A GENERATIVE MODEL FOR MOLECULAR DISTANCE GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Computing equilibrium states for many-body systems, such as molecules, is a long-standing challenge. In the absence of methods for generating statistically independent samples, great computational effort is invested in simulating these systems using, for example, Markov chain Monte Carlo. We present a probabilistic model that generates such samples for molecules from their graph representations. Our model learns a low-dimensional manifold that preserves the geometry of local atomic neighborhoods through a principled learning representation that is based on Euclidean distance geometry. We create a new dataset for molecular conformation generation with which we show experimentally that our generative model achieves state-of-the-art accuracy. Finally, we show how to use our model as a proposal distribution in an importance sampling scheme to compute molecular properties.

1 INTRODUCTION

Over the last few years many highly-effective deep learning methods generating small molecules with desired properties (e.g., novel drugs) have emerged (Gómez-Bombarelli et al., 2018; Segler et al., 2018; Dai et al., 2018; Jin et al., 2018; Bradshaw et al., 2019a; Liu et al., 2018; You et al., 2018; Bradshaw et al., 2019b). However, before any candidate molecule is synthesized, some of the following questions first need to be answered: What are the possible three-dimensional (3D) structures, also known as a *conformations*, of this molecule? And, if it is a drug, is it flexible enough to fit into the active site of the target protein? What are the properties (e.g., solubility or polarizability) of this molecule at a given temperature? All relevant conformations of this molecule need to be known to answer any of these questions (which are illustrated in Fig. 1) with computational methods.

Under a wide range of conditions, the probability $p(\mathbf{x})$ of a conformation $\mathbf{x}$, where $\mathbf{x}$ is the set of elements and Cartesian coordinates of the atoms, is governed by the Boltzmann distribution and is proportional to $\exp\{-E(\mathbf{x})/k_BT\}$, where $E(\mathbf{x})$ is its energy, k_B is the Boltzmann constant, and T is the temperature. To compute the expectation of some property $\mathcal{O}$ for a given molecule $\mathcal{G}$, one needs to average over all its conformations $\mathcal{M}_\mathcal{G}$: $\mathbb{E}_\mathcal{G}[\mathcal{O}] = \int_{\mathcal{M}_\mathcal{G}} \mathcal{O}(\mathbf{x})p(\mathbf{x})d\mathbf{x}$.

As the enumeration of all possible configurations is infeasible for all but the simplest molecules, one must instead attempt to sample them from their equilibrium distribution. The main approach is to start with one conformation and make small changes to it over time, e.g., by using Markov chain Monte Carlo (MCMC) or molecular dynamics (MD). These methods can be used to sample equilibrium states of molecules, but they become computationally expensive for large compounds (Shim & MacKerell, 2011; Ballard et al., 2015; De Vivo et al., 2016). Other stochastic approaches based on Euclidean distance geometry (EDG) exist in which distances between atoms are set to fixed idealized values (Havel, 2002; Blaney & Dixon, 2007). However, atomic distances are highly context and temperature dependent. Several methods based on statistical learning have also recently been developed to tackle the issue of generating conformations. However, they are geared towards proteins and their folding dynamics (Evans et al., 2018; AlQuraishi, 2019; Lemke & Peter, 2019; Ingraham et al., 2019; Noé et al., 2019). As a result, these models are not targeting a distribution over conformation but the most stable folded configuration. Further, they are either not transferable between molecules or are only applicable to linear ones.

This work includes the following key contributions:

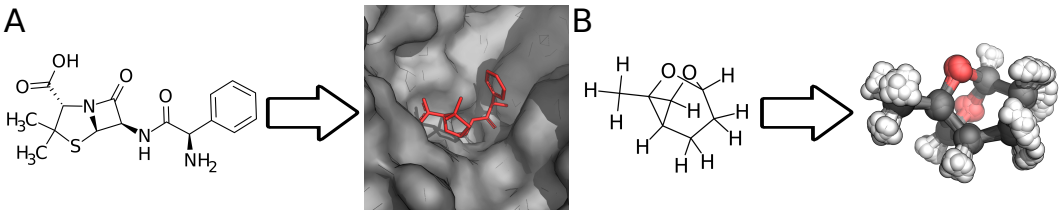

Figure 1: Two examples of the importance of conformations in molecular science. (A) 3D structure prediction of a drug in the active site of a protein (PDB 3A3I). (B) Calculation of molecular properties at a certain temperature. On the right, an overlay of multiple conformations is shown. Hydrogen (H), carbon (C), and oxygen (O) atoms are colored white, gray, and red, respectively.

- We introduce a novel probabilistic model for learning conformational distributions of molecules from their graph representations with graph neural networks.
- By combining a conditional variational autoencoder (CVAE) with an EDG algorithm we present the first work on sampling molecular conformations for molecules of arbitrary size and shape based on deep learning.
- We create a new, challenging benchmark dataset CONF17 for conformation generation, which is made publicly available.
- We develop a rigorous experimental approach for evaluating and comparing the accuracy of conformation generation methods based on the mean maximum deviation distance metric.
- We show how this generative model can be used as a proposal distribution in an importance sampling (IS) scheme to estimate molecular properties.

2 METHOD

Our goal is to build a statistical model that samples from a distribution over conformations of a molecule given its graph representation. The resulting samples could then be used to estimate molecular properties. First, we describe how a molecule's conformation can be represented by a set of pairwise distances between atoms and why this presentation is advantageous over one in Cartesian coordinates (Section 2.1). Second, we present a generative model in Section 2.2 that will sample from this distribution by generating sets of distances for a given molecular graph. Third, we explain in Section 2.3 how a set of predicted distances can be transformed into a molecular conformation in Cartesian coordinates with an EDG algorithm and why this transformation is necessary. Finally, we detail in Section 2.4 how our generative model can be used as a proposal distribution in an important sampling integration scheme to estimate molecular properties.

2.1 MOLECULAR GRAPHS AND DISTANCE GEOMETRY

In this study, a molecule is represented by an undirected graph which is defined as a tuple $\mathcal{G} = (V, E)$. $V = \{v_i\}_{i=1}^{N_v}$ is the set of nodes, where each v_i holds a node's attributes. Here, V represents the atoms in the molecule, with attributes for the element type. $E = \{(e_k, r_k, s_k)\}_{k=1}^{N_e}$ is the set of edges, where each e_k holds an edge's attributes, and r_k and s_k are the nodes an edge is connecting. Here, E represents the molecular bonds (and the auxiliary edges which are explained below) in the molecule.

We assume that a conformation can be represented by a set of distances $\mathbf{d} = \{d_k\}_{k=1}^{N_e}$, where $d_k \in \mathbb{R}$ is the distance between the atoms r_k and s_k in Cartesian coordinates. As the set of edges between the bonded atoms (E_{bond}) alone would not suffice to describe a conformation, we expand the graph by adding *auxiliary* edges. Auxiliary edges between atoms that are second neighbors in the original graph fix angles between atoms, and those between third neighbors fix dihedral angles (denoted E_{angle} and E_{dihedral}, respectively). In this work, E_{angle} consists of edges between all second neighbors in the original graph. Edges between third neighbors are added according to a heuristic (see Appendix A.1). This process of extending the graph is illustrated in Fig. 2, (A).

A key advantage of a representation in terms of distances is its invariance to rotation and translation; by contrast, Cartesian coordinates depend on the (arbitrary) choice of origin, for example. In addition, it reflects pair-wise physical interactions and their generally local nature. Auxiliary edges can be placed between higher-order neighbors depending on how far the physical interactions dominating the potential energy of the system reach.

For each molecule, we have a set of N_{GT} conformational samples $\{\mathbf{x}_j\}_{j=1}^{N_{\text{GT}}}$ from the ground-truth distribution, resulting in N_{GT} sets of distances $\{\mathbf{d}_j\}_{j=1}^{N_{\text{GT}}}$. In addition, we have a set of N graphs $\{\mathcal{G}_l\}_{l=1}^{N}$, each with $N_{\text{GT}}^{(l)}$ conformational samples. With this data, we will train a generative model which we detail in the following section.

2.2 Generative Model

We employ a CVAE (Kingma & Welling, 2014; Pagnoni et al., 2018) to model the distribution over distances $\mathbf{d}$ in a molecule. A CVAE first encodes the augmented graph representation of the molecule together with a set of distances into a latent space $\mathbf{z} \in \mathbb{R}^{N_v}$ with an encoder $q_\phi(\mathbf{z}|\mathbf{d}, \mathcal{G})$. Subsequently, the decoder $p_\theta(\mathbf{d}|\mathbf{z}, \mathcal{G})$ decodes $\mathbf{z}$ back into a set of distances. The CVAE model is illustrated in Fig. 2.2 (B), a graphical model is shown in Fig. 2.2 (C).

