Recent Advances in Materials Science Research Through Machine Learning

André F. Oliveira[1], Gabriel A. Pinheiro[2], Felipe V. Calderan[2], Antônio V. F. Bezerra[2], Piero A. L. Ribeiro[2], Juarez L. F. Da Silva[3], and Marcos G. Quiles[2]

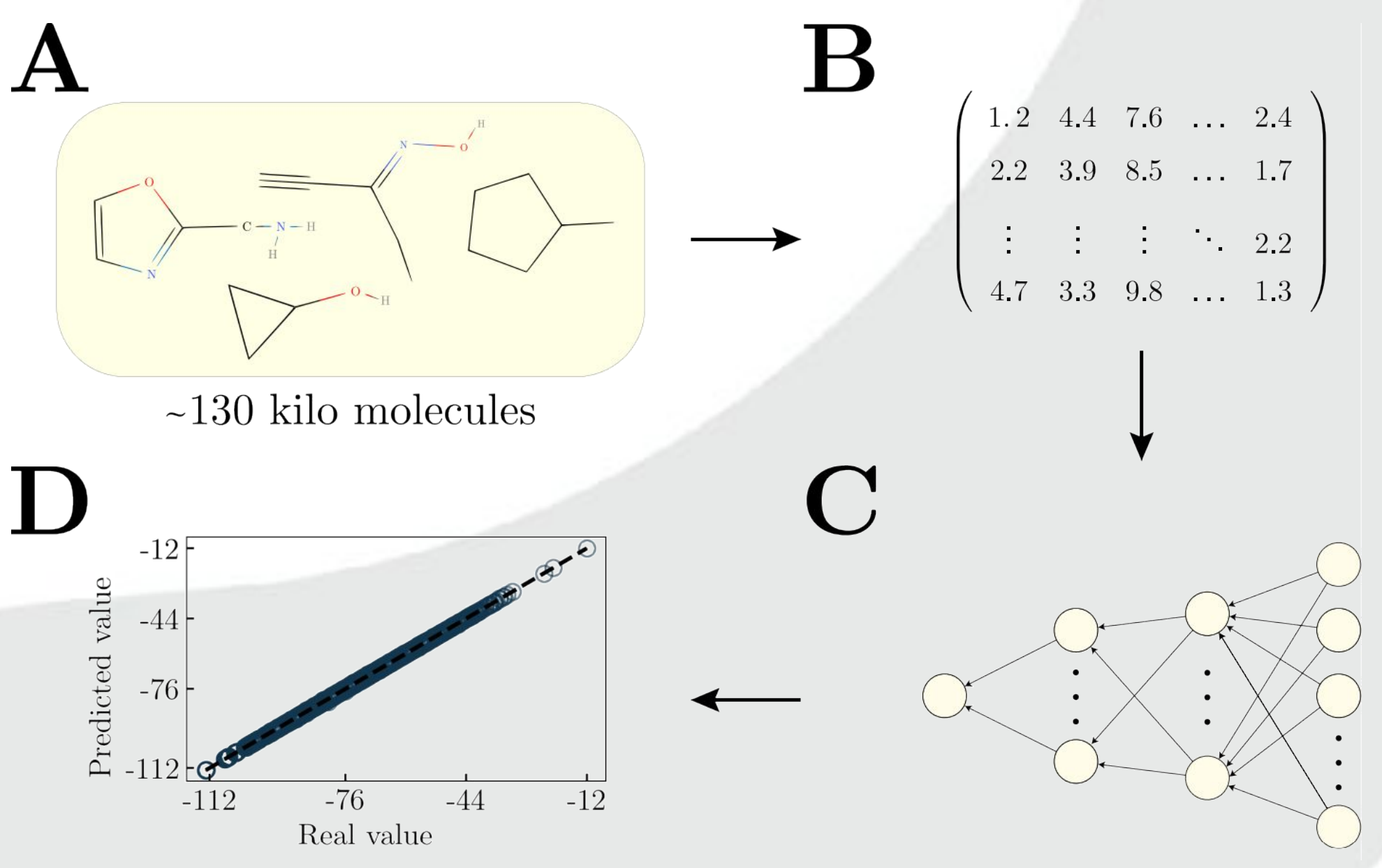
[1]Associate Laboratory for Computing and Applied Mathematics, National Institute for Space Research, São José dos Campos, SP, Brazil
[2]Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil
[3]São Carlos Institute of Chemistry, University of São Paulo, São Carlos, SP, Brazil

