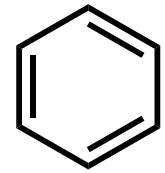
Molecular Property Prediction Based on Scarce Data Using a Novel Machine Learning Framework

Quantitative Structure-Property Relationships (QSPR) is the historical standard machine learning method in cheminformatics

QSPR Paradigm:



Molecule

Constitutional: number of specific atoms, functional group, bonds…

Topological: size, branching degree, ring structure…

Quantum Chemical: Frontier molecular orbitals, partial charge distribution, polarizability, hardness…

…

Hundreds of descriptors have been developed for QSPR models.

Conversion to a set of descriptors (X_i)

$$f(\mathbf{X}) = c_0 + \sum_{i=1}^{N} c_i X_i$$

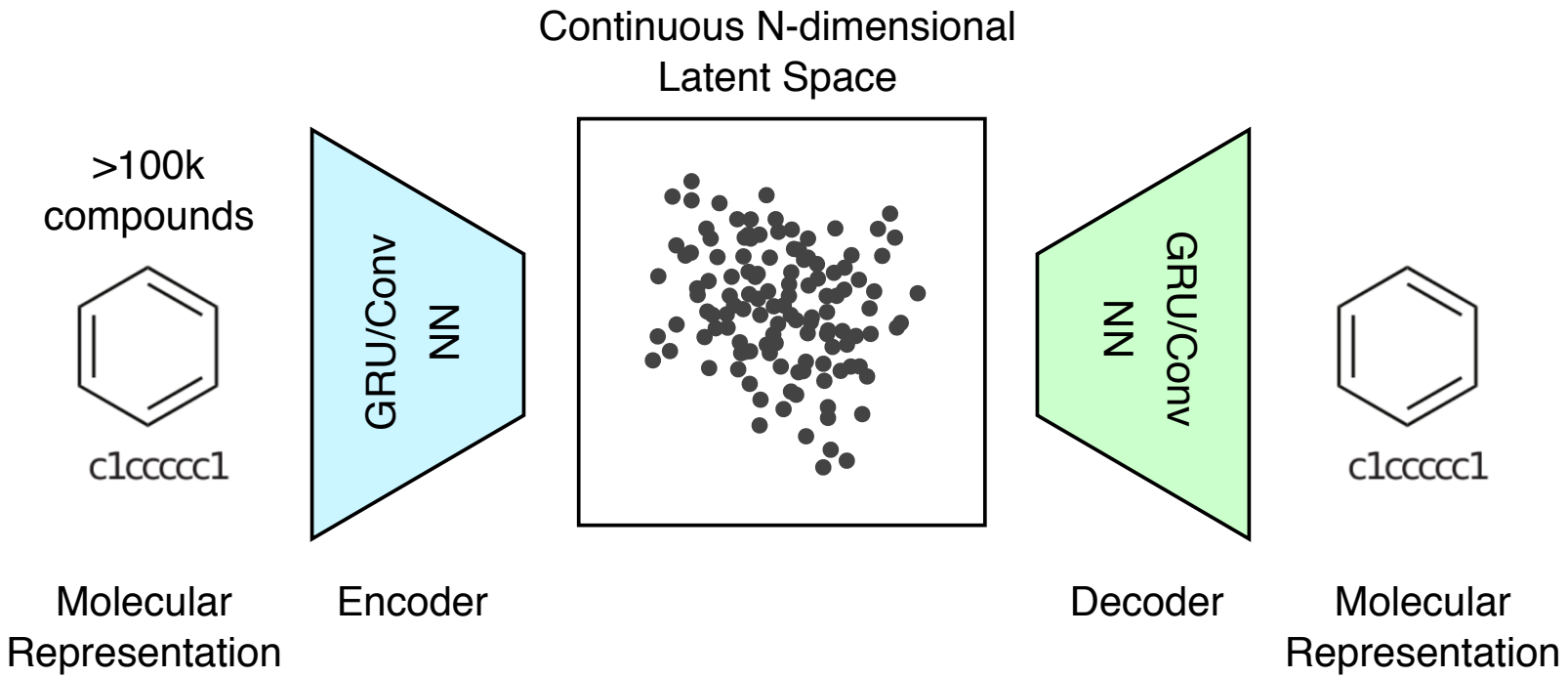
Calculation of property via linear regression

A particular QSPR model is derived by fitting the c_i parameters to reproduce data from existing databases.