The size of the latent space plays a central role in a VAE. A conformation has, in general, $3N_v - 6$ degrees of freedom: one degree of freedom per spacial dimension per atom minus three translational and three rotational degrees of freedom. Since it is the role of the encoder to compress the conformation into a lower-dimensional space, we opted for a latent space that is equal to the number of atoms in the molecule but smaller than $3N_v$ to ensure compression is taking place.

Here, $q_\phi(\mathbf{z}|\mathbf{d}, \mathcal{G})$ and $p_\theta(\mathbf{d}|\mathbf{z}, \mathcal{G})$ are normal distributions, the mean and variance of which are modeled by two graph neural networks (GNN), a well-studied technique that achieves state-of-the-art performance in representation learning for molecules (Kipf & Welling, 2017; Duvenaud et al., 2015; Kearnes et al., 2016; Schütt et al., 2017b; Gilmer et al., 2017; Kusner et al., 2017; Bradshaw et al., 2019a). In particular, we use a message passing neural network (MPNN) (Gilmer et al., 2017) with a multi-head attention mechanism (Veličković et al., 2018).

Below, we give a high-level overview of the model implementation; the full model (including all parameters) is available online.[1]. In the encoder $q_\phi(\mathbf{z}|\mathbf{d}, \mathcal{G})$, the graph $\mathcal{G}$ and the set of distances $\mathbf{d}$ are passed to a function F_{enc}, where each entry in $\mathbf{d}$ is concatenated with the respective edge feature. Subsequently, the edge and node features are individually transformed by multilayer perceptrons (MLPs) to generate new features forming the latent graph $\mathcal{G}_{\text{enc}}^{(0)}$. Then, the message passing scheme, denoted by MPNN_{enc}, is applied T times to obtain $\mathcal{G}_{\text{enc}}^{(T)} = (V_{\text{enc}}^{(T)}, E_{\text{enc}}^{(T)})$. Finally, the read-out function R_{enc} (an MLP) takes each $v_{\text{i, enc}}^{(T)}$ to generate a Gaussian distribution for the corresponding $\mathbf{z}_i$. The so-called reparametrization trick is employed to sample from $\mathbf{z}$. In summary,

$$\mathcal{G}_{\text{enc}}^{(0)} = F_{\text{enc}}(\mathcal{G}, \mathbf{d}), \quad \mathcal{G}_{\text{enc}}^{(t+1)} = \text{MPNN}_{\text{enc}}(\mathcal{G}_{\text{enc}}^{(t)}), \quad \mathbf{z} = R_{\text{enc}}(V_{\text{enc}}^{(T)}). \tag{1}$$

In the decoder $p_\theta(\mathbf{d}|\mathbf{z}, \mathcal{G})$, each entry in $\mathbf{z}$ is appended to the node features of the graph $\mathcal{G}$ and subsequently, the resulting node and edge features are individually transformed by MLPs to give the latent graph $\mathcal{G}_{\text{dec}}^{(0)}$. These operations are implemented in F_{dec}. The message passing scheme MPNN_{dec} is then applied T times. Finally, the read-out function R_{dec} (an MLP) returns the mean and variance of $\hat{\mathbf{d}}$ based on the edge features $E_{\text{dec}}^{(T)}$.

$$\mathcal{G}_{\text{dec}}^{(0)} = F_{\text{dec}}(\mathcal{G}, \mathbf{z}), \quad \mathcal{G}_{\text{dec}}^{(t+1)} = \text{MPNN}_{\text{dec}}(\mathcal{G}_{\text{dec}}^{(t)}), \quad \hat{\mathbf{d}} = R_{\text{dec}}(E_{\text{dec}}^{(T)}). \tag{2}$$

The sets of parameters in the encoder and decoder, ϕ and θ, respectively, are optimized by maximizing the evidence lower bound (ELBO):

$$L = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{d}, \mathcal{G})}[\log p_\theta(\mathbf{d}|\mathbf{z}, \mathcal{G})] - D_{\text{KL}}[q_\phi(\mathbf{z}|\mathbf{d}, \mathcal{G})||p_\theta(\mathbf{z}|\mathcal{G})], \tag{3}$$

where the prior $p_\theta(\mathbf{z}|\mathcal{G})$ consists of factorized Gaussians. The optimal values for the hyperparameters for the network dimensions, number of message passes, batch size, and learning rate of the Adam optimizer (Kingma & Ba, 2014) were tuned by maximizing the validation performance (ELBO) with a Bayesian optimizer.

[1] `https://figshare.com/s/1b42bf865bd78c457354`

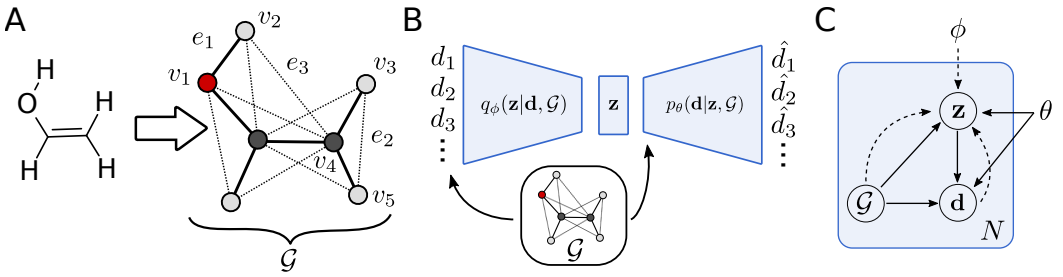

Figure 2: (A) Left: Structural formula of a molecule. Right: Augmentation of the molecular graph consisting of nodes representing atoms (circles, e.g., v_1) and edges representing molecular bonds (solid lines, e.g., $e_1 \in E_{\text{bond}}$) by auxiliary edges (dotted lines, e.g., $e_2 \in E_{\text{angle}}$ and $e_3 \in E_{\text{dihedral}}$). (B) Diagram of variational autoencoder for distances $\mathbf{d}$ between atoms in a conformation of molecule $\mathcal{G}$. (C) Graphical model of variational autoencoder: generative model $p_\theta(\mathbf{z}|\mathcal{G})p_\theta(\mathbf{d}|\mathbf{z},\mathcal{G})$ (solid lines) and variational approximation $q_\phi(\mathbf{z}|\mathbf{d},\mathcal{G})$ (dashed lines).

2.3 Conformation Generation through Euclidean Distance Geometry

To compute molecular properties, quantum-chemical methods need to be employed which require the input, i.e., the molecular system, to be in Cartesian coordinates.[2] Therefore, we use an EDG algorithm to translate the set of distances $\mathbf{d}$ (together with the chirality constraints) to a conformation $\mathbf{x}$ in Cartesian coordinates.

EDG is the mathematical basis for a geometric theory of molecular conformation. In the field of machine learning, Weinberger & Saul (2006) used it for learning image manifolds, Tenenbaum et al. (2000) for image understanding and handwriting recognition, Jain & Saul (2004) for speech and music, and Demaine et al. (2009) for music and musical rhythms. An EDG description of a molecular system consists of a list of lower and upper bounds on the distances between pairs of atoms $\{(d_{i,\text{min}}, d_{i,\text{max}})\}$ (chirality constraints are fixed, and thus not modeled by our method). Here, $p_\theta(\mathbf{d}|\mathbf{z},\mathcal{G})$ is used to model these bounds, namely we set the bounds to $\{(\mu[d_k] - \sigma[d_k], \mu[d_k] + \sigma[d_k])\}$, where $\mu[d_k]$ and $\sigma[d_k]$ are the mean and standard deviation for each distance d_k given by the CVAE. Then, an EDG algorithm determines Cartesian coordinates for each atom so that these bounds are fulfilled (see Appendix A.2 for details). Together with the corresponding chemical elements, we thus obtain a conformation $\mathbf{x}$.

