

MACHINE LEARNING APPROACH TO QUANTUM MECHANICS

**MEDICINE IS A SCIENCE OF
UNCERTAINTY AND AN ART OF
PROBABILITY.**

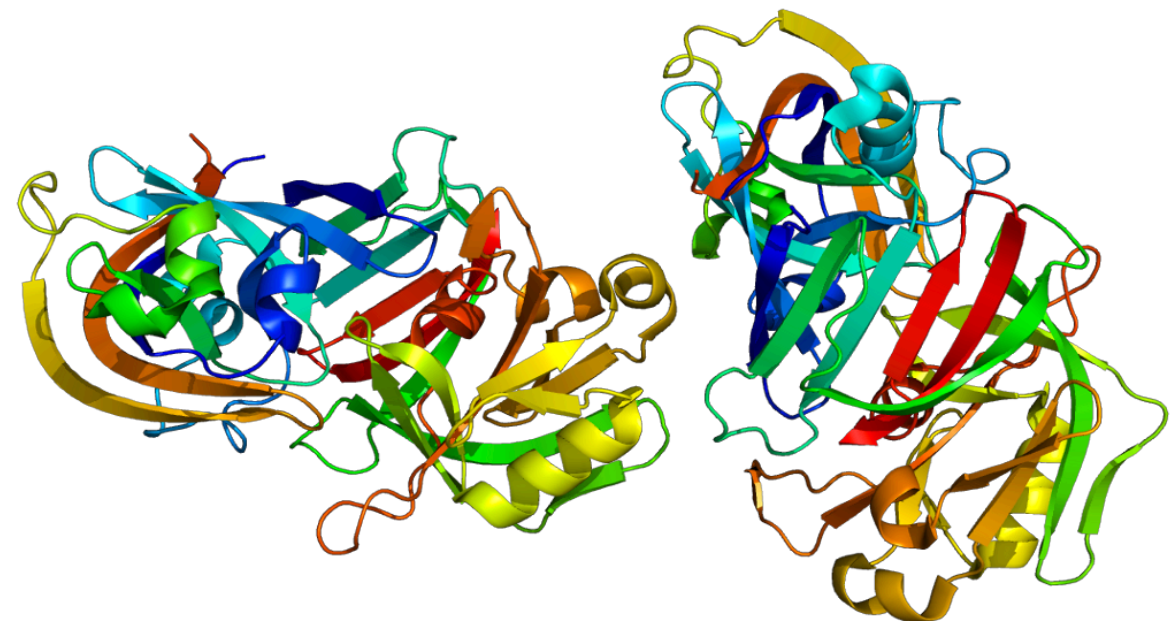
William Osler

COST TO DEVELOP NEW PHARMACEUTICAL DRUG NOW EXCEEDS \$2.87B



TAXOL: A SUCCESS STORY

- ▶ In the second half of the 20th century, researchers realized the need for systemic treatments for different types of cancer
- ▶ A new NCI institution funded such screening programs
- ▶ Welcome Taxol
- ▶ To date, Taxol is the best-selling cancer drug ever manufactured



IDENTIFY AND DESIGN

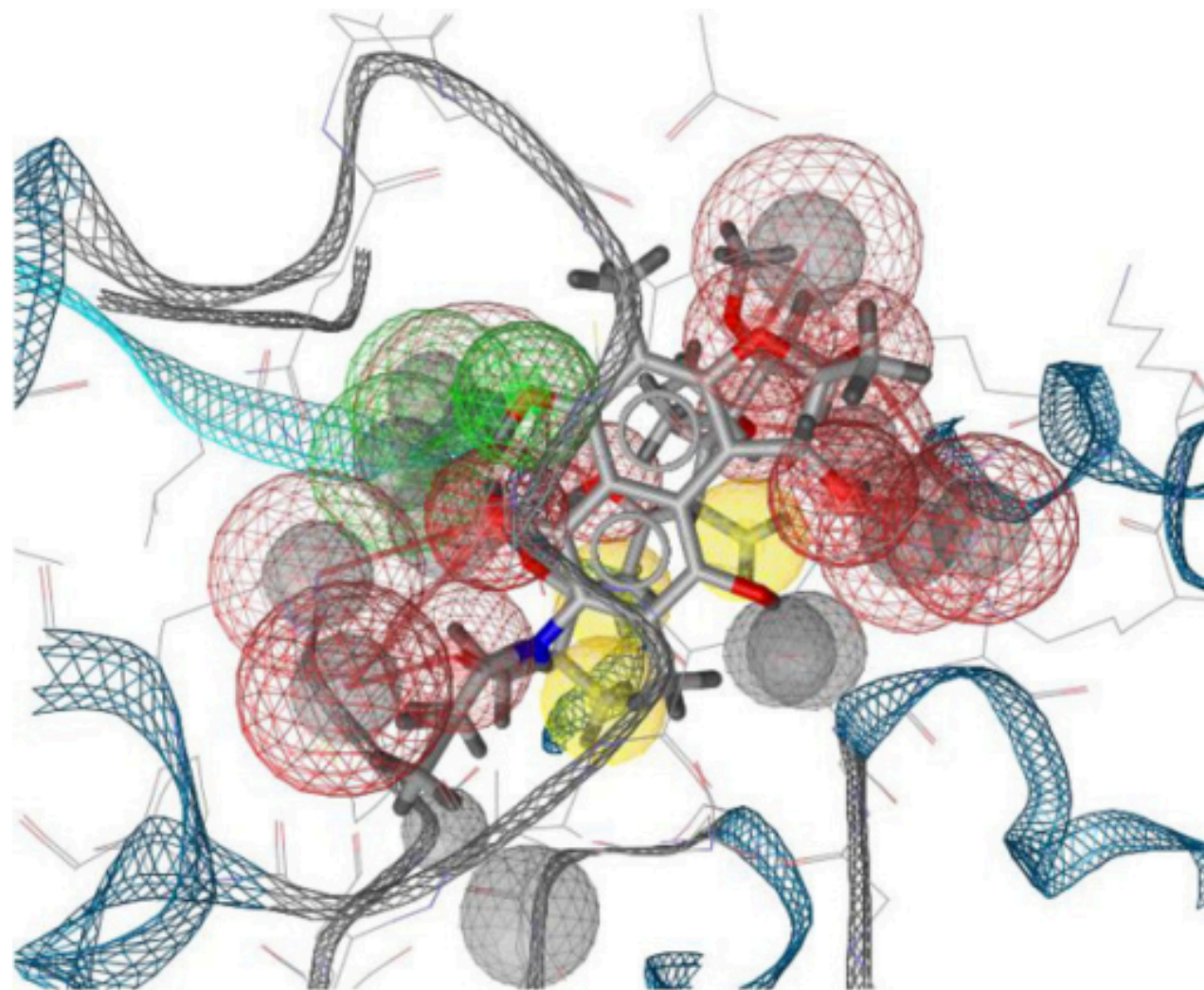
- ▶ Chemical & electrical properties
- ▶ Quantum mechanics
- ▶ Coulomb Matrix
- ▶ Computationally expensive

$$C_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & I \neq J \end{cases}$$

