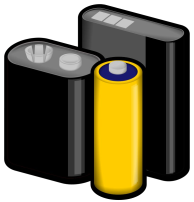
Inverse design of 3d molecular structures with conditional generative neural networks

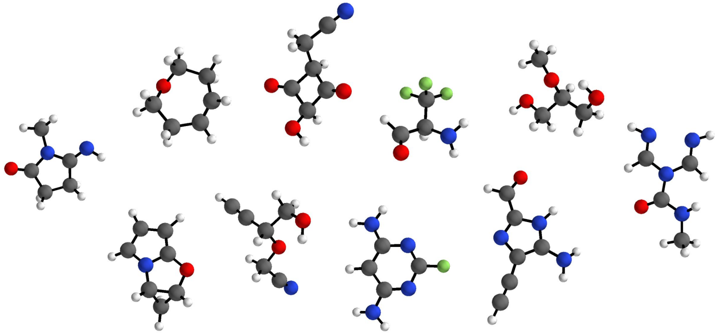
Niklas Gebauer

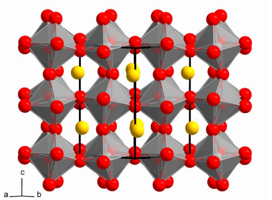
Technische Universität Berlin - Machine Learning Group

27th of February 2024

Quantum chemistry/Physics

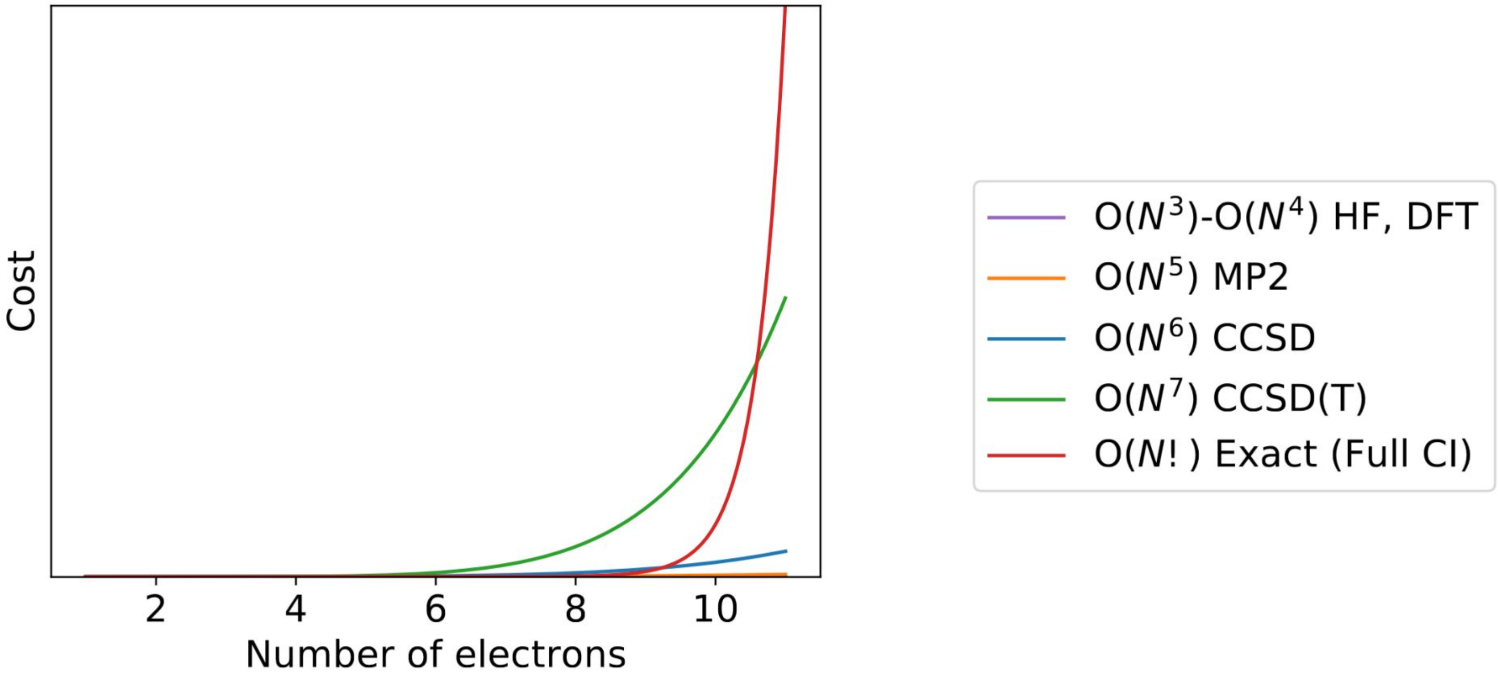




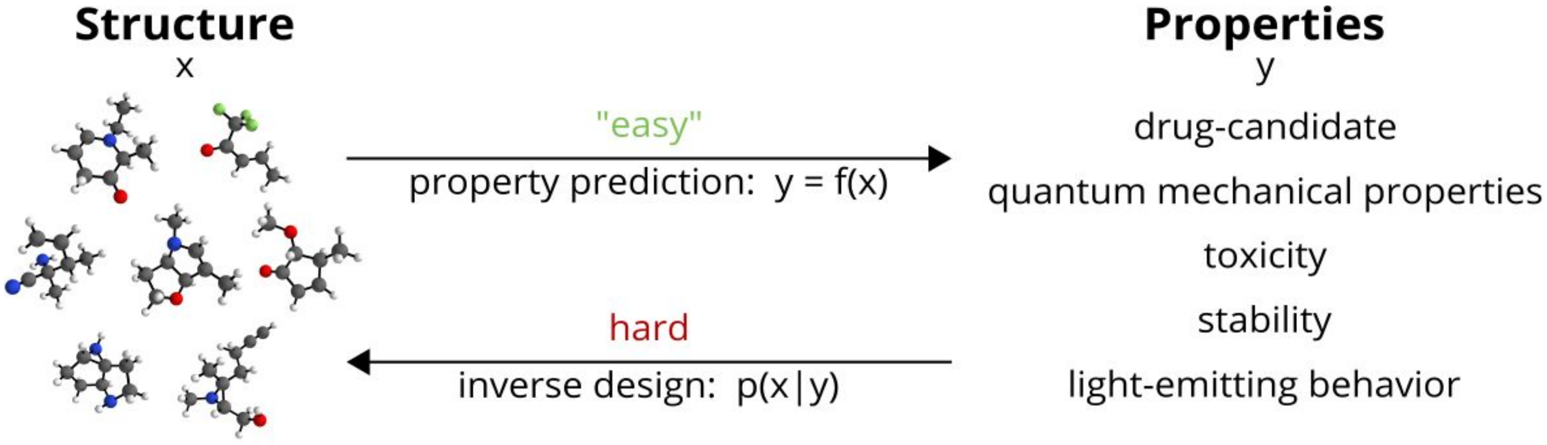


https://commons.wikimedia.org/wiki/
File:CaTiO3_perovskite_structure.png

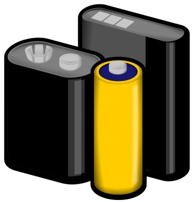
Computational complexity of QC



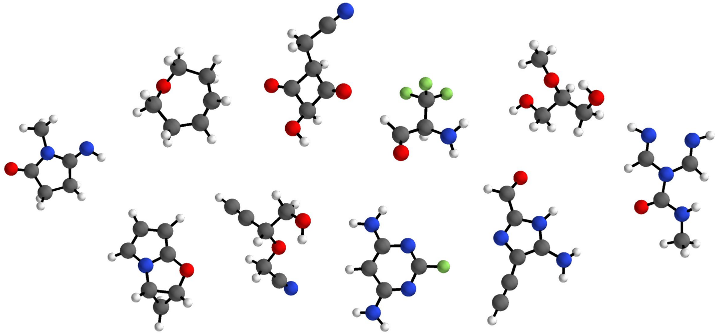
ML for molecules

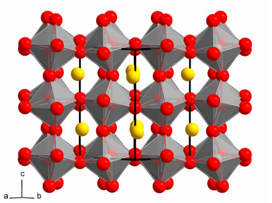


Quantum chemistry/Physics



<u>guided</u> exploration of chemical space





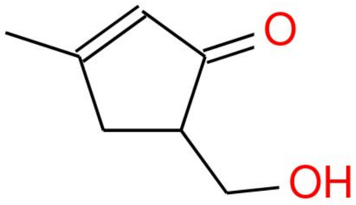
https://commons.wikimedia.org/wiki/File:CaTiO3_perovskite_structure.png