2.4 Calculation of Molecular Properties

To efficiently approximate $\mathbb{E}_\mathcal{G}[\mathcal{O}]$ we employ an IS integration scheme (Bishop, 2009) with our CVAE as the proposal distribution. We assume that we can readily evaluate the unnormalized probability of a conformation $\tilde{p}(\mathbf{x}) = \exp\{-E(\mathbf{x})/k_B T\}$, where the energy $E(\mathbf{x})$ is determined with a quantum-chemical method (e.g., density functional theory). Since the EDG algorithm is mapping the distribution $p_\theta(\mathbf{d}|\mathbf{z},\mathcal{G})$ to a point mass in $\mathbb{R}^{3N_v}$, the MC estimate for the resulting distribution $p_{\text{prop}}(\mathbf{x}|\mathcal{G})$ is given by a mixture of delta functions, each of which is centered at the $\mathbf{x}_i$ resulting from mapping $p_\theta(\mathbf{x}|\mathbf{z}_i,\mathcal{G})$ to $\mathbb{R}^{3N_v}$, where $\mathbf{z}_i\ p_\theta(\mathbf{z}|\mathcal{G})$, that is, $p_{\text{prop}}(\mathbf{x}|\mathcal{G}) \approx \frac{1}{N_{\text{sample}}} \sum_{i=1}^{N_{\text{sample}}} \delta(\mathbf{x} - \mathbf{x}_i)$. The IS estimator for the expectation of $\mathcal{O}$ w. r. t. $\tilde{p}(\mathbf{x})$ is

$$\hat{\mathbb{E}}_\mathcal{G}[\mathcal{O}] = \int_{\mathcal{M}_\mathcal{G}} \mathcal{O}(\mathbf{x})\tilde{p}(\mathbf{x})d\mathbf{x} = \int_{\mathcal{M}_\mathcal{G}} \mathcal{O}(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{p_{\text{prop}}(\mathbf{x}|\mathcal{G})} p_{\text{prop}}(\mathbf{x}|\mathcal{G})d\mathbf{x}. \tag{4}$$

The expectation of $\mathcal{O}$ w. r. t. the normalized version of $\tilde{p}(\mathbf{x})$ is then

$$\mathbb{E}_\mathcal{G}[\mathcal{O}] = \frac{\hat{\mathbb{E}}_\mathcal{G}[\mathcal{O}]}{\hat{\mathbb{E}}_\mathcal{G}[1]} \approx \frac{1}{Z_p} \sum_{i=1}^{N_{\text{sample}}} \mathcal{O}(\mathbf{x}_i)\tilde{p}(\mathbf{x}_i), \tag{5}$$

where $Z_p \approx \sum_{i=1}^{N_{\text{sample}}} \tilde{p}(\mathbf{x}_i)$. When dividing two delta functions we have assumed that they take some arbitrarily large finite value.

[2] Even though quantum-chemical methods require the input to be in Cartesian coordinates, calculated properties such as the energy are invariant to translation and rotation.

3 RELATED WORKS

The standard approach for generating molecular conformations is to start with one, and make small changes to it over time, e.g., by using MCMC or MD. These methods are considered the gold standard for sampling equilibrium states, but they are computationally expensive, especially if the system is large and the Hamiltonian is based on quantum-mechanical principles (Shim & MacKerell, 2011; Ballard et al., 2015; De Vivo et al., 2016).

A much faster but more approximate approach for conformation generation is EDG (Havel, 2002; Blaney & Dixon, 2007; Lagorce et al., 2009; Riniker & Landrum, 2015). Lower and upper distance bounds for pairs of atoms in a molecule are fixed values based on ideal bond lengths, bond angles, and torsional angles. These values are often extracted from crystal structure databases (Allen, 2002). These methods aim to generate a low-energy conformation, not to generate unbiased samples from the underlying distribution at a certain temperature.

There exist several machine learning approaches as well, however, they are mostly tailored towards studying protein dynamics. For example, Noé et al. (2019) trained Boltzmann generators on the energy function of proteins to provide unbiased, one-shot samples from their equilibrium states. This is achieved by training an invertible neural network to learn a coordinate transformation from a system's configurations to a latent space representation. Further, Lemke & Peter (2019) proposed a dimensionality reduction algorithm that is based on a neural network autoencoder in combination with a nonlinear distance metric to generate samples for protein structures. Both models learn protein-specific coordinate transformations that cannot be transferred to other molecules.

AlQuraishi (2019) introduced an end-to-end differentiable recurrent geometric network for protein structure learning based on amino acid sequences. Also, Ingraham et al. (2019) proposed a neural energy simulator model for protein structure that makes use of protein sequence information. In contrast to amino acid sequences, molecular graphs are, in general, not linear but highly branched and often contain cycles. This makes them unsuitable for recurrent networks.

Finally, Mansimov et al. (2019) presented a conditional deep generative graph neural network to generate molecular conformations given a molecular graph. Their goal is to predict the most likely conformation and not a distribution over conformations. Instead of encoding molecular environments in atomic distances, they work directly in Cartesian coordinates. As a result, the generated conformations showed significant structural differences compared to the ground-truth and required refinement through a force field, which is often employed in MD simulations.

We argue that our model has several advantages over the approaches reviewed above:

- It is a fast alternative to resource-intensive approaches based on MCMC or MD.
- Our principled representation based on pair-wise distances does not restrict our approach to any type of system (e.g., proteins) or any particular graph structure. In addition, it is extendable in a systematic fashion.
- Our model is transferable – it can extrapolate from only a few graphs to unseen ones.

4 THE CONF17 DATASET

The CONF17 dataset which we generated for this study is the first dataset for molecular conformations. It was derived from the ISO17 dataset (Schütt et al., 2017a) which consists of conformations of various molecules with the atomic composition $C_7H_{10}O_2$ drawn from the QM9 dataset (Ramakrishnan et al., 2014). These conformations were generated by *ab initio* molecular dynamics simulations at 500 Kelvin which generates trajectories of a single molecule covering a large variety of conformations. The CONF17 dataset consists of 127 distinct molecular graphs each with 3380 conformations on average. We split the dataset into multiple training and test splits, each consisting of 107 and 20 graphs, respectively (see Appendix A.1 for more details).[3]

In Fig. 3, (A), the structural formulae of a random selection of molecules from this dataset are shown. Most molecules feature highly-strained, complex 3D structures such as rings which are typical of

[3]The CONF17 dataset is publicly available online `https://figshare.com/s/1b42bf865bd78c457354`.

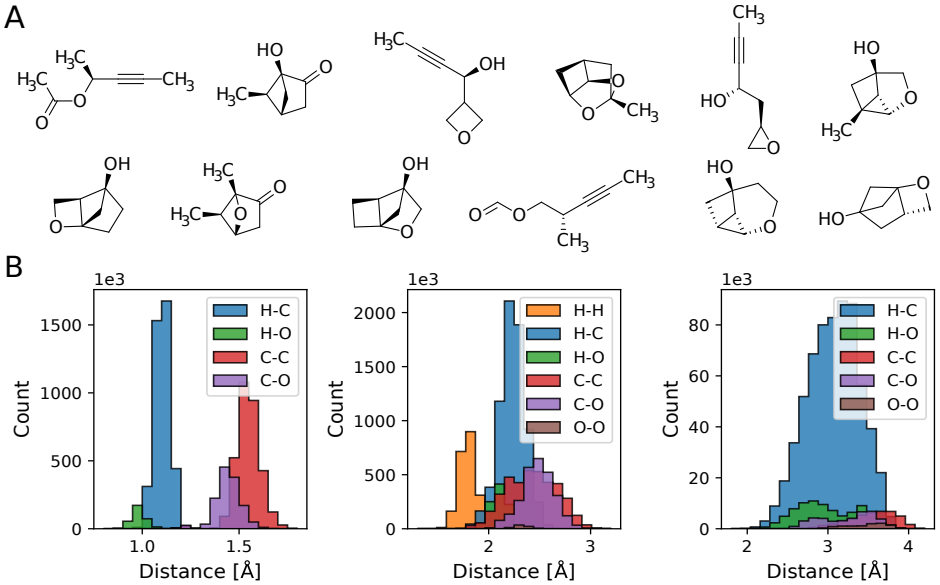

Figure 3: Overview of the CONF17 dataset. (A) Structural formulae of a random selection of molecules from the CONF17 dataset. (B) Distribution of distances (in Å) grouped by edge (from left to right: E_{bond}, E_{angle}, and E_{dihedral}) and vertex type (chemical element).

drug-like molecules. As a result, this is a challenging dataset in terms of structural complexity. In Fig. 3, (B), the frequency of distances (in Å) in the conformations are shown for each edge type. It can be seen that the marginal distributions of the edge distances are multimodal and highly context dependent.