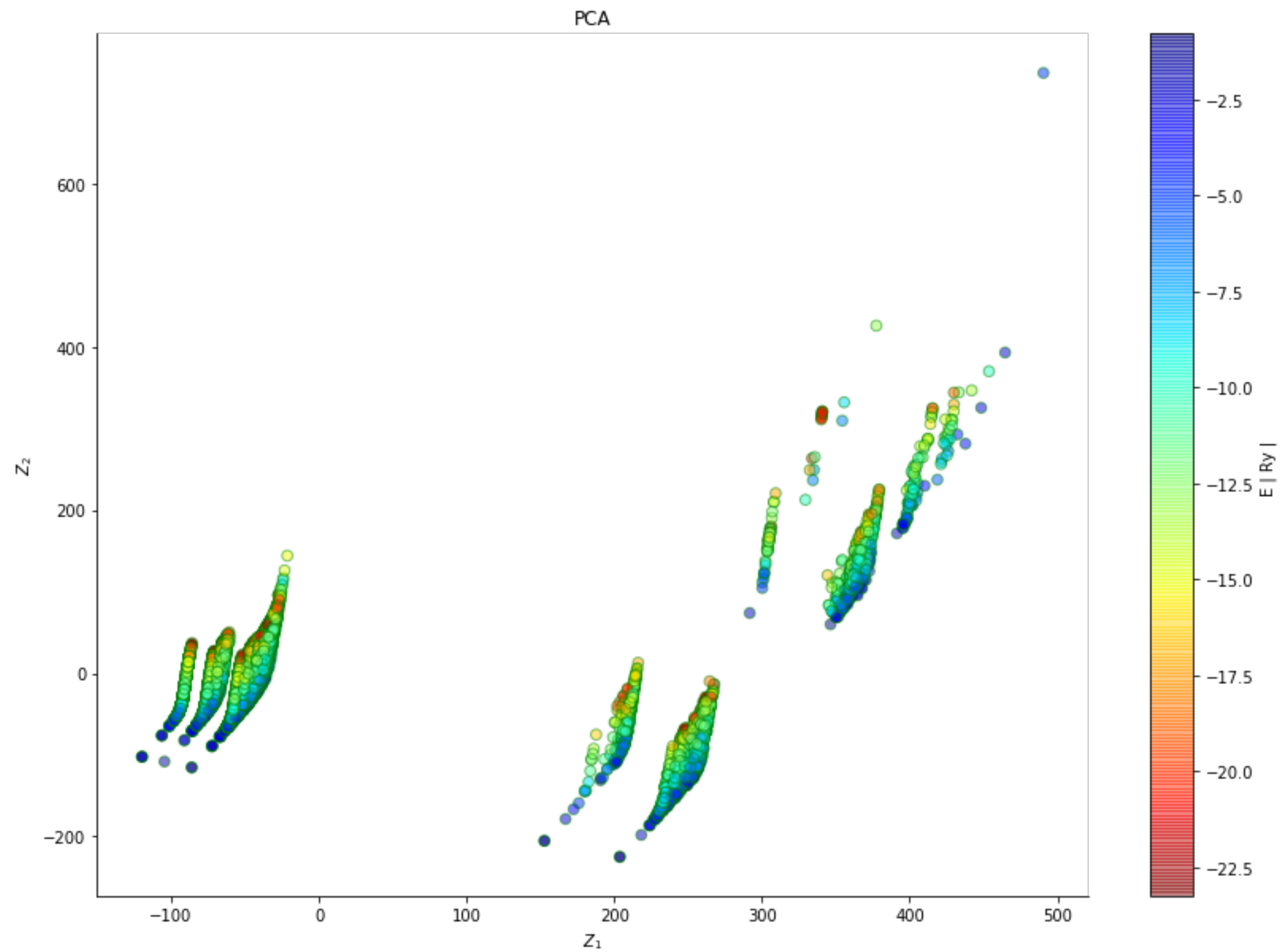
- ▶ Electronic state simulations via Density Functional Theory
- ▶ B. Himmetoglu (2016) [Tree based machine learning framework for predicting ground state energies of molecules](#)
- ▶ Restricted the atom/element set. I am not placing this constraint

