CS6208 : Advanced Topics in Artificial Intelligence
Graph Machine Learning

Lecture 9 : Molecular Science and Graph Generative Models

Semester 2 2022/23

Xavier Bresson

https://twitter.com/xbresson

Department of Computer Science
National University of Singapore (NUS)

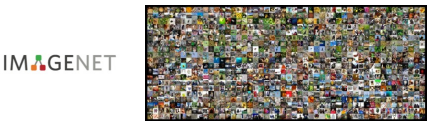
Outline

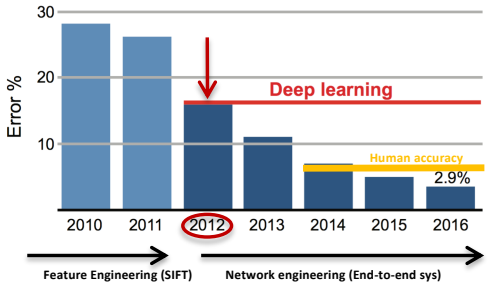
- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Outline

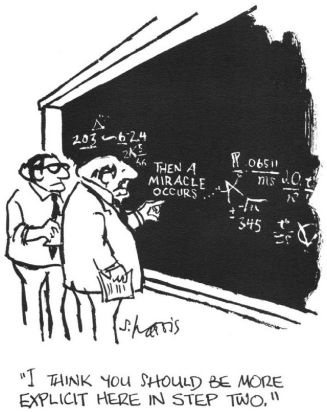
- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Deep Learning

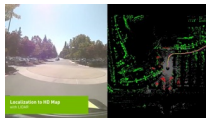




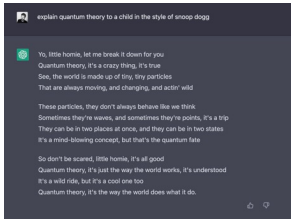
- The DL revolution – 2012 ImageNet
  - First breakthrough in computer vision (CV)
  - Then speech (SR) and language processing (NLP)



- DL works very well.. in practice but why?
  - High-dimensional data/curse of dimensionality
  - Characterization of energy landscapes/solutions
  - Generalization (feature learning, inductive bias, expressivity)



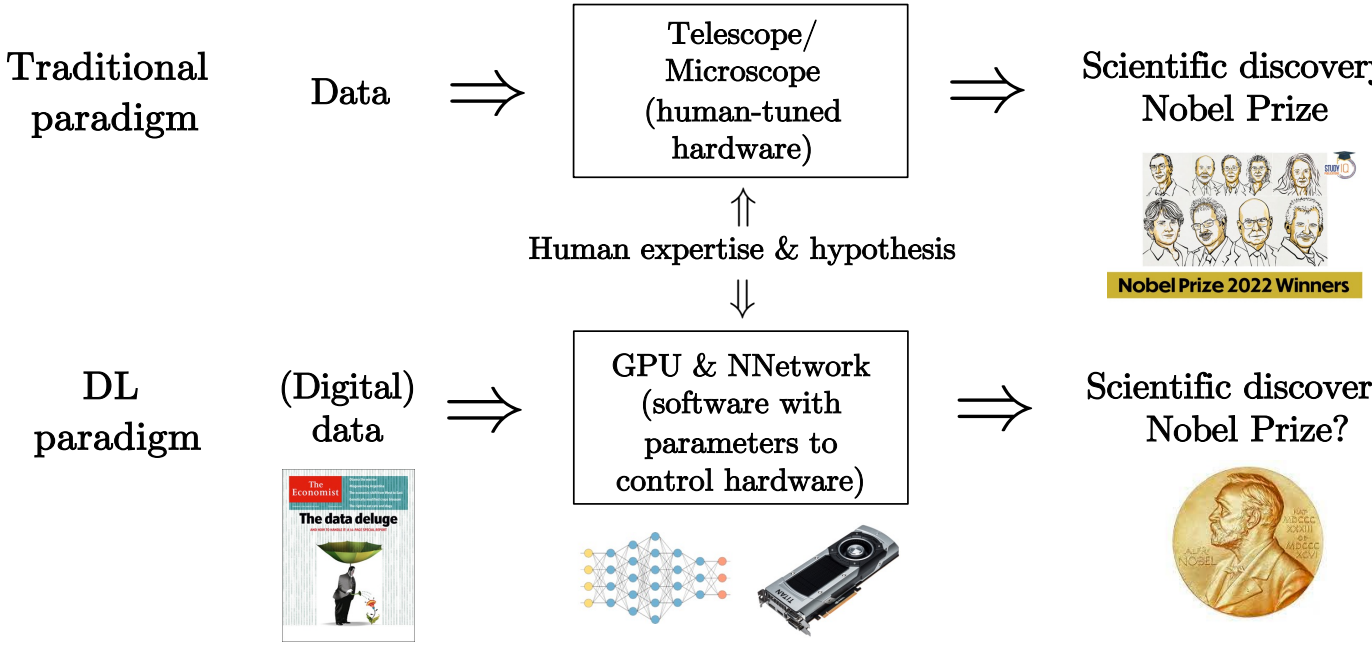
- DL products
  - Computer Vision : Face recognition, video surveillance, autonomous driving
  - Natural Language Processing : Machine translation, chatbot (ChatGPT)
  - Speech Recognition : Virtual assistants (Alexa/Siri/Google/Cortana)

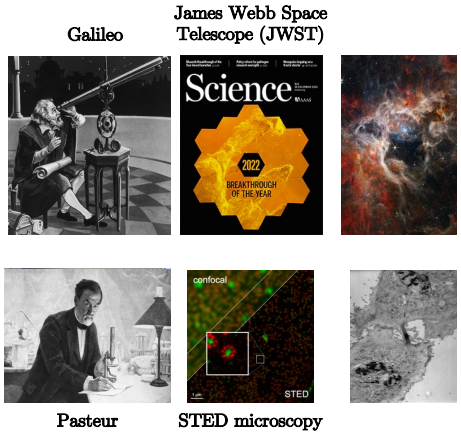


- DL beyond CV/NLP/SR for scientific discovery

DL for science

- DL has been applied to biology, physics, chemistry, material, medicine, engineering, etc
- Scientific tools s.a. telescope/microscope have served physics (Galileo's proof that earth is round and revolves around the sun, study of universe) and biology (Pasteur's discovery of microbes, study of cells, molecules).
- Can DL be the new telescope/microscope for scientific discovery?





DL for molecular science

- A promising application of DL and potentially breakthrough is in molecular science.
- Case studies
  - Drug discovery (halicin antibiotic)
  - De novo drug design (new molecules w/ optimized chemical property)
  - Protein folding (3D structure)
  - Protein-drug interaction (drugs with strong binding)
  - Generate protein with text prompt (ChatGPT for biology)

GNNs for DFT estimation

- Density Functional Theory (DFT) is one of the main workhorses to estimate molecular properties s.a. energy (Walter Kohn awarded Nobel Prize in Chemistry in '98 for DFT).
- DFT is O(N^3), N being the number of atoms vs. GNNs being O(N).

Neural Message Passing for Quantum Chemistry

Justin Gilmer [1] Samuel S. Schoenholz [1] Patrick F. Riley [2] Oriol Vinyals [3] George E. Dahl [1]

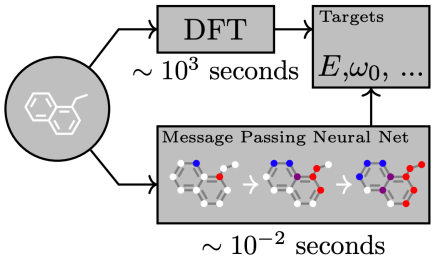
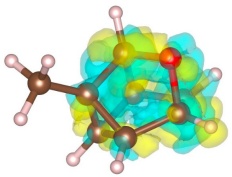
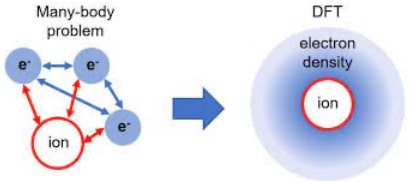
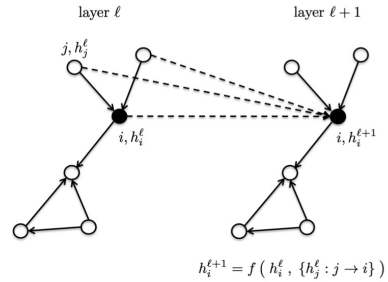


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.



(a) C_8H_8O (1.1%)





npj | computational materials

Explore content · About the journal · Publish with us

nature › npj computational materials › articles › article

Article | Open Access | Published: 23 August 2022

Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids

Peter Bjørn Jørgensen & Arghya Bhowmik

npj Computational Materials 8, Article number: 183 (2022) | Cite this article

3382 Accesses | 13 Altmetric | Metrics

Abstract

Electron density $\rho(\vec{\mathbf{r}})$ is the fundamental variable in the calculation of ground state energy with density functional theory (DFT). Beyond total energy, features and changes in $\rho(\vec{\mathbf{r}})$ distributions are often used to capture critical physicochemical phenomena in functional materials. We present a machine learning framework for the prediction of $\rho(\vec{\mathbf{r}})$. The model is based on equivariant graph neural networks and the electron density is predicted at special query point vertices that are part of the message-passing graph, but only receive messages. The model is tested across multiple datasets of molecules (QM9), liquid ethylene carbonate electrolyte (EC) and LixNiyMnzCo(1-y-z)O2 lithium ion battery cathodes (NMC). For QM9 molecules, the accuracy of the proposed model exceeds typical variability in $\rho(\vec{\mathbf{r}})$ obtained from DFT done with different exchange-correlation functionals. The accuracy on all three datasets is beyond state of the art and the computation time is orders of magnitude faster than DFT.

npj | computational materials

Explore content · About the journal · Publish with us

nature › npj computational materials › articles › article

Article | Open Access | Published: 21 May 2021

Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture

Cheol Woo Park, Mordechai Kornbluth, Jonathan Vandermause, Chris Wolverton, Boris Kozinsky & Jonathan P. Mailoa

npj Computational Materials 7, Article number: 73 (2021) | Cite this article

9838 Accesses | 36 Citations | 1 Altmetric | Metrics

Abstract

Recently, machine learning (ML) has been used to address the computational cost that has been limiting ab initio molecular dynamics (AIMD). Here, we present GNNFF, a graph neural network framework to directly predict atomic forces from automatically extracted features of the local atomic environment that are translationally-invariant, but rotationally-covariant to the coordinate of the atoms. We demonstrate that GNNFF not only achieves high performance in terms of force prediction accuracy and computational speed on various materials systems, but also accurately predicts the forces of a large MD system after being trained on forces obtained from a smaller system. Finally, we use our framework to perform an MD simulation of $Li_7P_3S_{11}$, a superionic conductor, and show that resulting Li diffusion coefficient is within 14% of that obtained directly from AIMD. The high performance exhibited by GNNFF can be easily generalized to study atomistic level dynamics of other material systems.