5 EXPERIMENTS

We assess the performance of our method, named Graph Distance Geometry (GRAPHDG), by comparing it with two state-of-the-art methods for molecular conformation generation: RDKIT (Riniker & Landrum, 2015), a classical EDG approach, and DL4CHEM (Mansimov et al., 2019), a machine learning approach. We trained GRAPHDG and DL4CHEM on three different training and test splits. We generated 3000 conformations with each method for molecules in a test set.

5.1 DISTRIBUTIONS OVER DISTANCES

We assessed the accuracy of the distance distributions of RDKIT, DL4CHEM, and GRAPHDG by calculating the maximum mean discrepancy (MMD) (Gretton et al., 2012) with respect to the ground-truth distribution. We compute the MMD using a Gaussian kernel, where we set the standard deviation to be the median distance between distances $\mathbf{d}$ in the aggregate sample. For this, we determined the distances in the conformations from the ground-truth and those generated by RDKIT and DL4CHEM. In addition to the MMD of the joint distribution of distances between heavy atoms (element C and O, H atoms are usually ignored), we also calculated the MMDs between the marginals of individual distances $p(d_i|\mathcal{G})$ and pair-wise distances $p(d_i, d_j|\mathcal{G})$. The results are summarized in Table 1. It can be seen that the samples from GRAPHDG are significantly closer to the ground-truth distribution than the other methods. RDKIT is slightly worse than GRAPHDG while DL4CHEM seems to struggle with the complexity of the molecules and the small number of graphs in the training set.

In Fig. 4, we showcase the accuracy of our model to the ground-truth by plotting the marginal distributions $p(d_i|\mathcal{G})$ for distances between heavy atoms in a molecule from the test set. It can seen that RDKIT consistently underestimates the marginal variances. This is because this method aims to predict the most stable conformation, i.e., the distribution's mode. In contrast, DL4CHEM often fails to predict the correct mean. For this molecule, GRAPHDG is the most accurate, predicting the

Table 1: Assessment of the accuracy of the distributions over conformations generated by three models compared to the ground-truth. We compare the distributions with respect to the marginals $p(d_k|\mathcal{G})$, $p(d_k, d_l|\mathcal{G})$, and the distribution over all edges between heavy atoms (elements C and O) $p(\{d_k\}|\mathcal{G})$. Two different metrics are used: median MMD between ground-truth conformations and generated ones, and mean ranking (1 to 3) based on the MMD. Reported are the results for molecules from the test set, averaged over three train-test splits. Standard errors are given in brackets.

	Median MMD			Mean Ranking		
	RDKIT	DL4CHEM	GRAPHDG	RDKIT	DL4CHEM	GRAPHDG
$p(d_k\|\mathcal{G})$	0.55 (0.01)	1.11 (0.01)	**0.38** (0.02)	1.71 (0.03)	2.74 (0.02)	**1.51** (0.03)
$p(d_k, d_l\|\mathcal{G})$	0.53 (0.01)	1.09 (0.01)	**0.34** (0.01)	1.66 (0.02)	2.92 (0.01)	**1.43** (0.02)
$p(\{d_k\}\|\mathcal{G})$	0.60 (0.01)	1.07 (0.03)	**0.44** (0.05)	1.58 (0.05)	2.90 (0.05)	**1.45** (0.02)

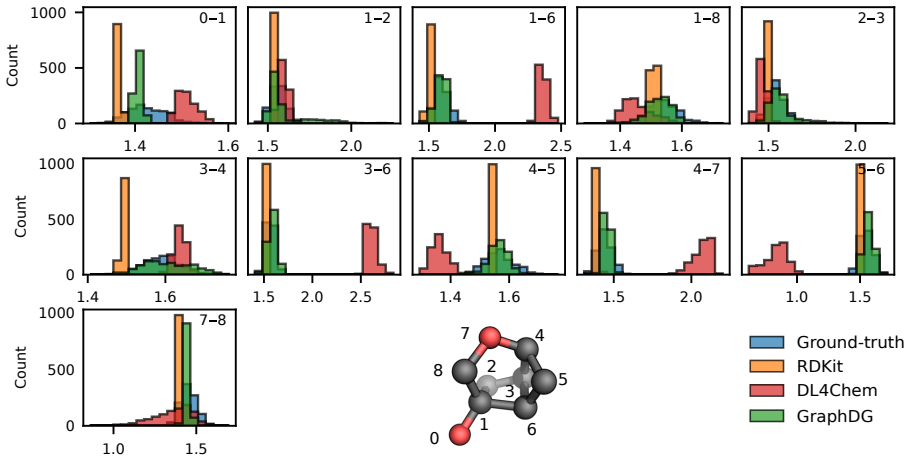

Figure 4: Marginal distributions $p(d_k|\mathcal{G})$ of ground-truth and predicted bond distances (in Å) between heavy atoms (elements C and O) in a molecule from the test set. The atoms connected by each edge d_k are indicated in each subplot (s_k–r_k). In the 3D structure of the molecule, carbon and oxygen atoms are colored gray and red, respectively. H atoms are omitted for clarity.

right mean and variance in most cases. Additional figures can be found in the Appendix A.4, where we also show plots for the marginal distributions $p(d_i, d_j|\mathcal{G})$.

5.2 GENERATION OF CONFORMATIONS

We passed the distances from our generative model to an EDG algorithm to obtain conformations. For 99.9% of the sets of distances, all triangle inequalities held. For 94% of the molecules, the algorithm succeeded which is 8 pp higher than the success rate we observed for RDKIT. In Fig. 5, an overlay of 50 conformations generated by the different methods is shown. It can be seen that RDKIT's conformations show too little variance, while DL4CHEM's structures are mostly invalid, which is due in part to its failure to predict the correct interatomic angles. Our method slightly overestimates the structural variance (see, for example, Fig. 5, top row, second column), but produces conformations that are the closest to the ground-truth.

5.3 CALCULATION OF MOLECULAR PROPERTIES

We estimate expected chemical properties for molecules from the test set with $N_{\text{sample}} = 50$ conformational samples. Due to their poor quality, we could not compute the energy $E(\mathbf{x})$ for conformations generated with DL4CHEM, and thus, this method is excluded from this analysis. In Table 2, it can be seen that RDKIT and GRAPHDG perform similarly well (see Appendix A.2 for computational details). However, both methods are still highly inaccurate for E_{elec} (in practice, an accuracy of less than 5 kJ/mol is required). Close inspection of the conformations shows that, even though GRAPHDG predicts the most accurate distances overall, the variances of certain strongly

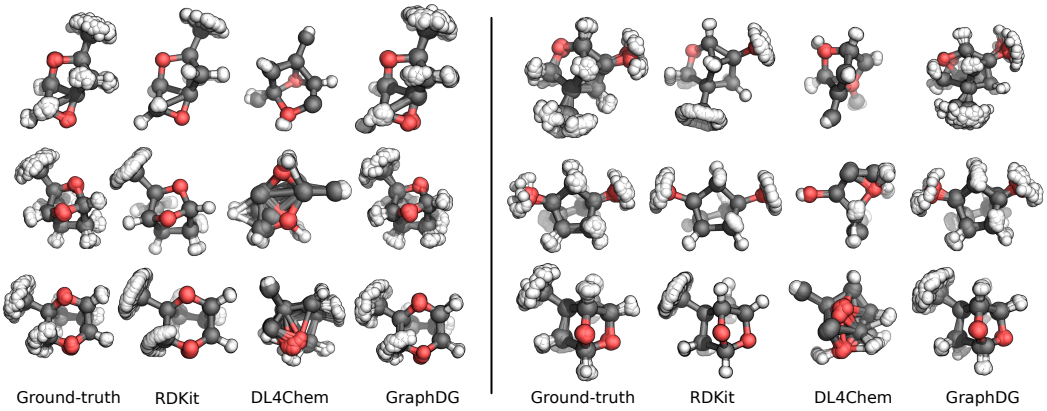

Figure 5: Overlay of 50 conformations from the ground-truth and three models of six random molecules from the test set. C, O, and H atoms are colored gray, red, and white, respectively.

Table 2: Median difference in average properties between ground-truth and RDKIT and GRAPHDG: total electronic energy E_{elec} (in kJ/mol), the energy of the HOMO and the LUMO ϵ_{LUMO} and ϵ_{LUMO}, respectively (in eV), and the dipole moment μ (in debye). Reported are the results for molecules from the test set, averaged over three train-test splits. Standard errors are given in brackets.