FEATURES

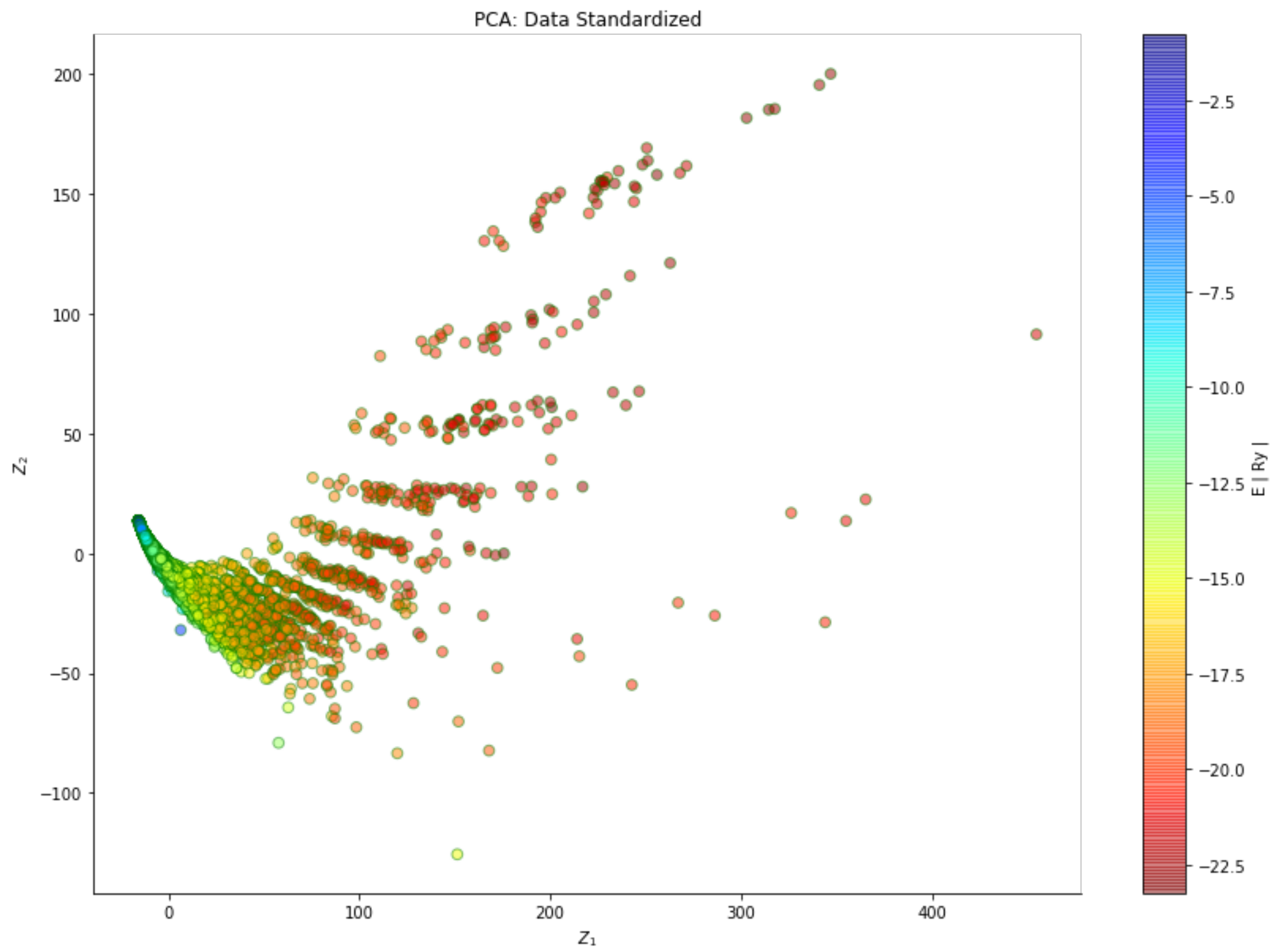
- ▶ Coulomb matrices
- ▶ Feature engineering
 - ▶ Multipoles
 - ▶ Valence electrons
 - ▶ Element counts
- ▶ 85,445 molecules with 1335 features



