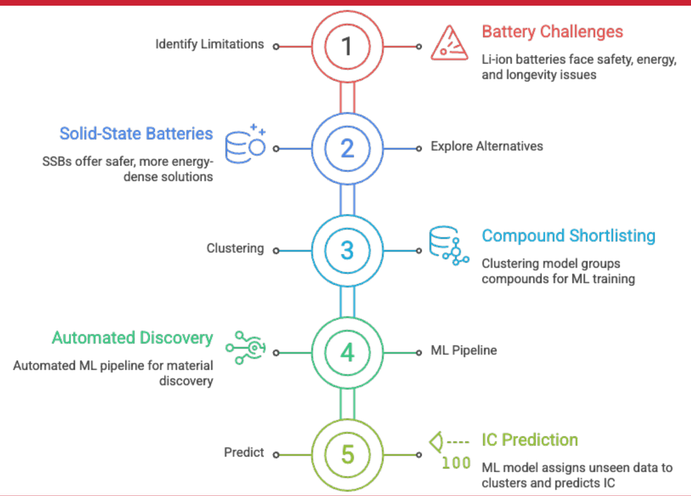
Molytics:
Entropy-Guided AutoML with Deep Clustering for Materials Science

Samarpan Mohanty [1], Sabyasachi Mohanty [2], Bijan Paul [1], Mona Bavarian [2]*

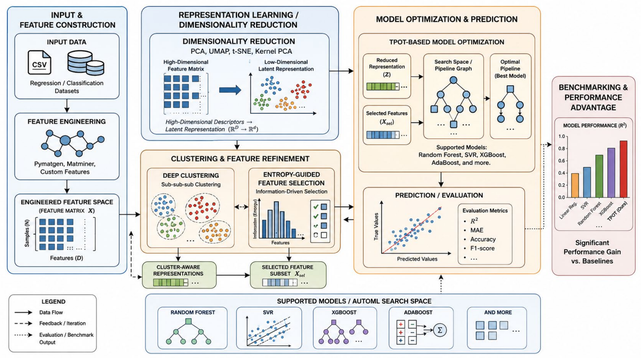
[1] School of Computing, University of Nebraska-Lincoln, Lincoln, NE, USA 68588

[2] Department of Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-8286, USA

Introduction



Pipeline Architecture and Modeling

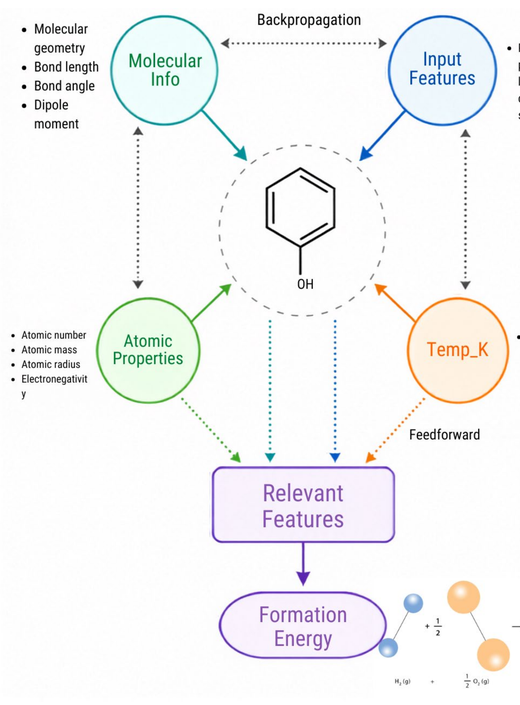


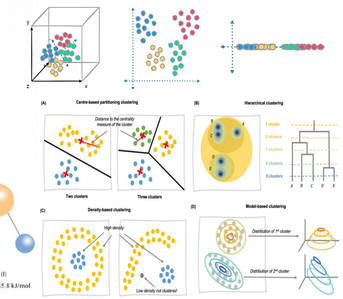
Modeling & Evaluation: TPOT automates the creation of ML pipelines through **genetic programming.**

- **Regressors:** XGBoost, Gradient Boosting, ElasticNet, SVR, KNN, Random Forest and etc,
- **Classifiers:** Random Forest, AdaBoost, Logistic Regression and etc,
- **Preprocessors:** Scaling (Standard, MinMax), Polynomial expansion, PCA and etc,
- **Selectors:** SelectKBest, SelectFwe, model-based importance, etc,

Each cluster is trained separately, and the best-performing pipeline is selected.

