RGCVAE: Relational Graph Conditioned Variational Autoencoder for Molecule Design

Supplementary Material

Davide Rigoni[1*], Nicolò Navarin[2] and Alessandro Sperduti[2]

[1]Molecular Modelling Section, Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Via Marzolo 5, Padova, 35131, Padova, Italy.
[2]Department of Mathematics, University of Padua, Via Trieste 63, Padova, 35131, Padova, Italy.

*Corresponding author(s). E-mail(s): davide.rigoni.1@unipd.it;
Contributing authors: nnavarin@math.unipd.it;
sperduti@math.unipd.it;

Abstract

This document contains the Appendices of the article *RGCVAE: Relational Graph Conditioned Variational Autoencoder for Molecule Design* submitted to the *Machine Learning journal, 2023.*

Keywords: RGCVAE, Relational GIN, VAE, Molecules Generation, Graph Generation

Table A1 Statistics about the QM9 and ZINC datasets.

Dataset	#Molecules	#Atoms	#Atom Types	#Bond Types
QM9	134K	9	4	3
ZINC	250K	38	9	3

Appendix A Dataset Details

As presented in the main article, Table A1 synthetically reports the main characteristics of the QM9 and ZINC datasets used for the evaluations. Both use the same number of bond types, but they differ in the number of atom types, i.e., 4 for QM9 and 9 for ZINC, and atom per molecule, i.e., at a maximum of 9 for QM9 and 38 for ZINC.

Appendix B Evaluating Molecule Properties

We have performed additional evaluations on the presented RGCVAE. Specifically, we have explored the properties of the molecules generated by the model as done for (Rigoni, Navarin, & Sperduti, 2020a, 2020b) using some common metrics used in this area of research:

1. *Natural Product (NP)* (Ertl, Roggo, & Schuffenhauer, 2008) which indicates how much the generated molecules' structural space is similar to the one covered by natural products;
2. *Solubility (Sol.)* which indicates how much a molecule is soluble in water, an important property for drugs;
3. *Synthetic Accessibility Score (*SAS*)* (Ertl & Schuffenhauer, 2009) which represents how easy (0) or difficult (100) it is to synthesize a molecule. This metric holds significant importance in acquiring an initial estimation of the synthesis complexity associated with the molecule; and
4. *Quantitative Estimation Drug-likeness (*QED*)* (Bickerton, Paolini, Besnard, Muresan, & Hopkins, 2012) which indicates in percentage how likely it is that the molecule is a good candidate to become a drug.

Table B2 presents the results obtained on the QM9 dataset, while Table B3 presents the results obtained on the ZINC dataset. The tables report the evaluations of the properties and the number of epochs used for the training of each model. Moreover, the last row reports the average values of the properties for each dataset.

On the QM9 dataset, the properties evaluation performed on the molecules generated by both RGCVAE - joint and RGCVAE - 2 stages, returned results close to those calculated on the dataset, supporting again the fact that our model does learn well the probability distributions of the input dataset molecules. Precisely, our model trained with the two-stage approach seems to generate molecules whose properties are slightly more similar to those in the dataset than those obtained by the model jointly trained. In fact, the RGCVAE - 2 stages model presents an *NP* value very close to that of the dataset.

Table B2 Average and standard deviation (where applicable) of different evaluation metrics on the QM9 dataset.

Model	↑%NP	↑%Sol.	↓%SAS	↑%QED	T.*Epoch	N.Epochs
Graph VAE	94.71 ±10.82	35.92 ±13.49	29.72 ±28.27	48.25 ±9.53	00h:01m	200
Regularized GVAE	95.77 ±9.26	39.38 ±14.52	30.58 ±24.69	48.79 ±7.83	00h:01m	150
CGVAE	93.80 ±5.62	28.62 ±12.38	10.28 ±16.11	47.91 ±7.04	00h:21m	10
CCGVAE	96.13 ±8.64	35.58 ±11.91	17.08 ±22.96	46.62 ±7.51	01h:30m	10
RGCVAE - joint	86.93 ±13.68	25.4 ±14.03	11.17 ±20.34	40.92 ±9.55	00h:01m	200
RGCVAE - 2 stages	88.76 ±13.03	24.22 ±14.24	13.59 ±22.97	40.32 ±9.88	00h:01m	200
QM9 Scores	88.52 ±17.75	27.91 ±13.76	21.86 ±22.88	46.12 ±7.76		

Table B3 Average and standard deviation (where applicable) of different evaluation metrics on the ZINC dataset.

