3DMolNet: A Generative Network for Molecular Structures

Vitali Nesterov, Mario Wieser, Volker Roth
Department of Mathematics and Computer Science
University of Basel, Switzerland
vitali.nesterov@unibas.ch

Recent advances in machine learning for quantum chemistry allow to predict the chemical properties of compounds and to generate novel molecules. Existing generative models mostly use a string- or graph-based representation [2, 8, 5, 4], but the precise three-dimensional coordinates of the atoms are usually not encoded. First attempts in this direction have been proposed, where autoregressive or GAN-based models generate atom coordinates [9, 1, 3]. Those either lack a smooth exploration of the compound space or cannot generalize to varying chemical compositions.

In this paper, we introduce the 3DMolNet, a generative network for 3-d molecular structures. Our model is based on the Variational autoencoder (VAE) [6, 13] for learning a low-dimensional latent representation of the molecules. The molecule representation consists of two core components, a nuclear charge matrix and an Euclidean distance matrix. Since some bond types have high variances in the bond lengths, an exact assignment of bond types often cannot be derived from the distances. For this reason, we additionally include an explicit bond matrix to our representation.

To overcome the permutation problem of the atom ordering, we use an InChI-based canonical ordering algorithm [14]. Due to the absence of canonical identifiers for most of the hydrogen atoms, we involve only heavy atoms in our representation. This is however not problematic, since the backbone structure of a molecule is the most important part and the hydrogen atoms can be easily added and fine-tuned in the post processing step.

In summary, the main contributions of this work are: first, a translation-, rotation-, and permutation-invariant molecule representation. Second, a VAE-based model to learn a continuous low-dimensional representation of the molecules which allow a generation of 3-d molecular structures in a one-shot fashion. Third, a demonstration on the QM9 dataset that 3DMolNet is able to reconstruct almost all chemical compositions with up to 9 heavy atoms while yielding a RMSD below 0.05 Å for the coordinate reconstructions.

1 Methods

1.1 Representation

A physical many-body system depends only on a set of nuclear charge numbers $Z_i \in \mathbb{N}$ and corresponding $\boldsymbol{R}_i \in \mathbb{R}^3$ coordinates. Our molecule representation involves this information with a nuclear charge matrix $\boldsymbol{C} \in \mathbb{N}^{N\times N}$ and an Euclidean distance matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$. Additionally, a bond matrix $\boldsymbol{B} \in \mathbb{N}^{M\times 3}$ is included to encode explicit bond types. The final molecule representation is a set of the three separate components $\{\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{B}\}$. We define the nuclear charge matrix with $C_{ij} = Z_i Z_j^\top$ and the distances matrix $D_{ij} = \|\boldsymbol{R}_i - \boldsymbol{R}_j\|_2^2$. The bond matrix $\boldsymbol{B}$ includes indices of connected atoms in the first and the second columns. The third column contains bond type numbers. To handle molecules of different sizes each of the $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{B}$ matrices are padded with zero rows and or columns to a fixed size.

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

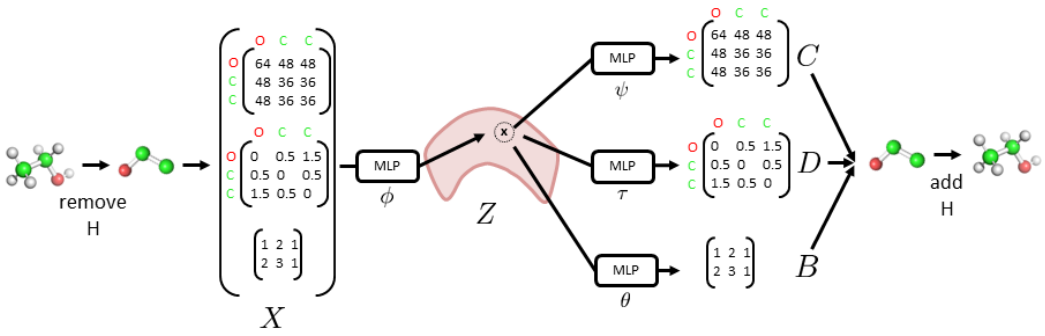


Figure 1: An illustration of the 3DMolNet framework. In the preprocessing step, we apply a canonical ordering and remove all hydrogen atoms. We then get the representation $\boldsymbol{X}$ as a concatenation of the nuclear charge $\boldsymbol{C}$, the euclidean distance $\boldsymbol{D}$, and the bond $\boldsymbol{B}$ matrices. Subsequently, we map $\mathbf{X}$ into a low-dimensional latent representation $\boldsymbol{Z}$ using an encoder neural network which is parametrized with ϕ. Here, (x) denotes the mean and the surrounding circle the variance of our estimate. To decode a molecule, we use three separate neural networks parametrized with ψ, τ, and θ which decode representation components $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{B}$ separately. In the post processing step, we recover the chemical and structural composition of the heavy atoms with a multidimensional scaling algorithm and add hydrogen atoms with Open Babel [10] and fine-tune with MOPAC [15].

