Optimizing Pre-Training Strategies for Chemical Models using VICReg

Peer Schliephacke
Darmstadt University of Applied Sciences
Faculty of Mathmatics and Natural Sciences & Faculty of Computer Science
Prof. Dr. Groos & Prof. Dr. Hergenröther

Motivation

Adapting self-supervised learning methods like VICReg (Variance-Invariance-Covariance Regularization) [1] to computational chemistry provides a novel way to integrate diverse molecular data types for improved representation learning. By combining graph, conformer, and chemical language modalities within a unified training framework, this work aims to enhance the quality and versatility of molecular encoders, contributing to the development of a chemical foundation model. Evaluating these encoders on challenging ADME prediction tasks and benchmarking them against established pre-training strategies addresses the need for more effective and generalizable molecular representations in drug discovery and chemical research.

Research Questions

i How do supervised model-agnostic pre-training strategies compare to self-supervised, model-specific pre-training strategies?
ii Can VICReg be successfully adapted to computational chemistry?
iii What experimental settings yield better downstream performance under VICReg?
iv Does VICReg outperform other pre-training strategies?

Methods

Data

- **Pre-training Data:** The largemix [2] dataset, containing 5 million molecules with labels from biochemical and quantum mechanical domains, is used for pre-training.
- **Downstream Task Data:** The public ADME dataset from *Fang et al.* [3] is used for drug property prediction. Train/test splits are determined by molecule similarity using the Butina clustering algorithm [4].

Evaluation

- **Linear Evaluation:** Pre-trained encoders are assessed by training a linear layer on frozen encoder representations.
- **Transfer Learning:** Pre-trained encoders are fine-tuned with a small MLP, using a lower learning rate for the encoder parameters.
- **Statistical Assessment:** Each evaluation is repeated 30 times with different parameter initializations to quantify uncertainty. Results are analyzed using repeated measures ANOVA and Tukey HSD post-hoc tests.

Experiments

- **No Pre-training:** Each encoder and an XGB [5] model are tuned on the downstream training split using 5-fold cross-validation based on Butina clustering. Final parameters are used for the encoders, with XGB as a benchmark.
- **Pre-training Comparison:** Each modality-specific encoder is pre-trained on largemix using supervised model-agnostic multitask training and modality-specific self-supervised pre-training. Node-attribute masking is used for GINs (Graph Isomorphism Networks) [6] and EGNNs (Equivariant Graph Neural Networks) [7]; LSTMs (Long short-term memory) networks [8] are pre-trained with masked language modeling (MLM) [9].
- **VICReg Experiments:** Various VICReg configurations are tested, including changes in data size and encoder weight initialization (random vs. pre-trained weights).

Multi-modal VICReg Architecture for different Molecule Modalities

Different molecular modalities are represented by chemical language, graph, and conformer representations. These modalities are then processed by modality-specific encoders.

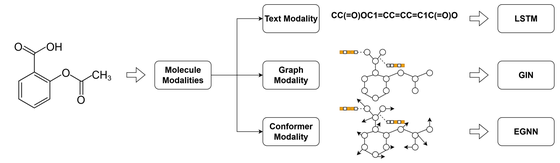

Figure 1. Various molecular modalities with corresponding encoders based on an Aspirin molecule [10]

VICReg can be adapted to computational chemistry using a multi-modal approach. A batch of SMILES strings serves as training data S and is processed by a modality-specific featurizer $\phi^{\cdot}$ to produce featurized inputs $S^{\cdot}$. Each modality-specific encoder generates a representation $Y^{\cdot}$ for the same molecules. These representations are expanded by projectors $P^{\cdot}$ to create embeddings $Z^{\cdot}$, which are used as inputs to the VICReg loss to align the different molecular views.

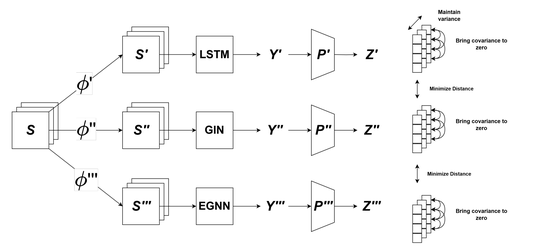

Figure 2. Adaption of VICReg to work with molecules by aligning different molecule modalities.

Results

The test MAE scores for each ADME downstream task is indicated by color. The supervised model-agnostic pre-training achieves the best overall results, followed by the LSTM with self-supervised MLM pre-training. Most pre-trained encoders perform at least as well as the XGB baseline across the downstream tasks.

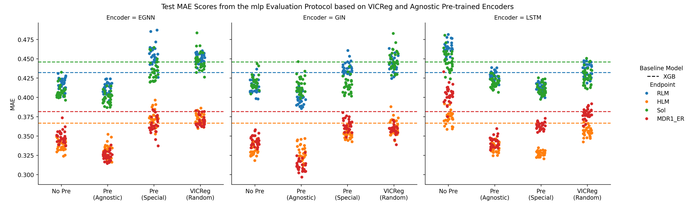

Figure 3. Comparison between the test MAE downstream task scores of differently pre-trained encoders.

Conclusion

Research Questions

i The supervised pre-training shows the best results alongside the LSTM pre-trained with the MLM approach.
ii The VICReg results show that an adaption to the field of computational chemistry is possible without encountering a representation collapse.
iii VICReg benefits from different experimental settings such as larger datasets or different encoder initializations.
iv The VICReg performance is generally inferior to the supervised pre-training results.

Summary

The pre-training experiments demonstrate that VICReg can effectively be adapted to molecular data. While supervised model-agnostic and MLM pre-training yield superior downstream performance, VICReg's ability to operate without labeled data and its inherent flexibility create significant opportunities for experimentation. This scalability with unlabeled data not only establishes a strong foundation for future research but also allows for a wide range of innovative approaches in the exploration of molecular data.

Future Works

Building on this work, next steps include an extensive ablation study to explore VICRegs full potential in the domain of chemoinformatics. This includes using an even larger pre-training set since VICReg doesn't need any labels, improved encoder complexity by using transformer based modality specific encoders, utilizing even more modalities such as the tabular representation of molecules, different batch sizes, varying regularization parameter settings, larger projector architectures, the use of weight-sharing, and encoder parameter settings. Different downstream task datasets could further evaluate the generalizability through classification benchmarks in biochemically distinct domains like material science.

References

[1] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.
[2] D. Beaini, S. Huang, J. A. Cunha, Z. Li, G. Moisescu-Pareja, O. Dymov, S. Maddrell-Mander, C. McLean, F. Wenkel, L. Müller, *et al.*, "Towards foundational models for molecular learning on large-scale multi-task datasets," *arXiv preprint arXiv:2310.04292*, 2023.
[3] C. Fang, Y. Wang, R. Grater, S. Kapadnis, C. Black, P. Trapa, and S. Sciabola, "Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective," *Journal of Chemical Information and Modeling*, vol. 63, no. 11, pp. 3263–3274, 2023.
[4] D. Butina, "Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 4, pp. 747–750, 1999.
[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
[6] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.
[7] V. G. Satorras, E. Hoogeboom, and M. Welling, "E (n) equivariant graph neural networks," in *International conference on machine learning*, pp. 9323–9332, PMLR, 2021.
[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
[10] Benjah-bmm27, "Aspirin-skeletal.svg," `https://de.m.wikipedia.org/wiki/Datei:Aspirin-skeletal.svg`, 2006. Public domain.

peer.schliephacke@stud.h-da.de