Structuring Latent Spaces of Variational Autoencoders for Molecular Design

Jan G. Rittig[1], Nils Graß[1], and Alexander Mitsos[3,1,2]

[1] Process Systems Engineering (AVT.SVT), RWTH Aachen University, Forckenbeckstraße 51, 52074 Aachen, Germany
[2] Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research IEK-10 – Energy Systems Engineering, Wilhelm-Johnen-Str., 52425 Jülich, Germany
[3] JARA Center for Simulation and Data Science (CSD), Aachen, Germany
Corresponding author, email: amitsos@alum.mit.edu

Abstract. Computer-aided design of molecules using variational autoencoders (VAE) has shown great promise, but achieving interpretability of the latent space remains a key challenge for guided generation of molecules with desired properties. We adapted and evaluated various disentanglement methods to chemically structure the latent space in VAE-based molecule generation. Using conditional VAEs enables applying physicochemical property constraints on the latent space and is thus promising for targeted molecular design.

Keywords: Generative graph machine learning · computer-aided molecular design · disentanglement.

1 Introduction

Generative machine learning has opened up new possibilities for designing molecules with desired properties [1, 3]. Variational autoencoders (VAEs) are a prominent generative ML approach that learn a continuous representation of the molecular space, referred to as molecular latent space. The molecular latent space allows strategic sampling of molecules through continuous optimization approaches, thereby facilitating the design of molecules with desired properties. VAEs have, for example, shown promising results for drug [7] and fuel design [15]. However, VAE latent spaces typically constitute a high dimensionality, e.g., $\mathbb{R}^{56}$ [10] and $\mathbb{R}^{72}$ [12], limiting chemical interpretability and optimization. Therefore, we evaluate various techniques for structuring VAE latent spaces regarding molecular properties.

2 Structuring variational autoencoders

To structure VAE latent spaces, unsupervised and supervised methods can be employed, also referred to as disentanglement methods, cf. [6]. Unsupervised

methods refine the VAE loss function to disentangle latent dimensions, i.e., decrease the shared information between multiple dimensions, without using property information about the molecules. Supervised methods, in contrast, utilize molecular information for structuring the VAE latent space during training.

We investigated both unsupervised, e.g., Kullback-Leibler (KL) loss adjustments [5], and supervised methods, such as conditional VAEs (CVAEs), cf. [13, 14], and structured VAEs (SVAEs), cf. [7], for disentangling structural characteristics and physicochemical properties of molecules in the latent space. Here, we focus on the results of the supervised methods for encoding a molecular target property p into the latent space relevant for molecular design. We refer to models as CVAEs when encoding property information in specific dimensions of the latent space, e.g., by simultaneously training the VAE and a property prediction MLP that considers only these dimensions as input. Thus, to achieve high prediction accuracy, the CVAE encodes p-related information from the molecular structure into these specific dimensions. SVAEs encourage the whole latent space to be informative about the property, e.g., by training an MLP for property prediction on the whole latent space or applying deep metric learning (DML). In both cases, an additional loss term accounting for the molecular property is added to the VAE loss. Specifically, we consider the following methods:

- **VAE**: a default VAE using β to control the KL loss;
- **CVAE**: a CVAE using a multilayer perceptron (MLP) to embed p into single latent space dimensions;
- **CVAE(M)**: a CVAE using an MLP to embed molar mass M into single latent space dimensions;
- **SVAE**: a SVAE using an MLP to embed p in the entire latent space;
- **SVAE**$_{\text{reg}}$: a SVAE using an MLP to embed p in the entire latent space, regularized with group lasso loss to encourage feature selection;
- **SVAE**$_{\text{DML}}$: a SVAE using DML with triplet loss to embed p in the entire latent space.

3 Case Study: Melting Point Encoding

As a case study, we aim to encode the melting point T_m into a molecular latent space of a VAE because it provides information about the aggregate state of a molecule that is critical for practical applications. We employ the Junction-Tree VAE (JT-VAE) [10], which uses molecular representations in the form of graphs and junction trees, i.e., graphs without loops, and allows molecular generation with chemical validity. We use the publicly available data set of melting points for 34,000 molecules from [2] and extend it with molecules from the ZINC database [9], which are composed of the same junction tree nodes as the initial set, resulting in 66,000 molecules. The data set is split into a 98% training and a 2% validation set by a stratified split based on the junction tree nodes and melting points. We train all models with the hyperparameters provided by Jin et al. [10]. We further use β annealing to ensure 90% reconstruction accuracy on the validation set, so that the different VAEs are comparable and the evaluation

of the property disentanglement focuses mostly on molecules that can be reconstructed. Investigating the effects of disentanglement on the reconstruction of molecules not included in training and validation would be interesting for future work. We repeat all training runs three times and average the results.