1.2 Canonical Identifier

A major obstacle for a learning algorithm is an arbitrary atom ordering. To overcome this problem we generate a unique atom ordering for molecules. There exist different canonicalization algorithms. We get the canonical identifiers (CIs) with an InChI-based algorithm implemented in the CDK package [14]. The generation of CIs involves a structure normalization, an InChI-based canonical labelling, and a tree traversal [11]. With this method, we get CIs for each heavy atom and only for explicit hydrogen atoms. Those are indicated isotops, dihydrogen and hydrogen ions, and hydrogen atoms attached to tetrahedral stereocentres with defined stereochemistry. Due to the absence of canonical labels for most of the hydrogen atoms, our representation involves only heavy atoms.

1.3 Model

As previously stated, our aim is to generate novel molecular compositions and corresponding 3-d structures. Therefore, we want to learn both, a low-dimensional and a rotation-, translation-, and permutation-invariant representation of molecules (see Figure 1). To do so, we reformulate the standard VAE by defining each representation component separately in a form of independent random variables. This leads to an extended parametric formulation:

$$L_{\text{VAE}} \geq \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{d},\boldsymbol{b})}[p(\boldsymbol{c},\boldsymbol{d},\boldsymbol{b} \mid \boldsymbol{z})] - \beta D_{\text{KL}}[q(\boldsymbol{z} \mid \boldsymbol{c},\boldsymbol{d},\boldsymbol{b}) \mid\mid p(\boldsymbol{z})]. \tag{1}$$

Encoder. The encoder is defined as the KL-divergence between the posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{c},\boldsymbol{d},\boldsymbol{b})$ and the prior $p(\boldsymbol{z})$ and is denoted as:

$$D_{\text{KL}}[q_\phi(\boldsymbol{z} \mid \boldsymbol{c},\boldsymbol{d},\boldsymbol{b}) \mid\mid p(\boldsymbol{z})],$$

where ϕ are the neural network parameters. We define the posterior as a Gaussian distribution and assume a Gaussian prior $p(\boldsymbol{z}) = \mathcal{N}(0, \boldsymbol{I})$.

Decoder. The decoder is defined as the the negative log-likelihood of $p(\boldsymbol{c},\boldsymbol{d},\boldsymbol{b} \mid \boldsymbol{z})$. As we assume conditional independence between $\boldsymbol{c}$, $\boldsymbol{d}$, and $\boldsymbol{b}$ we can express the joint distribution as follows:

$$\mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{d},\boldsymbol{b})}[p(\boldsymbol{c},\boldsymbol{d},\boldsymbol{b} \mid \boldsymbol{z})] = \mathbb{E}_{\boldsymbol{z}_\phi \sim q(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{d},\boldsymbol{b})}[p_\theta(\boldsymbol{c} \mid \boldsymbol{z}) \cdot p_\tau(\boldsymbol{d} \mid \boldsymbol{z}) \cdot p_\psi(\boldsymbol{b} \mid \boldsymbol{z})],$$

where ϕ, τ and θ denote neural network parameters of each respective decoder. We define each log-likelihood term to be the mean-absolute-error between a particular representation component and the reconstructed counterpart.

Geometry Loss. To further improve the quality of the EDM reconstruction, we additionally penalize the negative eigenvalues and the rank of the Gram matrix. To do so, we first define the geometric centering matrix as $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, where $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ is the all-ones vector. The Gram matrix is defined with $\boldsymbol{G} = -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$. The matrix $\boldsymbol{D}$ is an EDM, if the Gram matrix $\boldsymbol{G}$ is positive semi-definite. To enforce this property we penalize negative eigenvalues of the Gram matrix $\boldsymbol{G}$. Therefor, we first get eigenvalues $\boldsymbol{\lambda}$ with the eigenvalue decomposition of $\boldsymbol{G}$. We then apply the ReLU function to the negative of the eigenvalues $\bar{\boldsymbol{\lambda}} = \mathrm{ReLU}(-\boldsymbol{\lambda})$. The corresponding loss term is defined as follows:

$$L_{\mathrm{EV}} = \bar{\boldsymbol{\lambda}}^\top \bar{\boldsymbol{\lambda}}. \tag{2}$$