Model	↑%NP	↑%Sol.	↓%SAS	↑%QED	T.*Epoch	N.Epochs
Character VAE	80.82 ±12.83	29.60 ±17.60	31.11 ±30.14	38.70 ±10.63	00h:07m	100
Grammar VAE	80.99 ±11.40	50.24 ±33.65	26.75 ±33.14	25.42 ±14.91	00h:21m	100
Syntax Directed VAE	77.84 ±19.76	55.94 ±27.51	14.46 ±24.14	39.45 ±20.98	00h:24m	500
Junction Tree VAE	52.20 ±17.12	48.06 ±18.48	44.74 ±24.39	75.05 ±13.40	07h:55m	10
CGVAE	81.38 ±15.98	57.76 ±20.04	16.25 ±21.63	65.14 ±16.39	15h:30m	3
CCGVAE	94.28 ±10.08	63.54 ±20.11	19.95 ±23.10	52.41 ±16.52	21h:30m	3
RGCVAE - joint	86.62 ±13.13	40.74 ±12.37	36.59 ±32.22	43.01 ±7.56	00h:12m	200
RGCVAE - 2 stages	92.72 ±10.21	42.20 ±15.10	25.72 ±26.30	45.79 ±10.10	00h:12m	200
ZINC Scores	42.08 ±18.37	56.11 ±17.44	55.95 ±22.90	73.18 ±13.86		

When considering the ZINC dataset, both RGCVAE - joint and RGCVAE - 2 stages are still showing promising results according to those calculated on the dataset. However, *NP* and *QED* are far from the values exhibited by the molecule in the dataset, probably because of the high *Diversity* of its generated molecules (see Table 3 in the main paper). Consequently, since our models can generate substructures dissimilar from those in training while presenting good *Reconstruction* values, the generated molecules are likely to present also dissimilar properties. This behavior can be adjusted reducing the expressiveness of the model. We will explore this aspect in future works. On the contrary, Junction Tree VAE that have low *Diversity*, presents results closer to those of the dataset.

Table B4 follows the format of Table 5 of the main manuscript, reporting the property metrics results obtained in the ablation study, using either representation 1 or 3. For the sake of interpretability, we report in the first column of the table the numerical IDs corresponding to the three atom representations:

1. atom type, e.g., "**C**" for carbon;
2. atom type, total valence, and formal charge, e.g., "**C4(0)**" for carbon atom with total valence 4 and formal charge 0; and
3. atom type, total valence, formal charge, presence of chiral property e.g., "**O3(1)0**" for an oxygen atom with total valence 3, formal charge 1 and without the chiral property.

It is interesting to notice that the use of the histogram of valences drives the generation of molecules towards those that are simpler to synthesize (lower *SAS*) and it increases *Reconstruction*, *Novelty* and *Uniqueness* (see Table 5 in the main manuscript). However, it seems that the use of histograms slightly decreases *QED* values.

We report the molecules' properties results of our model as the λ_1 hyper-parameter values change in Table B5. From these results, we can see that in both QM9 and ZINC datasets, the *QED* values tend to decrease as the λ_1 hyper-parameter values increase.

Appendix C Hyper-parameter Optimization

We performed a grid search on the λ_1 hyper-parameter with the following values: $\{0.2, 0.1, 0.05, 0.01\}$ for both QM9 and ZINC datasets. Specifically, we have noticed that the *Reconstruction* performance of the model increases with low λ_1 values, but at the same time, *Uniqueness* and *Novelty* decrease. The opposite behaviour has been observed for values of λ_1 close to 1. We did not tune the hyper-parameter λ_2 that we fixed at 10, while we selected the learning rate from the following set of values: $\{0.005, 0.001, 0.0005\}$. The batch size is fixed at 100 for both datasets, while the maximum training epochs are 200 for ZINC and 300 for QM9, respectively. We have obtained the best performance in the validation set using $\lambda_1 = 0.05$, $\lambda_2 = 10$, and 0.001 as learning rate for QM9, and $\lambda_1 = 0.001$, $\lambda_2 = 10$, and 0.001 as learning rate for ZINC.

During the generation of new molecules, we sample the atoms and bonds according to the predicted probabilities, while during the *Reconstruction* of an input molecule, we used the *argmax* function on the returned probabilities.

Appendix D Implementation Details

In the following, we report the implementation details of our model. When it is not explicitly indicated, all the neural networks use the *leaky ReLU* activation function.

Table B4 This table presents the ablation study results obtained by RGCVAE - joint model. Notation and format are the same as used in Table 5 of the main manuscript.