Molecule representations

SMILES strings
[Weininger 1988]

CC1=CC(=O)C(CO)C1

molecular graph



spatial

atom positions

$r_1=(x,y,z)$ $r_2=(x,y,z)$ … $r_{19}=(x,y,z)$

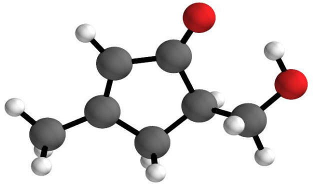
atom types

$Z_1=O$ $Z_2=C$ … $Z_{19}=O$

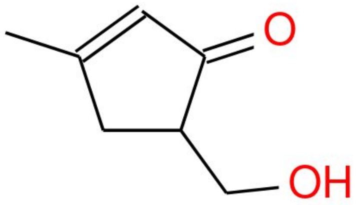
Molecule representations

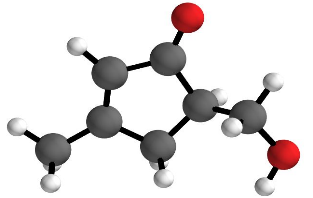
SMILES strings
[Weininger 1988]

CC1=CC(=O)C(CO)C1



molecular graph





spatial

atom positions

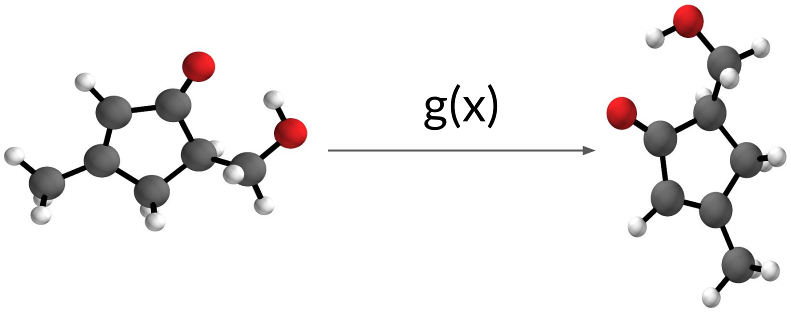
$r_1=(x,y,z)$ $r_2=(x,y,z)$ … $r_{19}=(x,y,z)$

atom types

$Z_1=O$ $Z_2=C$ … $Z_{19}=O$

Neural network for spatial representation

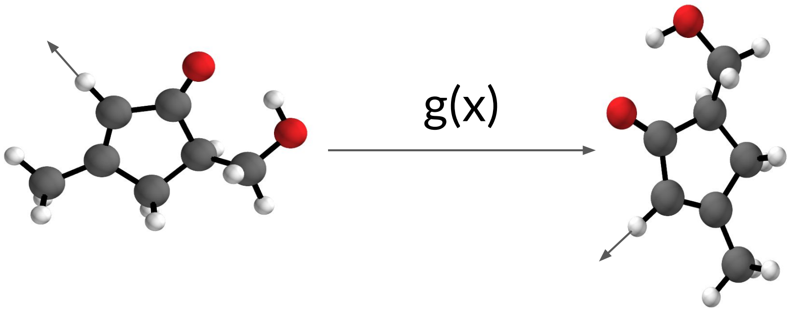
- invariant networks:
  - f(g(x)) = f(x)
  - features based on pairwise distances



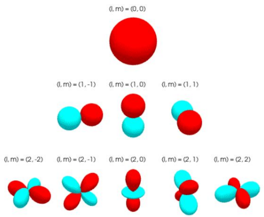
C O H

Neural network for spatial representation

- invariant networks:
  - f(g(x)) = f(x)
  - features based on pairwise distances



- equivariant networks:
  - f(g(x)) = g(f(x))
  - features based on direction vectors
  - specific layers required
  - spherical harmonics



https://commons.wikimedia.org/wiki/File:Real_Spherical_Harmonics_Figure_Table_Complex_Radial_Magnitude.gif

C O H

Approaches for 3d molecule generation

iterative

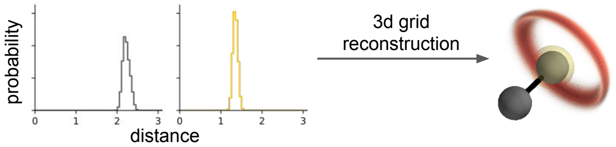
one-shot

Approaches for 3d molecule generation

iterative

one-shot

reconstruction from pairwise distances

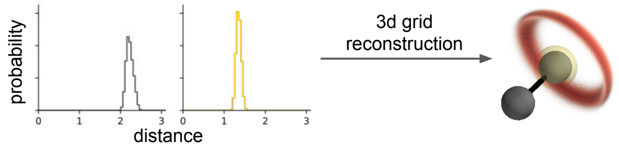


[Gebauer et al., 2019]

Approaches for 3d molecule generation

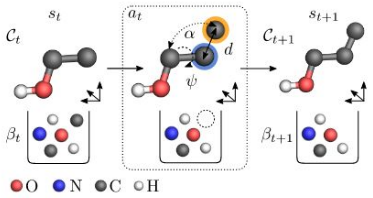
iterative

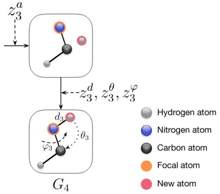
reconstruction from pairwise distances



[Gebauer et al., 2019]

reconstruction from internal coordinates





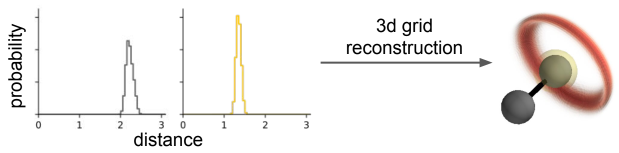
[Simm et al., 2020] [Luo & Ji, 2022]

one-shot

Approaches for 3d molecule generation

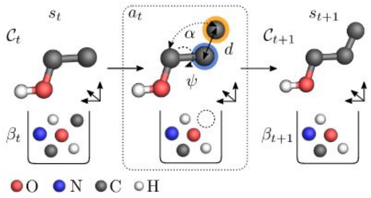
iterative

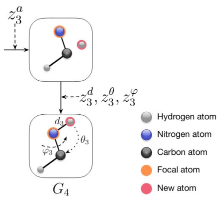
reconstruction from pairwise distances



[Gebauer et al., 2019]

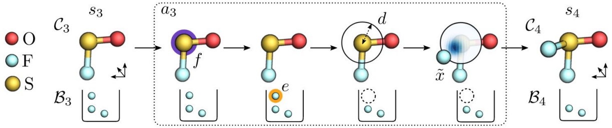
reconstruction from internal coordinates





[Simm et al., 2020] [Luo & Ji, 2022]

equivariant prediction of direction (spherical harmonics)



