Auto-Encoding Molecular Conformations

Robin Winter
Machine Learning Research
Bayer AG
Berlin
robin.winter[at]bayer.com

Frank Noé
Mathematics and Computer Science
Freie Universität Berlin
Berlin

Djork-Arné Clevert
Machine Learning Research
Bayer AG
Berlin
djork-arne.clevert[at]bayer.com

Abstract

In this work we introduce an Autoencoder for molecular conformations. Our proposed model converts the discrete spatial arrangements of atoms in a given molecular graph (conformation) into and from a continuous fixed-sized latent representation. We demonstrate that in this latent representation, similar conformations cluster together while distinct conformations split apart. Moreover, by training a probabilistic model on a large dataset of molecular conformations, we demonstrate how our model can be used to generate diverse sets of energetically favorable conformations for a given molecule. Finally, we show that the continuous representation allows us to utilize optimization methods to find molecules that have conformations with favourable spatial properties.

1 Introduction

Representing chemical matter in a meaningful and expressive way plays a crucial role when it comes to computer-aided modeling in the field of chemistry (Todeschini and Consonni, 2009). Recently, substantial progress has been made in many molecule-related tasks, such as bio- and physicochemical property prediction (Montanari et al., 2020), inverse design of desirable molecules (Gómez-Bombarelli et al., 2018; Winter et al., 2019; Le et al., 2020), retrosynthetic analysis and synthesis planning (Segler et al., 2018). Most of these advancements can be attributed to the use of deep neural networks that enable representation learning of chemical matter. While traditional methods are mostly based on features, generated by rule-based algorithms extracting structural information (e.g. extended-connectivity fingerprints), these novel methods are directly trained on a discrete but more comprehensive representation of molecules. In particular, Graph Neural Networks utilizing the molecular graph with atoms as nodes and bonds as edges (Duvenaud et al., 2015) or Recurrent Neural Networks utilizing the one-dimensional line notation of molecules, namely SMILES (Segler et al., 2018). Still, the underlying molecular representation of the aforementioned methods are limited in the sense that they do not reflect the spatial arrangement of the atoms in the molecule. However, many interesting molecular properties, such as its ability to fit inside a protein binding pocket, inducing a pharmacological effect, are driven by its possible (energetically stable) spatial arrangements (conformations). Recently, work has been done to apply neural networks directly on specific conformations of molecules to predict properties such as the equilibrium energy or the *HOMO-LUMO* gap (Schütt et al., 2018). Moreover, models have been proposed to generate molecular conformations for a given molecular graph (Mansimov et al., 2019; Simm and Hernández-Lobato,

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

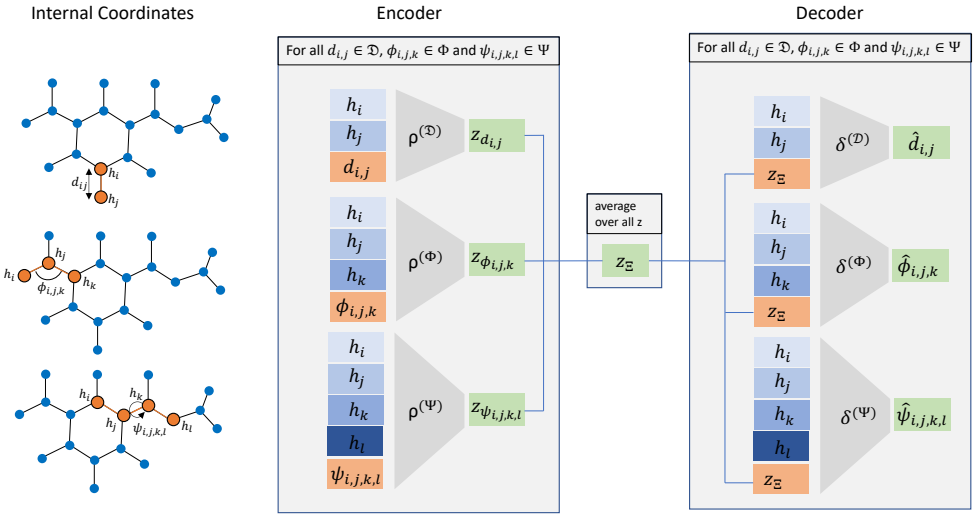

Figure 1: Model architecture of the conformation encoder (middle) and decoder (right). The encoding functions $\rho^{(\mathcal{D})}$, $\rho^{(\Phi)}$ and $\rho^{(\Psi)}$ encode their respective internal coordinates into a latent representation. The decoding functions $\delta^{(\mathcal{D})}$, $\delta^{(\Phi)}$ and $\delta^{(\Psi)}$ are conditioned on the averaged latent representations (conformer embedding) to reconstruct their respective internal coordinates, given a set of node embeddings $h_i \in \mathcal{H}$. On the left hand side, the definition of the internal coordinates, bond length $d_{i,j} \in \mathcal{D}$, bond angle $\phi_{i,j,k} \in \Phi$ and dihedral angle $\psi_{i,j,k,l} \in \Psi$, is visualized.