	RDKIT	GRAPHDG
E_{elec}	42.7 (4.3)	58.0 (21.0)
ϵ_{HOMO}	0.08 (0.04)	0.10 (0.05)
ϵ_{LUMO}	0.15 (0.03)	0.09 (0.05)
μ	0.29 (0.05)	0.33 (0.09)

constrained distances (e.g., triple bonds) are overestimated so that the energies of the conformations increase drastically.

6 LIMITATIONS

The first limitation of this work is that the CVAE can sample (with low probability) invalid sets of distances for which there exists no 3D structure. Second, the CONF17 dataset covers only a small portion of chemical space. Finally, a large set of auxiliary edges would be required to capture long range correlations (e.g., in proteins). Future work will address these points.

7 CONCLUSIONS

We presented GRAPHDG, a transferable, generative model that allows sampling from a distribution over molecular conformations. We developed a principled learning representation of conformations that is based on distances between atoms. Then, we proposed a new dataset set, CONF17, for molecular conformations. With this dataset, we show experimentally that conformations generated by GRAPHDG are closer to the ground-truth than those generated by other methods. Finally, we employ our model as a proposal distribution in an IS integration scheme to estimate molecular properties. While orbital energies and the dipole moments were predicted well, a larger and more diverse dataset will be necessary for meaningful estimates of electronic energies. Finally, our model could be trained on conformational distributions at different temperatures in a transfer learning-type setting.

REFERENCES

F. H. Allen. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci*, 58(3):380–388, 2002.

Mohammed AlQuraishi. End-to-End Differentiable Learning of Protein Structure. *Cell Systems*, 8 (4):292–301.e3, 2019.

Andrew J. Ballard, Stefano Martiniani, Jacob D. Stevenson, Sandeep Somani, and David J. Wales. Exploiting the potential energy landscape to sample free energy. *WIREs Comput. Mol. Sci.*, 5(3): 273–289, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 8 edition, 2009. ISBN 978-0-387-31073-2.

Jeffrey M. Blaney and J. Scott Dixon. Distance Geometry in Molecular Modeling. In *Reviews in Computational Chemistry*, pp. 299–335. John Wiley & Sons, Ltd, 2007. ISBN 978-0-470-12582-3.

John Bradshaw, Matt J. Kusner, Brooks Paige, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. In *International Conference on Learning Representations*, 2019a.

John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A Model to Search for Synthesizable Molecules. *arXiv:1906.05221*, 2019b.

Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. In *International Conference on Learning Representations*, 2018.

Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.*, 59(9):4035–4061, 2016.

Erik D. Demaine, Francisco Gomez-Martin, Henk Meijer, David Rappaport, Perouz Taslakian, Godfried T. Toussaint, Terry Winograd, and David R. Wood. The distance geometry of music. *Computational Geometry*, 42(5):429–454, 2009.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2224–2232. Curran Associates, Inc., 2015.

R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. F. G. Green, C. Qin, A. Zidek, A. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, and A. W. Senior. De novo structure prediction with deep-learning based scoring. In *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction*, 2018.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1263–1272, 2017.

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. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.*, 4(2):268–276, 2018.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Wolfgang Guba, Agnes Meyder, Matthias Rarey, and Jérôme Hert. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *J. Chem. Inf. Model.*, 56(1):1–5, 2016.

Timothy F. Havel. Distance Geometry: Theory, Algorithms, and Chemical Applications. In *Encyclopedia of Computational Chemistry*. American Cancer Society, 2002. ISBN 978-0-470-84501-1.

John Ingraham, Adam Riesselman, Chris Sander, and Debora Marks. Learning Protein Structure with a Differentiable Simulator. In *International Conference on Learning Representations*, 2019.

V. Jain and L. K. Saul. Exploratory analysis and visualization of speech and music by locally linear embedding. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pp. iii–984, 2004.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2323–2332, Stockholm, Sweden, 2018. PMLR.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput.-Aided Mol. Des.*, 30(8):595–608, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations*, 2017.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar Variational Autoencoder. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1945–1954, International Convention Centre, Sydney, Australia, 2017. PMLR.

David Lagorce, Tania Pencheva, Bruno O. Villoutreix, and Maria A. Miteva. DG-AMMOS: A New tool to generate 3D conformation of small molecules using Distance Geometry and Automated Molecular Mechanics Optimization for in silico Screening. *BMC Chem. Biol.*, 9:6, 2009.

Tobias Lemke and Christine Peter. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.*, 2019.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained Graph Variational Autoencoders for Molecule Design. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7795–7804. Curran Associates, Inc., 2018.

Elman Mansimov, Omar Mahmood, Seokho Kang, and Kyunghyun Cho. Molecular geometry prediction using a deep generative graph neural network. *arXiv:1904.00314 [physics, stat]*, 2019.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Artidoro Pagnoni, Kevin Liu, and Shangyan Li. Conditional Variational Autoencoder for Neural Machine Translation. *arXiv:1812.04405*, 2018.

John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.*, 105(22):9982–9985, 1996.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 1:140022, 2014.

Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 991–1001. Curran Associates, Inc., 2017a.

Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8:13890, 2017b.

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.*, 4(1): 120–131, 2018.

Jihyun Shim and Alexander D. MacKerell, Jr. Computational ligand-based rational design: Role of conformational sampling and force fields in model development. *Med. Chem. Commun.*, 2(5): 356–370, 2011.

Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: The Python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.*, 8(1), 2018.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

Florian Weigend. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.*, 8(9): 1057–1065, 2006.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming. *Int. J. Comput. Vision*, 70(1):77–90, 2006.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6410–6421. Curran Associates, Inc., 2018.

A Appendix

A.1 Conf17 Dataset

A.1.1 Dataset Generation

The ISO17 dataset (Schütt et al., 2017a) was processed in the following way. First, conformations in which the molecular connectivity was modified (i.e., bonds were broken or new ones are formed) were discarded. Second, the molecular graphs were augmented by adding auxiliary edges for reasons described in Section 2.1. Auxiliary edges between all second neighbors were added. This can lead to a slight over-specification of the system's geometry, however, this did not pose a problem in our experiments. In addition, auxiliary edges between third neighbors were added to fix dihedral angles. Since there are potentially many ways of specifying a dihedral angle in a molecular system, we resorted to the works of Riniker & Landrum (2015) and Guba et al. (2016) to decide where to place edges between third neighbors.

A.1.2 Input Features

Below we list the node and edges features in the Conf17 dataset.

Table 3: Node features.

Feature	Data Type	Dimension
atomic number	integer	1
chiral tag	one-hot (R, S, and N/A)	3

Table 4: Edge features.

Feature	Data Type	Dimension
kind	one-hot (indicating whether e is in E_{bond}, E_{angle}, or E_{dihedral})	3
stereo chemistry	one-hot (E, Z, Any, None, and N/A)	5
type	integer (single, double, triple or N/A)	1
is aromatic	binary	1
is conjugated	binary	1
is in ring of size	one-hot (3, 4, . . . , 9) and N/A	8

A.2 Computational Details

A.2.1 Quantum-Chemical Calculations

All quantum-chemical calculations were carried out with the `PySCF` program package (version 1.5) (Sun et al., 2018) employing the exchange-correlation density functional PBE (Perdew et al., 1996), and the def2-SVP (Weigend & Ahlrichs, 2005; Weigend, 2006) basis set.

Conformations generated by DL4CHEM did not succeed as some atoms were too close to each other. Self-consistent field algorithms in quantum-chemical software such as `PySCF` do not converge for such molecular structures.

With quantum-chemical methods we calculate a number of properties that concern the states of the electrons in the conformation. These are the total electronic energy E_{elec}, the energy of the electron in the highest occupied molecular orbital (HOMO in eV) ϵ_{HOMO}, the energy of the lowest unoccupied molecular orbital (LUMO in eV) ϵ_{LUMO}, and the norm of the dipole moment μ (in debye).

A.2.2 Eucledian Distance Geometry

We refer the reader to Havel (2002) for theory on EDG, algorithms, and chemical applications. In summary, the EDG procedure consists of the following three steps:

1. Bound smoothing: extrapolating a complete set of lower and upper limits on all the distances from the sparse set of lower and upper bounds.
2. Embedding: choosing a random distance matrix from within these limits, and computing coordinates that are a certain best-fit to the distances.
3. Optimization: optimizing these coordinates versus an error function which measures the total violation of the distance (and chirality) constraints.