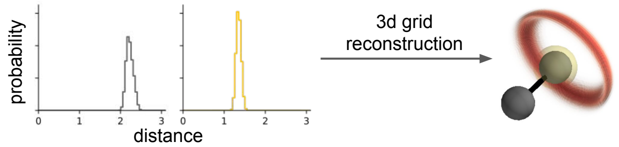
[Simm et al., 2021]

one-shot

Approaches for 3d molecule generation

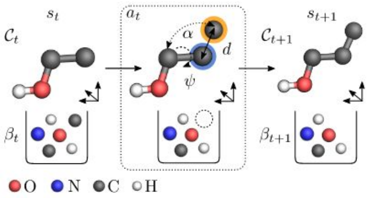
iterative

reconstruction from pairwise distances

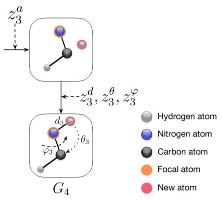


[Gebauer et al., 2019]

reconstruction from internal coordinates

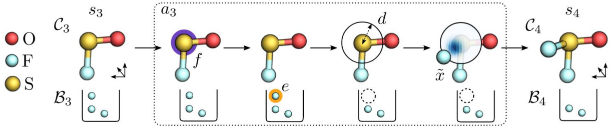


[Simm et al., 2020]



[Luo & Ji, 2022]

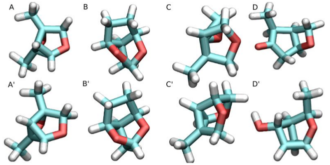
equivariant prediction of direction (spherical harmonics)



[Simm et al., 2021]

one-shot

generating valid euclidean distance matrices

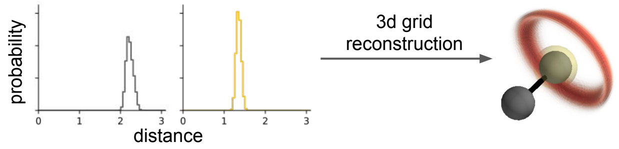


[Hoffmann & Noé, 2019]

Approaches for 3d molecule generation

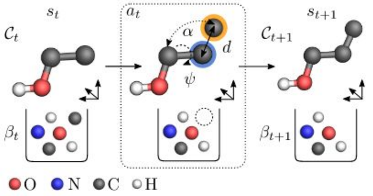
iterative

reconstruction from pairwise distances

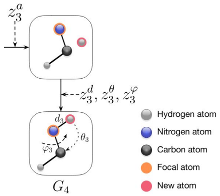


[Gebauer et al., 2019]

reconstruction from internal coordinates

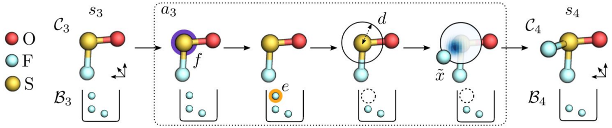


[Simm et al., 2020]



[Luo & Ji, 2022]

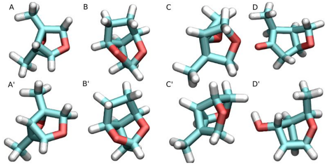
equivariant prediction of direction (spherical harmonics)



[Simm et al., 2021]

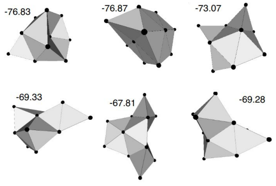
one-shot

generating valid euclidean distance matrices



[Hoffmann & Noé, 2019]

equivariant normalizing flow

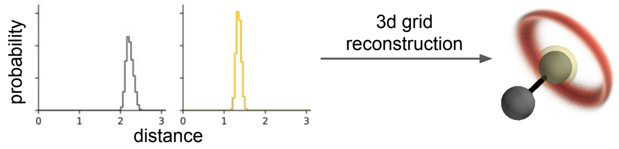


[Köhler et al., 2020]

Approaches for 3d molecule generation

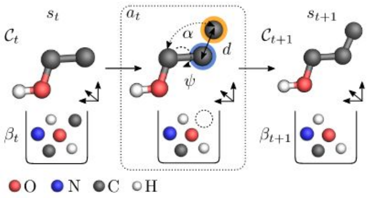
iterative

reconstruction from pairwise distances

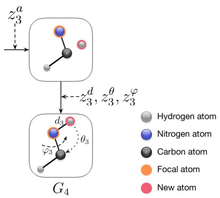


[Gebauer et al., 2019]

reconstruction from internal coordinates

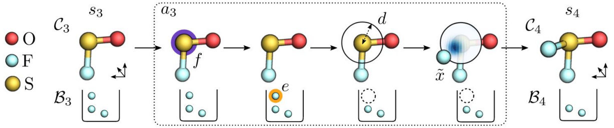


[Simm et al., 2020]



[Luo & Ji, 2022]

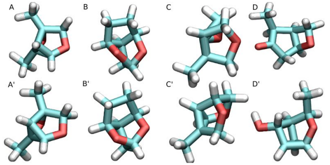
equivariant prediction of direction (spherical harmonics)



[Simm et al., 2021]

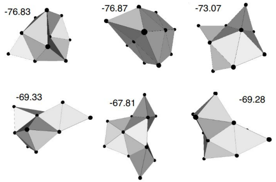
one-shot

generating valid euclidean distance matrices

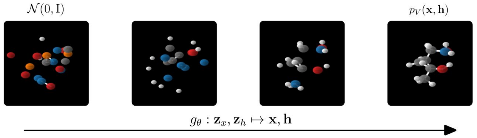


[Hoffmann & Noé, 2019]

equivariant normalizing flow



[Köhler et al., 2020]

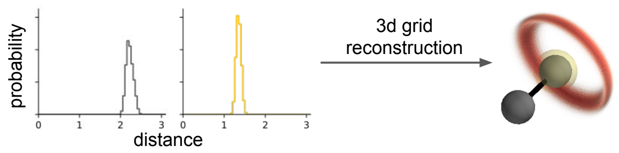


[Satorras et al., 2021]

Approaches for 3d molecule generation

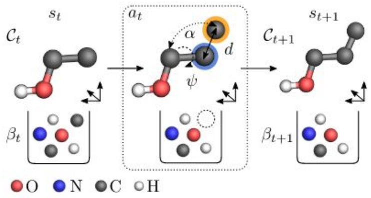
iterative

reconstruction from pairwise distances

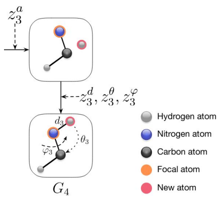


[Gebauer et al., 2019]

reconstruction from internal coordinates

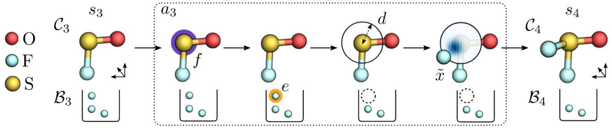


[Simm et al., 2020]



[Luo & Ji, 2022]

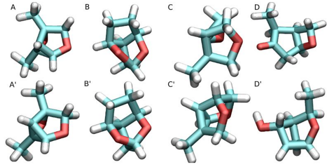
equivariant prediction of direction (spherical harmonics)



[Simm et al., 2021]

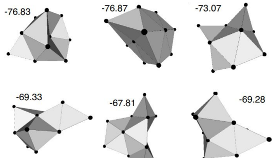
one-shot

generating valid euclidean distance matrices

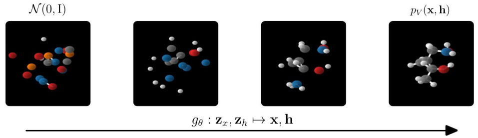


[Hoffmann & Noé, 2019]

equivariant normalizing flow

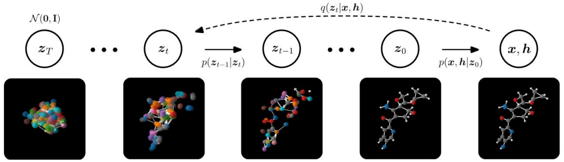


[Köhler et al., 2020]