Abstract

Machine learning algorithms have become an exciting tool for material design and discovery due to their time efficiency and recent results. These techniques often rely on learning a map function to predict physicochemical properties given the molecular structure. In this realm, our latest progress included the application of various ML paradigms. For instance, we introduced ML techniques to tackle the problem of property prediction via a fully supervised setting without depending on expensive quantum calculations for inference. We also proposed a framework that conducts representation learning in an unsupervised and semi-supervised setting. This is possible by applying a contrastive learning strategy on multiple molecular representations, reducing the need for human intervention. Finally, we developed a supervised algorithm that combines variational autoencoder and multilayer perceptron to perform property prediction and molecular design with targeted properties.

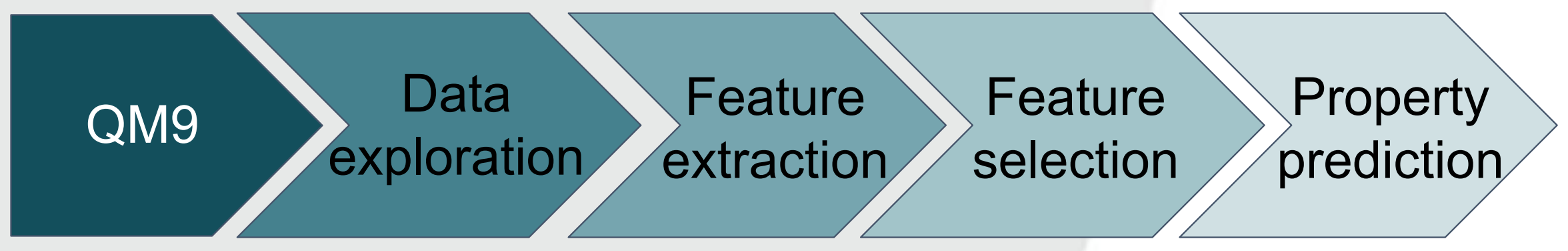
Supervised Learning

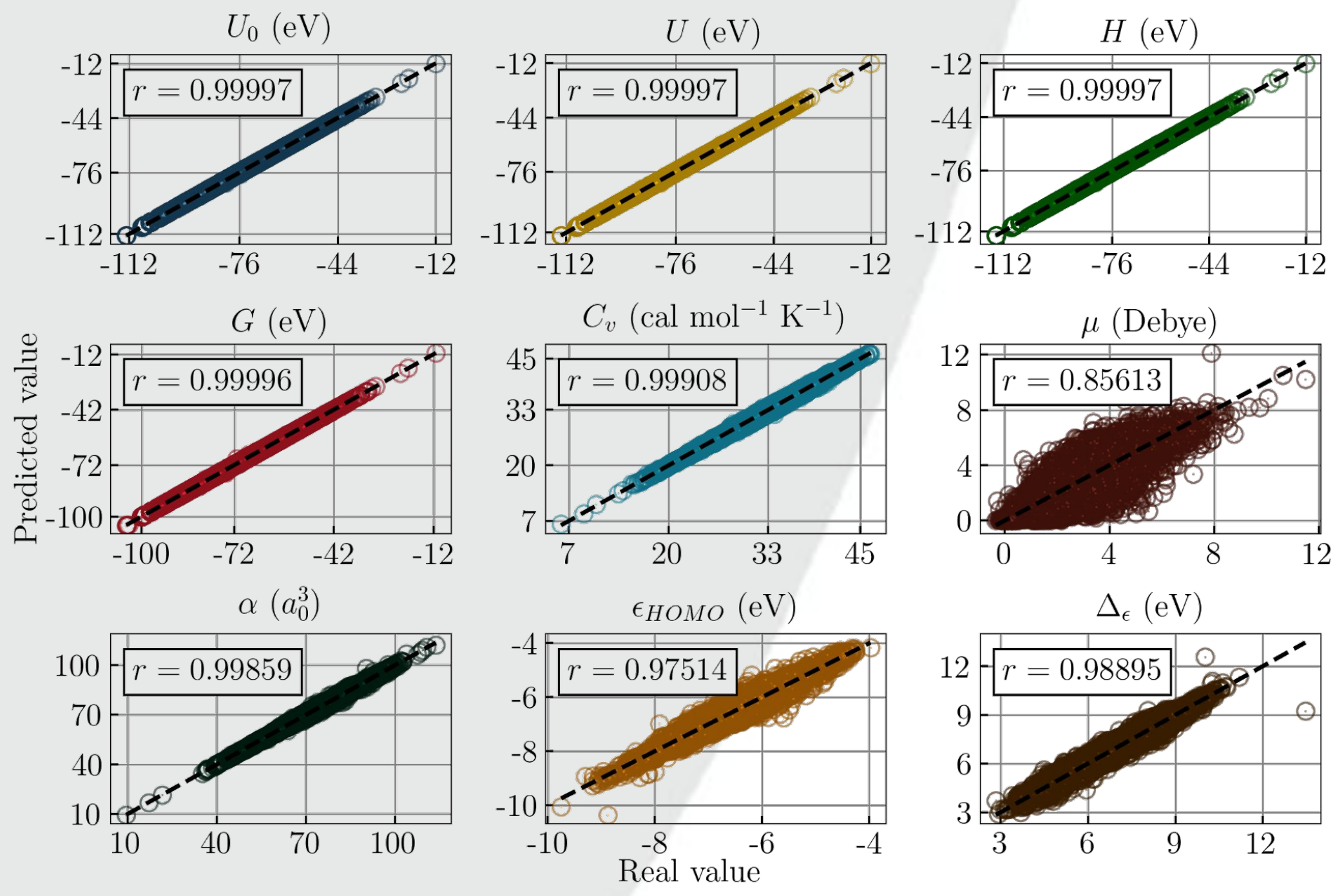
SMILES-based Descriptors for Property Prediction