GNNs for drug discovery

Cell

Volume 180, Issue 4, 20 February 2020, Pages 688-702.e13

Article

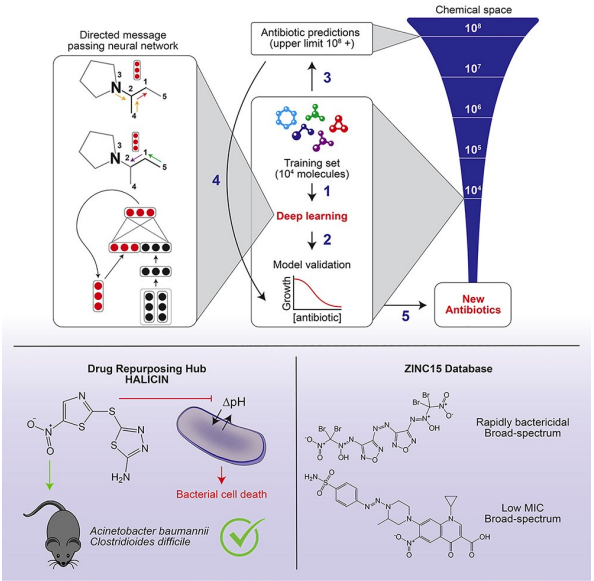
A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes [1, 2, 3], Kevin Yang [3, 4, 10], Kyle Swanson [3, 4, 10], Wengong Jin [3, 4], Andres Cubillos-Ruiz [1, 2, 5], Nina M. Donghia [1, 5], Craig R. MacNair [6], Shawn French [6], Lindsey A. Carfrae [6], Zohar Bloom-Ackermann [2, 7], Victoria M. Tran [2], Anush Chiappino-Pepe [5, 7], Ahmed H. Badran [2], Ian W. Andrews [1, 2, 5], Emma J. Chory [1, 2], George M. Church [5, 7, 8], Eric D. Brown [6], Tommi S. Jaakkola [3, 4] ... James J. Collins [1, 2, 5, 8, 9, 11]

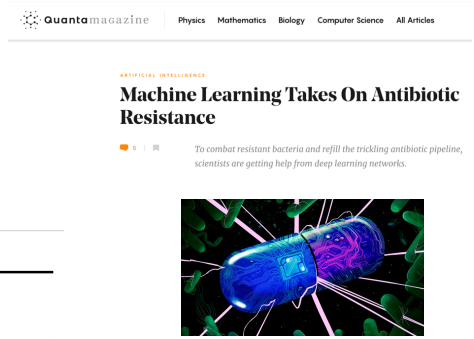
Highlights

- A deep learning model is trained to predict antibiotics based on structure
- Halicin is predicted as an antibacterial molecule from the Drug Repurposing Hub
- Halicin shows broad-spectrum antibiotic activities in mice
- More antibiotics with distinct structures are predicted from the ZINC15 database

https://www.sciencedirect.com/science/article/pii/S0092867420301021



Halicin was previously developed for anti-diabetic treatment.



GNNs for de novo drug design

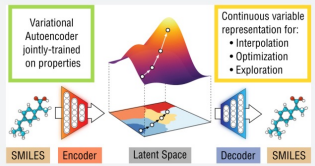
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,[†,#] Jennifer N. Wei,[‡,#] David Duvenaud,[¶,#] José Miguel Hernández-Lobato,[§,#] Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,[∇,∥] and Alán Aspuru-Guzik*[,‡,⊥]

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States
[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States
[¶]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada
[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.
[∇]Google Brain, Mountain View, California, United States
[∥]Princeton University, Princeton, New Jersey, United States
[⊥]Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Supporting Information

ABSTRACT: We report a method to convert discrete representations of molecules to and from a multidimensional continuous representation. This model allows us to generate new molecules for efficient exploration and optimization through open-ended spaces of chemical compounds. A deep neural network was trained on hundreds of thousands of existing chemical structures to construct three coupled functions: an encoder, a decoder, and a predictor. The encoder converts the discrete representation of a molecule into a real-valued continuous vector, and the decoder converts these continuous vectors back to discrete molecular representations. The predictor estimates chemical properties from the latent continuous vector representation of the molecule. Continuous representations of molecules allow us to automatically generate novel chemical structures by performing simple operations in the latent space, such as decoding random vectors, perturbing known chemical structures, or interpolating between molecules. Continuous representations also allow the use of powerful gradient-based optimization to efficiently guide the search for optimized functional compounds. We demonstrate our method in the domain of drug-like molecules and also in a set of molecules with fewer that nine heavy atoms.



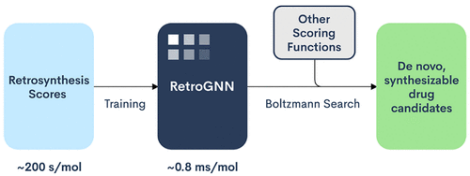
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software

Cheng-Hao Liu*, Maksym Korablyov, Stanisław Jastrzębski, Paweł Włodarczyk-Pruszyński, Yoshua Bengio, and Marwin Segler*

Cite this: J. Chem. Inf. Model. 2022, 62, 10, 2293–2300
Publication Date: April 22, 2022
https://doi.org/10.1021/acs.jcim.1c01476

Article Views 1798 | Altmetric 8 | Citations 1
LEARN ABOUT THESE METRICS



scientific reports

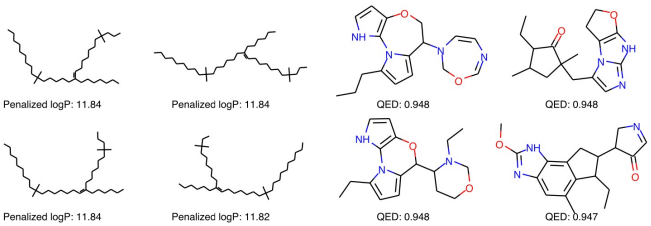
Explore content | About the journal | Publish with us

nature › scientific reports › articles › article

Article | Open Access | Published: 24 July 2019

Optimization of Molecules via Deep Reinforcement Learning

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare & Patrick Riley



(a) Optimization of penalized logP (b) Optimization of QED

Figure 2. Sample molecules in the property optimization task. (**a**) Optimization of penalized logP from MolDQN-bootstrap; note that the generated molecules are obviously not drug-like due to the use of a single-objective reward. (**b**) Optimization of QED from MolDQN-twosteps.

logP: molecular solubility
QED: Quantitative Estimate of Druglikeness

Sundin et al. Journal of Cheminformatics (2022) 14:86
https://doi.org/10.1186/s13321-022-00667-8

Journal of Cheminformatics

RESEARCH Open Access

Human-in-the-loop assisted de novo molecular design

Iiris Sundin[1*], Alexey Voronov[2*], Haoping Xiao[1], Kostas Papadopoulos[2,5], Esben Jannik Bjerrum[2,5], Markus Heinonen[1], Atanas Patronov[2,5], Samuel Kaski[1,3] and Ola Engkvist[2,4]

REVIEWS Drug Discovery Today • Volume 26, Number 6 • June 2021
ELSEVIER
Reviews • KEYNOTE REVIEW

Teaser This paper highlights the progress and challenges in de novo drug design using graph neural network technology.

Graph neural networks for automated *de novo* drug design

Jiacheng Xiong[1,2], Zhaoping Xiong[1,3], Kaixian Chen[1,2], Hualiang Jiang[1,2] and Mingyue Zheng[1,2]

[1] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China
[2] University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China
[3] Shanghai Institute for Advanced Immunochemical Studies, and School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China

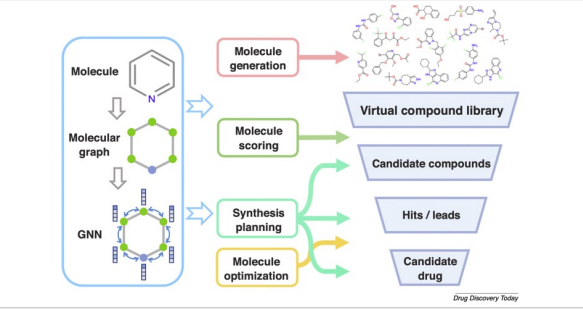
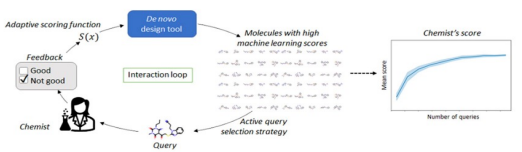


FIGURE 1
The applications of graph neural networks (GNNs) in all stages of automated *de novo* drug design.

Graphical Abstract



GNNs for molecular generation

- Categories of graph generative models : VAE, GAN, Diffusion Model

A Two-Step Graph Convolutional Decoder for Molecule Generation

Xavier Bresson
School of Computer Science and Engineering
NTU, Singapore
xbresson@ntu.edu.sg

Thomas Laurent
Department of Mathematics
Loyola Marymount University
tlaurent@lmu.edu

Abstract

We propose a simple auto-encoder framework for molecule generation. The molecular graph is first encoded into a continuous latent representation z, which is then decoded back to a molecule. The encoding process is easy, but the decoding process remains challenging. In this work, we introduce a simple two-step decoding process. In a first step, a fully connected neural network uses the latent vector z to produce a molecular formula, for example CO_2 (one carbon and two oxygen atoms). In a second step, a graph convolutional neural network uses the same latent vector z to place bonds between the atoms that were produced in the first step (for example a double bond will be placed between the carbon and each of the oxygens). This two-step process, in which a bag of atoms is first generated, and then assembled, provides a simple framework that allows us to develop an efficient molecule auto-encoder. Numerical experiments on basic tasks such as novelty, uniqueness, validity and optimized chemical property for the 250k ZINC molecules demonstrate the performances of the proposed system. Particularly, we achieve the highest reconstruction rate of 90.5%, improving the previous rate of 76.7%. We also report the best property improvement results when optimization is constrained by the molecular distance between the original and generated molecules.

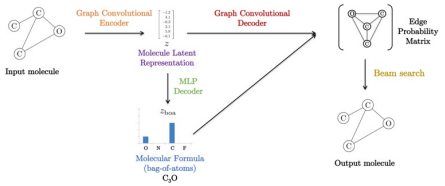


Figure 1: Proposed auto-encoder. The encoder reduces the molecular graph to a latent vector z. The decoder uses a MLP to produce a molecular formula and a graph convolutional network classifies each bond between the atoms given by the molecular formula. Finally, a beam search generates a valid molecule.

Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation

Jiaxuan You[1]* jiaxuan@stanford.edu
Bowen Liu[2]* liubowen@stanford.edu
Rex Ying[1] rexying@stanford.edu
Vijay Pande[3] pande@stanford.edu
Jure Leskovec[1] jure@cs.stanford.edu
[1]Department of Computer Science, [2]Department of Chemistry, Stanford University, Stanford, CA, 94305 [3]Department of Bioengineering

Abstract

Generating novel graph structures that optimize given objectives while obeying some given underlying rules is fundamental for chemistry, biology and social science research. This is especially important in the task of molecular graph generation, whose goal is to discover novel molecules with desired properties such as drug-likeness and synthetic accessibility, while obeying physical laws such as chemical valency. However, designing models to find molecules that optimize desired properties while incorporating highly complex and non-differentiable rules remains to be a challenging task. Here we propose Graph Convolutional Policy Network (GCPN), a general graph convolutional network based model for goal-directed graph generation through reinforcement learning. The model is trained to optimize domain-specific rewards and adversarial loss through policy gradient, and acts in an environment that incorporates domain-specific rules. Experimental results show that GCPN can achieve 61% improvement on chemical property optimization over state-of-the-art baselines while resembling known molecules, and achieve 184% improvement on the constrained property optimization task.

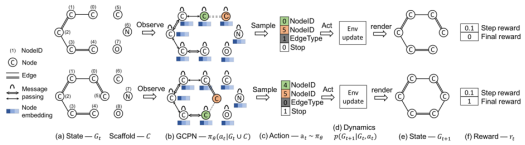


Figure 1: An overview of the proposed iterative graph generation method. Each row corresponds to one step in the generation process. **(a)** The state is defined as the intermediate graph G_t, and the set of scaffold subgraphs defined as C is appended for GCPN calculation. **(b)** GCPN conducts message passing to encode the state as node embeddings then produce a policy π_θ. **(c)** An action a_t with 4 components is sampled from the policy. **(d)** The environment performs a chemical valency check on the intermediate state, and then returns **(e)** the next state G_{t+1} and **(f)** the associated reward r_t.

Equivariant Diffusion for Molecule Generation in 3D

Emiel Hoogeboom [*1] **Victor Garcia Satorras** [*1] **Clément Vignac** [*2] **Max Welling** [1]

Abstract

This work introduces a diffusion model for molecule generation in 3D that is equivariant to Euclidean transformations. Our E(3) Equivariant Diffusion Model (EDM) learns to denoise a diffusion process with an equivariant network that jointly operates on both continuous (atom coordinates) and categorical features (atom types). In addition, we provide a probabilistic analysis which admits likelihood computation of molecules using our model. Experimentally, the proposed method significantly outperforms previous 3D molecular generative methods regarding the quality of generated samples and efficiency at training time.