2019) or chemical formula (Hoffmann and Noé, 2019).
In this work we propose a novel model that converts a molecular conformation to and from a fixed-sized latent representation (conformation embedding), independent of the number of atoms and bonds of a molecule. Moreover, training the model in a probabilistic setting, we show that we can model the conformational distributions of molecules. We demonstrate that sampling from this model results into a diverse set of energetically reasonable conformers. Finally, combining the conformation embedding with a continuous molecular structure embedding, we demonstrate how molecules can be optimized with respect to spatial properties.

2 Methods

2.1 Representing Molecular Conformations

The three-dimensional arrangement of atoms of a molecule can be represented in many different ways. Annotating each atom with a Cartesian coordinate is probably the most straight-forward way, however, does not reflect a molecules invariance to rigid translations and rotations. In this work we utilize the *internal coordinate representation*, also known as *Z-matrix*. In this notation, a molecules spatial arrangement (conformation) Ξ is defined by the set of distances $\mathcal{D} = \{d_1, \ldots, d_{N_\mathcal{D}}\}$ between bonded atoms (bond length), the angles $\Phi = \{\phi_1, \ldots, \phi_{N_\Phi}\}$ of three connected atoms (bond angles) and the torsion angles (dihedral angles) $\Psi = \{\psi_1, \ldots, \psi_{N_\Psi}\}$ of three consecutive bonds (see Figure 1). This representation is invariant to rotations and rigid translations and can always be transformed to and from Cartesian coordinates.

2.2 Conformation Autoencoder

The overall goal of the proposed model is to find functions f_Θ and g_Θ that map a conformation $\Xi_\mathcal{G}$ of a molecule $\mathcal{G}$ to and from a fixed-sized latent representation $z_\Xi \in \mathbb{R}^{F_z}$, respectively. This introduces two major challenges. Firstly, the model has to map molecules with a different number of atoms and bonds to the same fixed-sized space. Secondly, f_Θ has to invariant with respect to the ordering of

atoms in the input. Our proposed model consist of a conformation-independent and a conformation-dependent part. The conformation-independent part comprises a Graph Neural Network utilizing the molecular graph to extract node-level features for a given molecule. The conformation-dependent part utilizes these extracted node-level features either to encode the internal coordinates of a specific molecular conformation into a latent representation (conformation embedding) or to reconstruct a conformation by predicting the internal coordinates of sets of connected atoms, given their respective node features and a conformation embedding. The whole model is trained on minimizing the reconstruction error of the internal coordinates Ξ for a given molecule:

$$\mathcal{C}_\Xi = \frac{1}{N_\mathcal{D}} \sum_{d \in \mathcal{D}} \|d, \hat{d}\|_2^2 + \frac{1}{N_\Phi} \sum_{\phi \in \Phi} \|\phi, \hat{\phi}\|_2^2 + \frac{1}{N_\Psi} \sum_{\psi \in \Psi} \min((\|\psi, \hat{\psi}\|_2^2, 2\pi - \|\psi, \hat{\psi}\|_2^2), \tag{1}$$

where the last term accounts for the periodicity of the dihedral angle. Our proposed model can easily be extended to a probabilistic generative model by employing the ideas from Kingma and Welling (2013), effectively defining the model as a variational auto encoder.

2.2.1 Molecular Graph Encoder

In this work we define the conformation-independent molecular graph as an undirected Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertices or nodes $v_i \in \mathcal{V}$ and edges $e_{ij} = (v_i, v_j) \in \mathcal{E}$ connecting v_i and v_j. Where nodes $v_i \in \mathbb{R}^{F_v}$ represent atoms and encode atom features such as element type and charge. Edges $e_{ij} \in \mathbb{R}^{F_e}$ represent bonds between atoms and encode the bond type (i.e. single, double, triple or aromatic bond).
We utilize a Graph Neural Network (GNN) to extract a node-level representation of a molecular graph. Given a molecular graph with initial node and edge features defined by the atoms and bonds of the molecule, a GNN iteratively updates node embeddings by aggregating localized information of connected nodes respectively. In particular, we utilized edge-conditioned graph convolution (EConv) (Simonovsky and Komodakis, 2017) and Graph Attention Network (GAT) (Veličković et al., 2017) layers. To aggregate information about higher-order neighbours, we combine one EConv (to encode edge information) with multiple consecutive GAT layers:

$$\mathbf{h}_i^l = \text{GAT}^{l-1} \circ \ldots \circ \text{GAT}^1 \circ \text{EConv}(\mathbf{h}_i^0). \tag{2}$$

where $\mathbf{h}_i^0 = v_i \in \mathbb{R}^{F_v}$ are the atom features of the input molecular graph.