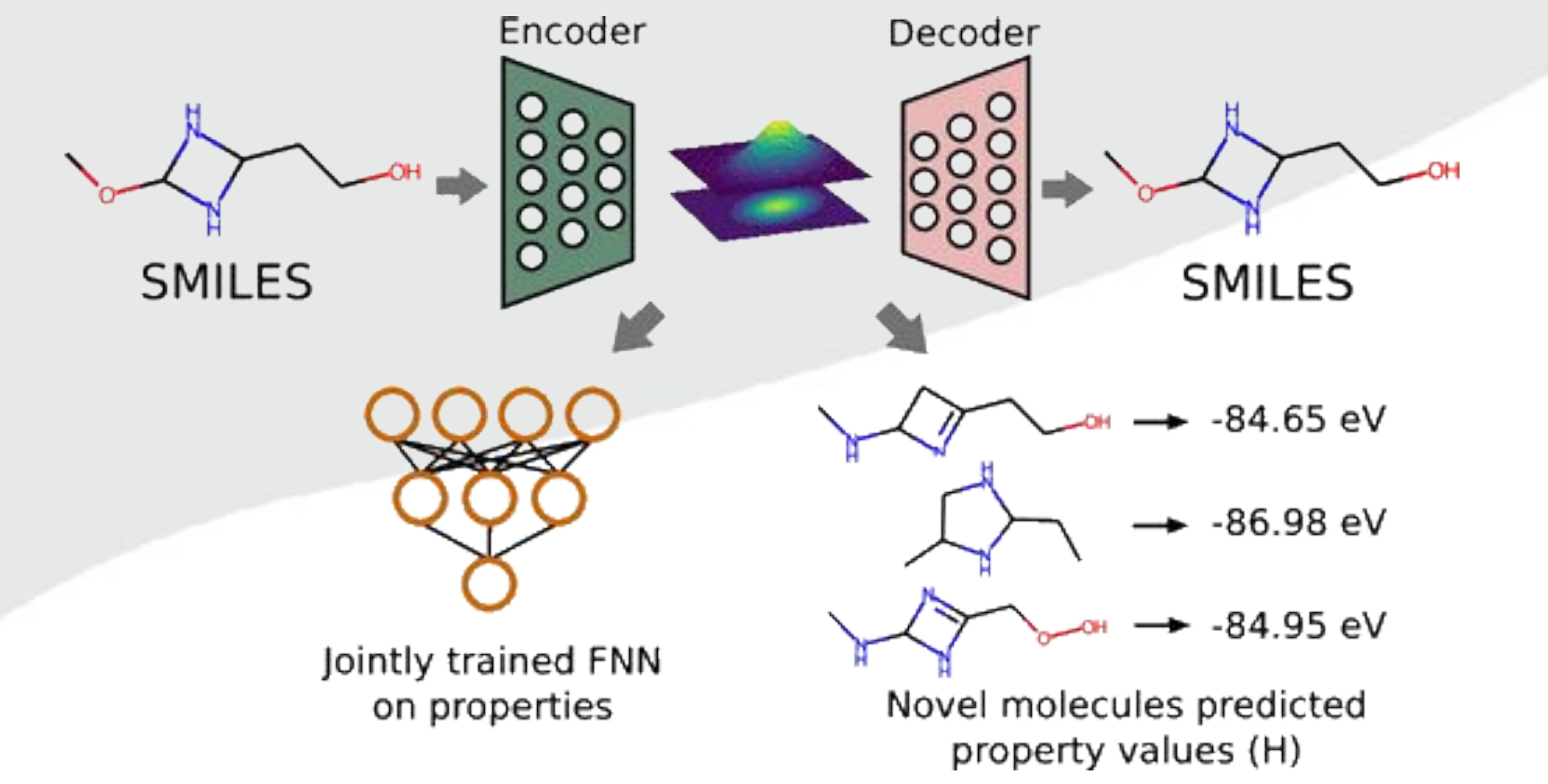
We proposed a **supervised** approach to predict molecular properties without relying on molecular geometry. For this, we used a collection of **molecular descriptors based on the simplified molecular-input line-entry system** (SMILES) string as input for our multi-layered perceptron model. Furthermore, we study several feature selection approaches for reducing out-of-sample errors.