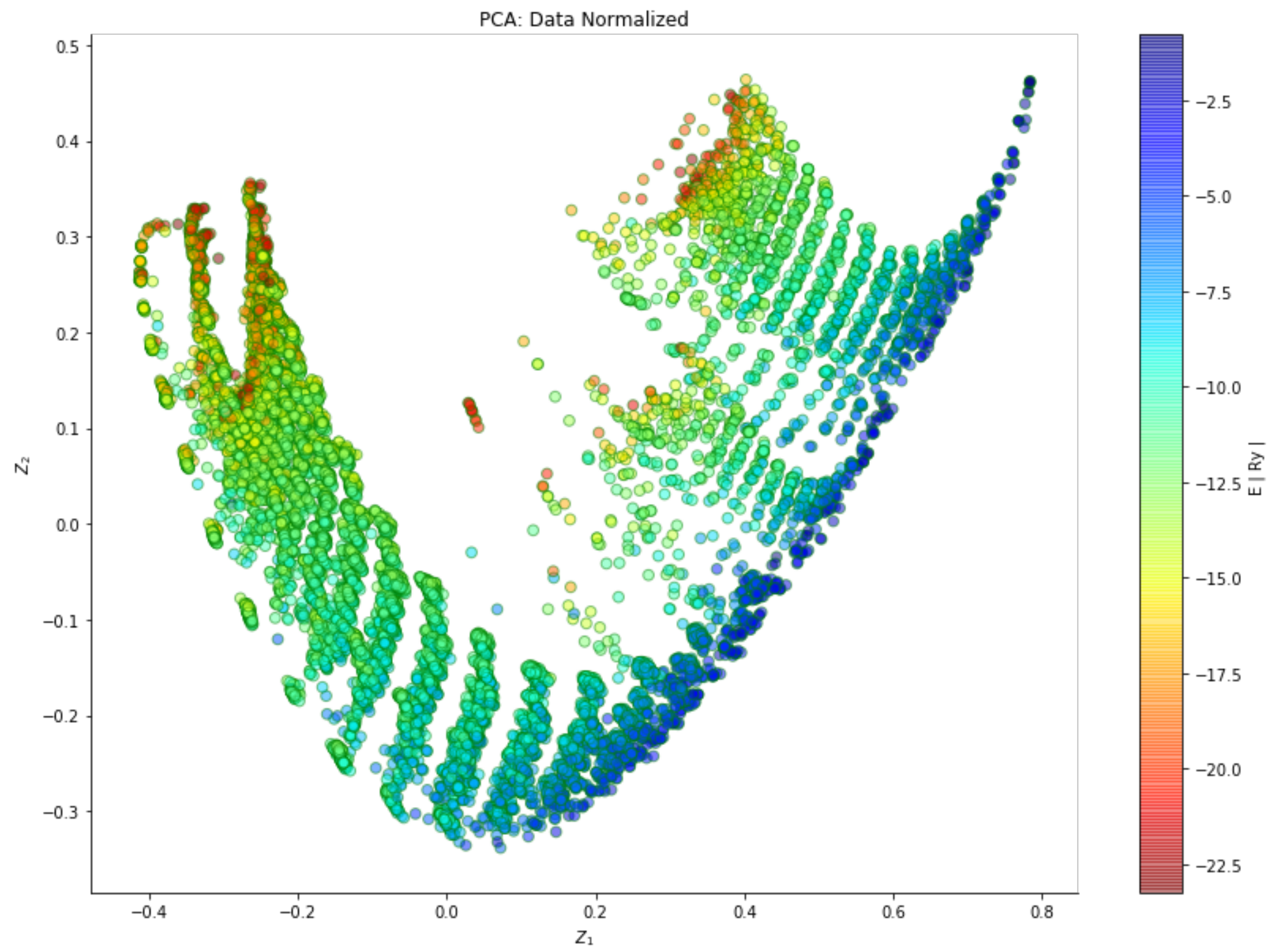
VISUALIZATION: PCA



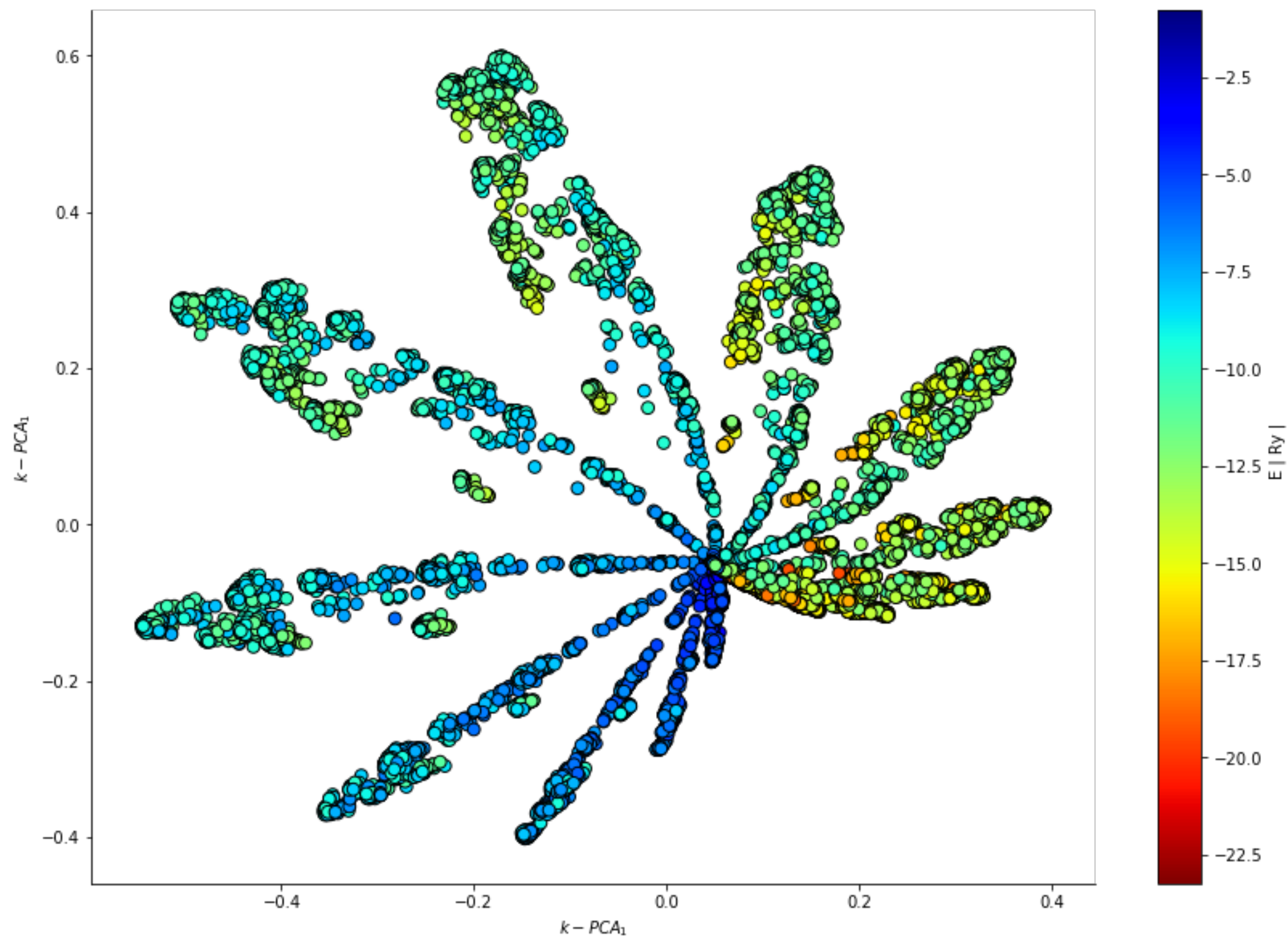
VISUALIZATION: PCA (STANDARDIZED)



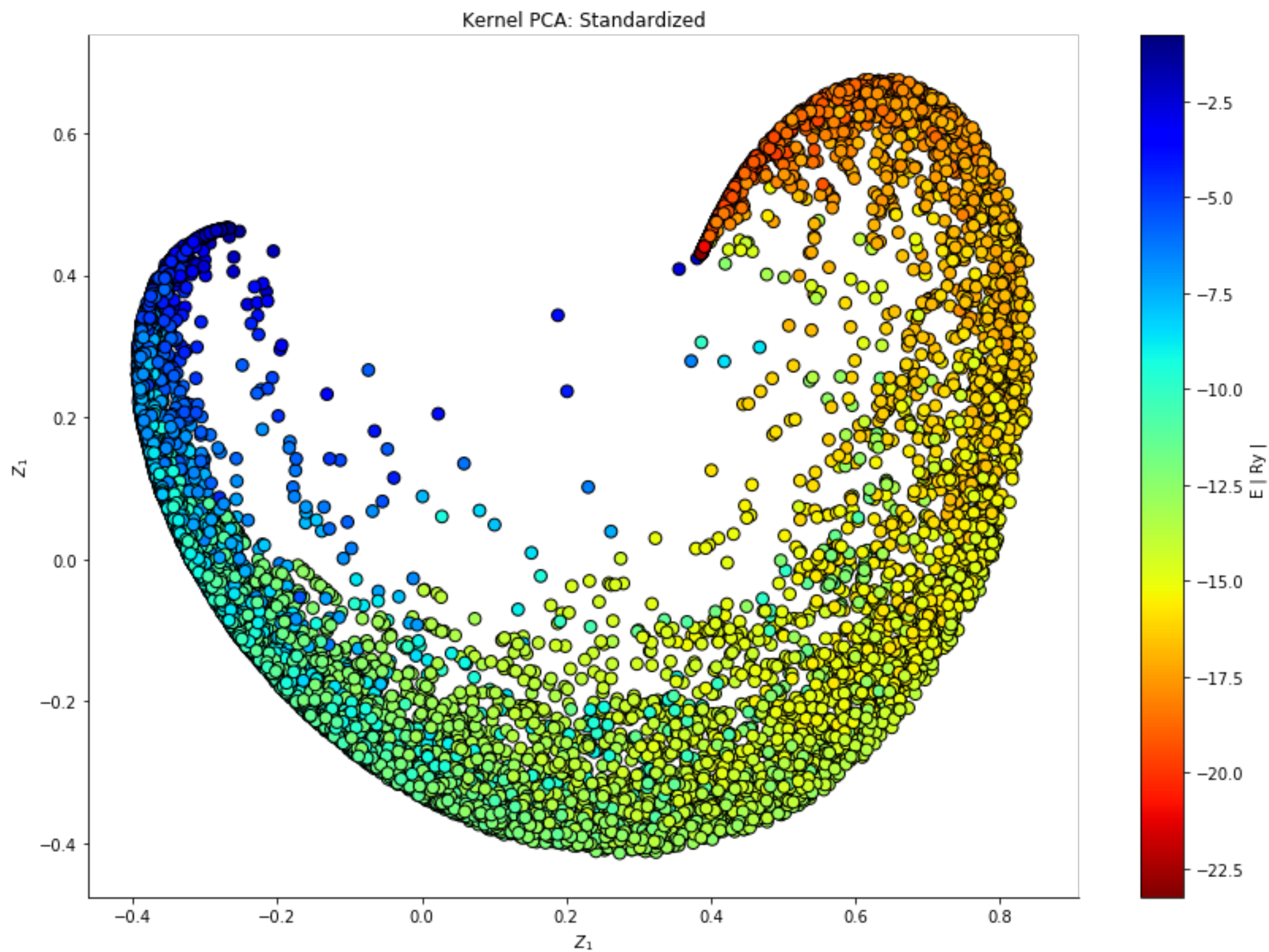
VISUALIZATION: PCA (NORMALIZED)