It's interesting to note that this is exactly the now obsolete paradigm that once dominated image recognition.

Molecular Property Prediction Based on Scarce Data Using a Novel Machine Learning Framework

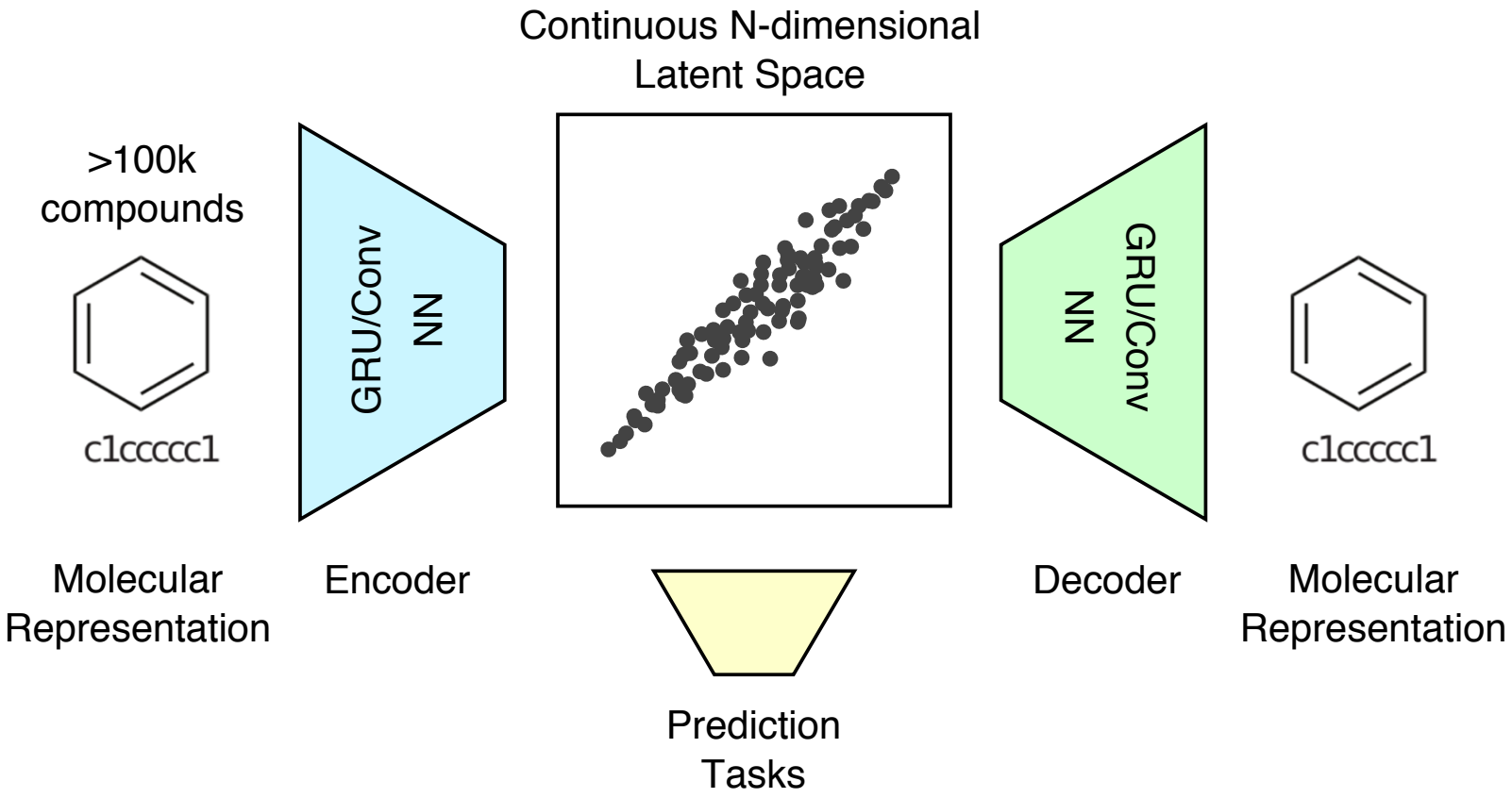
Auto-Encoder Paradigm:



Idea: By forcing the molecular representation through a low-dimensional vector space, the autoencoder “learns” a data-efficient representation of chemical structures.