1. Introduction

Modern deep learning methods are starting to make an important impact on molecular sciences. Behind the success of Alphafold in protein folding prediction (AlQuraishi, 2019), an increasing body of literature develops deep learning models to analyze or synthesize (in silico) molecules (Simonovsky & Komodakis, 2018; Gebauer et al., 2019; Klicpera et al., 2020; Simm et al., 2021).

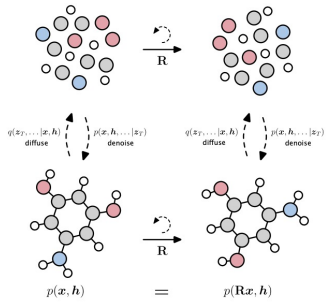


Figure 1. Overview of the EDM. To generate a molecule, a normal distributed set of points is denoised into a molecule consisting of atom coordinates $\boldsymbol{x}$ in 3D and atom types $\boldsymbol{h}$. As the model is rotation equivariant, the likelihood is preserved when a molecule is rotated by $\mathbf{R}$.

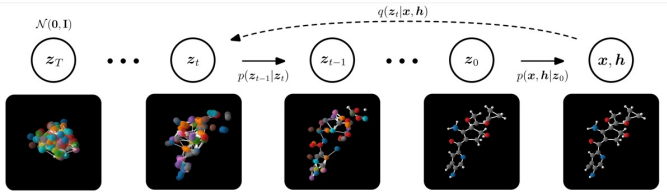


Figure 2. Overview of the Equivariant Diffusion Model. To generate molecules, coordinates $\boldsymbol{x}$ and features $\boldsymbol{h}$ are generated by denoising variables $\boldsymbol{z}_t$ starting from standard normal noise $\boldsymbol{z}_T$. This is achieved by sampling from the distributions $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ iteratively. To train the model, noise is added to a datapoint $\boldsymbol{x}, \boldsymbol{h}$ using $q(\boldsymbol{z}_t|\boldsymbol{x}, \boldsymbol{h})$ for the step t of interest, which the network then learns to denoise.

GNNs for protein folding/structure

AlphaFold: a solution to a 50-year-old grand challenge in biology

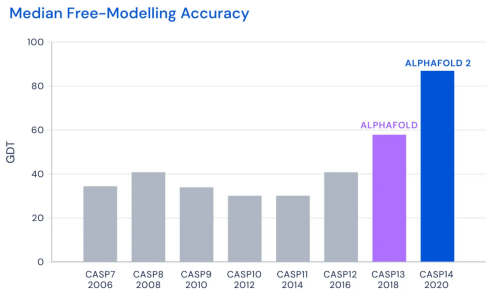
SHARE

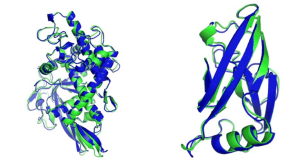
AUTHORS

TAt The AlphaFold team

Proteins are essential to life, supporting practically all its functions. They are large complex molecules, made up of chains of amino acids, and what a protein does largely depends on its unique 3D structure. Figuring out what shapes proteins fold into is known as the "protein folding problem", and has stood as a grand challenge in biology for the past 50 years. In a major scientific advance, the latest version of our AI system AlphaFold has been recognised as a solution to this grand challenge by the organisers of the biennial Critical Assessment of protein Structure Prediction (CASP). This breakthrough demonstrates the impact AI can have on scientific discovery and its potential to dramatically accelerate progress in some of the most fundamental fields that explain and shape our world.

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

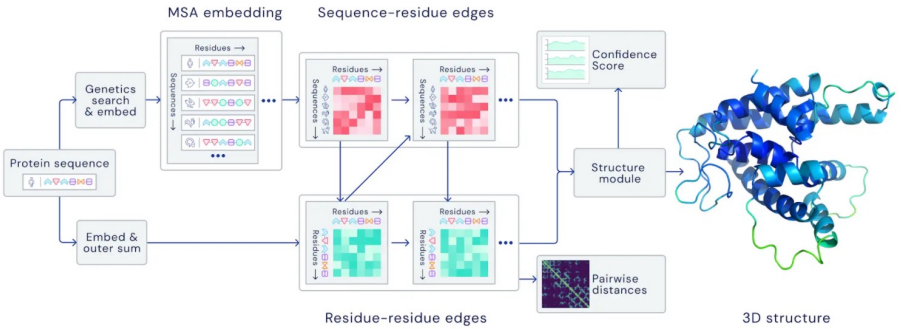




T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction



Graph Transformer

Science RESEARCH ARTICLES

Cite as: M. Baek *et al.*, *Science* 10.1126/science.abj8754 (2021).

Accurate prediction of protein structures and interactions using a three-track neural network

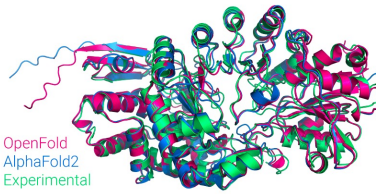
Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[2,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,21,7], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

RoseTTaFold

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization

Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, Mohammed AlQuraishi



OpenFold

GNNs for protein-drug discovery & design

Structure-based drug discovery with deep learning

Rıza Özçelik[1,2+], **Derek van Tilborg**[1,2+], **José Jiménez-Luna**[3], **and Francesca Grisoni**[1,2*]

[1]Eindhoven University of Technology, Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven, Netherlands.
[2]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Netherlands.
[3]Microsoft Research Cambridge, Cambridge, United Kingdom.

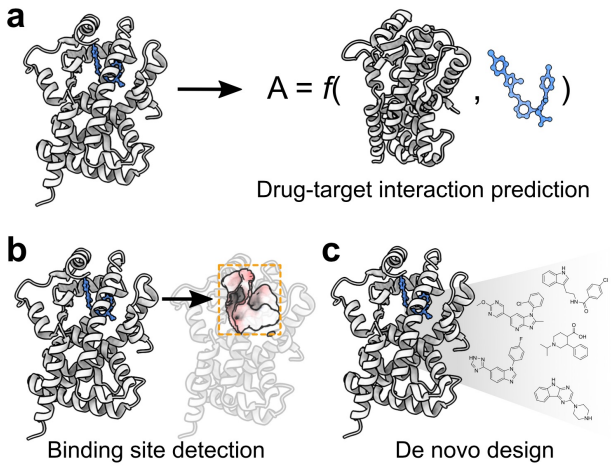
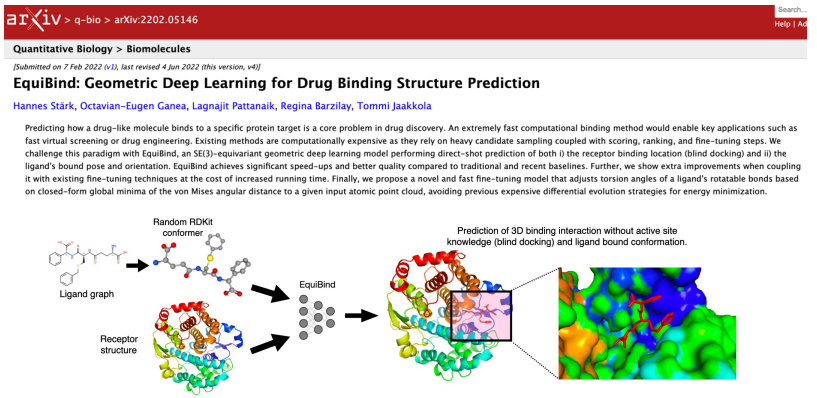
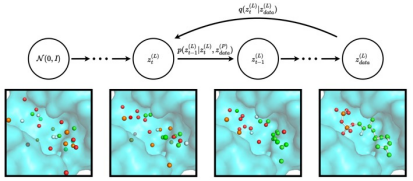


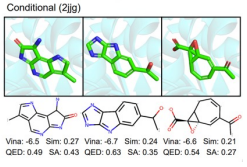
Figure 1. Structure-based drug discovery tasks discussed in this review: (a) *drug-target interaction prediction*, which aims to predict the affinity between a protein and a ligand, using the structural information of both molecular entities; (b) *binding site detection*, which aims to identify 'druggable' cavities in the protein structure, (c) *de novo design*, aiming to design bioactive molecules from scratch using the information of a protein target.



Structure-based Drug Design with Equivariant Diffusion Models

Arne Schneuing[1*], **Yuanqi Du**[2*], **Charles Harris**[3], **Arian Jamasb**[3], **Ilia Igashov**[1], **Weitao Du**[4], **Tom Blundell**[3], **Pietro Lió**[3], **Carla Gomes**[2], **Max Welling**[5], **Michael Bronstein**[6] **& Bruno Correia**[1]
[1]École Polytechnique Fédérale de Lausanne, [2]Cornell University, [3]University of Cambridge, [4]USTC, [5]Microsoft Research AI4Science, [6]University of Oxford



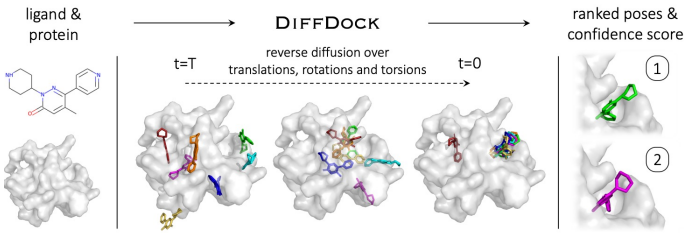


DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking

Gabriele Corso*, **Hannes Stärk***, **Bowen Jing***, **Regina Barzilay** & **Tommi Jaakkola**
CSAIL, Massachusetts Institute of Technology

Abstract

Predicting the binding structure of a small molecule ligand to a protein—a task known as *molecular docking*—is critical to drug design. Recent deep learning methods that treat docking as a regression problem have decreased runtime compared to traditional search-based methods but have yet to offer substantial improvements in accuracy. We instead frame molecular docking as a *generative* modeling problem and develop DiffDock, a diffusion generative model over the non-Euclidean manifold of ligand poses. To do so, we map this manifold to the product space of the degrees of freedom (translational, rotational, and torsional) involved in docking and develop an efficient diffusion process on this space. Empirically, DiffDock obtains a 38% top-1 success rate (RMSD<2Å) on PDBBind, significantly outperforming the previous state-of-the-art of traditional docking (23%) and deep learning (20%) methods. Moreover, while previous methods are not able to dock on computationally folded structures (maximum accuracy 10.4%), DiffDock maintains significantly higher precision (21.7%). Finally, DiffDock has fast inference times and provides confidence estimates with high selective accuracy.