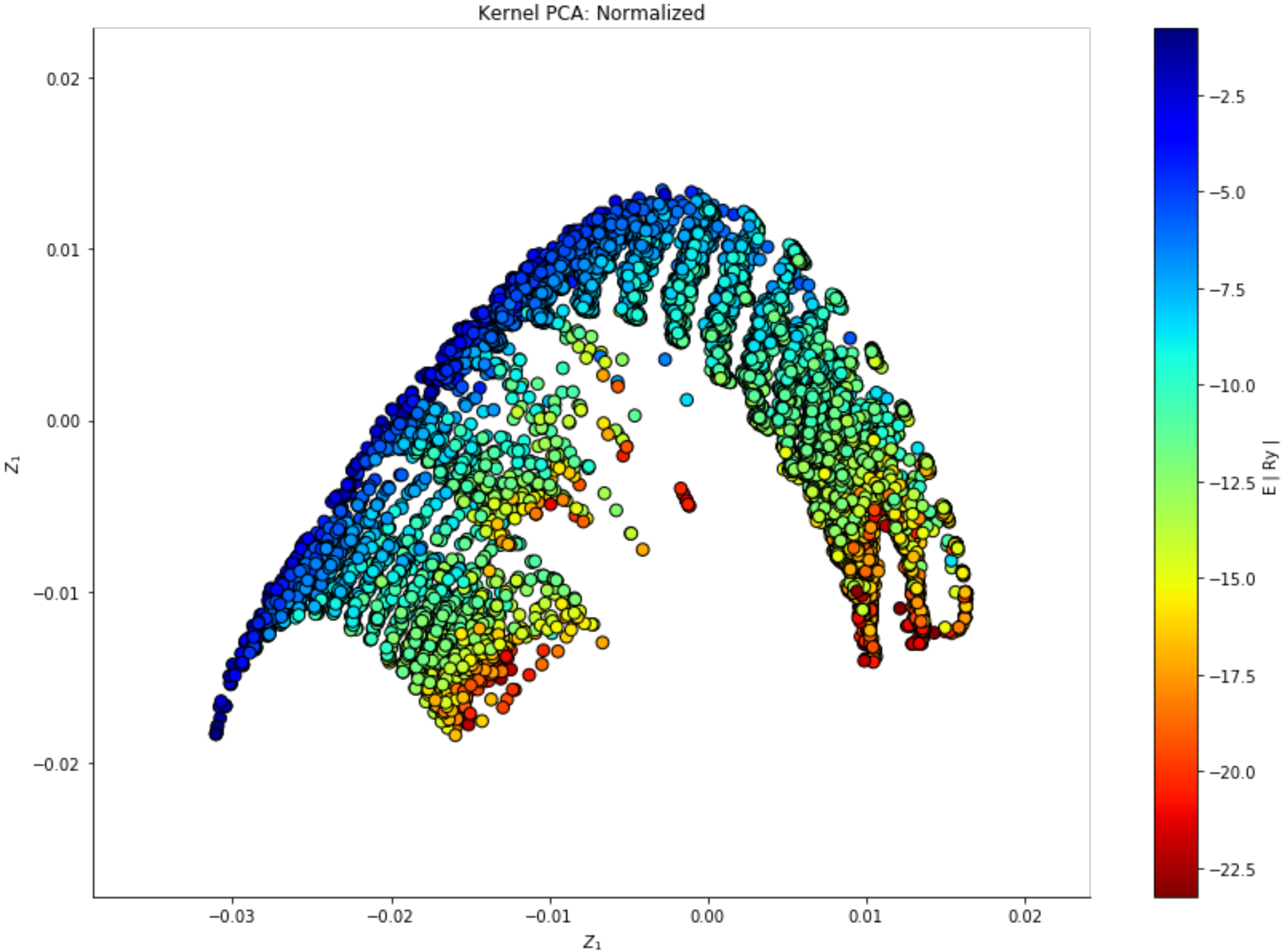
VISUALIZATION: KERNEL PCA



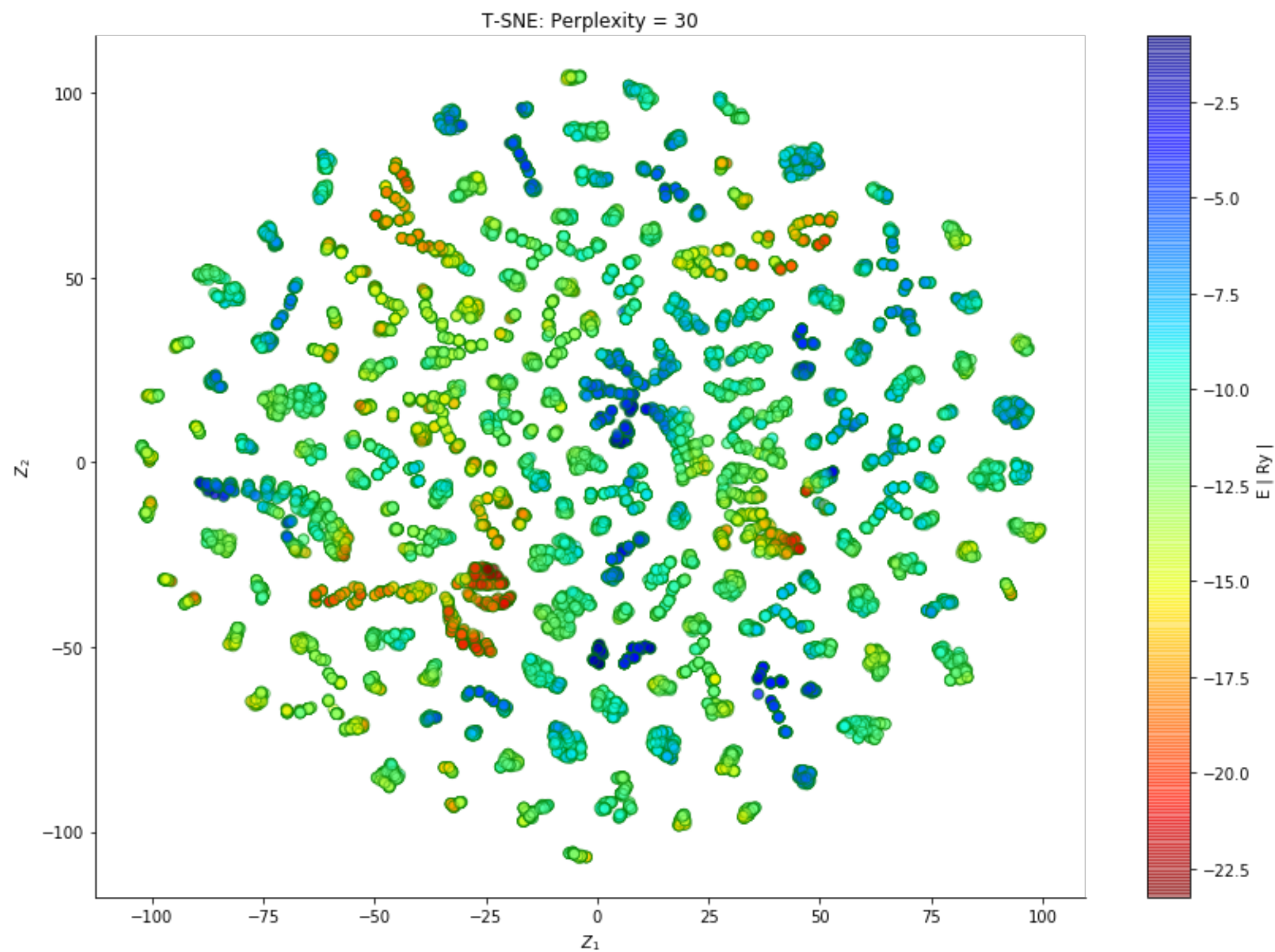