Atom Rep.	Hist.	Encoding Network	QM9				ZINC			
			↑%NP	↑%Sol.	↓%SAS	↑%QED	↑%NP	↑%Sol.	↓%SAS	↑%QED
1	✖	RGIN	94.09 ±10.49	34.66 ±13.31	28.48 ±28.58	46.93 ±7.07	78.43 ±7.87	32.77 ±8.54	51.89 ±27.15	37.41 ±2.39
1	✔	RGIN	85.07 ±14.57	29.57 ±12.78	17.65 ±23.61	42.36 ±9.66	93.05 ±10.77	42.69 ±14.56	21.15 ±25.66	46.93 ±9.93
3	✖	RGIN	94.30 ±9.63	31.48 ±14.87	25.13 ±28.88	45.25 ±7.75	78.27 ±8.01	33.0 ±8.65	49.78 ±28.21	37.21 ±2.43
3	✔	RGIN	86.93 ±3.53	25.40 ±14.03	11.17 ±20.34	40.92 ±9.55	86.62 ±13.13	40.74 ±12.37	36.59 ±32.33	43.01 ±7.56
3	✔	GIN	91.48 ±11.32	26.77 ±13.81	16.25 ±24.26	41.77 ±8.49	83.6 ±11.46	41.21 ±10.92	36.91 ±34.96	42.46 ±6.98
3	✔	RGCNC	85.30 ±14.31	29.16 ±13.60	16.85 ±23.26	41.28 ±9.60	90.54 ±11.37	41.04 ±14.27	26.01 ±29.85	45.24 ±9.68
Properties' Scores			88.52 ±17.75	27.91 ±13.76	21.86 ±22.88	46.12 ±7.76	42.08 ±18.37	56.11 ±17.44	55.95 ±22.90	73.18 ±13.86

Table B5 This table presents the ablation study results obtained by the RGCVAE model. Notation and format are the same as used in Table 6 of the main manuscript.

λ_1	QM9				ZINC			
	↑%NP	↑%Sol.	↓%SAS	↑%QED	↑%NP	↑%Sol.	↓%SAS	↑%QED
0.01	91.47 ±11.87	28.58 ±13.77	13.07 ±21.33	42.74 ±8.94	86.62 ±13.13	40.74 ±12.37	36.59 ±32.22	43.01 ±7.56
0.05	86.93 ±13.68	25.04 ±14.03	11.17 ±20.34	40.92 ±9.55	92.72 ±9.21	44.53 ±18.12	6.16 ±14.83	43.22 ±13.43
0.1	84.27 ±14.1	25.75 ±13.67	11.77 ±20.02	40.12 ±10.08	87.51 ±11.72	45.79 ±16.13	18.69 ±26.78	42.36 ±10.92
0.2	78.67 ±14.87	29.49 ±14.19	17.85 ±23.09	40.57 ±10.08	88.68 ±10.2	47.82 ±19.31	4.03 ±12.79	39.82 ±14.69
Properties' Scores	88.52 ±17.75	27.91 ±13.76	21.86 ±22.88	46.12 ±7.76	42.08 ±18.37	56.11 ±17.44	55.95 ±22.90	73.18 ±13.86

D.1 Encoder

In our implemented model we used the following values for the encoder parameters: $K = 5$, $s_h = 70$, $s_{lt} = 70$, $d_p = 1$, $\epsilon^{(k)} = 0 \; \forall k \in [1, K]$. Moreover, $\text{MLP}^{(k)}_{\Phi_e(u,v)}$ is implemented as a linear layer followed by *leaky ReLU*, $\text{MLP}^{(k)}$ is a multi-layer perceptron[1] (MLP) with only one hidden layer and batch normalization before the *leaky ReLU* activation, while MLP_μ and MLP_Σ are feed-forward neural networks with *leaky ReLU* activation. In particular, MLP_Σ clamps the activation to be at a maximum of 2.5.