We use the EDG implementation found in RDKIT (Riniker & Landrum, 2015) with default settings.

A.3 Generation of Conformations

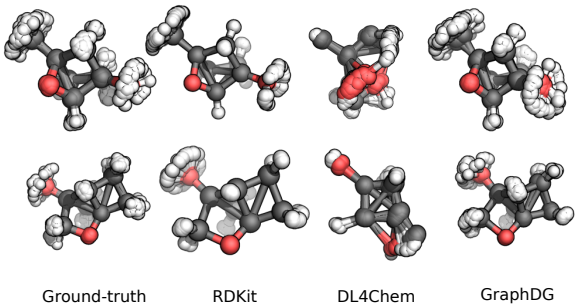

Figure 6: Overlay of 50 conformations from the ground-truth, RDKIT, DL4CHEM, and GRAPHDG of two random molecules from the test set. C, O, and H atoms are colored gray, red, and white, respectively.

A.4 DISTRIBUTIONS OVER DISTANCES

Below, the marginal distributions of the distances in a variety of molecules is shown.

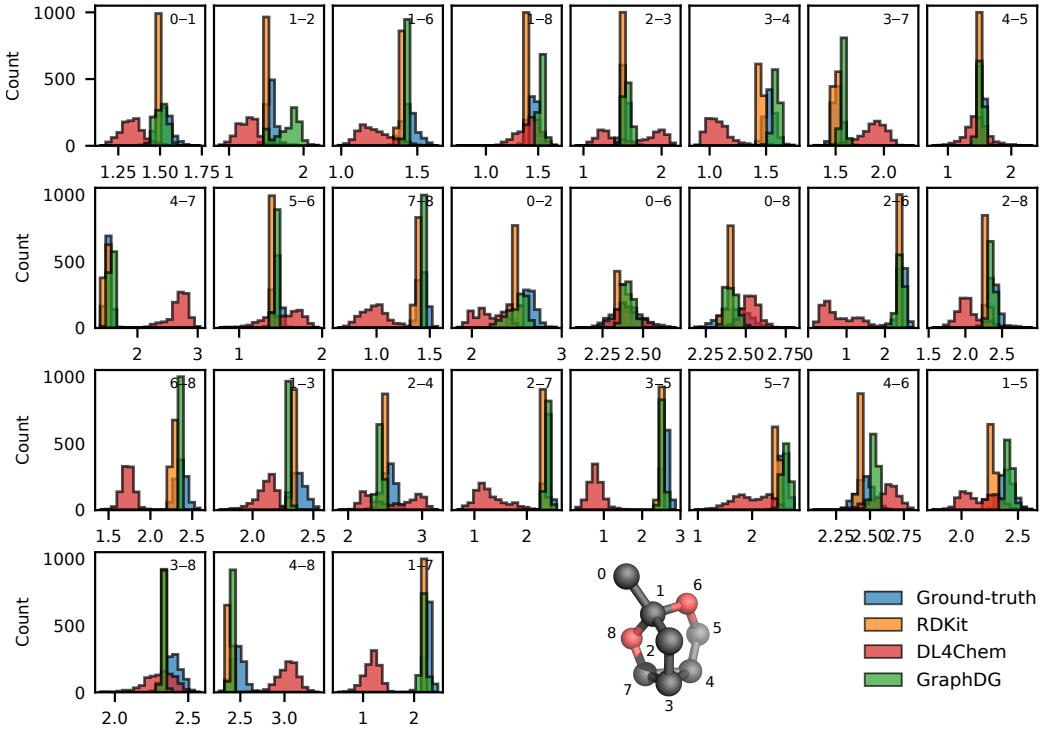

Figure 7: Marginal distributions $p(d_k|\mathcal{G})$ of ground-truth and predicted distances (in Å) between heavy atoms (elements C and O) in a molecule from the test set. The atoms connected by each edge d_k are indicated in each subplot (s_k–r_k). In the 3D structure of the molecule, carbon and oxygen atoms are colored gray and red, respectively. H atoms are omitted for clarity.

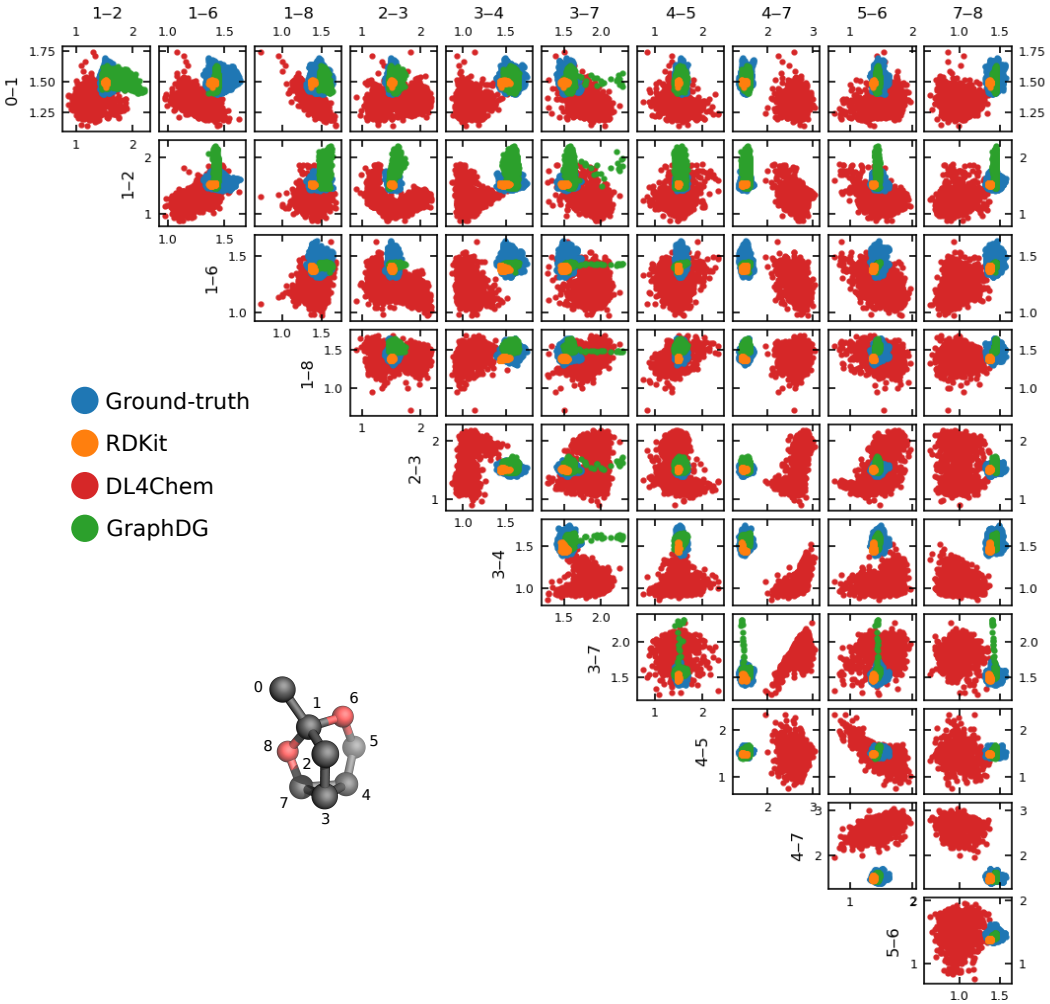

Figure 8: Marginal distributions $p(d_i, d_j|\mathcal{G})$ of ground-truth and predicted distances in a molecule (sample 3D structure shown for illustration) from the test set (in Å). Here, d_i and d_j are restricted to edges representing bonds between heavy atoms (elements C and O). In the 3D structure of the molecule, carbon and oxygen atoms are colored gray and red, respectively. H atoms are omitted for clarity.