Framework steps

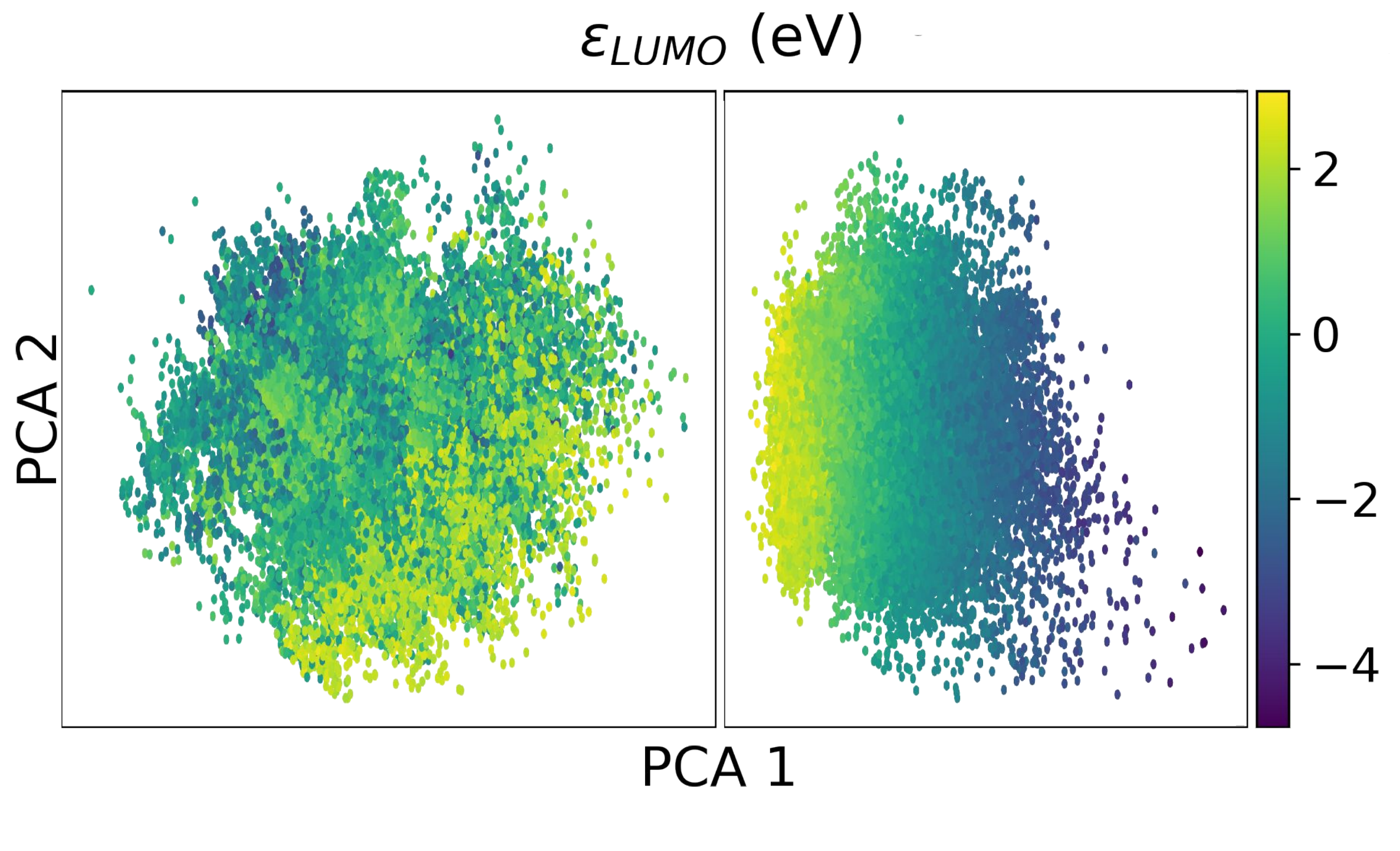

QM9 → Data exploration → Feature extraction → Feature selection → Property prediction



