SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry

Mario Krenn[1,2,3*], **Florian Häse**[1,2,3,4], **AkshatKumar Nigam**[2], **Pascal Friederich**[1,5], **Alán Aspuru-Guzik**[1,2,3,6 *]

[1]Department of Chemistry, University of Toronto, Canada.
[2]Department of Computer Science, University of Toronto, Canada.
[3]Vector Institute for Artificial Intelligence, Toronto, Canada.
[4]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, USA.
[5]Institute of Nanotechnology, Karlsruhe Institute of Technology, Germany.
[6]Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, Canada

1 Introduction

Objects of interest in the natural sciences can often be expressed as graphs with additional domain-specific semantic constraints. Examples are structures of molecules in chemistry (element-dependent bond limitations), quantum optical experiments in physics (component dependent connectivity) or DNA and RNA in biology (nucleobase-dependent connectivity). These constraints pose major challenges for generative models, as their violation leads to invalid results. A popular research question is: *How to design generative models for semantically constrained graphs?* [1, 2].

Here we ask a related, but conceptually different question: *How can we represent the information encoded in a semantically constrained graph in a simple, robust, deterministic, domain-independent, model-independent way?* An answer to this question would allow us to use, as a direct input, our representation into existing (and even future) models without any model-dependent adaptation, and thus has the potential to be transferable across a spectrum of applications.

In this work, we present SELFIES (SELF-referencIng Embedded Strings), a sequence-based representation of semantically constrained graphs that aims to fulfill these criteria. At the heart of SELFIES is a formal Chomsky type-2 grammar [3], which is augmented with two self-referencing, recursive functions to ensure the generation of syntactically and semantically valid graphs.

We show that SELFIES can be used as a direct input to deep learning models, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), in the domain of chemistry and quantum optics. In all experiments tested, 100% of the obtained SELFIES were valid – even for entirely random sequences.

1.1 Related Work

The application of VAEs in chemistry has seen a rapid evolution of robustness. In the first application of VAEs in chemistry [4], chemical compounds were represented with SMILES [5]. Even though they are fragile, *i.e.* small variations often lead to invalid molecules, SMILES are still one of the standard representations used today. DeepSMILES [6] is an effort to extend SMILES by introducing a more robust encoding of rings and branches of molecules.

*Correspondence to: mario.krenn@utoronto.ca, alan@aspuru.com; Source: https://github.com/aspuru-guzik-group/selfies

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

To improve the robustness of molecular representations, parse trees have been employed to formally derive SMILES strings, denoted as GrammarVAE [7]. GrammarVAE uses the well-defined grammar of SMILES which has been defined to *construct* all possible graph structures of chemical elements – a class which contains much more than just all valid molecules. We use SMILES, DeepSMILES and the deterministic rewriting system of GrammarVAE as our baselines.

Further improvements on graphs in chemistry using VAEs have been achieved in [8, 9]. A general semantically constrained graphs combined with regularization demonstrated high validity of the decoded molecules from a VAE [1]. Other advances involve, for example, [10, 11, 2, 12]. Their objective was to adapt the generative models itself to increase validity of the generated graphs, especially in the settings of VAEs. Our motivation is different, we propose a new valid representation which does not require adaptation of deep learning models.

2 Robust representation of semantically constrained graphs

We take advantage of a **formal grammar** to derive *words*, which will represent semantically valid graphs. A formal grammar is a tuple $G(V, \Sigma, R, S)$, where $v \in V$ are non-terminal symbols that are replaced using rules, $r \in R$, into non-terminal or terminal symbols $t \in \Sigma$. S is a start symbol. When the resulting string only consists of terminal symbols, the derivation of a new word is completed [13]. The SELFIES representation is a Chomsky type-2, context-free grammar with self-referencing functions for valid generation of branches in graphs. The rule system is shown in Table 1.