[Satorras et al., 2021]

equivariant diffusion



[Hoogeboom et al., 2022]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

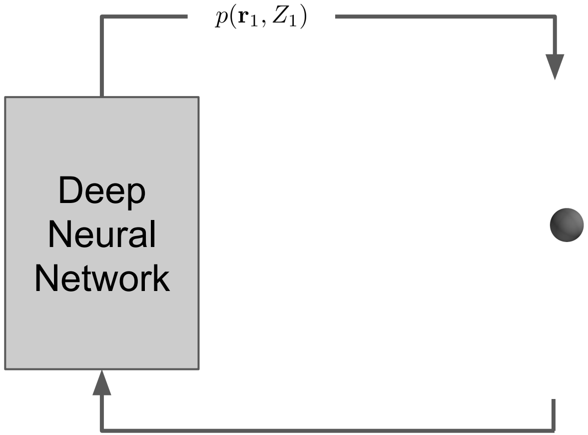
Deep
Neural
Network

G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

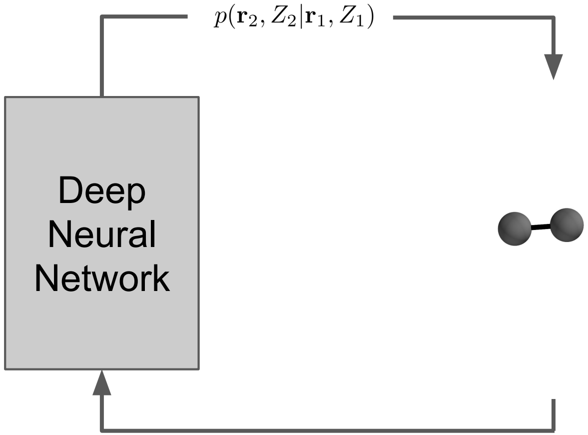


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

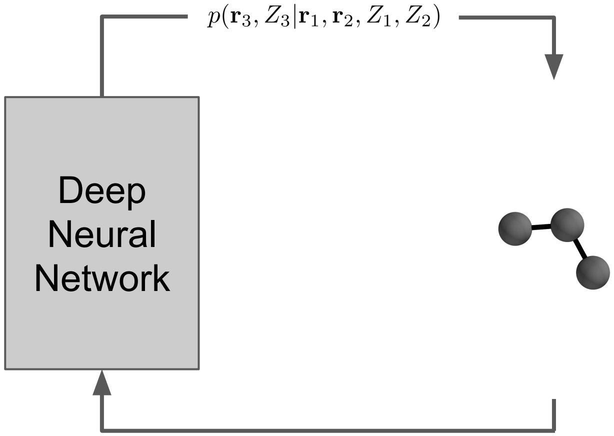


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

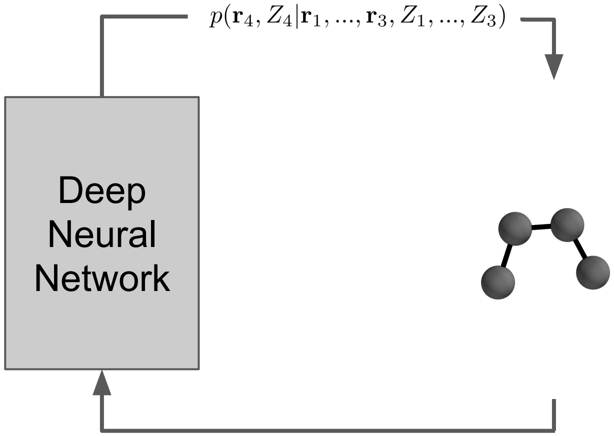


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

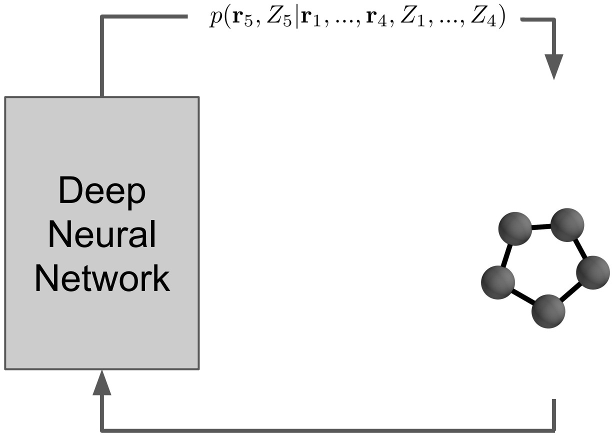


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

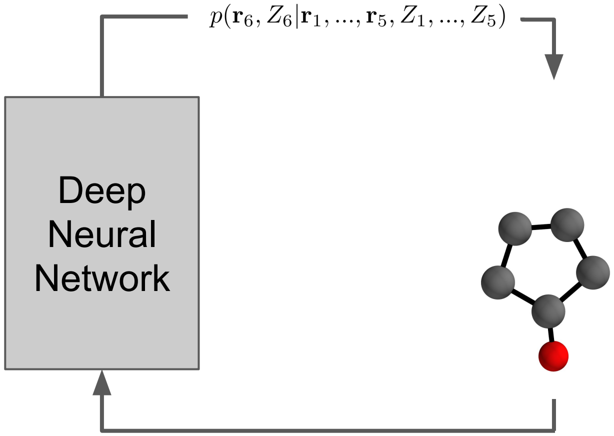


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

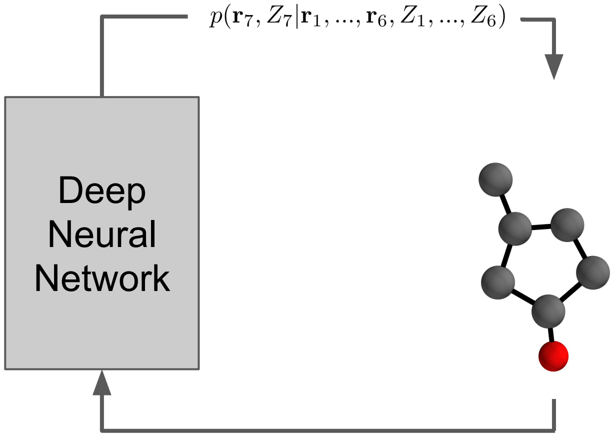
$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$



G-SchNet
[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

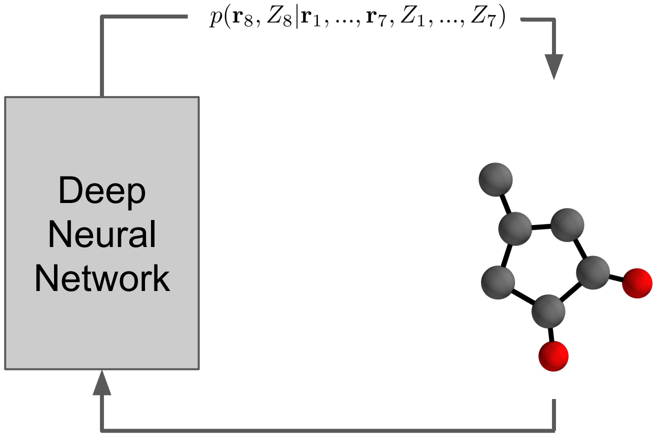


G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$



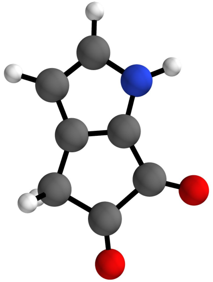
G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