Molecule Design with Tailored Properties via a Variational Autoencoder

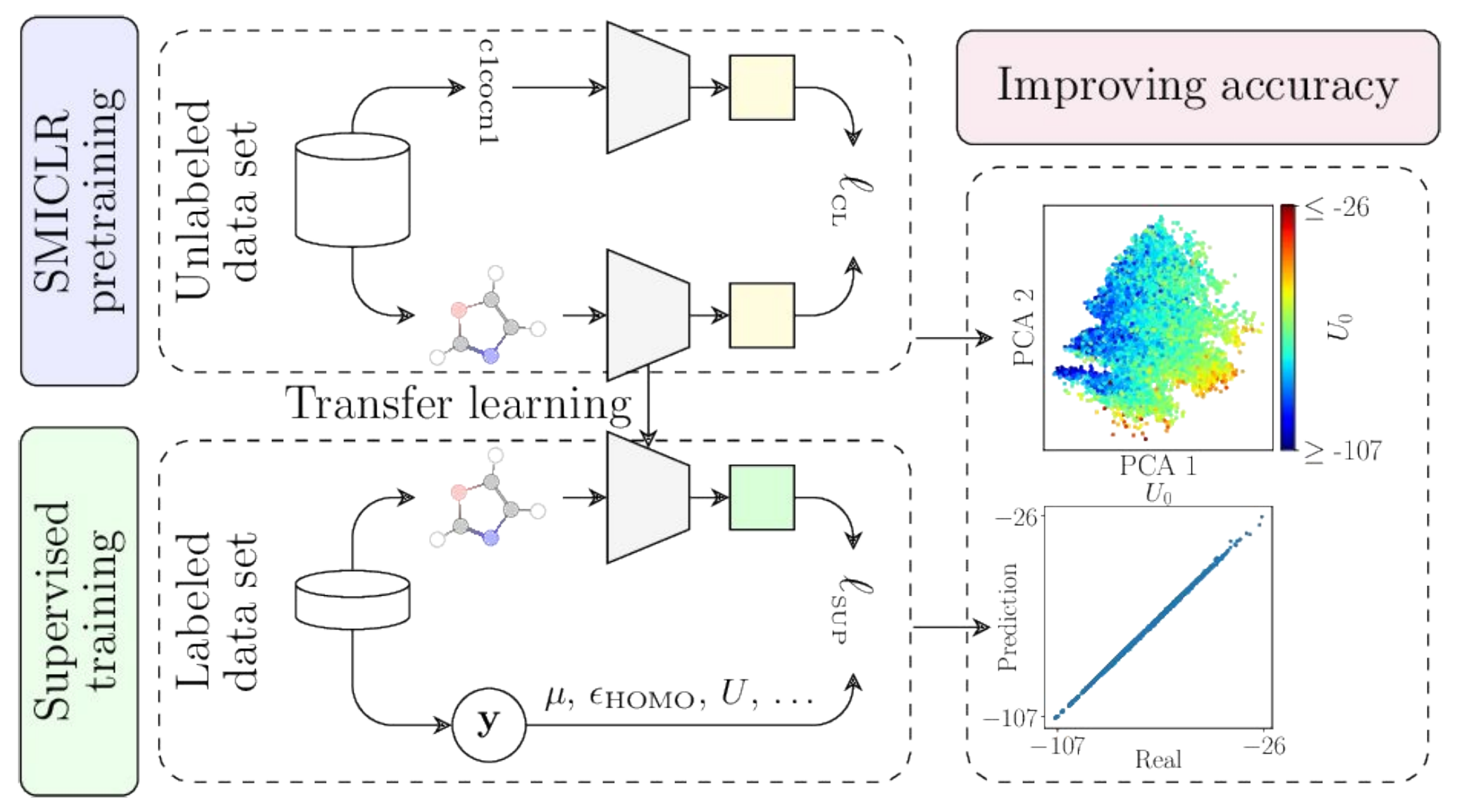

Encoder
Decoder
SMILES
SMILES
Jointly trained FNN on properties
-84.65 eV
-86.98 eV
-84.95 eV
Novel molecules predicted property values (H)

We united the prediction of molecular properties and the design of novel molecules under a single molecular representation. Our ML algorithm changed the default **Grammar Variational Autoencoder** (GVAE) by attaching an MLP to the latent space, thus incorporating property information into the training procedure and generating a **supervised** version of the GVAE model. The biased latent space learned by our approach demonstrated the ability to perform both molecular property prediction and produce novel molecules with desired properties.