Benchmarking supervised methods – Figure 1 shows the single dimension variable prediction (SDVP) results of all supervised methods for the prediction of the melting point T_m. The SDVP trains one MLP for each latent dimension to predict T_m, thereby indicating whether the dimension is informative for the target property, cf. [4]. Compared to the baseline of β-VAE, all supervised methods embed T_m into the latent space to some extent. The CVAE achieves the highest SDVP scores in the first two dimensions with average R^2 values of 0.56 and 0.34, whereas the latent dimensions z_i with $i \geqslant 6$ exhibit lower SDVP scores than the other methods. This indicates that the CVAE successfully disentangles the melting point information in the latent dimensions. The SVAE_{DML} implementation based on deep metric learning demonstrated only a minor improvement over the baseline. Conversely, the other SVAEs, i.e., standard SVAE and SVAE_{reg}, show improved informativeness for the melting point in all latent dimensions, indicating high entanglement of the entire latent space. Notably, the regularization applied in the form of a group lasso loss did not improve over standard SVAEs.

Further, the CVAE(M) trained on the molar mass M is also informative in regard to T_m, see Figure 1. On the most informative dimensions, the CVAE(M) exhibits similar SDVP values to the best SVAE trained on the melting point. This raises a potential challenge when using a CVAE for molecule design with multiple properties that relate to similar structural characteristics of a molecule and are correlated. To avoid embedding redundant information into the latent

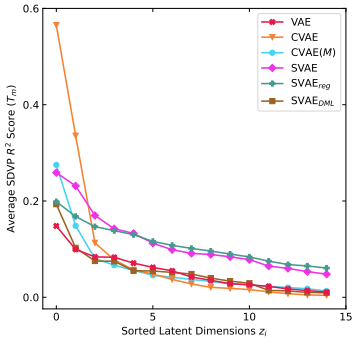

Fig. 1. Comparison of supervised disentanglement methods evaluated using single dimension variable prediction with chemical properties.

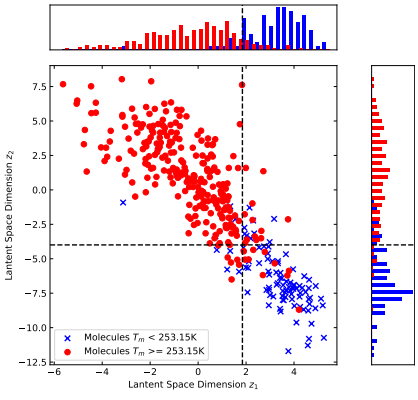

Fig. 2. Conditional VAE embedding of the test set in conditional dimensions of latent space. Histogram of distributions left and right. Optimal thresholds as black lines.

space, i.e., entanglement, CVAEs with multiple properties will have to share latent dimensions and allocate separate dimensions for further, non-shared structural characteristics. This may become increasingly challenging as the number of selected properties increases.

Constrained molecule generation – Furthermore, we test the best performing method, the CVAE, for constrained molecular generation. That is, molecular design often involves optimizing a certain property while applying constraints to other properties. For example, in fuel design, a necessary constraint is to ensure liquid aggregate state of a fuel component at ambient temperature, i.e., that the melting point is below 253.15 K, see, e.g., [8]. We thus used the training set to determine thresholds t_1, t_2 for the two most informative latent dimensions of the CVAE to classify molecules with a melting point T_m below or above 253.15 K.