VISUALIZATION: KERNEL PCA (STANDARDIZED)



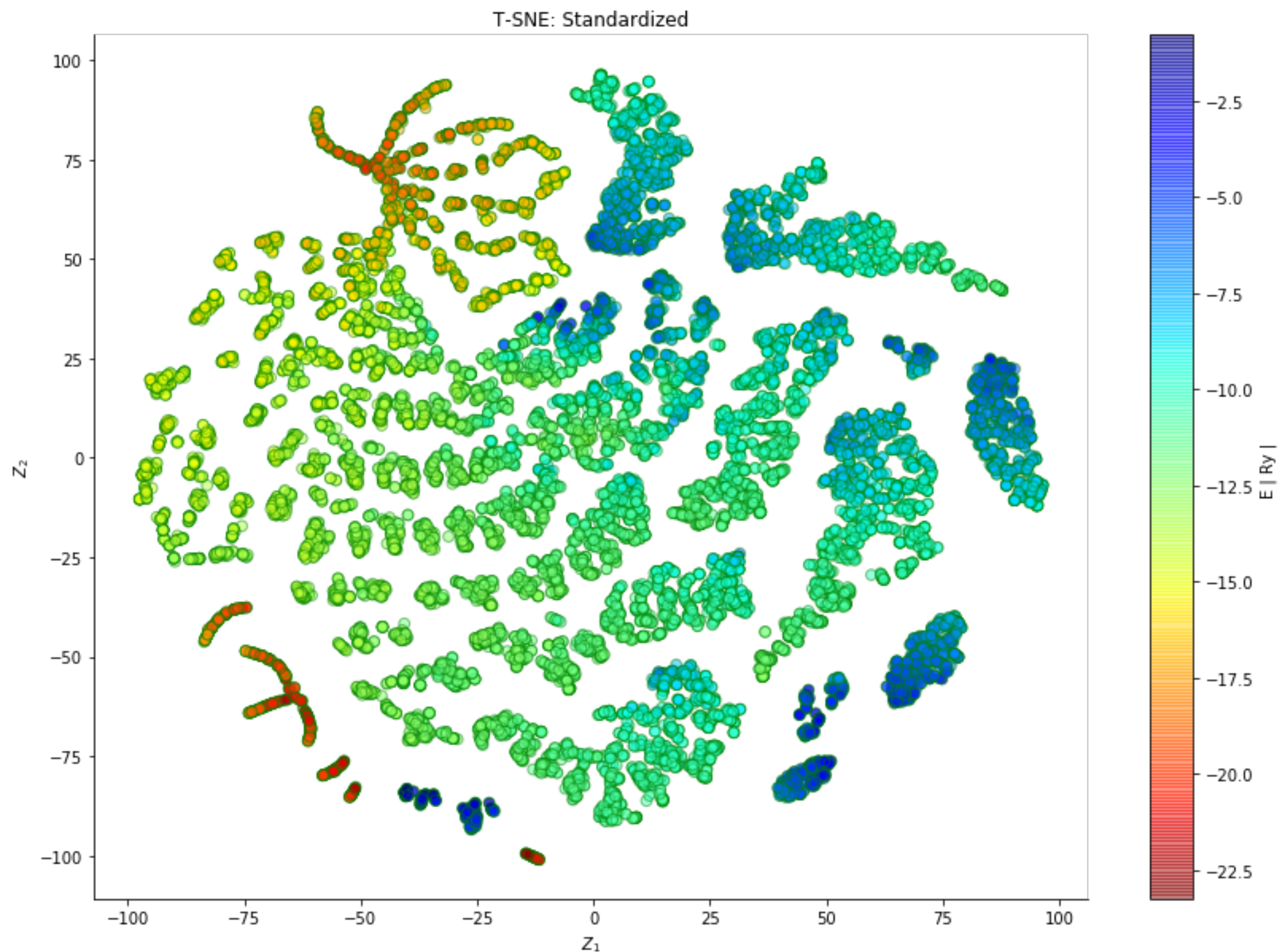
VISUALIZATION: KERNEL PCA (NORMALIZED)