Generate protein with text prompt

- Inspired by breakthrough multi-modal image-language models s.a. Midjourney, OpenAI's DALL-E, Google Brain's Imagen and Stable Diffusion:

Midjourney image from the prompt "swimming pool filled with a galaxy on a moonlit night"

An image generated with DALL-E 2 based on the text prompt "Teddy bears working on new AI research underwater with 1990s technology"

An image generated by Stable Diffusion based on the text prompt "a photograph of an astronaut riding a horse"

Generate protein with text prompt

Illuminating protein space
with a programmable generative model

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier,
Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer,
Andrew Beam, Gevorg Grigoryan

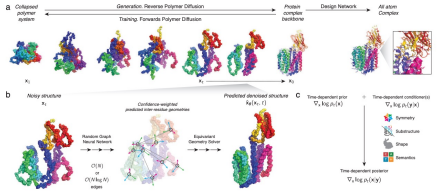
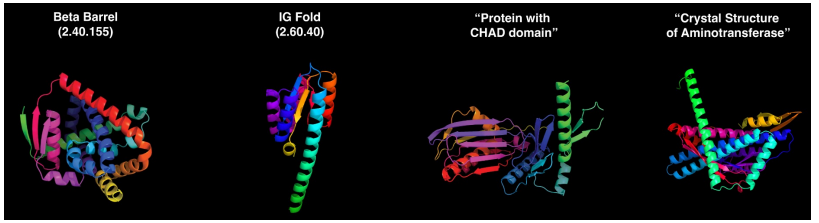


Figure 1: ***Chroma* is a generative model for proteins and protein complexes that combines a structured diffusion model for protein backbones with scalable molecular neural networks for backbone synthesis and all-atom design. a**, A correlated diffusion process with chain and ra-

https://www.generatebiomedicines.com/chroma

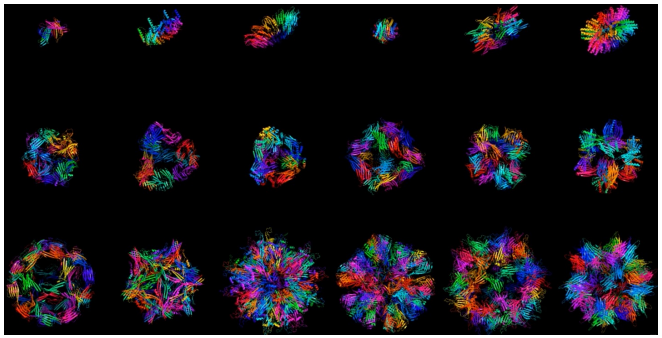
Programming proteins with Chroma

Semantic conditioning

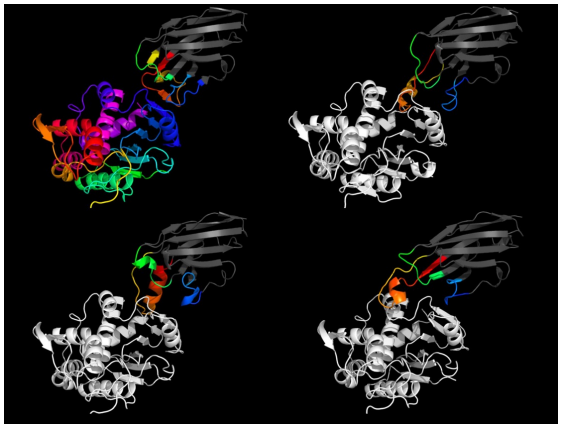


Chroma: "Generate a protein with CHAD domain"
(with small GPT trained on protein captioning)

Making giants



Symmetry groups



Protein infilling

Generate protein with text prompt

Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

Joseph L. Watson[‡1,2], David Juergens[‡1,2,3], Nathaniel R. Bennett[‡1,2,3], Brian L. Trippe[‡2,4], Jason Yim[‡2,6], Helen E. Eisenach[‡1,2], Woody Ahern[‡1,2,7], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,9], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,8], Isaac Sappington[1,2,8], Susana Vázquez Torres[1,2,8], Anna Lauko[1,2,8], Valentin De Bortoli[9], Emile Mathieu[10], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[12], David Baker*[1,2,11]

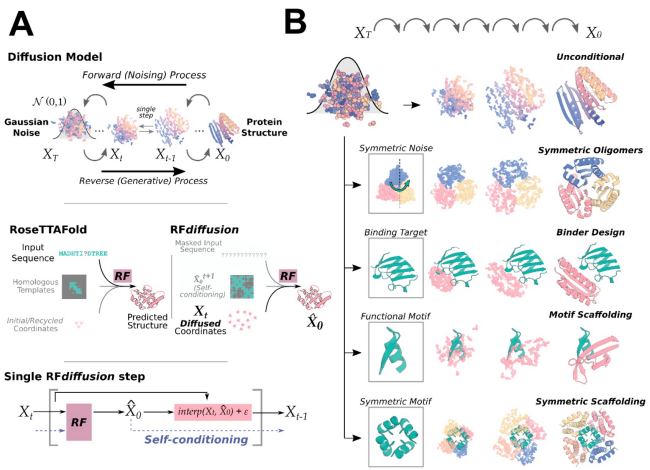


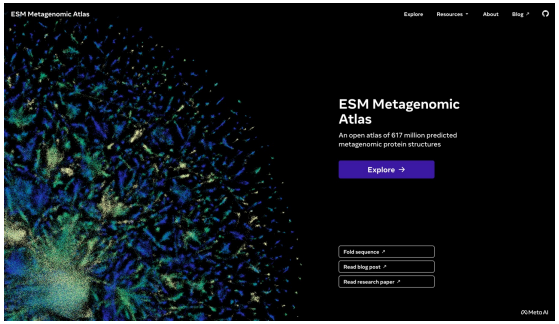
Figure 1: RF*diffusion* is a denoising diffusion probabilistic model with RoseTTAFold **fined-tuned as the denoising network. A)** Top panel: Diffusion models for proteins are trained

RoseTTaFold Diffusion: "Generate a protein that binds to X"

Evolutionary-scale prediction of atomic level protein structure with a language model

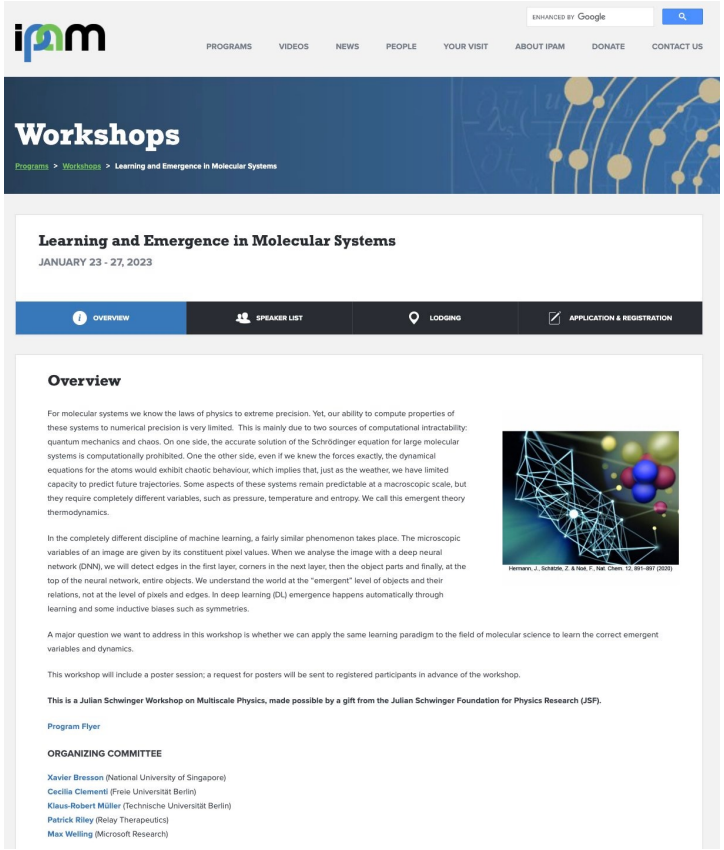
Zeming Lin[1 2 *] Halil Akin[1 *] Roshan Rao[1 *] Brian Hie[1 3 *] Zhongkai Zhu[1] Wenting Lu[1] Nikita Smetanin[1] Robert Verkuil[1] Ori Kabeli[1] Yaniv Shmueli[1] Allan dos Santos Costa[4] Maryam Fazel-Zarandi[1] Tom Sercu[1 †] Salvatore Candido[1 †] Alexander Rives[1 5 † ‡]

This is the first large-scale structural characterization of metagenomic proteins, with more than 617 million structures. The atlas reveals more than 225 million high confidence predictions, including millions whose structures are novel in comparison with experimentally determined structures, giving an unprecedented view into the vast breadth and diversity of the structures of some of the least understood proteins on earth.



Metagenomic Atlas
https://esmatlas.com

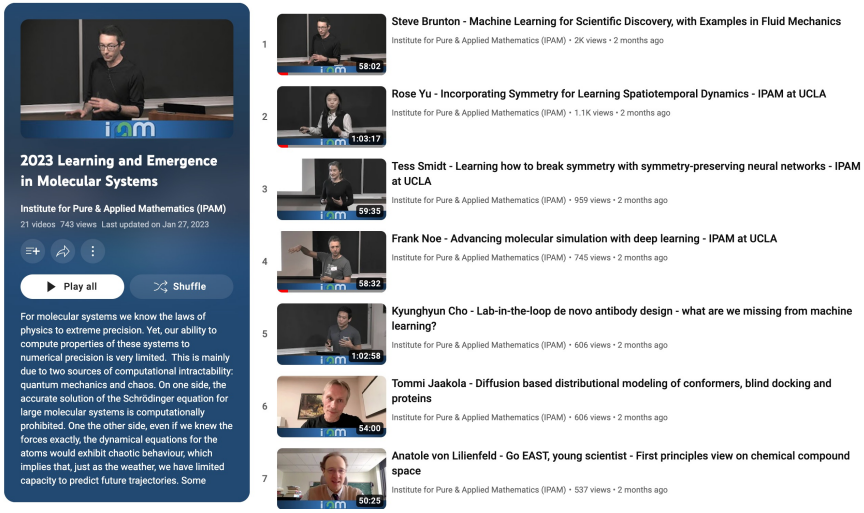
UCLA workshop on GNNs and molecular science



http://www.ipam.ucla.edu/programs/workshops/learning-and-emergence-in-molecular-systems

Speaker List

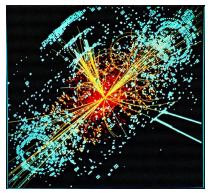
Mohammed AlQuraishi (Harvard Medical School)
Xavier Bresson (National University of Singapore)
Steve Brunton (University of Washington)
Stefan Chmiela (Technische Universität Berlin)
Kyunghyun Cho (New York University)
Cecilia Clementi (Freie Universität Berlin)
Bruno Correia (École Polytechnique Fédérale de Lausanne (EPFL))
Kyle Cranmer (University of Wisconsin-Madison)
Payel Das (IBM Research)
Ron Dror (Stanford University)
Rafael Gomez-Bombarelli (Massachusetts Institute of Technology)
Tommi Jaakkola (Massachusetts Institute of Technology)
Frank Noe (Freie Universität Berlin)
Kristin Persson (University of California, Berkeley (UC Berkeley))
Patrick Riley (Relay Therapeutics)
Jutta Rogal (New York University)
Tess Smidt (Massachusetts Institute of Technology)
Alexandre Tkatchenko (University of Luxembourg)
Mark Tuckerman (New York University)
Anatole von Lilienfeld (University of Toronto)
Max Welling (Microsoft Research)
Rose Yu (University of California, San Diego (UCSD))
Maxim Zlatdinov (Oak Ridge National Laboratory)



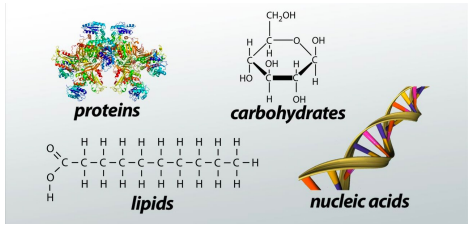
https://www.youtube.com/playlist?list=PLHyI3Fbmv0SeEiaFaO2XHMUHdZnZZiqa8

GNNs for science/biology

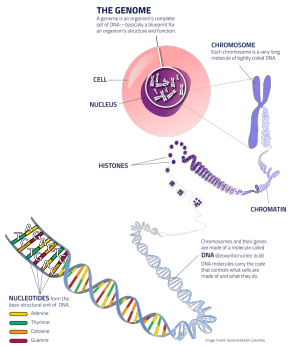
Time



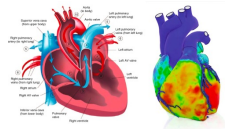
Particles
Physics



Molecules
Chemistry



Genome
Genetics



Body
Medicine

GNNs will increase its impact in science by replacing human engineered heuristics with learned features, replacing combinatorial objects with continuous vectors & accelerating estimation of system properties.

