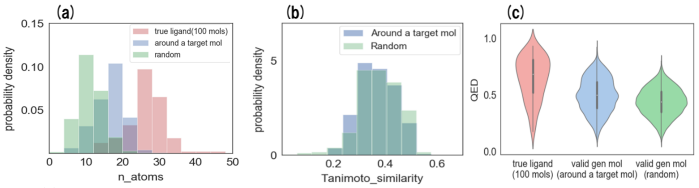


pink : true ligand green : generated molecule rigid body alignment

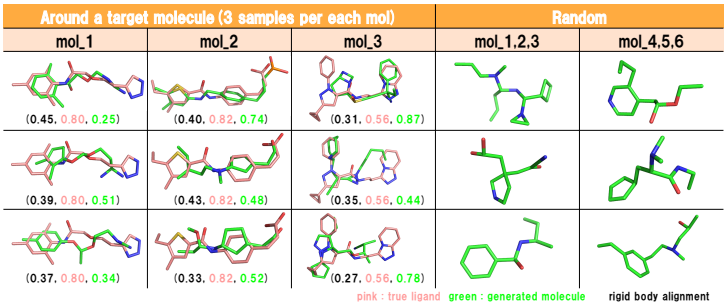
- We succeeded in reproducing 3D molecules using 3DCNN, although it's a challenging task to reproduce molecules without any assumptions
- Improvement of fitting is desirable to reproduce molecules without any assumptions

➢Variational Autoencoder

Fitting constraint	Sampling	# of valid molecules	Valid samples
w/o	Around a target molecule	100/100	4,052/10,000
w/o	Random	100/100	4,922/10,000



(a) Distribution of the number of atoms in true ligands and valid generated molecules
(b) Distribution of Tanimoto similarity using RDKit FP between a target molecule and generated one.
(c) QED (Quantitative Estimate of Drug-likeness) for test molecules and valid generated ones



pink : true ligand green : generated molecule rigid body alignment

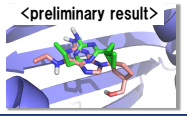
values in parentheses represent similarity, QED of a target molecule, and QED of a generated one, respectively

- Molecules similar to the target molecule were successfully generated
- Relatively small molecules are generated (especially in random case)
- The use of larger data set is considered to leads to highly diverse molecules

Future work

1. Larger dataset for acquiring more diverse molecules
2. Improvement of "Density to Structure"
3. Wasserstein GANs
4. Application to protein-ligand systems
   - binding pocket is also taken into account by GANs

<preliminary result>



PDBid : 5u98
pink : true ligand
green : generated molecule

Conclusion

- We confirmed proof of concept on molecular generation based on 3DCNN.
- We succeeded in generating 3D molecules directly both around the target ones (posterior distribution) and based on random noise (prior distribution).
- There is room to improve the method of atom fitting and bond adding. The use of even larger data set is strongly desirable to generate a variety of molecules.

Reference

1. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2017**, 57, 942–957.
2. Sunseri, J.; Koes, D. R. libmolgrid: GPU Accelerated Molecular Gridding for Deep Learning Applications. arXiv preprint arXiv:1912.04822 **2019**.
3. Francoeur, P. G.; Tomohide, M.; Koes, D. R. 3D Convolutional Neural Networks and a CrossDocked Dataset for Structure-Based Drug Design. ChemRxiv. Preprint (https://doi.org/10.26434/chemrxiv.11833323.v1)

Acknowledgement

This research was supported by R01GM108340 from the National Institute of General Medical Sciences, Toray Industries, Inc., and contributions from NVIDIA Corporation, the University of Pittsburgh Center for Research Computing, and the Extreme Science and Engineering Discovery Environment (XSEDE).