Since the atom coordinates exists in a maximally 3-dimensional embedding, we additionally penalize larger than $k = 3$ rank of the Gram matrix $\boldsymbol{G}$. For this, we sort the eigenvalues in descending order $\boldsymbol{\lambda} = [\lambda_1 \geq \lambda_2 \geq \ldots \lambda_N]$. The rank loss is obtained with

$$L_{\mathrm{R}} = \sum_{i=k+1}^{N} \lambda_i^2. \tag{3}$$

Overall Training Objective. The total loss function is given with:

$$L_{\mathrm{total}} = L_{\mathrm{VAE}} + L_{\mathrm{EV}} + L_{\mathrm{R}}. \tag{4}$$

1.4 Recovering Molecular Structures

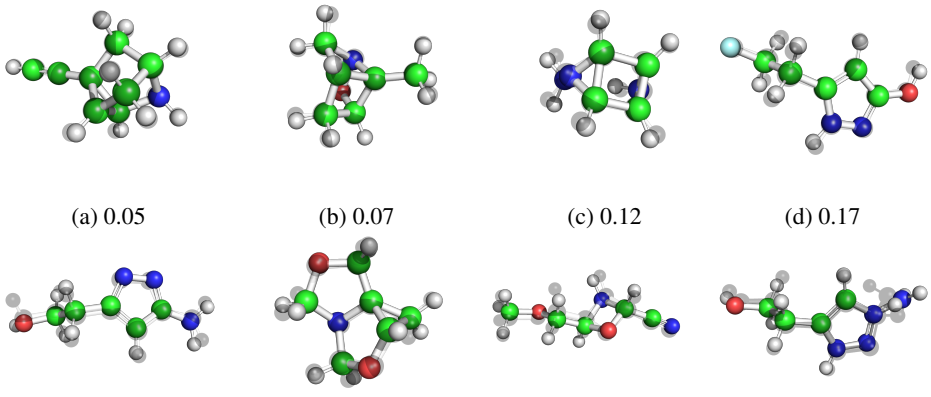
To obtain the atom and coordinate pairs we first symmetrize the generated nuclear charge matrix $\boldsymbol{C}$ and the distance matrix $\boldsymbol{D}$. The diagonal of the distance matrix $\boldsymbol{D}$ is set to zero. The reconstructed floating point values of the bond matrix $\boldsymbol{B}$ are rounded to integer values. Subsequently, we recover a set of the nuclear charge numbers and the corresponding coordinates $\{Z_i, \boldsymbol{R}_i\}$. To do so, we use the classical multidimensional scaling (MDS) algorithm. We then use Open Babel to read-in the molecular structure by setting types and coordinates of the atoms and assign bonds and corresponding bond types. Lastly, we reconstruct a complete molecule by adding hydrogen atoms with Open Babel. To get initial positions of hydrogen atoms we use an efficient force-filed method with Open Babel [10] and fine-tune by using a semi-empirical *ab initio* approximation with MOPAC [15]. Given the coordinates of the added and the target hydrogen atoms, we use the Hungarian algorithm [7] to minimize the assignment costs.

2 Experiments

Dataset and Experimental Setup. For our experiments, we use the QM9 dataset [12], which includes 133,885 small organic molecules and consist of up to nine heavy atoms (C, O, N, and F). Each molecule has geometric, energetic, electronic, and thermodynamic properties obtained from the Density Functional Theory (DFT) calculations.

The 3DMolNet consists of a single encoder and three decoder networks for the generation of the nuclear charge matrix $\boldsymbol{C}$, the distance matrix $\boldsymbol{D}$, and the bond matrix $\boldsymbol{B}$. The encoder and each decoder have a standard fully connected architecture. The latent space is set to 64 dimensions. The compression parameter β in Equation 1 is reduced after each epoch.

We train the 3DMolNet on a set of 50K randomly selected molecules from the QM9 dataset. For validation 5K molecules are randomly selected. The remaining molecules are used for evaluation as a test set. We use Open Babel for validity check of the valences. To compare molecular geometries, we apply Procrustes analysis and calculate the root-mean-square deviation (RMSD) of pair-wise atomic coordinates between generated and ground-truth structures.