Partially reproduced from M. Welling

Space

Outline

- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Molecule generation

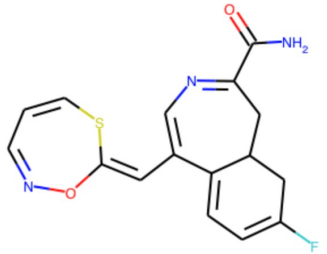
- Goal is to design a neural network that can
  - Auto-encode molecules
  - Generate novel molecules
  - Produce molecules with optimized chemical property

A Two-Step Graph Convolutional Decoder for Molecule Generation

Xavier Bresson
School of Computer Science and Engineering
NTU, Singapore
xbresson@ntu.edu.sg

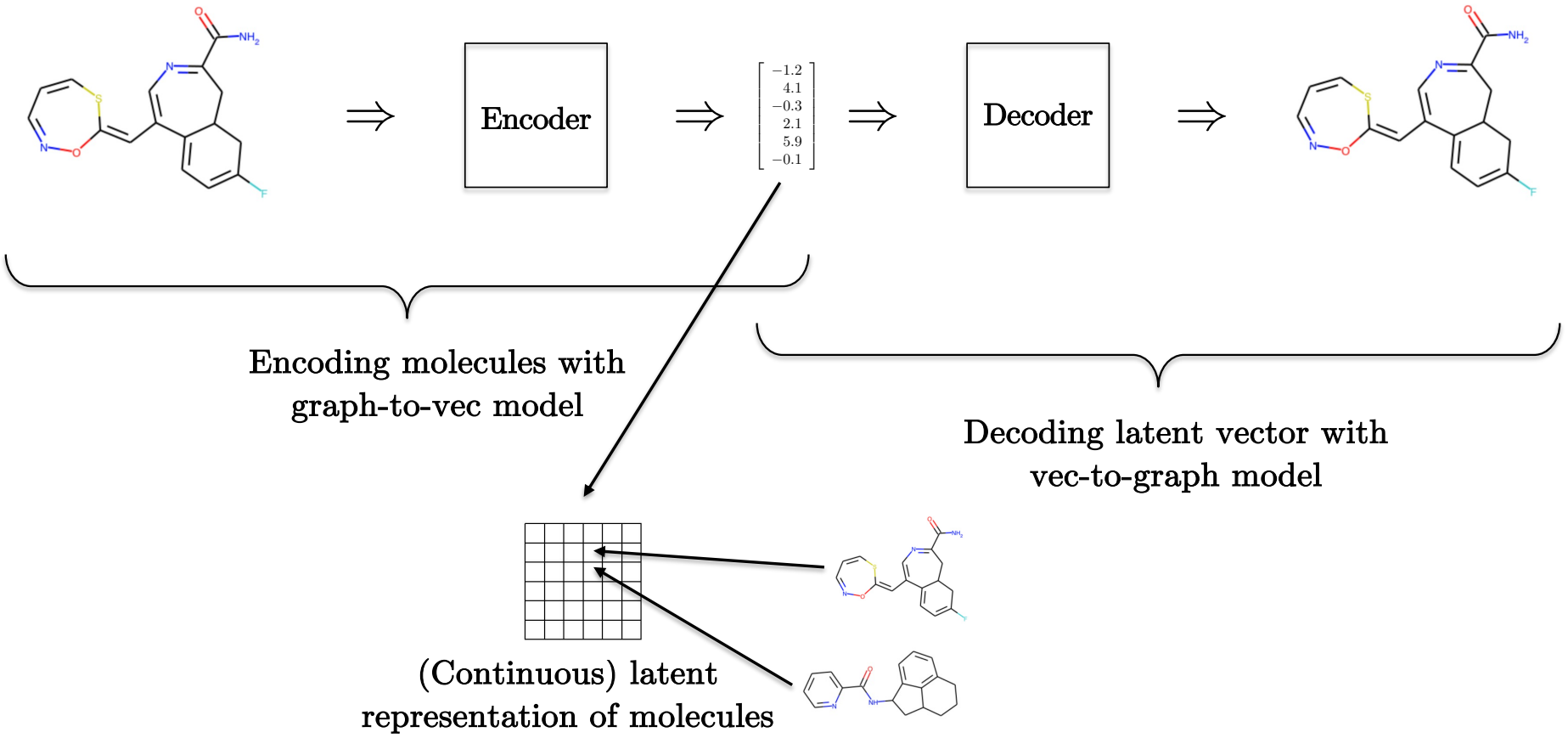
Thomas Laurent
Department of Mathematics
Loyola Marymount University
tlaurent@lmu.edu

ArXiv : https://arxiv.org/pdf/1906.03412.pdf



Graph auto-encoder

- Graph-to-graph Model :



Outline

- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Graph encoder

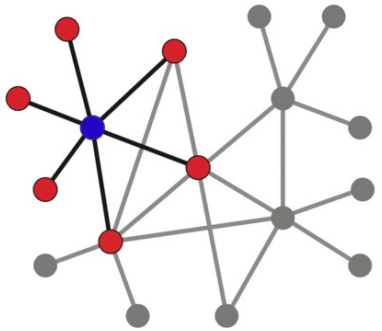
- GNNs have been used to encode molecules into a continuous vectorial space.
  - GNNs used for regression[1,2] to predict molecular properties (1-2 orders of magnitude faster than solving Schrodinger equation w/ DFT).

Any favorite GNN s.a. Graph ConvNets, Graph Attention Nets, Graph RNNs

$$h_i = f_{\text{node}}(\{h_j\}, j \in \mathcal{N}(i)) \in \mathbb{R}^d$$

$$z = g_{\text{graph}}(\{h_i\}, i \in V) \in \mathbb{R}^d$$

Reduce function : Sum or Mean

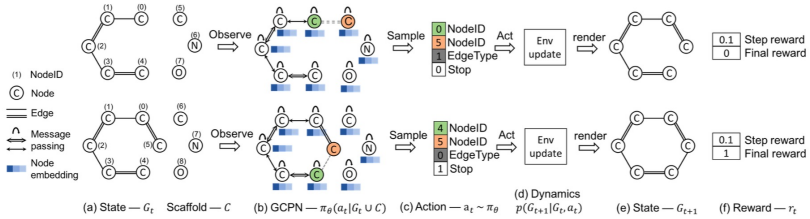


$$\Rightarrow \quad z = \begin{bmatrix} -1.2 \\ 4.1 \\ -0.3 \\ 2.1 \\ 5.9 \\ -0.1 \end{bmatrix} \in \mathbb{R}^d$$

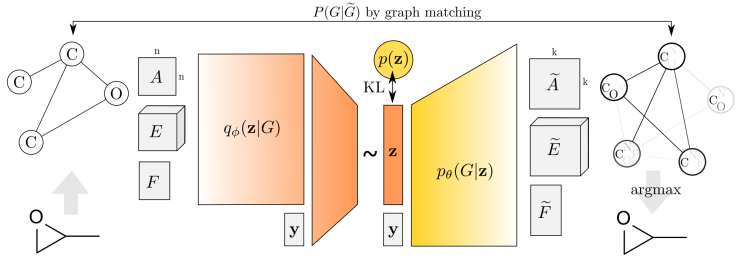
[1] Duvenaud, Maclaurin, Iparraguirre, Bombarell, Hirzel, Aspuru-Guzik, Adams, Convolutional networks on graphs for learning molecular fingerprints, 2015
[2] Gilmer, Schoenholz, Riley, Vinyals, Dahl, Neural message passing for quantum chemistry, 2017

Decoder for graph generation

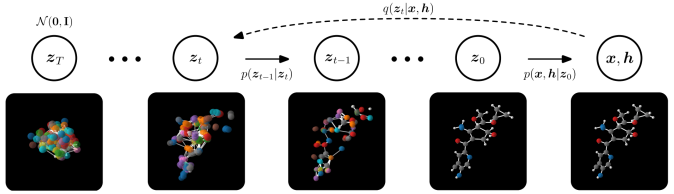
- Encoding is easy. Decoding is the challenging part!
- Three approaches :
  - Step-by-step auto-regressive models : Sequential generation of molecules (atom-by-atom)
    - Jin-et.al, 2018[1], You-Leskovec-et.al, 2018[2], etc
  - One-shot models : Generation of all atoms and bonds in a single pass
    - Simonovsky, Komodakis, 2018[3], De Cao, Kipf, 2018[4], etc
  - Step-by-step generative models : Sequential denoising of molecules
    - Hoogeboom-et.al, 2022[5], Xu-et.al, 2022[6]



You-Leskovec-et.al, 2018



Simonovsky, Komodakis, 2018

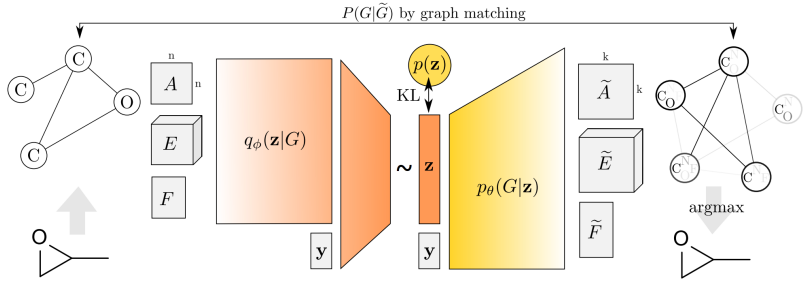


Hoogeboom-et.al, 2022

[1] Jin, Barzilay, Jaakkola, Junction Tree Variational Autoencoder for Molecular Graph Generation, 2018
[2] You, Liu, Ying, Pande, Leskovec, Graph convolutional policy network for goal-directed molecular graph generation, 2018
[2] Simonovsky, Komodakis, GraphVAE: Towards generation of small graphs using variational autoencoders, 2018
[4] De Cao, Kipf, MolGAN: An implicit generative model for small molecular graphs, 2018
[5] Hoogeboom, Satorras, Vignac, Welling, Equivariant diffusion for molecule generation in 3d, 2022
[6] Xu, Yu, Song, Shi, Ermon, Tang, Geodiff: A geometric diffusion model for molecular conformation generation, 2022

One-shot decoder

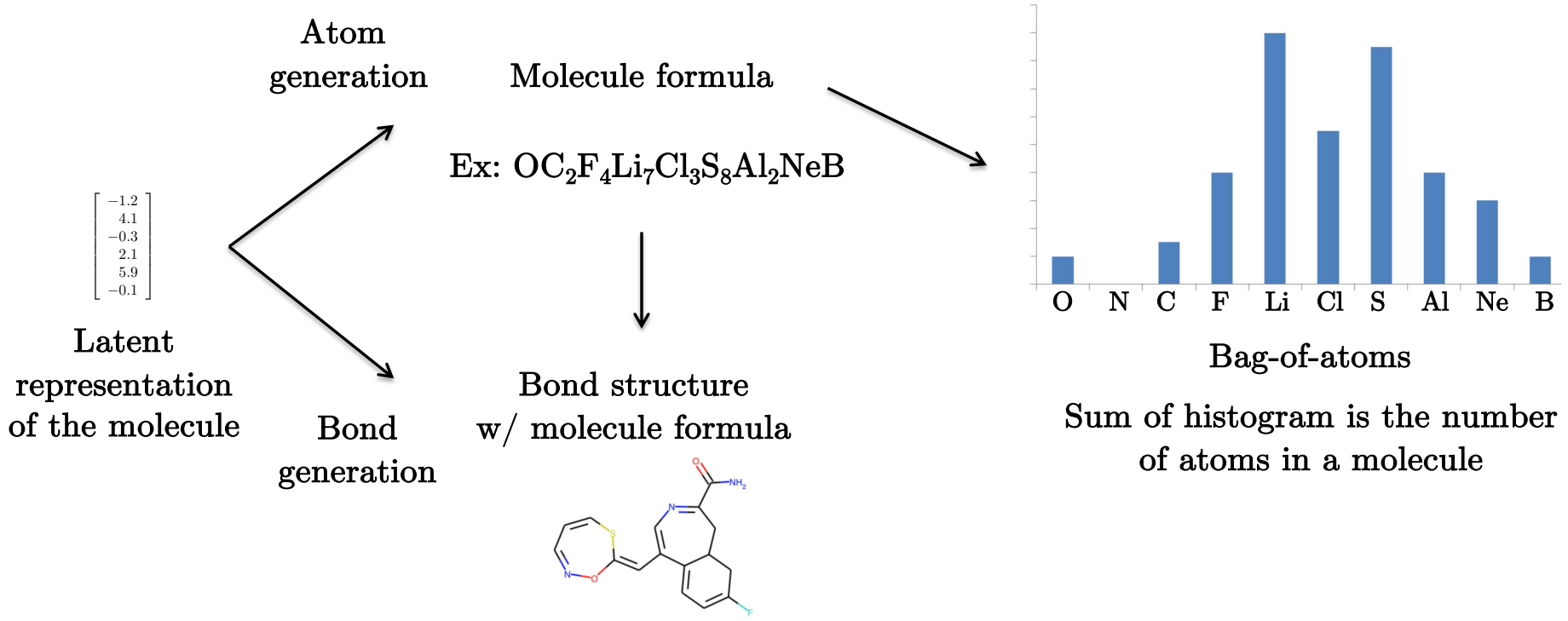
- A challenge with one-shot decoder is to generate molecules of different sizes.
  - It is hard to generate simultaneously
    - The number of atoms
    - The type of atoms
    - The bond structures between the atoms
  - Techniques[1,2] generated molecules with a fixed size (the size of the largest molecule).