Molecular Property Prediction Based on Scarce Data Using a Novel Machine Learning Framework

Auto-Encoder Paradigm:



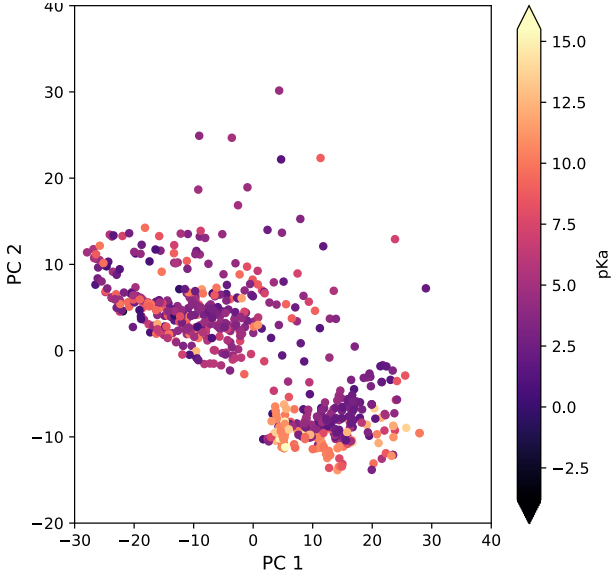
Potential: Joint training of decoding and predictor task reorganizes the latent space to put chemically similar molecules in the same neighborhood. Multi-objective training on correlated properties has the effect of improving the prediction of both.

Molecular Property Prediction Based on Scarce Data Using a Novel Machine Learning Framework

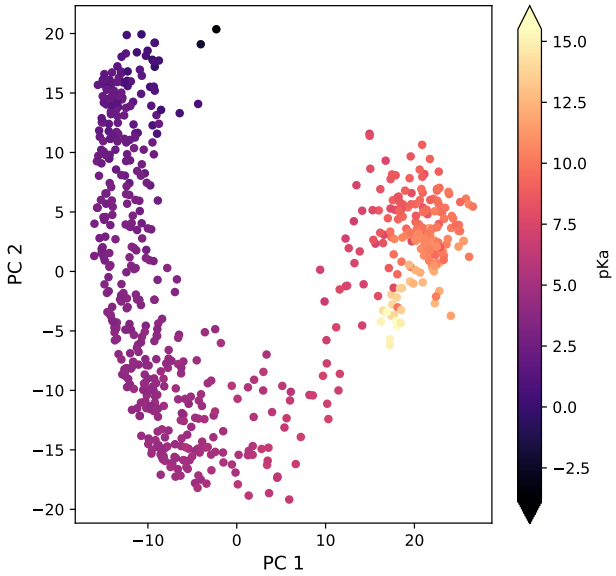
Latent Space Restructuring:
joint-training on pKa (aq) and ΔG_s prediction tasks (N~600) show clear reorganization of latent space.