VISUALIZATION: T-SNE



VISUALIZATION: T-SNE (STANDARDIZED)



MODELS

- ▶ Linear Regression
- ▶ Boosted Regression Models
- ▶ Ensemble Models

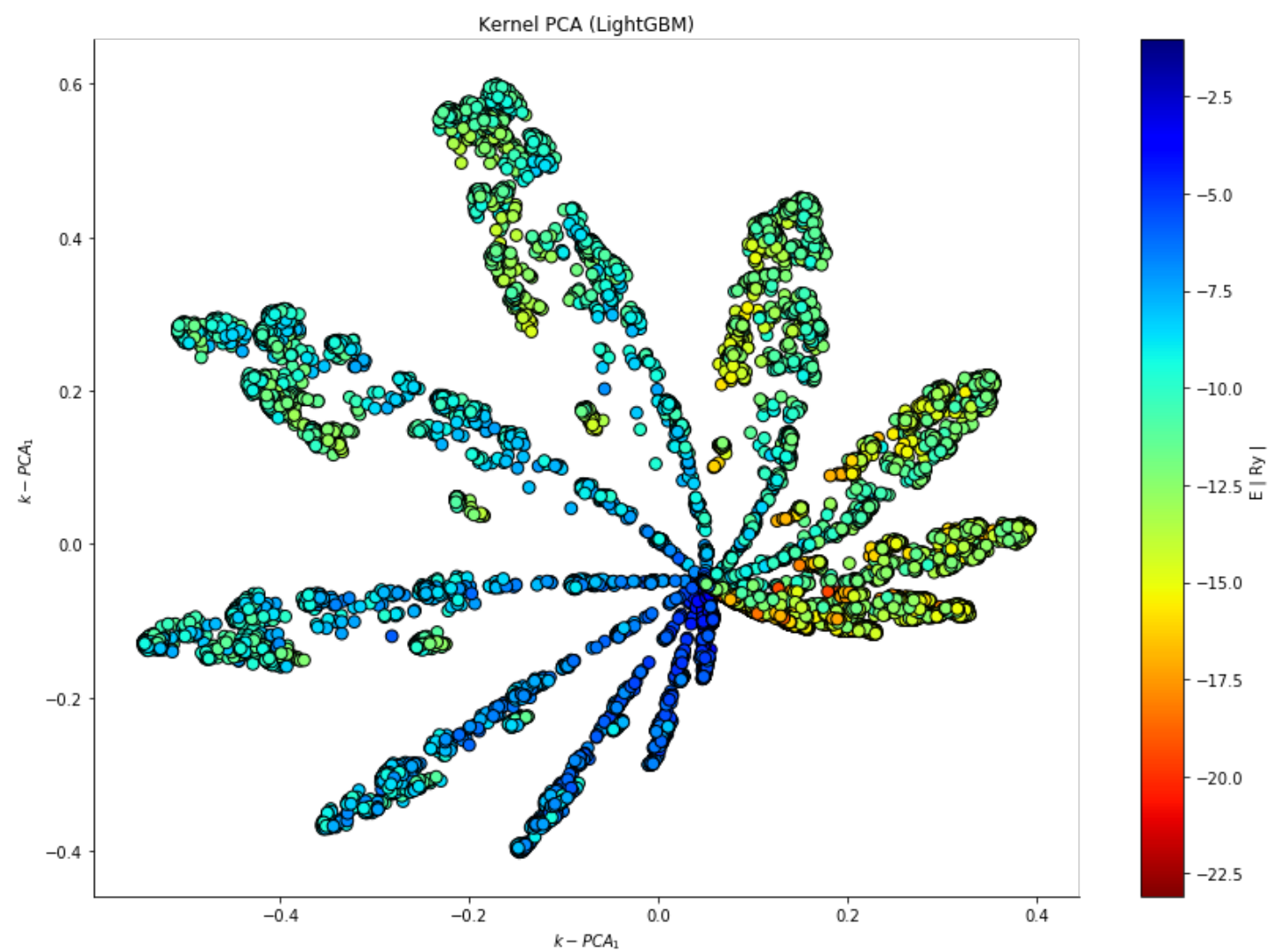
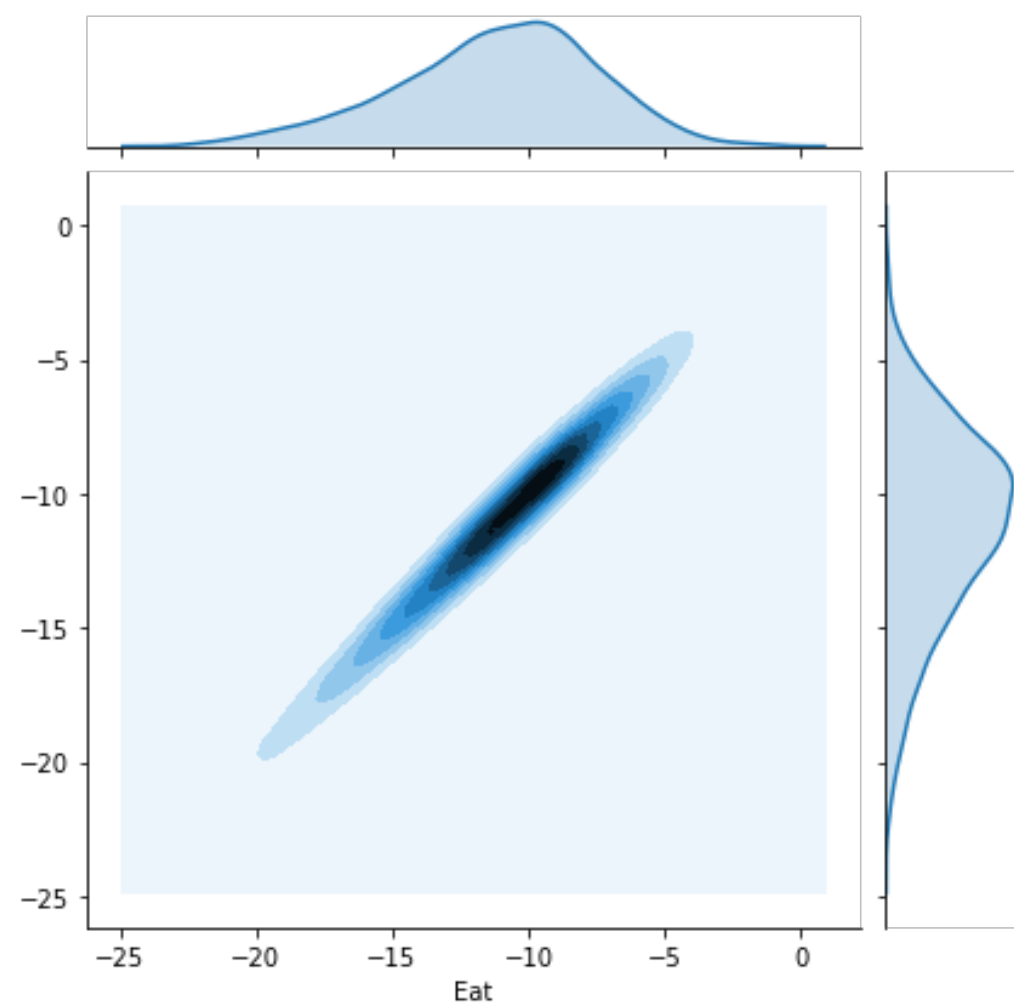
PARAMETER TUNING