	Vertices			Branches			Rings	
	$\mathbf{A}_0$	$\mathbf{A}_1$	$\mathbf{A}_n$	$\mathbf{A}_{n+1}$		$\mathbf{A}_{n+m}$	$\mathbf{A}_{n+m+1}$	$\mathbf{A}_{n+m+p}$
$\mathbf{X}_0 \rightarrow$	ϵ	$\mid t_{0,1}\,\mathbf{X}_{h_{0,1}}$	$\ldots \mid t_{0,n}\,\mathbf{X}_{h_{r,0}}$	$\mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{0,1}})\,\mathbf{X}_{j_{0,1}}$	$\ldots \mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{0,m}})\,\mathbf{X}_{j_{0,m}}$		$\mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{0,1}}$	$\ldots \mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{0,p}}$
$\mathbf{X}_1 \rightarrow$	ϵ	$\mid t_{1,1}\,\mathbf{X}_{h_{1,1}}$	$\ldots \mid t_{1,n}\,\mathbf{X}_{h_{r,1}}$	$\mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{1,1}})\,\mathbf{X}_{j_{1,1}}$	$\ldots \mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{1,m}})\,\mathbf{X}_{j_{1,m}}$		$\mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{1,1}}$	$\ldots \mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{1,p}}$
					$\ldots$			
$\mathbf{X}_r \rightarrow$	ϵ	$\mid t_{r,1}\,\mathbf{X}_{h_{r,1}}$	$\ldots \mid t_{r,n}\,\mathbf{X}_{h_{r,n}}$	$\mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{r,1}})\,\mathbf{X}_{j_{r,1}}$	$\ldots \mid \mathrm{B}(\mathbf{N}, \mathrm{X}_{i_{r,m}})\,\mathbf{X}_{j_{r,m}}$		$\mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{r,1}}$	$\ldots \mid \mathrm{R}(\mathbf{N})\,\mathbf{X}_{k_{r,p}}$
$\mathbf{N} \rightarrow$	0	$\mid 1$	$\ldots \mid \mathrm{n}$	$\mid \mathrm{n{+}1}$	$\ldots \mid \mathrm{n{+}m}$		$\mid \mathrm{n{+}m{+}1}$	$\ldots \mid \mathrm{n{+}m{+}p}$

Table 1: Grammar of SELFIES, with recursion and $\mathbf{S} \rightarrow \mathbf{X}_0$.

In SELFIES, $V = \{\mathbf{X}_0, \ldots, \mathbf{X}_r, \mathbf{N}\}$ are non-terminal symbols or states. The states $\mathbf{X}_i$ restrict the subsequent edge to a maximal multiplicity of i; the maximal edge multiplicity of the generated graphs is r. The symbol $\mathbf{N}$ represents a numerical value, which acts as argument for the two self-referencing functions. $\Sigma = \{t_{0,1}, \ldots, t_{r,n}\}$ are terminal symbols. The derivation rule set R has exactly $(n + m + p + 1) \times (r + 2)$ elements, corresponding to n rules for vertex production, m rules for producing branches, p rules for rings and r non-terminal symbols in V. The subscripts $h_{a,b}$, $i_{a,b}$, $j_{a,b}$ and $k_{a,b}$ have values from 1 to r, and encode the actual domain-specific constraints. The semantic and syntactical constraints are encoded into the rule vectors, which guarantees strong robustness. There are $n + m + p + 1$ rule vectors $\mathbf{A}_i$, each with a dimension $(r + 2)$.

Self-referencing functions for syntactic validity – In order to account for **syntactic validity** of the graph, we augment the context-free grammar with **branching functions** and **ring functions**. $B(\mathbf{N}, X_i)$ is the branching function, that recursively starts another grammar derivation with subse-

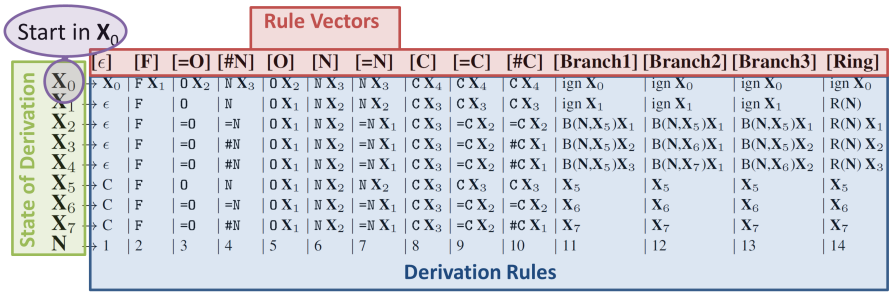

Figure 1: Derivation rules of SELFIES for molecules in the QM9 dataset.