Deep
Neural
Network



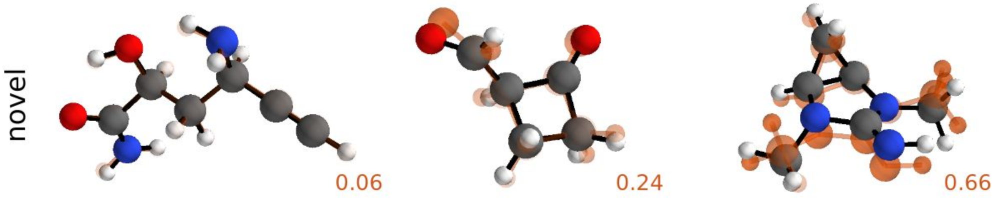
G-SchNet

[Gebauer et al., 2019]

Autoregressive approach

$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i)$$

Deep Neural Network



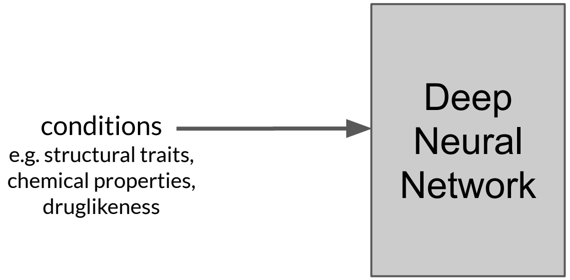
- sampling of valid, stable small organic molecules
- captures characteristics of the training data

G-SchNet

[Gebauer et al., 2019]

Autoregressive approach - conditional

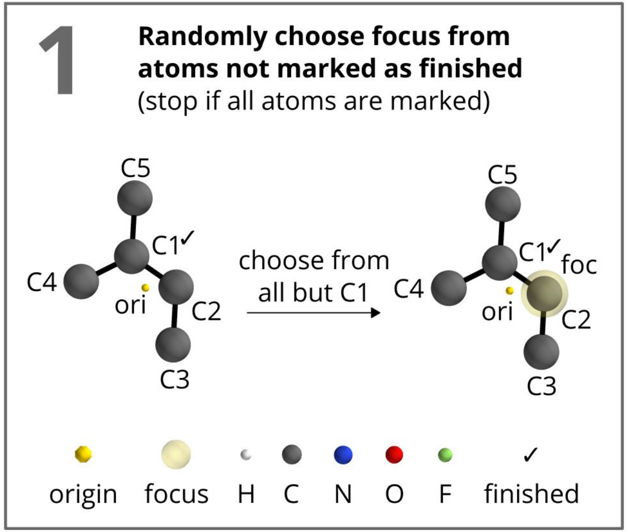
$$p(\mathbf{r}_1, ..., \mathbf{r}_n, Z_1, ..., Z_n | \lambda_1, ..., \lambda_k) = \prod_{i=0}^{n-1} p(\mathbf{r}_{i+1}, Z_{i+1} | \mathbf{r}_1, ..., \mathbf{r}_i, Z_1, ..., Z_i, \lambda_1, ..., \lambda_k)$$



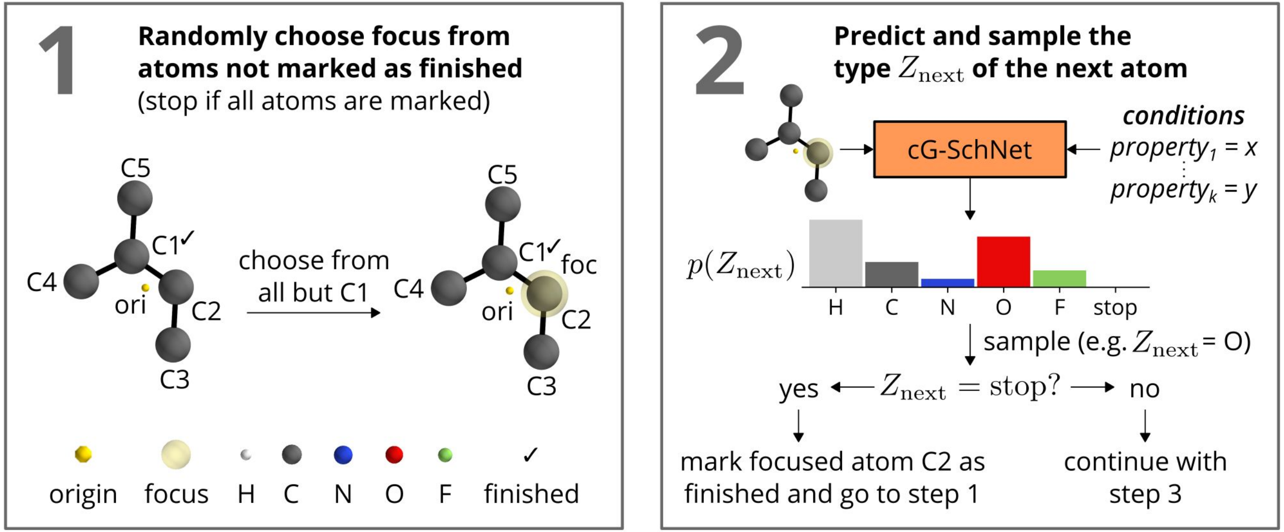
cG-SchNet

[Gebauer et al., 2022]