[1] Simonovsky, Komodakis, GraphVAE: Towards generation of small graphs using variational autoencoders, 2018
[2] De Cao, Kipf, MolGAN: An implicit generative model for small molecular graphs, 2018

Proposed decoder

- We propose to disentangle these 3 problems :

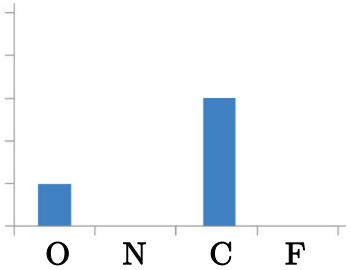


Atom decoder

- We decode the latent representation of the molecule with a Multi-Layer Perceptron (MLP) to produce the histogram over the atoms in a molecule :

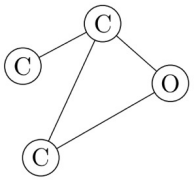
$$\begin{bmatrix} -1.2 \\ 4.1 \\ -0.3 \\ 2.1 \\ 5.9 \\ -0.1 \end{bmatrix}$$

MLP



Latent representation of the molecule

Bag-of-atoms

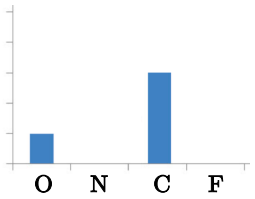


OC_3

Bond decoder

- The "IKEA" model :
  - The bag-of-atoms indicates what atoms are in the molecule (IKEA pieces),
  - The atoms are assembled with a GNN (IKEA assembly instructions).

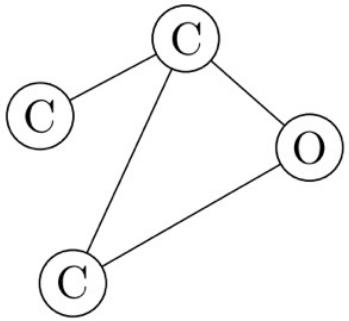
Bag-of-atoms
OC_3



Latent
representation
of the molecule

$$\begin{bmatrix} -1.2 \\ 4.1 \\ -0.3 \\ 2.1 \\ 5.9 \\ -0.1 \end{bmatrix}$$

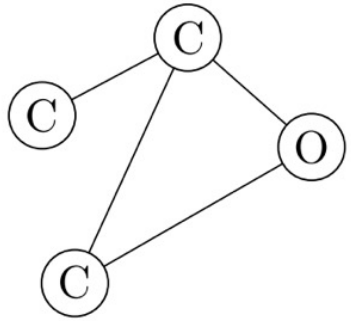
Graph
Net

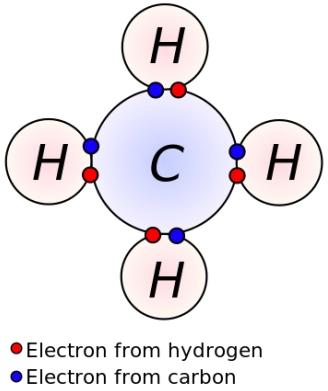


Decoded
molecule

Beam search for chemical valency

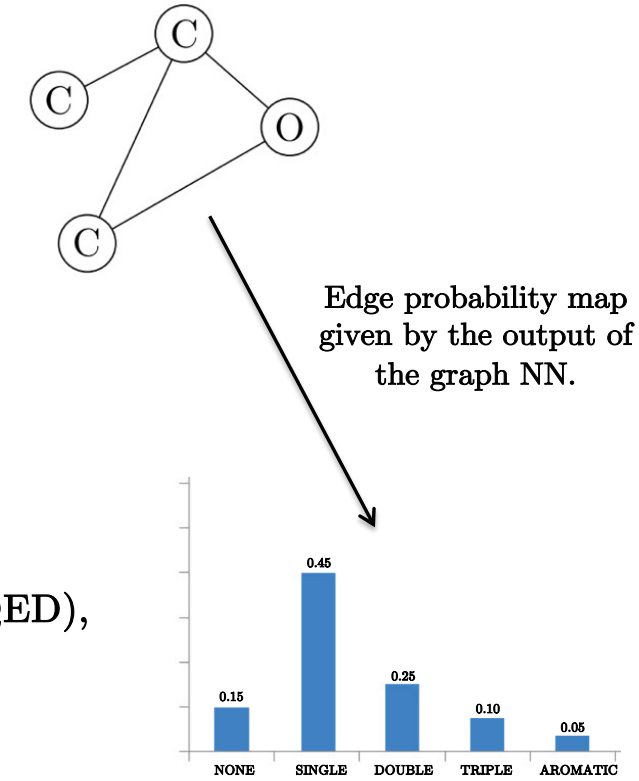
- The one-shot model may not produce a chemically valid molecule.
  - Violation of atom valency (maximum number of electrons in the outer shell of the atom that can participate of a chemical bond).
- We use beam search to produce valid molecules.





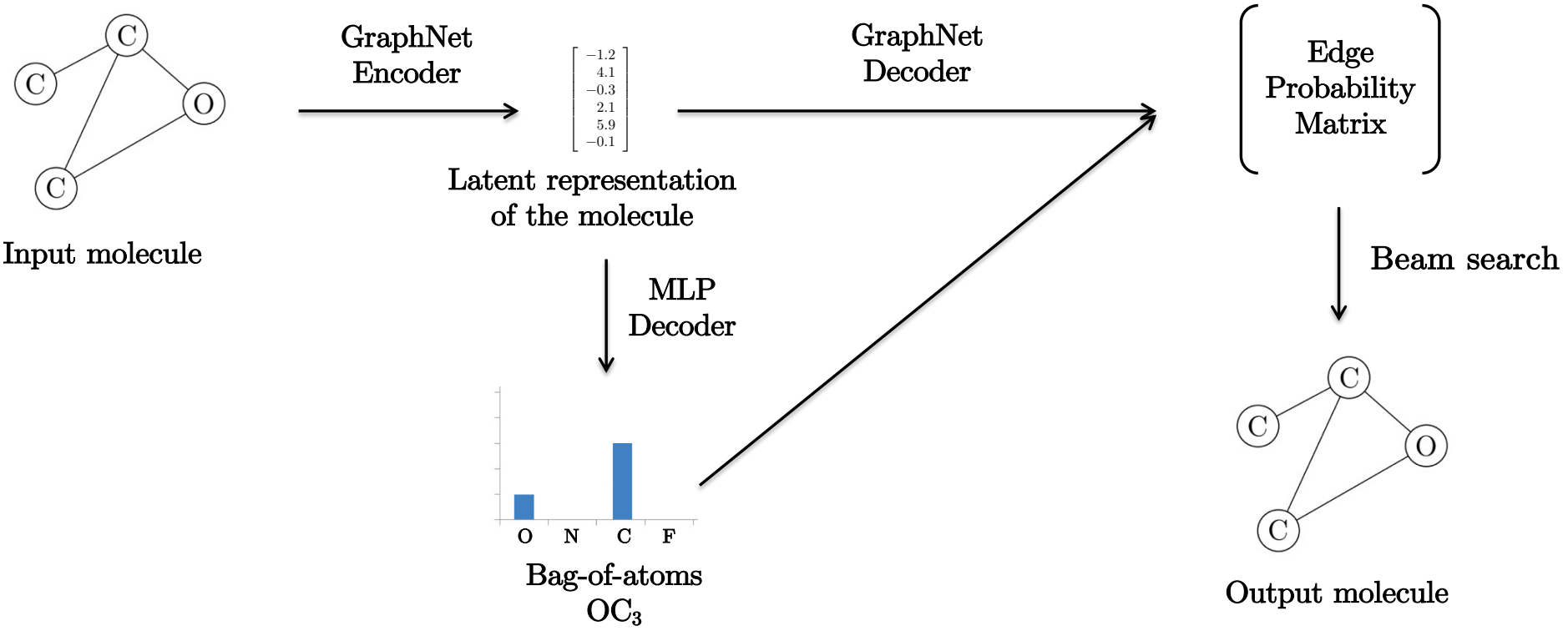
Beam search for chemical valency

- Beam search :
  - Start with a random edge.
    - Select the next edges that
      - have the largest probability (or Bernouilli sampling),
      - are connected to selected edges,
      - do not violate valency.
  - Repeat for a number of different initializations.
  - Select the molecule that maximizes
    - The product of edge probabilities or,
    - The chemical property to be optimized s.a. druglikeness (QED), constrained solubility (logP), etc.



Proposed auto-encoder

- Molecular auto-encoder system :

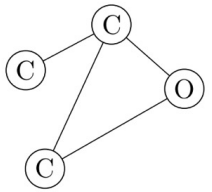


Outline

- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Encoder description

- We use graph ConvNet[1] :



Input molecule

Graph ConvNet Encoder

$$\begin{bmatrix} -1.2 \\ 4.1 \\ -0.3 \\ 2.1 \\ 5.9 \\ -0.1 \end{bmatrix}$$

Latent representation z of the molecule

Node and edge representations

$$\begin{cases} h_i^{\ell+1} = h_i^\ell + \mathrm{ReLU}\Big(\mathrm{BN}\Big(W_1^\ell h_i^\ell + \sum_{j\sim i} \eta_{ij}^\ell \odot W_2^\ell h_j^\ell\Big)\Big) \\ e_{ij}^{\ell+1} = e_{ij}^\ell + \mathrm{ReLU}\Big(\mathrm{BN}\Big(V_1^\ell e_{ij}^\ell + V_2^\ell h_i^\ell + V_3^\ell h_j^\ell\Big)\Big) \end{cases} \quad \text{with} \quad \eta_{ij}^\ell = \frac{\sigma(e_{ij}^\ell)}{\sum_{j'\sim i}\sigma(e_{ij'}^\ell) + \varepsilon}$$

Dense attention

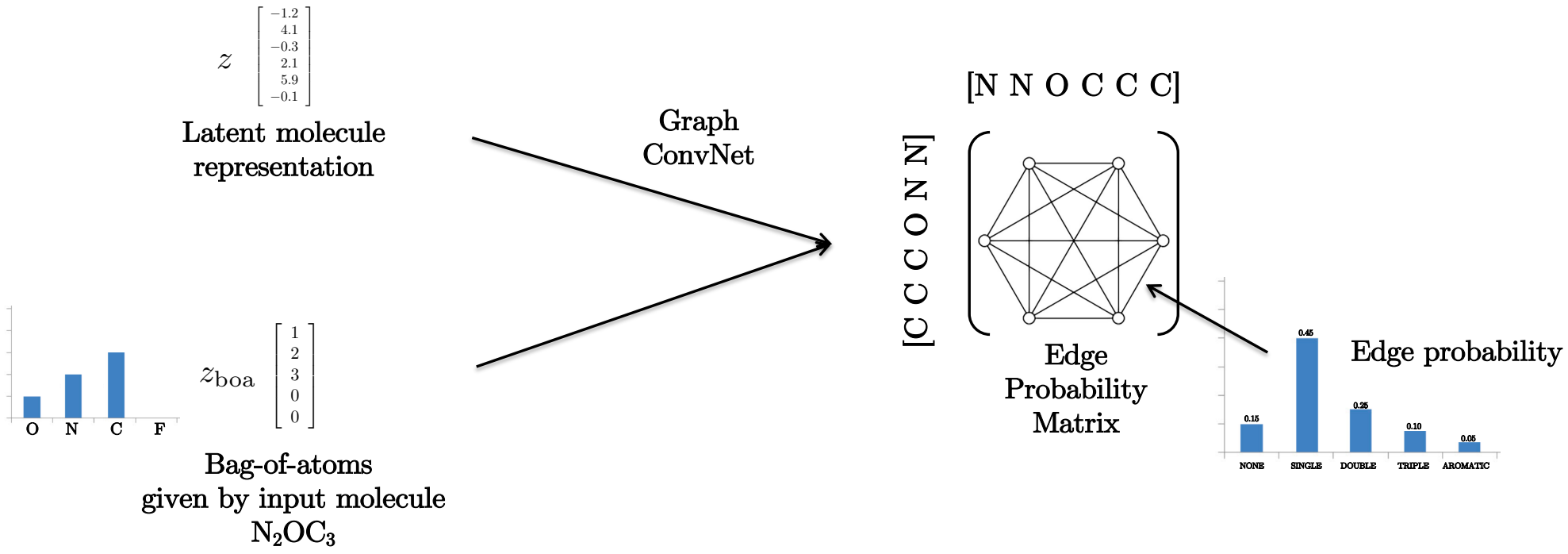
Graph representation

$$z = \sum_{i,j=1}^{N} \sigma\left(Ae_{ij}^L + Bh_i^L + Ch_j^L\right) \odot We_{ij}^L$$