(a) 0.05 (b) 0.07 (c) 0.12 (d) 0.17

(e) 0.24 (f) 0.39 (g) 0.42 (h) 0.55

Figure 2: An illustration of novel molecules discovered with 3DMolNet. The colored atoms and bonds represents generated structures and the translucent plots are relaxed counterparts. The molecules are labeled with a RMSD values for the generated and relaxed structures.

Reconstruction Accuracy. We evaluate the reconstruction accuracy of the nuclear charge numbers, the atom coordinates, and the bond matrix. Our experiments show that the chemical compositions are exactly reconstructed in more than 99 % of the cases. The bond matrix is exactly reconstructed in about 98 % of the cases. To evaluate the reconstructed geometries, we only accept molecules with exactly reconstructed atoms, bonds, and bond types. The reconstruction of heavy atom coordinates yields a RMSD of 0.05 Å. By using a semi-empirical *ab initio* method for optimization of the hydrogen coordinates, a RMSD of 0.16 Å is achieved. In comparison, Generative Graph Neural Networks (GGNN) [9] achieved a RMSD of 0.37 Å, G-SchNet [1] yields a RMSD of 0.18 Å for the reconstruction of heavy atom coordinates and 0.23 Å including hydrogen atoms. A comparison with EDMNet [3] in terms of the reconstruction accuracy of the coordinates cannot be directly done, due to the nature of the chosen GAN-based architecture. Beyond that, a fair comparison is not possible, since the EDMNet is trained only on a sub set of molecules of the same chemical composition.

Molecule Discovery. For discovery of novel molecules within a QM9 dataset, we randomly sample from the latent space around the mean regions of the molecules from the training set. To identify a novel chemical composition, first, we generate a set of canonical SMILES strings for the entire QM9 dataset. We then compare the canonical SMILES string of the generate molecule with the QM9 references. A molecule is accepted and is considered as novel if no identical canonical SMILES string is found within the QM9 dataset. Molecules with invalid bond and bond types are rejected. We were able to identify more than 20K novel molecules with new chemical compositions (see Figure 2). To investigate the quality of the generated molecular structures we relaxed the structures with MOPAC and computed the pair-wise atom distances yielding in a RMSD of 0.32 Å. To put our results into relation, for G-SchNet a median of around 0.3 Å is reported, being a comparable result achieved with our experiments.

3 Conclusion

In this paper, we introduced the 3DMolNet which allow an efficient generation of novel 3-d molecular structures of a variable size and chemical composition. To this end, we introduced a translation-, rotation-, and permutation-invariant molecule representation involving a canonical ordering of the atom and coordinate pairs. We used a VAE as a basis for our model, which allows to generate molecules in a one-shot fashion and explore molecular domains within a continuous low-dimensional representation of the chemical space. We achieved a high reconstruction precision of the atom coordinates, which is below 0.05 Å and is in the range of a typical spatial quantization error of common chemical descriptors. Furthermore, our model almost perfectly reconstructs exact chemical compositions and bond types on a QM9 test set.

References

[1] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*, 2019.

[2] Rafael Gomez-Bombarelli, Jennifer N. Wei, David Duvenaud, Jose Miguel Hernandez-Lobato, Benjamin Sanchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. In *ACS Central Science*. 2018.

[3] Moritz Hoffmann and Frank Noé. Generating valid euclidean distance matrices. *arXiv:1910.03131*, 2019.

[4] David Janz, Jos van der Westhuizen, Brooks Paige, Matt J Kusner, and José Miguel Hernández-Lobato. Learning a generative model for validity in complex discrete structures. *International Conference on Learning Representations*, 2018.

[5] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2018.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[7] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[8] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, 2017.

[9] Elman Mansimov, Omar Mahmood, Seokho Kang, and Kyunghyun Cho. Molecular geometry prediction using a deep generative graph neural network. *Scientific reports*, 9(1):1–13, 2019.

[10] Noel O'Boyle, Michael Banck, Craig James, Chris Morley, Tim Vandermeersch, and Geoffrey Hutchison. Open Babel: An open chemical toolbox. 2011.

[11] Noel M O'Boyle. Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. *Journal of cheminformatics*, 4(1):22, 2012.

[12] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. In *Scientific Data*. 2014.

[13] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[14] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk): an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 2003.

[15] James J. P. Stewart. MOPAC2016. *Stewart Computational Chemistry*, 2016.