2.2.2 Conformation Encoder

To this end, we define the molecular conformation representation learning task as a learning task on *sets*. In particular, our proposed model learns to extract a latent representation z_Ξ of a set of internal coordinates Ξ for a given molecule:

$$z_\Xi = f_\Theta(\mathcal{H}, \Xi) = f_\Theta(\mathcal{H}, \mathcal{D}, \Phi, \Psi), \tag{3}$$

with the permutation invariant function f_Θ, parameterized by a neural network and conditioned on the node embeddings $\mathcal{H} = \{h_1, \ldots h_N\}$, extracted by the molecular graph encoder defined in the previous section. Following Zaheer et al. (2017), we define the permutation invariant function f_Θ as

$$\begin{aligned} z_\Xi = f_\Theta(\mathcal{H}, \Xi) &= \sigma\left(\sum_{\xi \in \Xi} \rho(\mathcal{H}, \xi)\right) = \frac{1}{N_\Xi} \sum_{\xi \in \Xi} \rho_\Theta(\mathcal{H}, \xi) \\ &= \frac{1}{N_\mathcal{D} + N_\Phi + N_\Psi} \left(\sum_{d \in \mathcal{D}} \rho_\Theta^{(\mathcal{D})}(\mathcal{H}, d) + \sum_{\phi \in \Phi} \rho_\Theta^{(\Phi)}(\mathcal{H}, \phi) + \sum_{\psi \in \Psi} \rho_\Theta^{(\Psi)}(\mathcal{H}, \psi) \right). \end{aligned} \tag{4}$$

The functions $\rho_\Theta^{(\mathcal{D})}$, $\rho_\Theta^{(\Phi)}$ and $\rho_\Theta^{(\Psi)}$ are defined as feed-forward neural networks that take bond lengths, bond angles and dihedral angles as input respectively. Additionally, to put an internal coordinate into context of the graph, ρ_Θ is conditioned on corresponding node embeddings $\mathcal{H}$ as well. This means, if $\rho_\Theta^{(\Phi)}$ encodes the angle $\phi_{i,j,k}$ between atoms v_i and v_k connected by atom v_j, function $\rho_\Theta^{(\Phi)}$ takes also node embeddings h_i, h_j and h_k as argument (see Figure 1). Most importantly, equation (4) is invariant to the order of internal coordinates and node indices and the dimensionality of the resulting z_Ξ is invariant of the size of the input molecule.

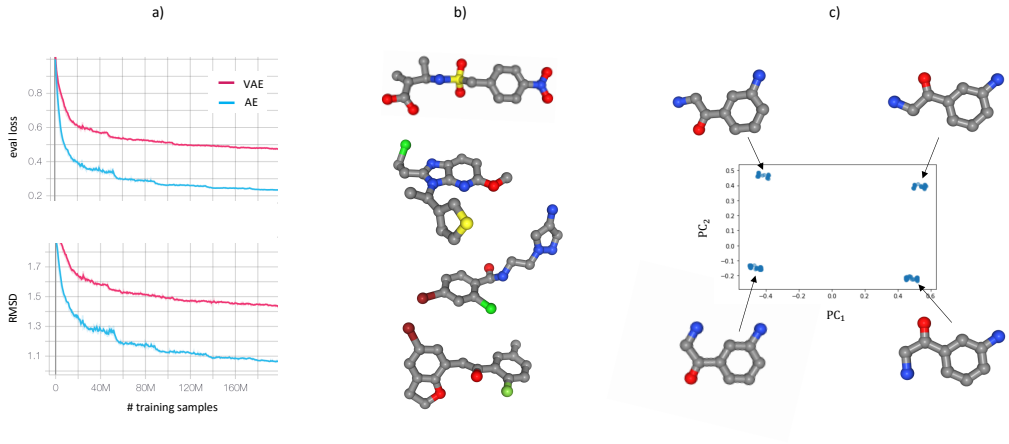

Figure 2: a) Learning curves for the autoencoder (AE) and variational autoencoder (VAE) during training, with evaluation loss as defined in (1) and root means squared deviation (RMSD) between predicted and input conformations on a holdout set. b) Four example conformations generated by our proposed model. c) First two principle components of the latent representation (conformation embedding) of 200 conformations with a corresponding representative conformation for each cluster.