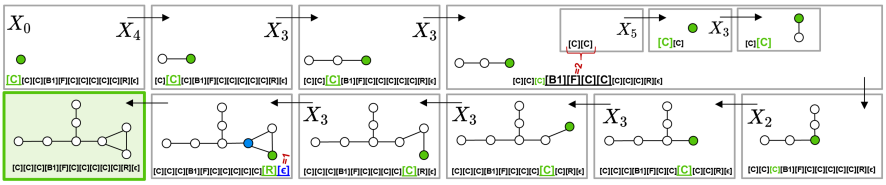

Figure 2: Derivation of a molecule with a recursive branch generation in step 4.

quent $\mathbf{N}$ SELFIES symbols in state X_i. After the full derivation of a new word (which is a graph), the branch function returns the graph, and connects it to the current vertex. The ring function $R(\mathbf{N})$ establishes edges between the current vertex and the $(\mathbf{N}+1)$-th last derived vertex. Both the branching and ring functions have access to the SELFIES string and the derived string, thus are self-referencing.

Rule vectors for semantic validity – To incorporate **semantic validity**, we denote A_i as the i-th vector of rules, with dimension $d_{A_i} = |V| = r + 2$. The **conceptual idea** is to interpret a symbol of a SELFIES string, $s_i \in \{0, \ldots, n+m+p\}$ as an index of a rule vector, A_{s_i}. In the derivation of a symbol, the rule vector is defined by the symbol of the SELFIES string (external state) while the specific rule is chosen by the non-terminal symbol (internal state). Thereby, we can encode semantic information into the rule vector A_i, which is *memorized* by the internal state during derivation.

Algorithmic derivation of grammar from data, and validity guarantees – Domain-specific grammars can be derived algorithmically directly from data, without any domain knowledge. Let T be the set of different types of vertices (such as C, O, N, ... in chemistry). We use a dataset to get the types of vertices T_i, and their maximal degrees D_i ($D_i = \mathrm{maxdeg}(T_i)$ – in chemistry, the $D_\mathrm{O} = \mathrm{maxdeg}(\mathrm{O}) = 2$, $D_\mathrm{C} = \mathrm{maxdeg}(\mathrm{C}) = 4$). Let $M = \max_i \mathrm{maxdeg}(T_i)$ be the maximal degree of the dataset. Starting from Table 1 (I) we identify the rule vectors $\mathbf{A}_i$, (II) define the non-terminal symbols $\mathbf{X}_j$, and (III) define the rules R.

I $\mathbf{A}_1 \ldots \mathbf{A}_n$ (vertices rules) consist of T_i with a potential multiedge connection γ up to D_i (in chemistry, D_O=2, thus we have two rule vectors for O, one with single edge $\gamma = 1$, one with double connection $\gamma = 2$). $\mathbf{A}_{n+1} \ldots \mathbf{A}_{n+m}$ represent branch rules. A branch forms connections to two vertices, thus we have maximally $(M-1)$ branch rules, (combinations of $(M-l, l)$ represent the maximal connectivity to the two branches). $\mathbf{A}_{n+m+1} \ldots \mathbf{A}_{n+m+p}$ denote ring rules, in a generic case $p = 1$ is sufficient.

II non-terminals $X_1 \ldots X_r$, with $r = M$, constrain the number of edges to connect two vertices.

III Rule $r_{i,j}$ for $\mathbf{A}_i \in \{\mathbf{A}_1 \ldots \mathbf{A}_n\}$ and $\mathbf{X}_j \in \{\mathbf{X}_1 \ldots \mathbf{X}_r\}$ can consist of a terminal and non-terminal symbol. The terminal consists of a T_i (given by $\mathbf{A}_i$) and a edge-multiplicity $\mu = \min(j, \gamma)$. The corresponding nonterminal symbol is $\mathbf{X}_{M-\mu}$ (if $M - \mu = 0$, no non-terminal will be added). Note that constraints are satisfied due to the min operation in μ. Rules in state $\mathbf{X}_j$ for rings are $R(N)\mathbf{X}_{j-1}$, and for branches are $B(N, X_i)\mathbf{X}_{j-i}$.

The edge-multiplicity $\mu = \min(j, \gamma)$ is responsible for the semantic constraint of local degrees being satisfied. This is the most immediate constraint in many applications for physical sciences, which allows for 100% validity. More complex, non-local constraints could be implemented by more complex grammars, such as explicit context-sensitive type-1 grammars.