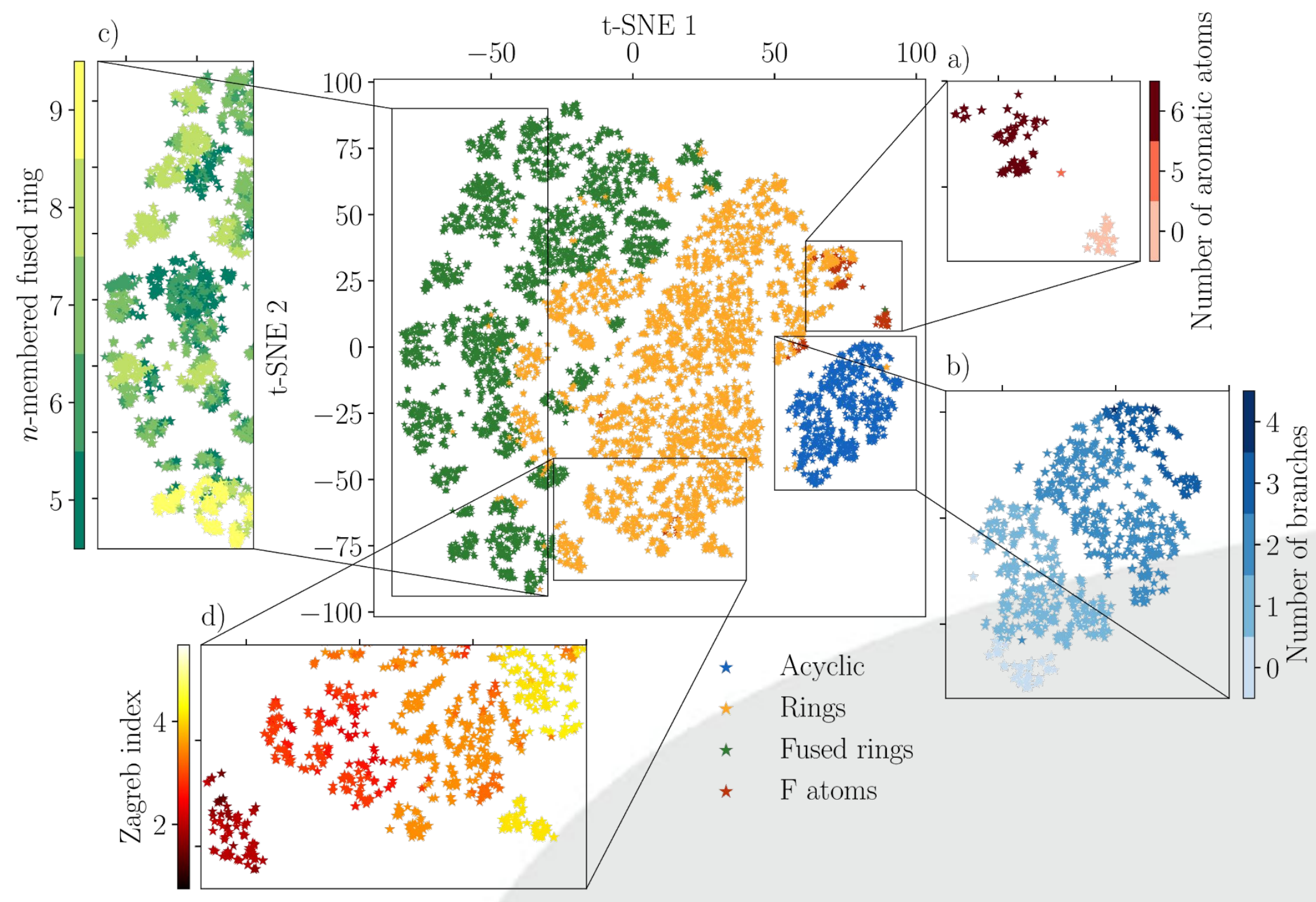


Semi-supervised Learning

Molecular Representation Learning via Contrastive Learning



We introduced SMILES Contrastive Learning (SMICLR), a multimodal **contrastive learning framework** for molecules. SMICLR jointly trains a graph and text encoder to embed two popular molecular representations, namely, the 3D molecular graph and SMILES string. Moreover, our approach incorporates data augmentation strategies to produce the correlated views for each data modality. We extensively evaluate our framework for **semi-supervised** tasks.



Conclusions and Perspectives

1. Our SMILES-based approach delivered nine times more accuracy than the result achieved with the sorted Coulomb matrix, a descriptor that depends on geometrical information. Also, our method showed better accuracies by using larger training sets and the complete set of features.
2. Our property-directed design approach for molecules demonstrated a positive effect when measuring reconstruction accuracy, prior validity, and the percentage of unique molecules.
3. SMICLR achieved superior performance over the supervised baseline and semi-supervised methods that achieved state-of-the-art results. Thus, SMICLR highlighted potential to enhance prediction tasks by effectively incorporating unlabeled data from the chemical space and relying less on labeled data.

References

1. Pinheiro, Gabriel A., et al. "Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset." The Journal of Physical Chemistry A 124.47 (2020): 9854-9866.
2. Oliveira, André F., Juarez LF Da Silva, and Marcos G. Quiles. "Molecular Property Prediction and Molecular Design Using a Supervised Grammar Variational Autoencoder." Journal of Chemical Information and Modeling 62.4 (2022): 817-828.
3. Pinheiro, Gabriel A., Juarez LF Da Silva, and Marcos G. Quiles. "SMICLR: Contrastive Learning on Multiple Molecular Representations for Semisupervised and Unsupervised Representation Learning." Journal of Chemical Information and Modeling (2022).

Acknowledgements

CINE CENTER FOR INNOVATION ON NEW ENERGIES
FAPESP
USP
IQSC
UNIFESP
INPE