2.2.3 Conformation Decoder

To reconstruct a molecular conformation back from its latent representation, we train three additional neural networks $\delta_\Theta^{(\mathcal{D})}$, $\delta_\Theta^{(\Phi)}$ and $\delta_\Theta^{(\Psi)}$ for each type of internal coordinate respectively (see Figure 1). Each neural network is conditioned on the conformation embedding and takes the node embeddings of the corresponding internal coordinate as input. For example, to predict the bond length $d_{i,j}$ between atoms v_i and v_j, the network takes h_i, h_j and z_Ξ as input.

3 Results and Discussion

We trained our model on the public PubChem3D dataset (Bolton et al., 2011), which comprises molecules (organic, up to 50 heavy atoms) with multiple conformations generated by the forcefield software OMEGA (Hawkins et al., 2010). Upon convergence, our model is able to predict internal coordinates for a given molecule that result into conformations that are similar (with respect to the RMSD) to the input conformations (see Figure 2). To quantitatively analyze how energetically reasonable the reconstructed conformations are, we calculated their internal energy with the MMFF94 forcefield (Halgren, 1996) as implemented in the Python package RDKit (Landrum et al., 2006). The median energetic difference between the input and reconstructed conformation is approximately $80\,\mathrm{kcal/mol}$, which corresponds to small deviations from local minimas, without e.g. clashes of atoms (see example molecules in Figure 2). Moreover, since the model does not only reconstruct any possible conformation for a molecule but is trained on reconstructing a specific input conformation, differences between these conformations have to be encoded in the latent representation. On the right side of Figure 2, we show this for a simple example of a small molecule in four different conformations.

As described in Section 2.2, we can easily extend the proposed model to a variational autoencoder, which can be used to sample conformations from the learned distribution. A major challenge in conformation generation is to efficiently sample diverse conformers. Therefore, we analyzed the average interconformer RMSD (icRMSD) for a set of 200 sampled conformers per molecule for the holdout set. Comparing the icRMSD of our proposed model with a state-of-the-art conformation generation algorithm ETKDG (Riniker and Landrum, 2015) as implemented in RDKit, we see a similar performance, with our model having a slightly higher average icRMSD of $0.07\,\text{Å}$.

Since the proposed model gives means to directly infer conformations for a given molecule, it is possible to optimize molecules in the continuous conformation embedding with respect to spatial properties. When combined with a latent representation of the molecular structure (Winter et al., 2019), optimization of molecules can even be performed with respect to both the molecular graph and its conformation. As a proof of principle, we optimized molecules with respect to a combination of

the conformation-independent *quantitative estimate of drug-likeness* (QED) score (values between 0 and 1) (Bickerton et al., 2012) and the conformation-dependent property *asphericity* (Todeschini and Consonni, 2009) (values between 0 and 1), which quantifies a molecules deviation from a spherical shape. We utilized the genetic *Particle Swarm Optimization* algorithm (Kennedy and Eberhart, 1995), to optimize both latent representations at the same time. Starting from the already drug-like molecule aspirin with a combined score of 0.76, we could already find after 50 iterations molecules with a score of 1.82. In general this method could also be used to optimize molecules for other interesting spatial properties, such as fitting pharmacophores or the shape of known bio active molecule.

References

Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*; John Wiley & Sons, 2009; Vol. 41.

Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling physico-chemical admet endpoints with multitask graph convolutional networks. *Molecules* **2020**, *25*, 44.

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

Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science* **2019**, *10*, 8016–8024.

Le, T.; Winter, R.; Noé, F.; Clevert, D.-A. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chem. Sci.* **2020**, –.

Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. **2018**, *555*, 604–610.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems. 2015; pp 2224–2232.

Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* **2018**, *4*, 120–131.

Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.

Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Scientific reports* **2019**, *9*, 1–13.

Simm, G. N.; Hernández-Lobato, J. M. A generative model for molecular distance geometry. *arXiv preprint arXiv:1909.11459* **2019**,

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

Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**,

Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; pp 3693–3702.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903* **2017**,

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; Smola, A. J. Deep sets. Advances in neural information processing systems. 2017; pp 3391–3401.

Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J., et al. PubChem3D: a new resource for scientists. *Journal of cheminformatics* **2011**, *3*, 32.

Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of chemical information and modeling* **2010**, *50*, 572–584.

Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **1996**, *17*, 490–519.

Landrum, G., et al. RDKit: Open-source cheminformatics. **2006**,

Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science* **2019**, *10*, 1692–1701.

Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **2012**, *4*, 90–98.

Kennedy, J.; Eberhart, R. Particle swarm optimization. Proceedings of ICNN'95-International Conference on Neural Networks. 1995; pp 1942–1948.