3 Application on chemical graphs

A concrete alphabet for the application in chemistry is in Figure 1, which we use to represent molecules in the benchmark dataset QM9 [14, 15]. The derivation of a molecular graph in Figure 2.

Coverage of the chemical space – Table 1 covers a large range of organic molecules, and it is easy to extend the grammar (for example, using the algorithm in section 2). We added ring and branch function which take more than one sequence for deriving a number $\mathbf{N}$ (enabling rings and branches of up to $\mathbf{N} = 8.000$). To increase the coverage further, we have added one additional generic rule vector, which has no semantic restrictions (i.e. $D_i \to \infty$), but satisfies syntactic degrees. Unspecified vertices will be derived in this way. Thereby, we encoded and decoded all 72 million molecules

	bitflip	random string		VAE			GAN
	Validity	Validity	Length	Validity	Reconstruction	Diversity	Diversity
SMILES	18.1%	2.8%	1.4	71.9%	66.2%	5.9%	18.5%
DeepSMILES	38.2%	3.0%	1.4	81.4%	79.8%	67.3%	-
GrammarVAE	9.5%	17.2%	1.0	34.0%	84.0%	4.0%	-
SELFIES	**100%**	**100%**	**9.4**	**100%**	**98.2%**	**82.9%**	**78.9%**

Table 2: Results for bitflip (starting from valid graph), random sequence, VAE and GAN

from PubChem (the most complete collection of synthesized molecules) with ≤ 500 SMILES chars, demonstrating coverage of the space of chemical interest.

Validity after mutations, and random strings – In the first experiment, we test random mutations (starting from a valid molecule in QM9) and entirely random strings. We evaluate the resulting validity using RDKit [16]. Results are shown in Table 2. While the best competing representation has less than 40% validity and less than 20% respectively, SELFIES always produce valid molecules. Furthermore, the resulting valid molecules from random strings are significantly larger (in SMILES chars) than for other representations.

Application in a Variational Autoencoder – We demonstrate the robustness and practicality of SELFIES for molecule generation in VAE. We evaluate the performance of each representation based on the *reconstruction quality* (per-character matching between the input and output), *validity* (fraction of valid molecules), and *diversity* (fraction of valid molecules with different SMILES strings). The encoder is a fully connected NN with three layers, the decoder is a RNN. The hyperparameters for each representations are Bayesian optimized [17, 18]. We show in Table 2 that SELFIES has 100% validity, and the highest diversity and reconstruction quality.

For an extended test, we add a third neural network (which is connected to the latent space) for a regression task. We train it in tandem with the VAE to predict graph properties (partition coefficient logP [19]). The prediction quality of all representations is similar (r^2=0.97, except of GrammarVAE which has r^2=0.92). For inverse design, apart of high prediction quality, a diverse latent space is essential. Thus we investigate the density of valid diverse molecules by sampling latent space points (within certain σ around the center, stopping after 20 samples didn't produce new instances). We show in Fig. 3a that SELFIES VAE contains 100 times more valid diverse molecules than others.

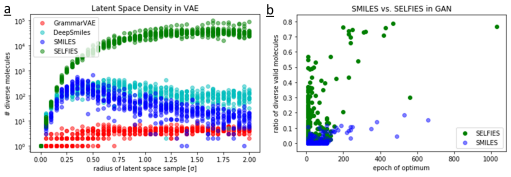

Figure 3: Diversity in VAE and GAN

Application in a Generative Adversarial Network – For chemistry (QM9), SELFIES outperform SMILES in GAN, see Fig. 3b. We train GANs (fully connected, with 200 different hyperparameter settings) to generate diverse molecules. Sampling 10k times, SELFIES produced 7889 different valid molecules (SMILES only 1855). Diversity is the critical metric of particular interest in molecular design. Also in this regard, SELFIES outperform all other available sequence-based graph representations, while also showing *model-independence*.

4 Conclusion and Future Work

SELFIES is a robust, general-purpose representation of graphs with domain-specific constraints. It enables the application of new deep learning methods in the natural sciences, such as chemistry, without the necessity to adapt models with domain-specific constraints. It is straight forward to apply SELFIES to other domains by deriving the grammar in section 3.3 algorithmically. We have applied SELFIES also in quantum optics, where we find – similar to chemistry – 100% validity of the generated quantum optics experimetal graphs [20, 21], and outperforming native representations which have similar difficulties as SMILES in chemistry (in particular in cases of cavities).