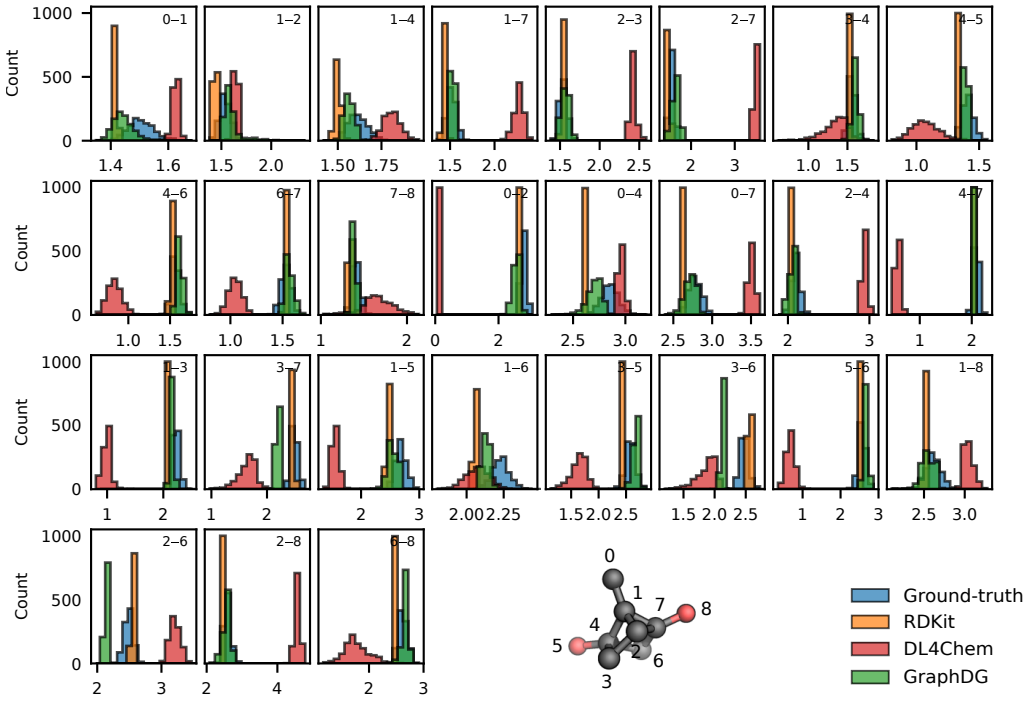

Figure 9: See caption of Fig. 7

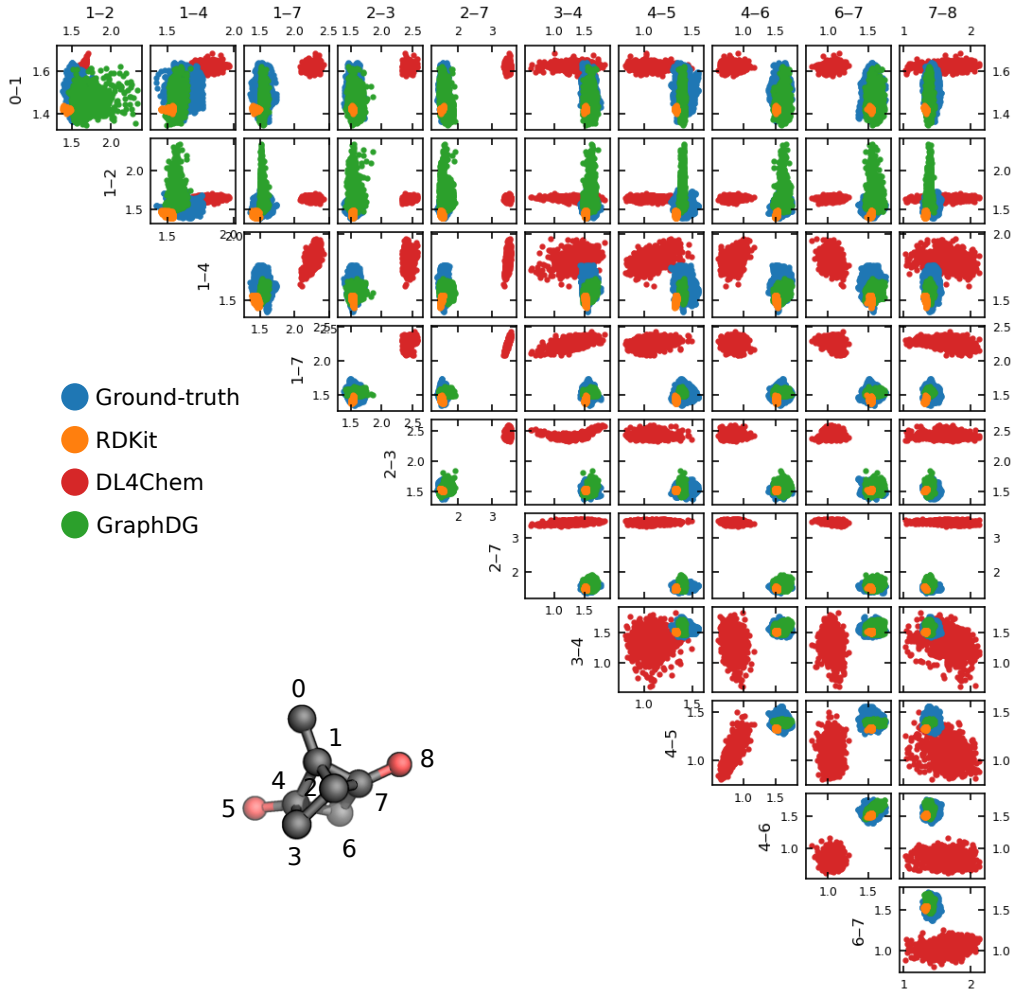

Figure 10: See caption of Fig. 8

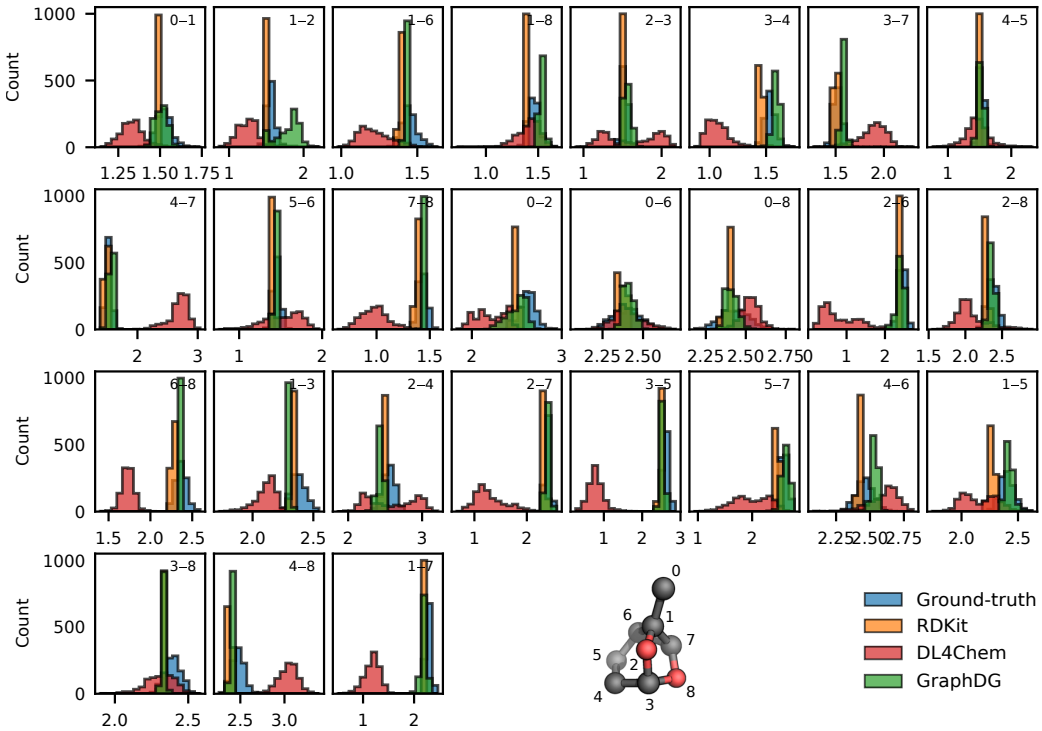

Figure 11: See caption of Fig. 7