D.2 Decoder

As explained in the main paper, we adopted the procedure provided in (Rigoni et al., 2020a), where the atom type assignment process is conditioned by the previously assigned atom types. Let $\boldsymbol{\alpha}^u_0$ be the histogram where all the valences are 0, and $t \in \{1, \ldots, m\}$, then each atom type τ_t is predicted according to the following equations:

$$
\begin{aligned}
\boldsymbol{\alpha}^d_t &= \boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}^u_{t-1}, \\
\boldsymbol{e}_t &= K(\boldsymbol{z}_t, \boldsymbol{\alpha}^d_t, \boldsymbol{\alpha}^u_{t-1}), \\
\boldsymbol{r}_t &= \text{Concat}(\boldsymbol{z}_t, \boldsymbol{e}_t), \\
\tau_t &= \text{Sample}_{\text{type}}(F(\boldsymbol{r}_t), \boldsymbol{\alpha}^d_t), \\
\boldsymbol{\alpha}^u_t &= \text{Update}(\tau_t, \boldsymbol{\alpha}^u_{t-1}), \\
\boldsymbol{\alpha}_t &= \text{Sample}_{\text{distr}}(\mathcal{H}, \boldsymbol{\alpha}^u_t),
\end{aligned}
$$

where $\boldsymbol{\alpha}^d_t$ is the difference histogram, $\boldsymbol{\alpha}^u_t$ is the updated histogram, $K(\boldsymbol{z}_t, \boldsymbol{\alpha}^d_t, \boldsymbol{\alpha}^u_t)$ is a function that maps the inputs to a new representation $\boldsymbol{e}_t$, $\text{Concat}(\boldsymbol{z}_t, \boldsymbol{e}_t)$ is the concatenation function, $F(\boldsymbol{r}_t)$ is a function that computes a probability distribution on the atom types, $\text{Sample}_{\text{type}}(F(\boldsymbol{r}_t), \boldsymbol{\alpha}^d_t)$ samples the atom type from the probabilities computed by F masking all the atoms whose valences have a zero-value in the histogram α^d_t. $\text{Update}(\tau_t, \boldsymbol{\alpha}^u_t)$ is a function that updates the histogram $\boldsymbol{\alpha}^u_t$ with the valence of the sampled $\text{Sample}_{\text{distr}}(\mathcal{H}, \boldsymbol{\alpha}^u_t)$: at training time this function returns the histogram $\boldsymbol{\alpha}^v_0$. At generation time, it samples from $\mathcal{H}$ a new histogram $\boldsymbol{\alpha}_{t+1}$ with at least m atoms, such that $\boldsymbol{\alpha}^u_t$ is compatible with $\boldsymbol{\alpha}_{t+1}$. Notice that in our implementation, in generation we have kept the same training behaviour for the function $\text{Sample}_{\text{distr}}(\mathcal{H}, \boldsymbol{\alpha}^u_t)$.

In our atom decoding implementation, the function K is a linear transformation followed by a *tanh* activation function, F is an MLP with one hidden layer of the same size as the input, and *leaky ReLU* as activation function, and $S_n = 120$. We apply a batch normalization layer to all the nodes' representation at the end of the atoms decoding.

[1] It preserves the input dimension.

Table E6 This table reports the best results obtained in the QM9, ZINC, and MOSES datasets.

Metric	QM9	ZINC	MOSES
Reconstruction	94.8 ± 22.2	79.7 ± 40.2	84.35 ± 35.85
Validity	99.9 ± 3.5	99.9 ± 1.2	100 ± 0
Novelty	96.4 ± 18.7	100.0 ± 0.0	100 ± 0
Uniqueness	94.0	61.1	78.13
Diversity	87.7 ± 19.2	98.9 ± 5.0	97.69 ± 7.81
↑%NP	86.93 ± 13.68	86.62 ± 13.13	91.78 ± 11.47
↑%Sol.	25.40 ± 14.03	40.74 ± 12.37	39.32 ± 13.93
↓%SAS	11.17 ± 20.34	36.59 ± 32.22	22.41 ± 8.34
↑%QED	40.92 ± 9.55	43.01 ± 7.56	42.71 ± 9.5

Regarding the edge decoding implementation, we use the following parameter values: $dim(s_v) = 190$ and $dim(\boldsymbol{\phi}_{(u,v)}) = 570$. C and L are implemented as two MLP networks with two hidden layers of dimension 590 and 190, respectively, and *leaky ReLU* as an activation function. In particular, C uses the *sigmoid* activation as the last activation function, while L uses *softmax* activation function in the output.

D.3 Optimization

The optimization function F is implemented as a feed-forward neural network without hidden layers and with *Leaky_ReLU* activation function. Q_p^1 and Q_p^2 are both feed-forward networks without hidden layers for each $p \in \Theta_p$. Regarding the optimization search on ZINC, we set the gradient ascent step to 0.01, we always fix the histogram α_0 for each node decoding iteration and we use the $softmax$ function to predict the type of edges and nodes.