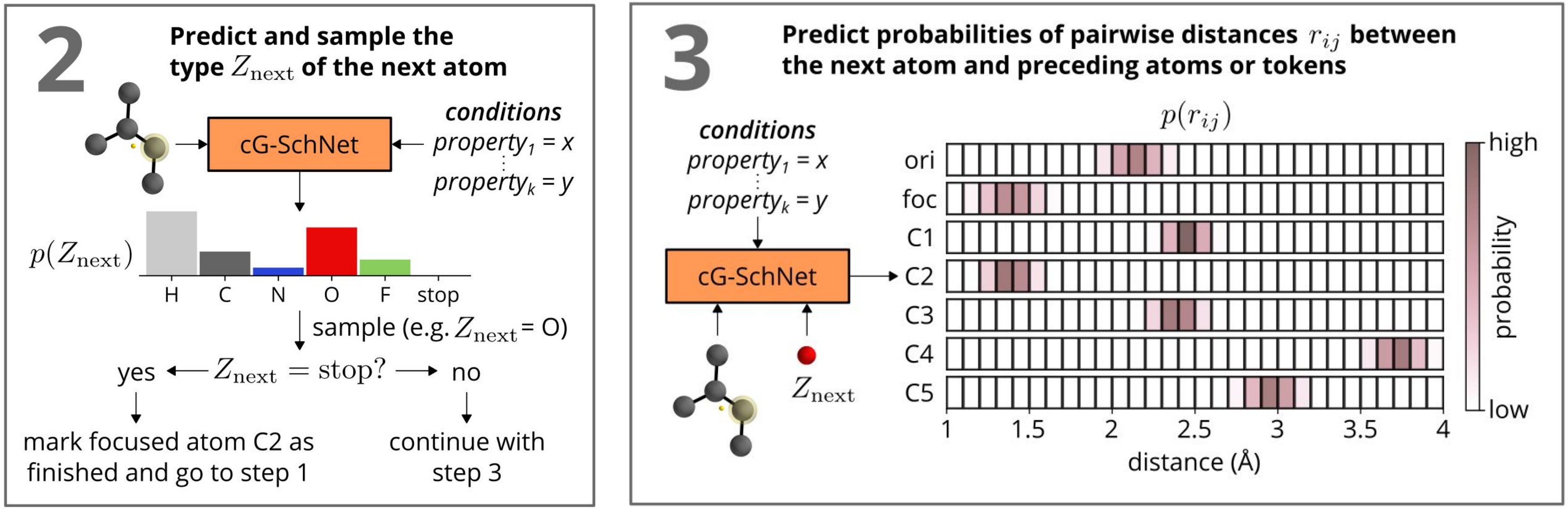
Atom placement loop with cG-SchNet



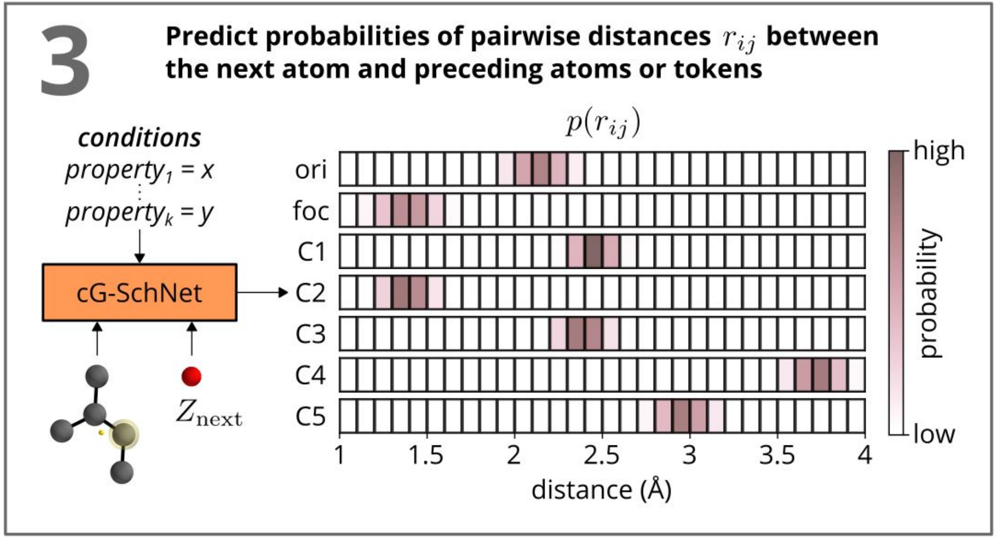
Atom placement loop with cG-SchNet

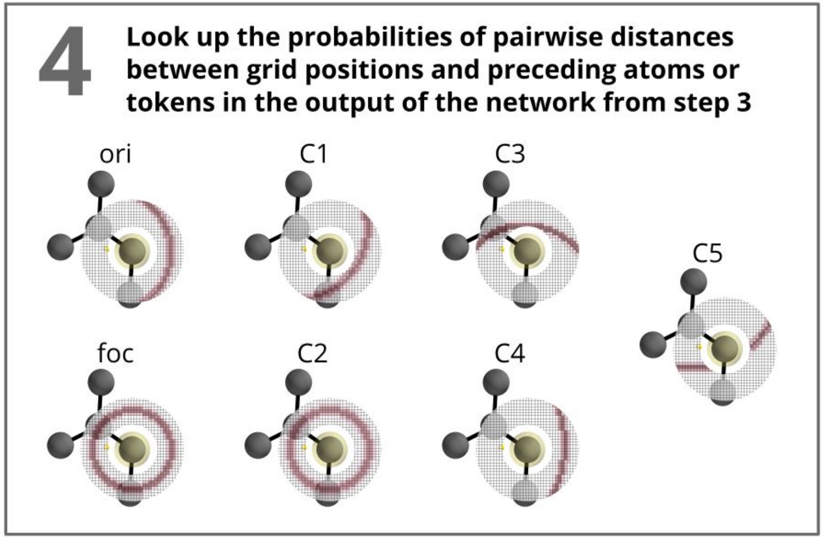


Atom placement loop with cG-SchNet



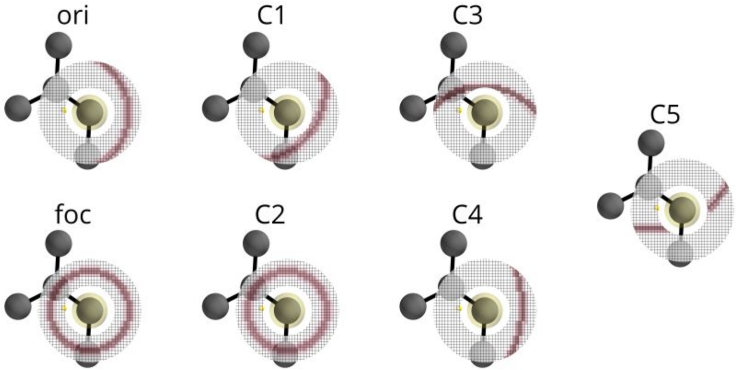
Atom placement loop with cG-SchNet





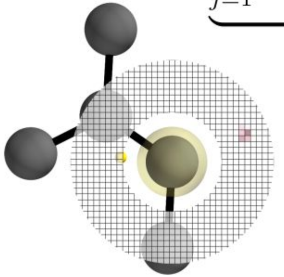
Atom placement loop with cG-SchNet

4 **Look up the probabilities of pairwise distances between grid positions and preceding atoms or tokens in the output of the network from step 3**



5 **Multiply probabilities at each position to reconstruct a distribution for the position $\mathbf{r}_{\text{next}}$ of the next atom**

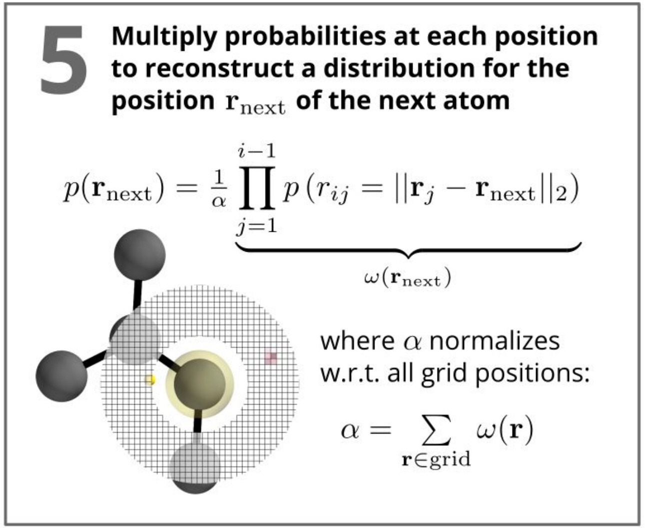
$$p(\mathbf{r}_{\text{next}}) = \frac{1}{\alpha} \underbrace{\prod_{j=1}^{i-1} p\left(r_{ij} = ||\mathbf{r}_j - \mathbf{r}_{\text{next}}||_2\right)}_{\omega(\mathbf{r}_{\text{next}})}$$

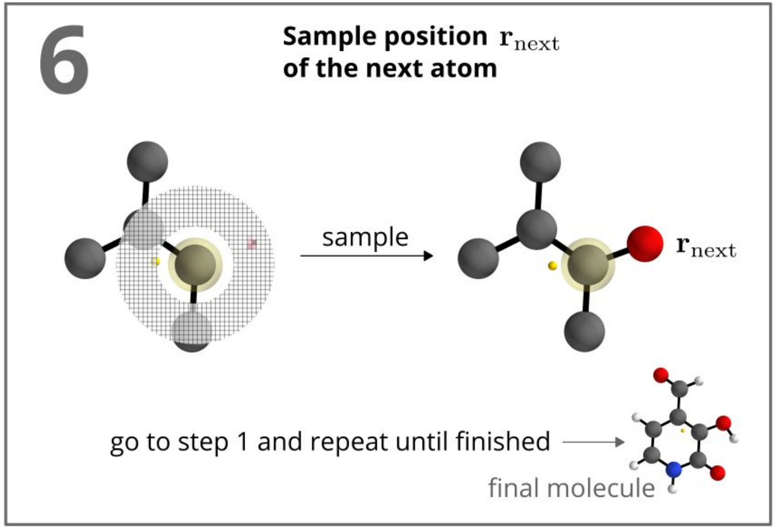


where α normalizes w.r.t. all grid positions:

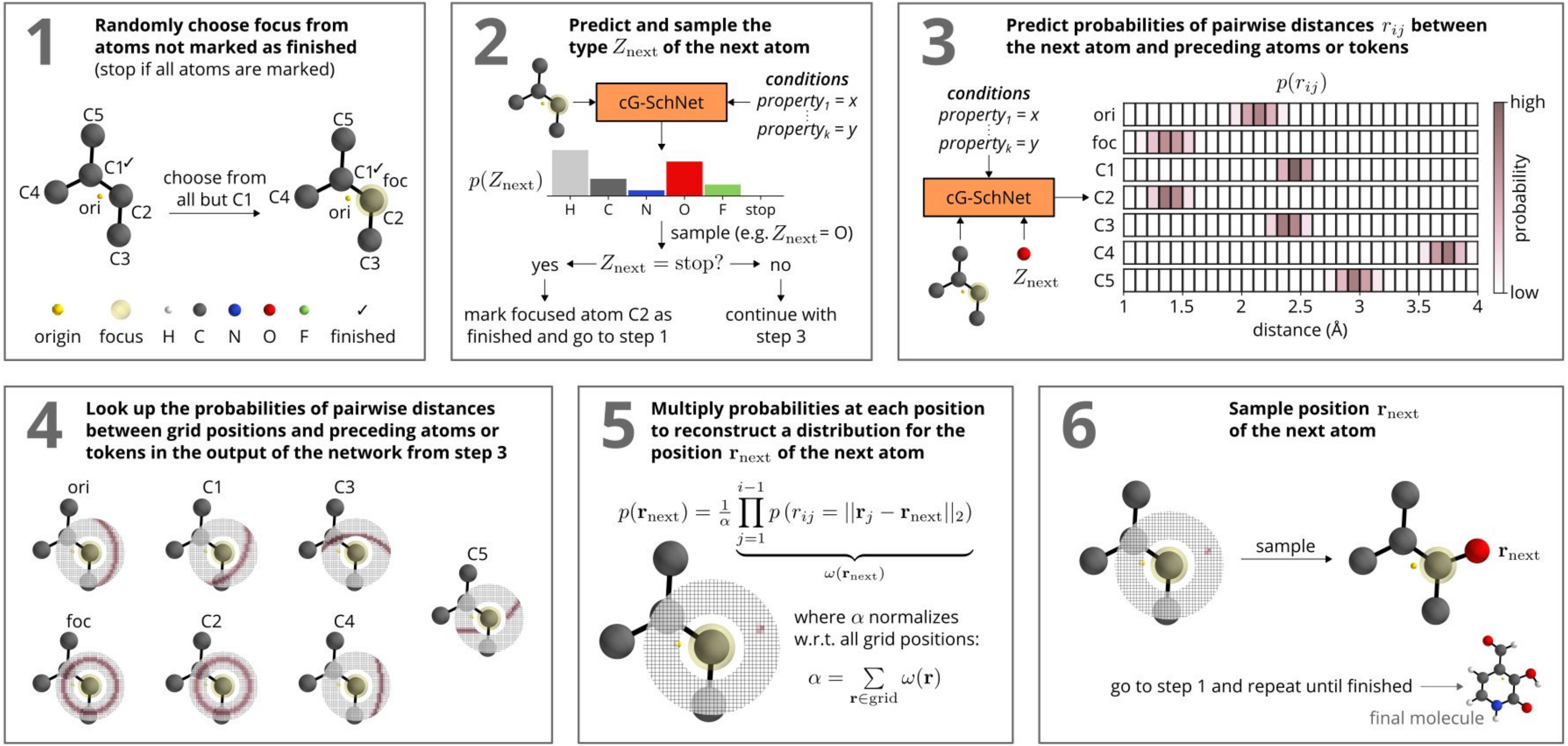
$$\alpha = \sum_{\mathbf{r}\in\text{grid}} \omega(\mathbf{r})$$