We conclude by stressing that a 100% valid latent space is essential for model interpretation [22, 23, 24], in particular for interpreting the internal representations [25] in a scientific context [26].

Acknowledgments

The authors thank Theophile Gaudin for useful discussions. A. A.-G. acknowledges generous support from the Canada 150 Research Chair Program, Tata Steel, Anders G. Froseth, and the Office of Naval Research. We acknowledge supercomputing support from SciNet. M.K. acknowledges support from the Austrian Science Fund (FWF) through the Erwin Schrödinger fellowship No. J4309. F.H. acknowledges support from the Herchel Smith Graduate Fellowship and the Jacques-Emile Dubois Student Dissertation Fellowship. P.F. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 795206.

References

[1] T. Ma, J. Chen and C. Xiao, Constrained generation of semantically valid graphs via regularizing variational autoencoders. *Advances in Neural Information Processing Systems* 7113–7124 (2018).

[2] Y. Li, O. Vinyals, C. Dyer, R. Pascanu and P. Battaglia, Learning deep generative models of graphs. *arXiv:1803.03324* (2018).

[3] N. Chomsky, Three models for the description of language. *IRE Transactions on information theory* **2**, 113–124 (1956).

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

[5] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31–36 (1988).

[6] N. O'Boyle and A. Dalke, DeepSMILES: An Adaptation of SMILES for Use in machine-learing chemical structures. *ChemRxiv* (2018).

[7] M.J. Kusner, B. Paige and J.M. Hernández-Lobato, Grammar variational autoencoder. *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 1945–1954 (2017).

[8] H. Dai, Y. Tian, B. Dai, S. Skiena and L. Song, Syntax-directed variational autoencoder for structured data. *arXiv:1802.08786* (2018).

[9] W. Jin, R. Barzilay and T. Jaakkola, Junction tree variational autoencoder for molecular graph generation. *arXiv:1802.04364* (2018).

[10] M. Simonovsky and N. Komodakis, Graphvae: Towards generation of small graphs using variational autoencoders. *International Conference on Artificial Neural Networks* 412–422 (2018).

[11] Q. Liu, M. Allamanis, M. Brockschmidt and A. Gaunt, Constrained graph variational autoencoders for molecule design. *Advances in Neural Information Processing Systems* 7795–7804 (2018).

[12] B. Samanta, A. De, G. Jana, P.K. Chattaraj, N. Ganguly and M. Gomez-Rodriguez, NeVAE: A Deep Generative Model for Molecular Graphs. *arXiv:1802.05283* (2018).

[13] J.E. Hopcroft, R. Motwani and J.D. Ullman, Introduction to Automata Theory, Languages, and Computation (3rd Edition). (2006).

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

[15] L. Ruddigkeit, R. Van Deursen, L.C. Blum and J.L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling* **52**, 2864–2875 (2012).

[16] G. Landrum and others, RDKit: Open-source cheminformatics. *Journal of chemical information and modeling* (2006).

[17] T.G. authors, GPyOpt: A Bayesian Optimization framework in python. *http://github.com/SheffieldML/GPyOpt* (2016).

[18] F. Häse, L.M. Roch and A. Aspuru-Guzik, Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chemical science* **9**, 7642–7655 (2018).

[19] S.A. Wildman and G.M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of chemical information and computer sciences* **39**, 868–873 (1999).

[20] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz and A. Zeilinger, Automated search for new quantum experiments. *Physical review letters* **116**, 090405 (2016).

[21] A.A. Melnikov, H.P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger and H.J. Briegel, Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences* **115**, 1221–1226 (2018).

[22] K.T. Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks. *Nature communications* **8**, 13890 (2017).

[23] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter and T. Unterthiner, Interpretable Deep Learning in Drug Discovery. *arXiv:1903.02788* (2019).

[24] F. Häse, I.F. Galván, A. Aspuru-Guzik, R. Lindh and M. Vacher, How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chemical Science* **10**, 2298–2307 (2019).

[25] T.Q. Chen, X. Li, R.B. Grosse and D.K. Duvenaud, Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems* 2610–2620 (2018).

[26] R. Iten, T. Metger, H. Wilming, L. Del Rio and R. Renner, Discovering physical concepts with neural networks. *arXiv:1807.10300* (2018).