Appendix E Generalization of the Approach

We have validated our RGCVAE - joint model on the MOSES (Polykovskiy et al., 2018) dataset to assess its ability to generalize to larger sets of molecules. MOSES is formed by $1,588,234$ examples of training, $167,503$ of validation, and 5000 of test, thus MOSES is approximately 704.3% times bigger than ZINC.

Table E6 reports the results achieved in the QM9, ZINC, and MOSES datasets. The experiments on MOSES are performed using the same hyperparameters adopted for ZINC, i.e., no model selection is performed on this dataset. RGCVAE generalizes to larger datasets and its performances improve with more data. Over ZINC, the *Reconstruction* increased by almost 5% points, while the *Uniqueness* increased by more than 17%. In addition, there are also larger improvements in the metrics related to the properties of the generated molecules.

In conclusion, RGCVAE can generalize to a larger set of molecules demonstrating strong improvements, especially in its ability to generate diverse new

molecules, i.e., the *Uniqueness*. Undoubtedly, a dedicated model selection for this dataset has the potential to improve further the results.

Appendix F Stereochemistry

Stereochemistry is a branch of chemistry that deals with the three-dimensional arrangement of atoms in molecules and the spatial relationships between different molecular structures. It is concerned with the study of stereoisomers, which are molecules with the same molecular formula and connectivity of atoms but differ in their spatial arrangement. Stereochemistry plays a crucial role in various fields, including drug design, biochemistry, and materials science. The spatial arrangement of atoms in a molecule can significantly influence its biological activity, reactivity, and physical properties. Researchers and chemists need to consider stereochemistry when designing molecules for specific purposes to ensure desired interactions and properties.

One of the key aspects of the stereochemistry is the chirality. Chirality is a concept that pertains to the spatial arrangement of atoms in a molecule, rendering it non-superimposable on its mirror image. Molecules exhibiting chirality often contain a chiral centre, typically a carbon atom bonded to four different substituents. Enantiomers, the two non-identical mirror-image forms of a chiral molecule, have distinct chemical and physical properties, including their interaction with plane-polarized light. Chirality is of paramount importance in fields such as drug development and biochemistry, as the biological activity of enantiomers can significantly differ. The study of chirality enhances our understanding of molecular structure and the unique properties associated with asymmetrical configurations in chemical compounds.

Acknowledgments

The authors acknowledge the HPC resources of the Department of Mathematics, University of Padua, made available for conducting the research reported in this paper. This research was partly supported by the SID/BIRD 2020 project "Deep Graph Memory Networks", University of Padua. Davide Rigoni has performed part of the work during his time in Fondazione Bruno Kessler.

Declarations

Ethics approval Not applicable.
Authors contributions All authors contributed equally to the manuscript.
Availability of data and material The authors made all data and material available.
Code availability Code can be found at: https://github.com/drigoni/RGCVAE.

Founding This research was partly supported by the SID/BIRD 2020 project "Deep Graph Memory Networks", University of Padua.

Conflict of interest The authors declare that there are no conflicts of interest.
Consent to participate The authors have consented to the submission of this manuscript to the journal.

Consent for publication The authors have consented to the publication of this manuscript to the journal.

References

Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L. (2012). Quantifying the chemical beauty of drugs. *Nature chemistry*, *4*(2), 90–98, https://doi.org/10.1038/nchem.1243

Ertl, P., Roggo, S., Schuffenhauer, A. (2008). Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.*, *48*(1), 68–74, https://doi.org/10.1021/CI700286X Retrieved from https://doi.org/10.1021/ci700286x

Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics*, *1*, 8, https://doi.org/10.1186/1758-2946-1-8 Retrieved from https://doi.org/10.1186/1758-2946-1-8

Polykovskiy, D., Zhebrak, A., Sánchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., ... Zhavoronkov, A. (2018). Molecular sets (MOSES): A benchmarking platform for molecular generation models. *CoRR*, *abs/1811.12823*, , Retrieved from http://arxiv.org/abs/1811.12823 1811.12823

Rigoni, D., Navarin, N., Sperduti, A. (2020a). Conditional constrained graph variational autoencoders for molecule design. *2020 IEEE symposium series on computational intelligence, SSCI 2020, canberra, australia, december 1-4, 2020* (pp. 729–736). IEEE. Retrieved from https://doi.org/10.1109/SSCI47803.2020.9308554

Rigoni, D., Navarin, N., Sperduti, A. (2020b). A systematic assessment of deep learning models for molecule generation. , 547–552, Retrieved from https://www.esann.org/sites/default/files/proceedings/2020/ES2020-141.pdf