Atom placement loop with cG-SchNet

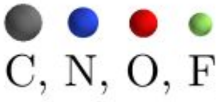




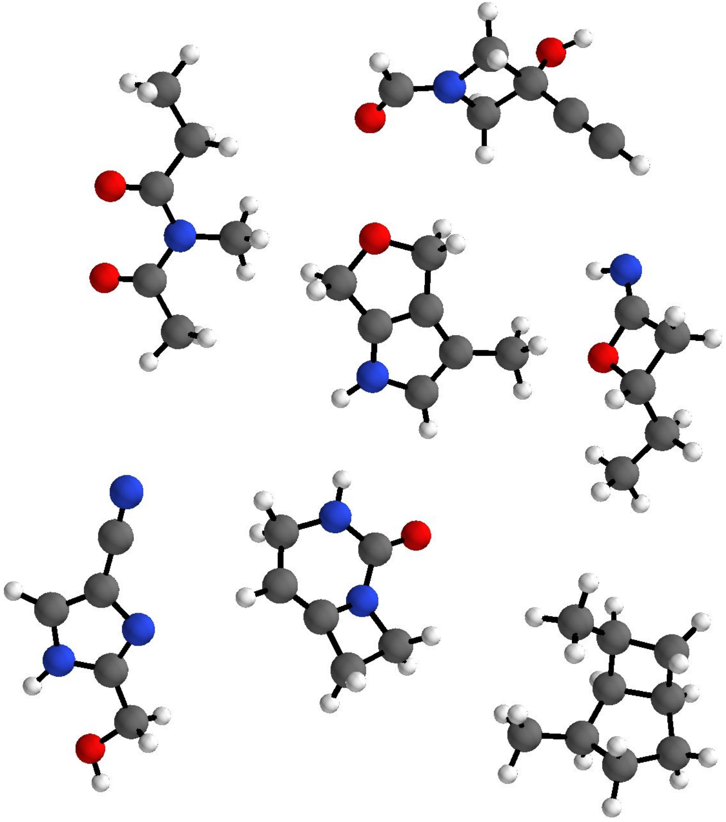
Atom placement loop with cG-SchNet



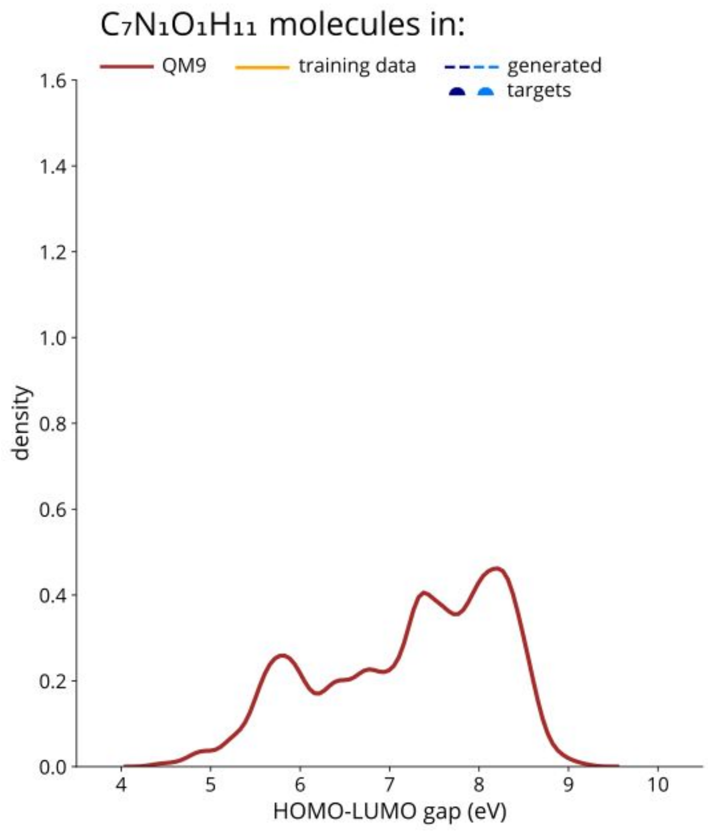
QM9

- ~130k stable molecules
- up to 9 heavy atoms 

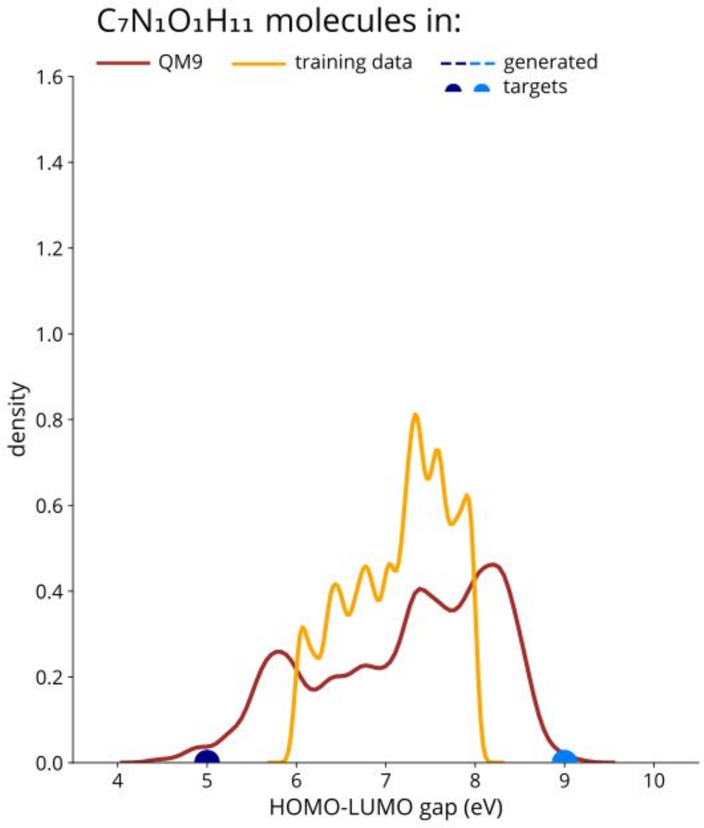
- 55k for training



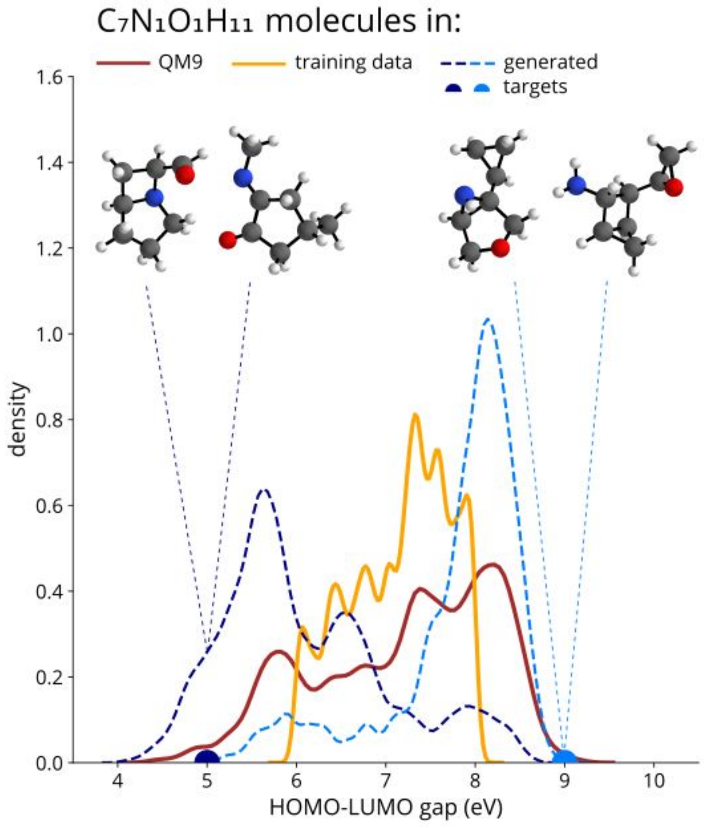
Results: Generalization across compositions



Results: Generalization across compositions



Results: Generalization across compositions



Challenges

- no real-world success story yet
- limited data availability
  - benchmark data sets have no practical applications
  - more interesting data owned by companies
  - data costly to compute
- mostly applied to:
  - small structures
  - non-periodic systems
- targeting many properties is difficult

Conclusion

- diverse applications require guided exploration of chemical space
- cG-SchNet for the conditional generation of 3d molecules
  - deep neural network respecting local and global symmetries
  - generates stable molecules capturing statistics of the training data
  - generalizes to unseen conditions to sample target-dependent candidates
  - allows for data-efficient, guided exploration of small organic compounds
- generative models for 3d structures are gaining traction
  - ignoring spatial component is a limiting factor
  - data availability is comparatively bad
  - models are still limited

➢ increase focus on targeted exploration (conditioning/biasing)

➢ application to crystal structures and larger molecules

Thank you!

Questions?

Literature

N. W. A. Gebauer, M. Gastegger, and K. T. Schütt, "Generating equilibrium molecules with deep neural networks," Machine Learning for Molecules and Materials, NeurIPS Workshop, 2018.

N. W. A. Gebauer, M. Gastegger, and K. T. Schütt, "Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules," in Advances in Neural Information Processing Systems 32, pp. 7564–7576, 2019.

N. W. A. Gebauer, M. Gastegger, S. S. P. Hessmann, K.-R. Müller, and K. T. Schütt, "Inverse design of 3d molecular structures with conditional generative neural networks," Nature Communications, vol. 13, p. 973, 2022.

M. Hoffmann and F. Noé, "Generating valid Euclidean distance matrices," *arXiv preprint arXiv:1910.03131*, 2019.

R. P. Joshi, N. W. A. Gebauer, M. Bontha, M. Khazaieli, R. M. James, J. B. Brown, and N. Kumar, "3D-Scaffold: A deep learning framework to generate 3D coordinates of drug-like molecules with desired scaffolds," The Journal of Physical Chemistry. B, *125*(44):12166–12176, 2021.

J. Köhler, L. Klein, and F. Noé, "Equivariant flows: exact likelihood generative learning for symmetric densities," *International Conference on Machine Learning*. PMLR, 2020.

Q. Liu, M. Allamanis, M. Brockschmidt, and A. Gaunt, "Constrained graph variational autoencoders for molecule design," in Advances in Neural Information Processing Systems, pp. 7795-7804, 2018.

Y. Luo and S. Ji, "An autoregressive flow model for 3d molecular geometry generation from scratch," in International Conference on Learning Representations, 2022.

R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Scientific data, *1*, 140022, 2014

L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17," Journal of Chemical Information and Modeling, vol. 52, no. 11, pp. 2864–2875, 2012.

V. G. Satorras, E. Hoogeboom, F. B. Fuchs, I. Posner, and M. Welling, "E (n) equivariant normalizing flows for molecule generation in 3D," *arXiv preprint arXiv:2105.09016*, 2021.

K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet – a deep learning architecture for molecules and materials," The Journal of Chemical Physics, vol. 148, no. 24, p. 241722, 2018.

G. Simm, R. Pinsler, and J.M. Hernandez-Lobato, "Reinforcement learning for molecular design guided by quantum mechanics," Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research 119:8959-8969, 2020.

G. Simm, R. Pinsler, G. Csányi, and J. M. Hernández-Lobato, "Symmetry-aware actor-critic for 3D molecular design," in International Conference on Learning Representations, 2021.

D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," Journal of Chemical Information and Computer Sciences, vol. 28, no. 1, pp. 31–36, 1988.