Errors:
pKa+ΔG_s MUE ~ 0.5
pKa MUE ~1.2

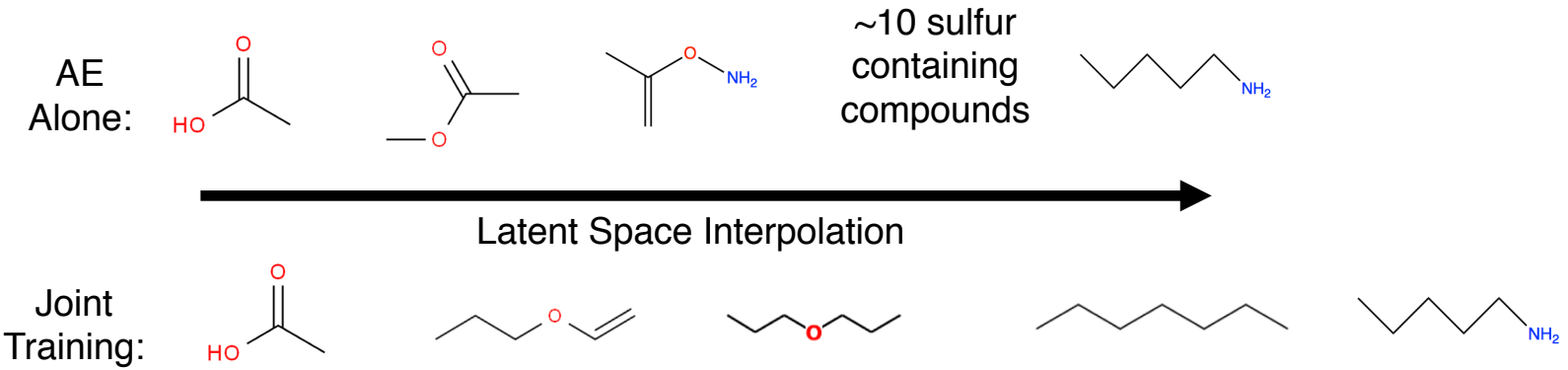
Autoencoder Alone



Joint Training



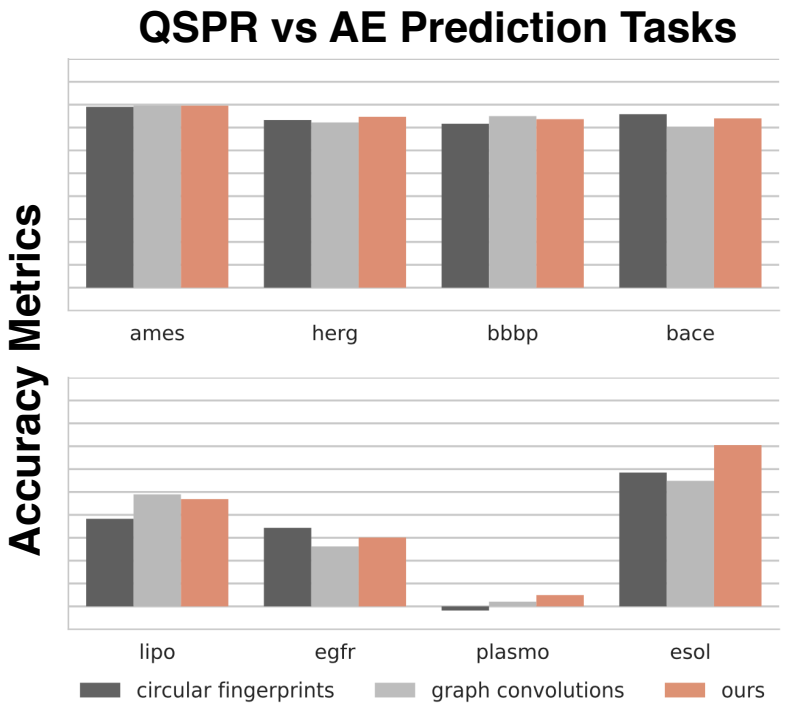
Inverse Design with Respect to Figure of Merit:



Molecular Property Prediction Based on Scarce Data Using a Novel Machine Learning Framework

Proposed activities:

- To date, autoencoder based predictions outperform QSPR based models in most head-to-head tasks.
- Autoencoder based models are particularly interesting for scarce data applications, where multiple data sources can be combined.
- Establish a partnership to develop autoencoders that are useful for predicting safety figures of merit. (e.g., SADT, TD_{24})
- Develop training data sets for validation and comparison.
- Deliver novel targets via inverse design to industrial collaborators.



Winter et al. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Arxiv* **2018**

Gómez-Bombarelli, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276.