Feature Section & Clustering





Feature Selection

100+ features → only relevant ones kept

Filters:

- Mutual Information (CORR based)
- Variance Threshold
- Random Forest importance

Results:

- Reduced overfitting
- Better interpretability

Clustering Strategy

Multi-resolution hierarchy

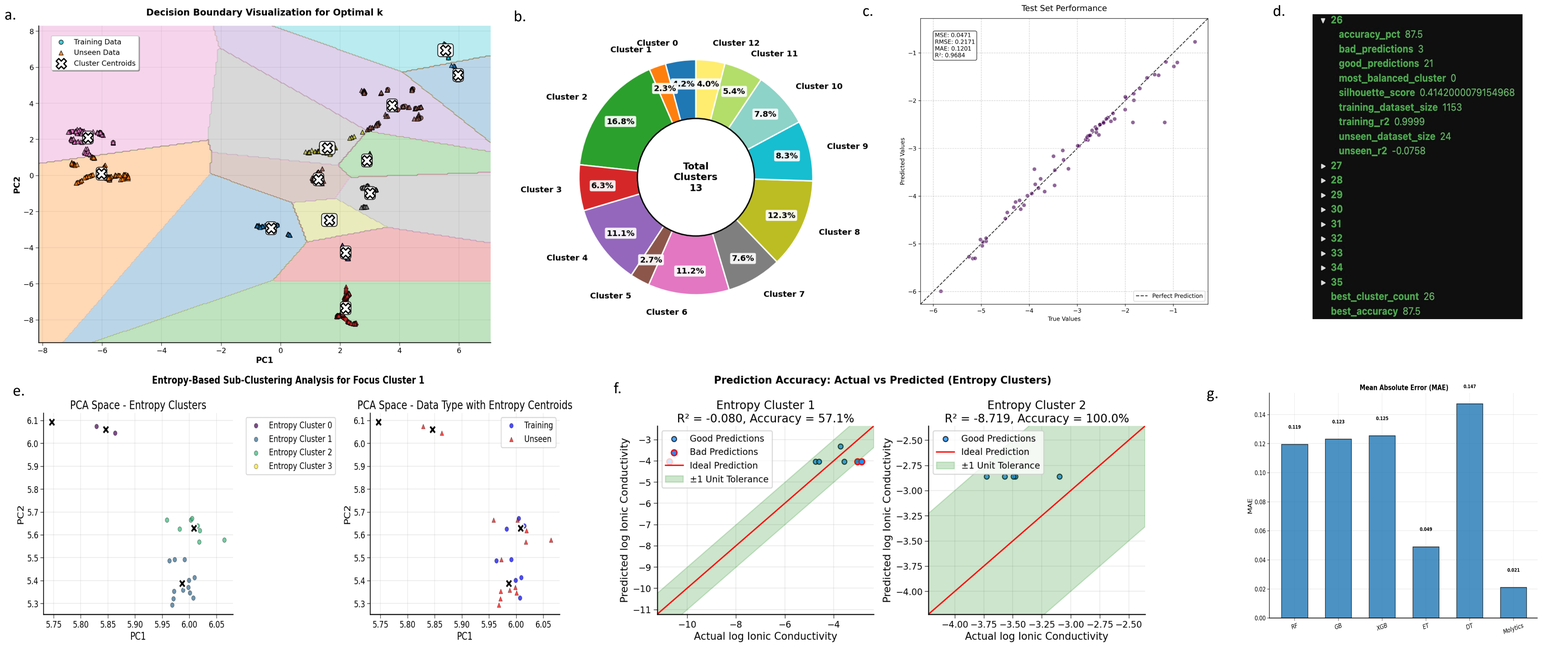
Steps:

- KMeans, DBSCAN, Agglomerative
- Sub-clustering by entropy & diversity

Benefits:

- Better local accuracy
- Clear cluster boundaries (UMAP/t-SNE)

Results



a.) PCA projection of decision boundaries across optimal clusters, b.) Cluster composition breakdown with percentage coverage, c.) Actual vs Predicted performance on test set d.) Console summary of per-cluster accuracy and best-performing regions, e1.) PCA visualization of entropy-based sub-clusters, e2.) PCA with training/test data labeled by entropy cluster, f.) Cluster 1: Predicted vs Actual with accuracy threshold, g.) Mean Absolute Error (MAE) comparison across different models

Conclusion

We present a **robust, modular AutoML framework** for predicting **arbitrary material properties**.

- **Robust AutoML framework** for predicting diverse material properties
- **75–91% accuracy** with multi-strategy clustering across datasets
- **Up to 100% accuracy** achieved via deep sub-clustered local models

Created a Web GUI for easy access.

Acknowledgement

This work was supported by the Nebraska Public Power District through the Nebraska Center for Energy Sciences Research at the University of Nebraska-Lincoln.

NEBRASKA CENTER FOR ENERGY SCIENCES RESEARCH