[1] Bresson, Laurent, Residual gated graph convnets, 2017

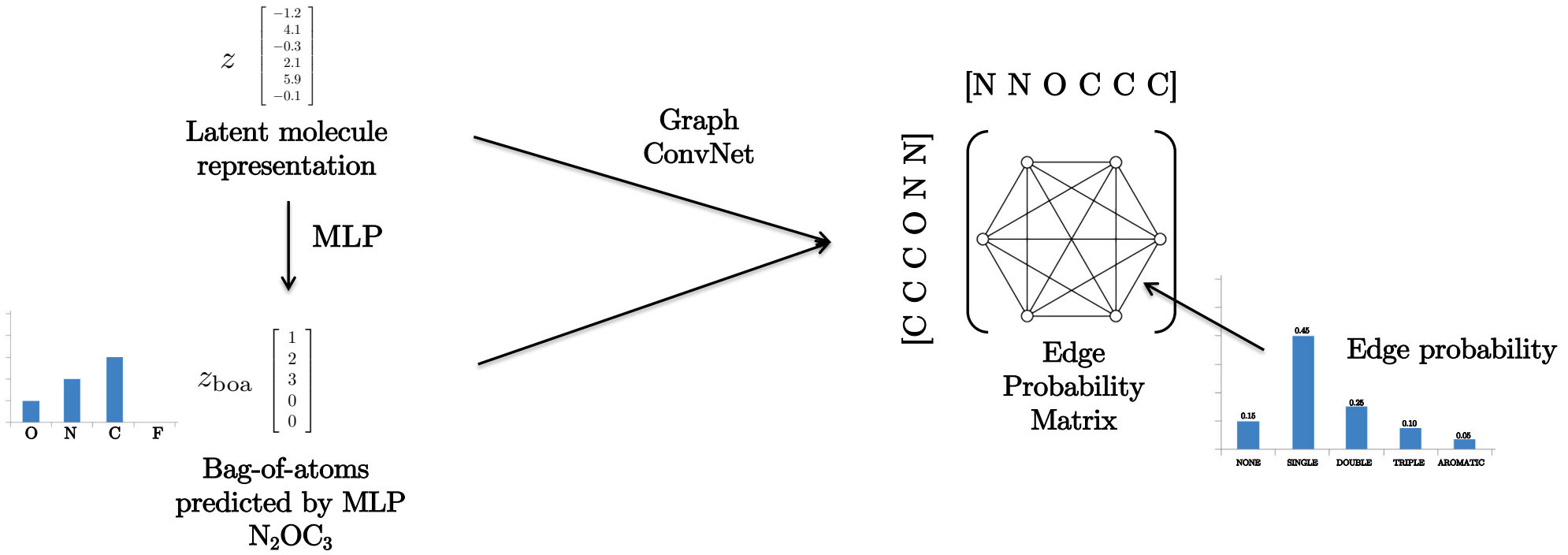
Bond decoding during training

- Given the latent encoding z of the molecule and the bag-of-atoms z_{boa}, we use another graph ConvNet to decode the bonds between the atoms :



Bond decoding at test time

- The bag-of-atoms of the input molecule is predicted by a MLP :

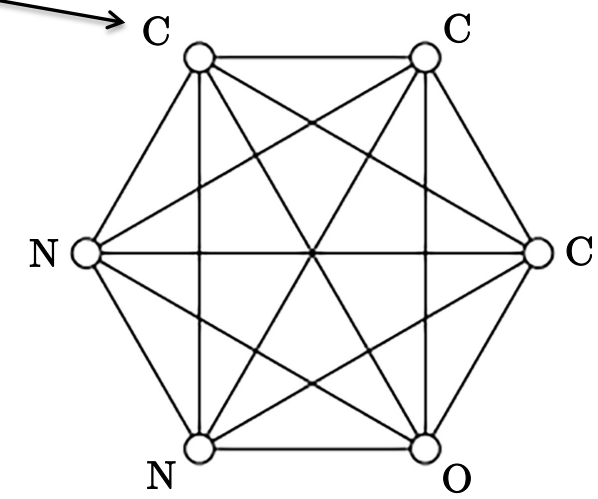


Breaking atom symmetry

- The bond decoder starts with a fully connected graph with the atom type z_{ato} on each node.
- This is not enough for the GNN to be able to differentiate the 3 atoms of Carbon and the 2 atoms of Nitrogen!
  - We break this symmetry by introducing positional features z_{pos} , which will differentiate several atoms of the same type.
  - We concatenate this positional feature with the atom type z_{ato} to form the input node feature of the decoder.

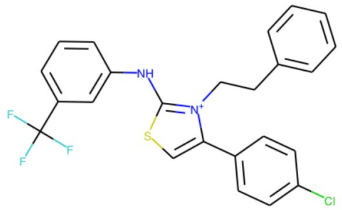
$$h_i^{\ell=0} = \begin{bmatrix} z_{\mathrm{ato}\,i} \\ z_{\mathrm{pos}\,i} \end{bmatrix}$$

$$h_i^{\ell=0} = \left[\; z_{\mathrm{ato}\,i} \; \right]$$



Atom positional encoding

- We need to order the atoms.
- We use the SMILE[1] representation of molecules to order the atoms.
  - A SMILE is a sequence of string characters that encodes atoms and bonds of a molecule.

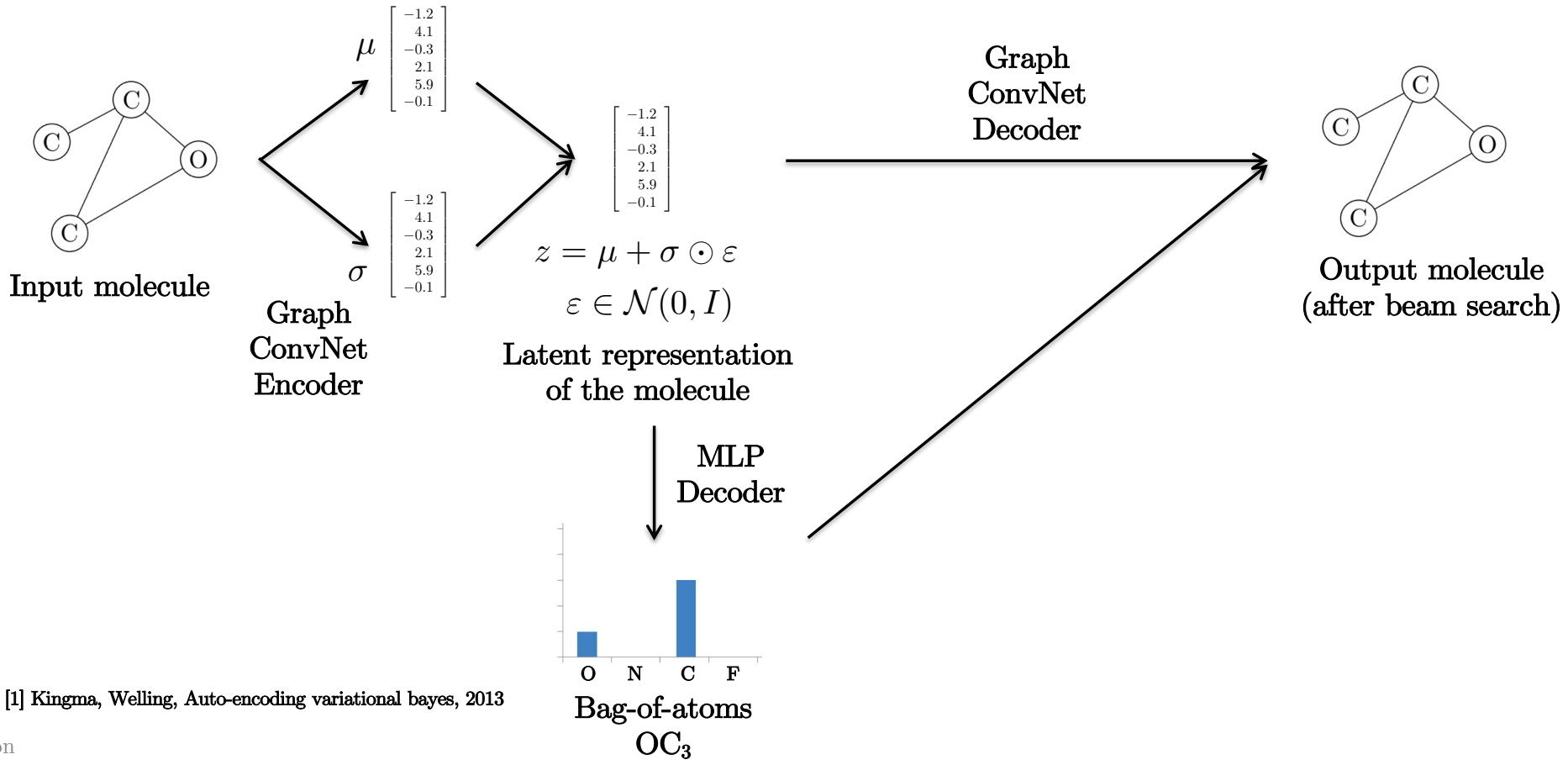
SMILE () = FC(F)(F)C1=CC=CC(NC2=[N+](CCC3=CC=CC=C3)...
...C(C3=CC=C(Cl)C=C3)=CS2)=C1

(F, 1) (C, 1) (F, 2) (F, 3) (C, 2) . . .

[1] Weininger, SMILES, a chemical language and information system, 1988

Variational auto-encoder

- We use the VAE formulation[1] to improve the molecule generation "by filling the latent space" :



[1] Kingma, Welling, Auto-encoding variational bayes, 2013

Molecular generative loss

- The generative loss is composed of
  - Cross-entropy loss for edge probability,
  - Cross-entropy loss for bag-of-atoms probability,
  - Kullback–Leibler divergence for the VAE Gaussian distribution.

$$L = \lambda_e \sum_e \hat{p}_e \log p_e + \lambda_a \sum_a \hat{p}_a \log p_a - \frac{\lambda_{vae}}{2} \sum_k (1 + \log \sigma_k^2 - \mu_k^2 - \sigma_k^2)$$

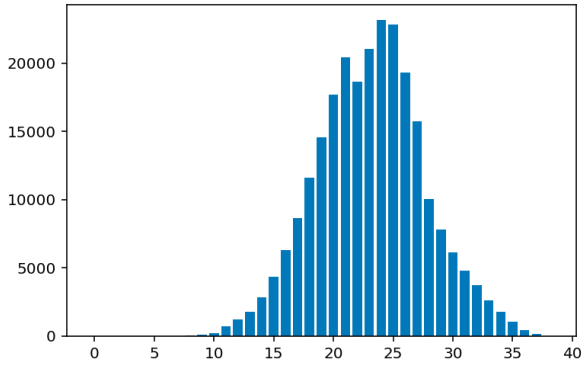
- No matching process is necessary between input and output molecules because the same atom ordering is used (with the SMILE representation).