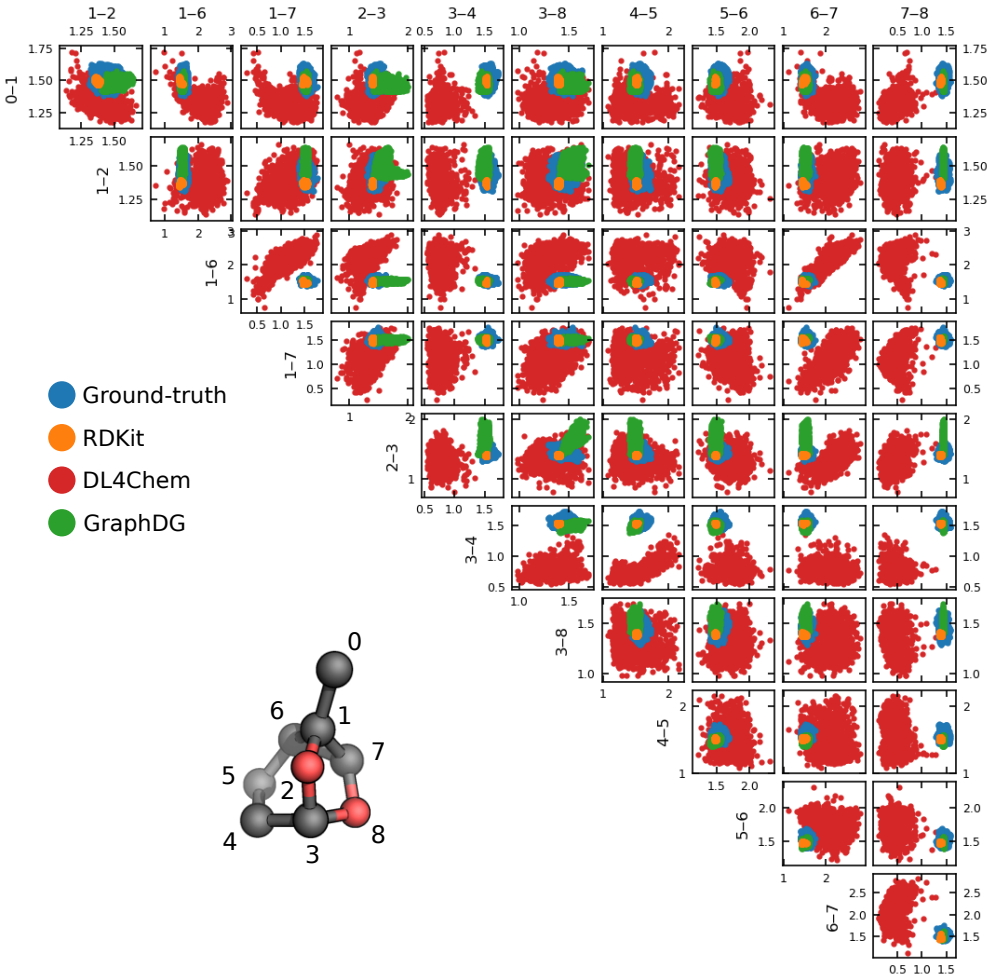

Figure 12: See caption of Fig. 8

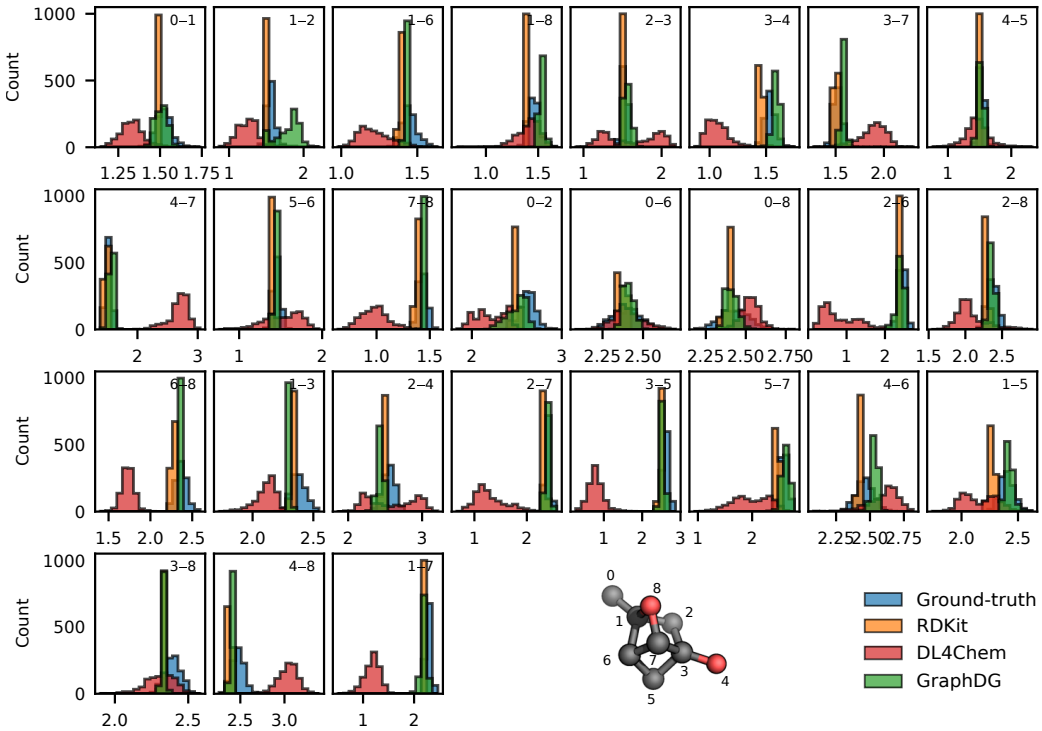

Figure 13: See caption of Fig. 7

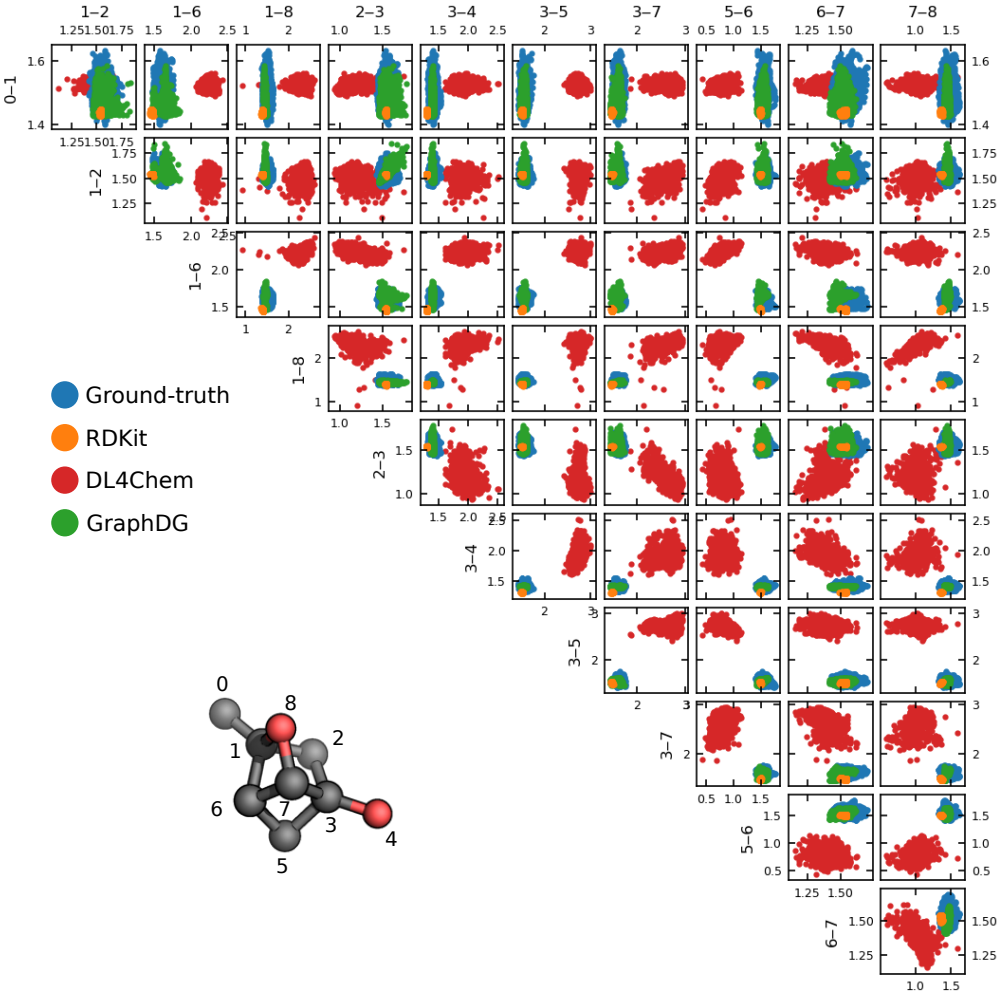

Figure 14: See caption of Fig. 8