Figure 2 shows a scatter plot of the molecular embeddings of the validation set in the two latent space dimensions. The embeddings are grouped by melting points T_m above (indicated in red) or below 253.15 K (indicated in blue) based on the labeled data. Histograms on the top and right show the distributions of both sets along each latent space dimension. The thresholds t_1 and t_2, which split the latent space dimensions z_1 and z_2, are indicated as dashed lines, creating four separate quadrants. Most molecules with $T_m < 253.15\,K$ are embedded in the lower right quadrant (95.69 %), whereas most molecules with $T_m \geqslant 253.15\,K$ are in the upper left quadrant (95.42 %), indicating that the thresholds successfully split the two classes. Furthermore, we observe that both latent dimensions are entangled, with a Pearson correlation coefficient of -0.84, which is a result of the JT-VAE architecture with the junction tree and molecular feature latent space, cf. [11]. Overall, the CVAE enables considering melting point constraints in the generation of molecules, thereby facilitating targeted molecular design.

4 Conclusion and Outlook

We utilize several methods to structure molecular latent spaces of VAEs with respect to molecular properties. We find that conditional VAEs can encode molecular properties into single latent dimensions, enabling consideration of property constraints in molecular generation with VAEs. Future work could target the encoding of properties into latent spaces of other generative ML approaches such as diffusion- and flow-based models [1]. Further, encoding multiple physicochemical properties into generative ML-based latent spaces similar to [13] will be highly relevant for molecular design which typically involves a multitude of property constraints and objectives.

Acknowledgments. This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466417970 – within the Priority Programme "SPP 2331: Machine Learning in Chemical Engineering". Simulations were performed with computing resources granted by RWTH Aachen under project "thes1332".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anstine, D.M., Isayev, O.: Generative models as an emerging paradigm in the chemical sciences. Journal of the American Chemical Society **145**(16), 8736–8750 (2023). https://doi.org/10.1021/jacs.2c13467
2. Bell, C., Cortes-Pena, Y.R., Contributors: Chemicals: Chemical properties component of chemical engineering design library (ChEDL) (2016–2023), https://github.com/CalebBell/chemicals
3. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., Jensen, K.F.: Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science **12**(5), e1608 (2022). https://doi.org/10.1002/wcms.1608
4. Cameron Eastwood, Jakob Verbeek, Christopher K. I. Williams, Yann Ollivier: A framework for the quantitative evaluation of disentangled representations. In: International Conference on Learning Representations, ICLR (2018)
5. Chen, R.T.Q., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders, http://arxiv.org/pdf/1802.04942v5
6. Du, Y., Guo, X., Wang, Y., Shehu, A., Zhao, L.: Small molecule generation via disentangled representation learning. Bioinformatics (Oxford, England) **38**(12), 3200–3208 (2022). https://doi.org/10.1093/bioinformatics/btac296
7. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science **4**(2), 268–276 (2018). https://doi.org/10.1021/acscentsci.7b00572
8. Heisterberg, M., Nørager, V., Hussaini, M.Y.: Computational analysis of the effect of fuel properties on diesel engine combustion characteristics. Fuel **87**(14-15), 3196–3205 (2008). https://doi.org/10.1016/j.fuel.2020.119243
9. Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S., Coleman, R.G.: Zinc: A free tool to discover chemistry for biology. Journal of Chemical Information and Modeling **52**(7), 1757–1768 (2012). https://doi.org/10.1021/ci3001277
10. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. 35th International Conference on Machine Learning, ICML 2018 **5**, 3632–3648 (2018), http://arxiv.org/pdf/1802.04364v4
11. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation, http://arxiv.org/pdf/1802.04364v4
12. Kajino, H.: Molecular hypergraph grammar with its application to molecular optimization. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3183–3191. PMLR (09–15 Jun 2019), 10.48550/arXiv.1809.02745
13. Lim, J., Ryu, S., Kim, J.W., Kim, W.Y.: Molecular generative model based on conditional variational autoencoder for de novo molecular design. Journal of cheminformatics **10**(1), 31 (2018). https://doi.org/10.1186/s13321-018-0286-7
14. Ramchandran, S., Tikhonov, G., Lönnroth, O., Tiikkainen, P., Lähdesmäki, H.: Learning conditional variational autoencoders with missing covariates. Pattern Recognition **147**, 110113 (2024). https://doi.org/10.1016/j.patcog.2023.110113
15. Rittig, J.G., Ritzert, M., Schweidtmann, A.M., Winkler, S., Weber, J.M., Morsch, P., Heufer, K.A., Grohe, M., Mitsos, A., Dahmen, M.: Graph machine learning for design of high–octane fuels. AIChE Journal **69**(4), e17971 (2023). https://doi.org/10.1002/aic.17971