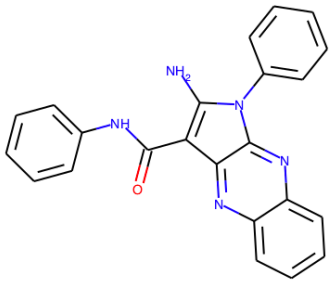
Outline

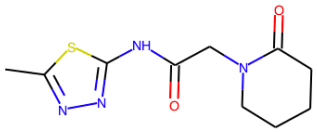
- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

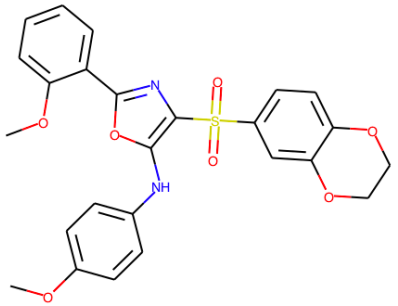
Dataset

- ZINC dataset
  - 250k drug-like molecules
  - Up to 38 heavy atoms (excluded hydrogen)









Training

- Training process
  - Mini-batch of 50 molecules
  - Learning rate is decreased by 1.25 after each epoch if training loss does not decrease by 1%.
  - Learning stops when LR is less than 10^{-6}.
  - Training takes 28 hours on a single Nvidia 1080Ti GPU.

Numerical experiments

- Molecule reconstruction
  - How many molecules are correctly decoded?
- Molecule novelty
  - Beyond memorization – how many molecules sampled from the learned distribution are not in the training set?
- Molecule optimization
  - How much property improvement can we obtain when optimizing in the latent space?
  - The chemical property is here the constrained solubility of molecules.

Baseline techniques

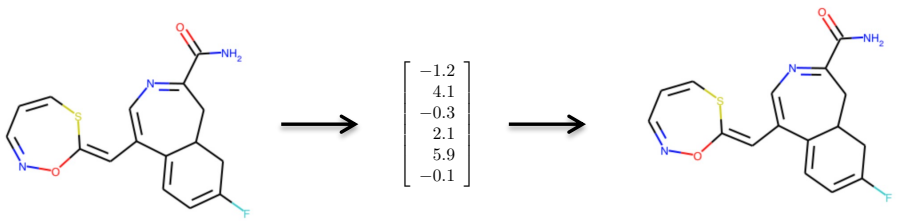
- VAE + SL + AR :
  - JT-VAE : Jin, Barzilay, Jaakkola, Junction Tree Variational Autoencoder for Molecular Graph Generation, 2018

- GAN + RL + AR :
  - GCPN : You, Liu, Ying, Pande, Leskovec, Graph convolutional policy network for goal-directed molecular graph generation, 2018

Molecule reconstruction

Method	Reconstruction	Validity
CVAE [Gomez-Bombarelli et al., 2016]	44.6%	0.7%
GVAE [Kusner et al., 2017]	53.7%	7.2%
SD-VAE [Dai et al, 2018]	76.2%	43.5%
GraphVAE [Simonovsky, Komodakis, 2018]	-	13.5%
JT-VAE (SL) [Jin et al, 2018]	76.7%	**100.0%**
GCPN (GAN+RL) [You et al, 2018]	-	-
OURS (VAE+SL)	**90.5%**	**100.0%**

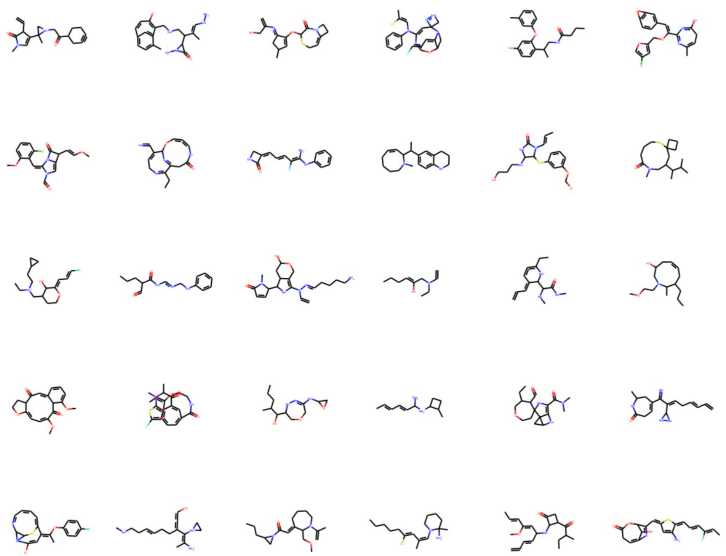
Table 1: Percentage of successful reconstruction of 250k ZINC molecules.



Molecule novelty

Method	Novelty	Uniqueness
JT-VAE (SL) [Jin et al, 2018]	**100.0%**	**100.0%**
GCPN (GAN+RL) [You et al, 2018]	-	-
OURS (VAE+SL)	**100.0%**	**100.0%**

Table 2: Sample 5000 molecules from learned prior distribution.



Molecule optimization task #1

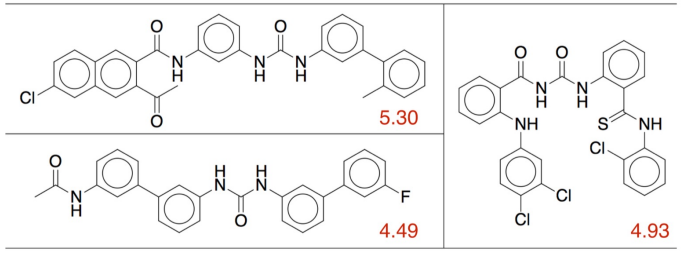
- Molecule optimization :
  - Goal is to maximize the constrained solubility of the training molecules.
  - Optimization is done by gradient ascent in the latent space of molecules.
  - Following JT-VAE, we report the top 3 optimized molecules :

Method	1st	2nd	3rd	Mean
ZINC	4.52	4.30	4.23	4.35
CVAE, Gómez-Bombarelli et al. [2018]	1.98	1.42	1.19	1.53
GVAE, Kusner et al. [2017]	2.94	2.89	2.80	2.87
SD-VAE, Dai et al. [2018]	4.04	3.50	2.96	3.50
JT-VAE, Jin et al. [2018]	5.30	4.93	4.49	4.90
OURS (VAE+SL)	5.24	5.10	5.06	5.14
GCPN (GAN+RL), You et al. [2018]	**7.98**	**7.85**	**7.80**	**7.88**

Molecule optimization task #1

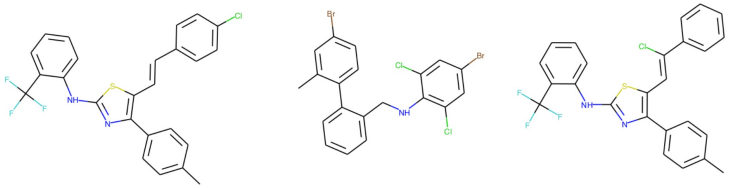
- Top 3 optimized molecules :

JT-VAE (VAE+SL)



Mean is 4.90

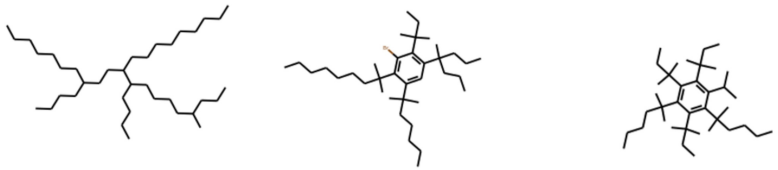
OURS (VAE+SL)



5.24 5.14 5.06

Mean is 5.14

GCPN (GAN+RL)



7.98 7.48 7.12

Mean is 7.52

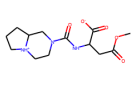
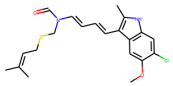
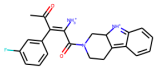
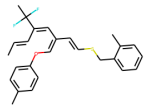
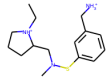
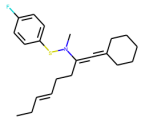
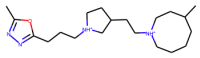
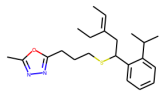
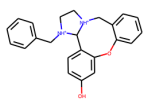
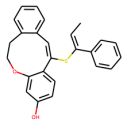
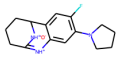
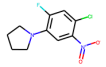
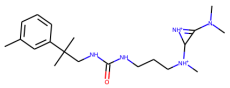
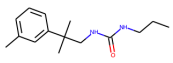
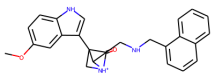
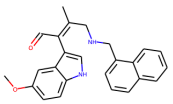
Molecule optimization task #2

- Constrained optimization :
  - Goal is to maximize the constrained solubility of the 800 test molecules with the lowest value.
  - The optimization of the chemical property is constrained by the similarity between the original molecule and the new generated molecule.
  - Following JT-VAE, we report property improvements w.r.t. molecule similarity δ :

	JT-VAE [Jin et al, 2018] (SL)			GCPN [You et al, 2018] (GAN+RL)			OURS (VAE+SL)		
δ	Improvement	Similarity	Success	Improvement	Similarity	Success	Improvement	Similarity	Success
0.0	1.91 ± 2.04	0.28 ± 0.15	97.5%	4.20 ± 1.28	**0.32 ± 0.12**	**100.0%**	**5.24 ± 1.55**	0.18 ± 0.12	**100.0%**
0.2	1.68 ± 1.85	0.33 ± 0.13	97.1%	4.12 ± 1.19	**0.34 ± 0.11**	**100.0%**	**4.29 ± 1.57**	0.31 ± 0.12	98.6%
0.4	0.84 ± 1.45	**0.51 ± 0.10**	83.6%	2.49 ± 1.30	0.47 ± 0.08	**100.0%**	**3.05 ± 1.46**	**0.51 ± 0.10**	84.0%
0.6	0.21 ± 0.71	**0.69 ± 0.06**	46.4%	0.79 ± 0.63	0.68 ± 0.08	**100.0%**	**2.46 ± 1.27**	0.67 ± 0.05	40.1%

Table 7: Molecule optimization results.

Molecule optimization task #2

Molecule similarity 0.0				
	4: -8.38	4: 2.19	77: -5.81	77: 4.75
Molecule similarity 0.2				
	143: -5.01	143: 3.54	136: -5.06	136: 3.10
Molecule similarity 0.4				
	604: -2.94	604: 3.39	103: -5.40	103: 0.88
Molecule similarity 0.6				
	89: -5.64	89: 0.94	782: -2.57	782: 2.44

Outline

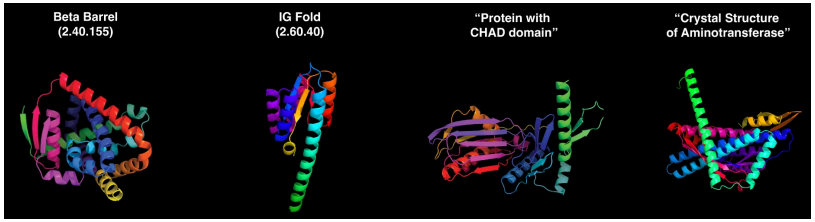
- Deep learning for molecular science
- Graph generative model with VAE
  - Auto-encoder
  - Encoder and decoder
  - System description
  - Numerical experiments
- Conclusion

Conclusion

- We propose an efficient VAE for atoms and bonds decoding.
  - Single-shot reconstruction + beam search
  - Simple/fast (GPU parallelizable)
  - Alternative to step-by-step techniques
- We report highest VAE accuracy on ZINC dataset for
  - Molecule reconstruction
  - Molecule optimization of constrained solubility property
- Comparing VAE+SL vs GAN+RL :
  - GAN+RL generates better molecules (outside the training statistics)
  - VAE+SL generates better optimized molecules similar to the original ones
  - GAN+RL generates optimized molecules with 100% success
- Limitation
  - Large molecular generation

Conclusion

- Graph generative models are still in their early development.
- Like for image/text, there is a huge potential to develop pre-trained graph molecular networks from all available datasets and then prompt the generation of new molecules for
  - Drug-like molecules, ligands, proteins, RNAs with optimized chemical, biological, potency, toxicity, pharmacokinetics properties
  - Material science with optimized mechanical, thermal, electrical, optical, chemical and environmental properties



Prompt: Generate a protein with CHAD domain

The end / selfie time


strymon