- ▶ Literature
- ▶ Grid Search
- ▶ Manual Tuning
 - ▶ Trial and error/intuition

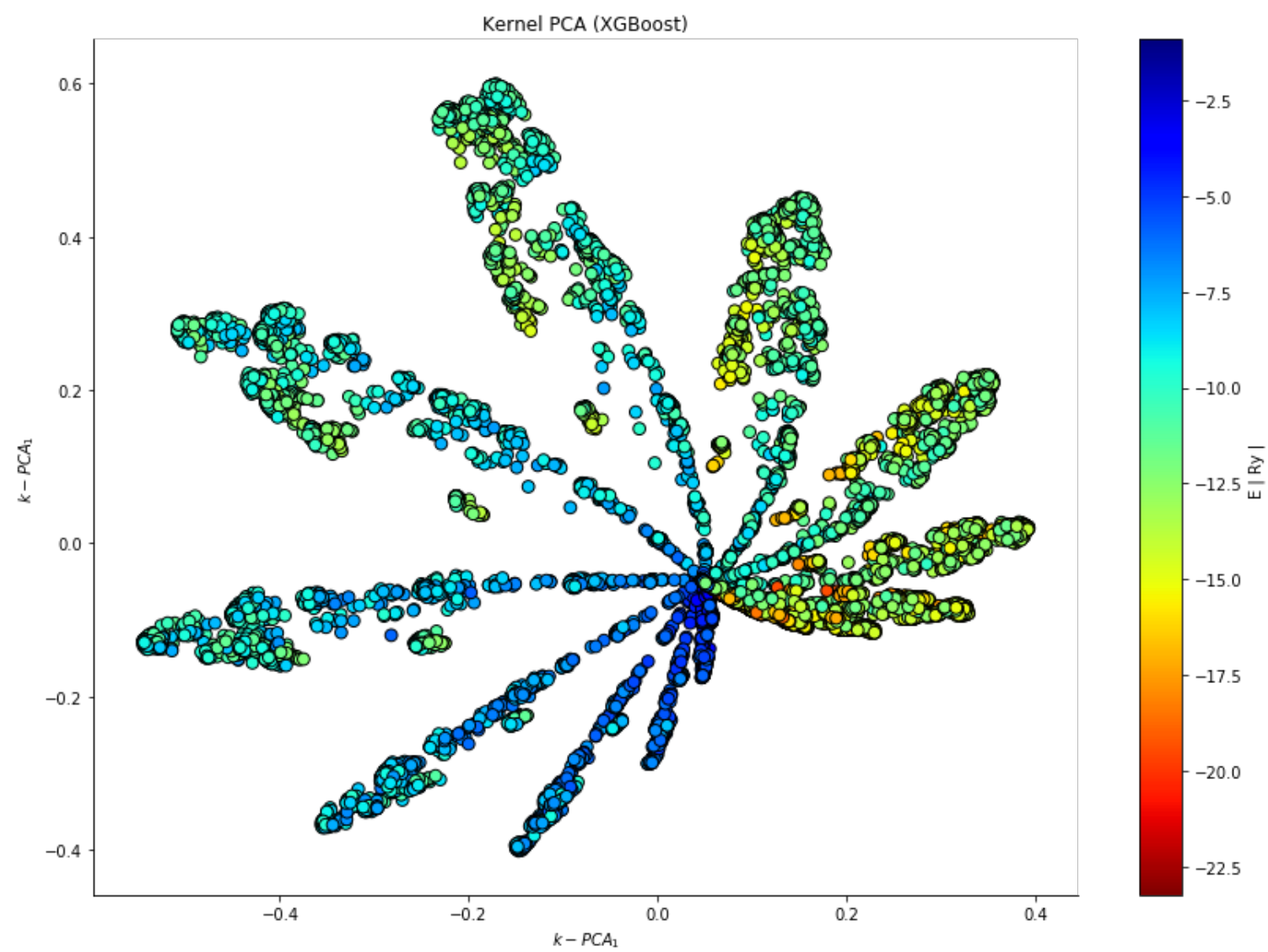
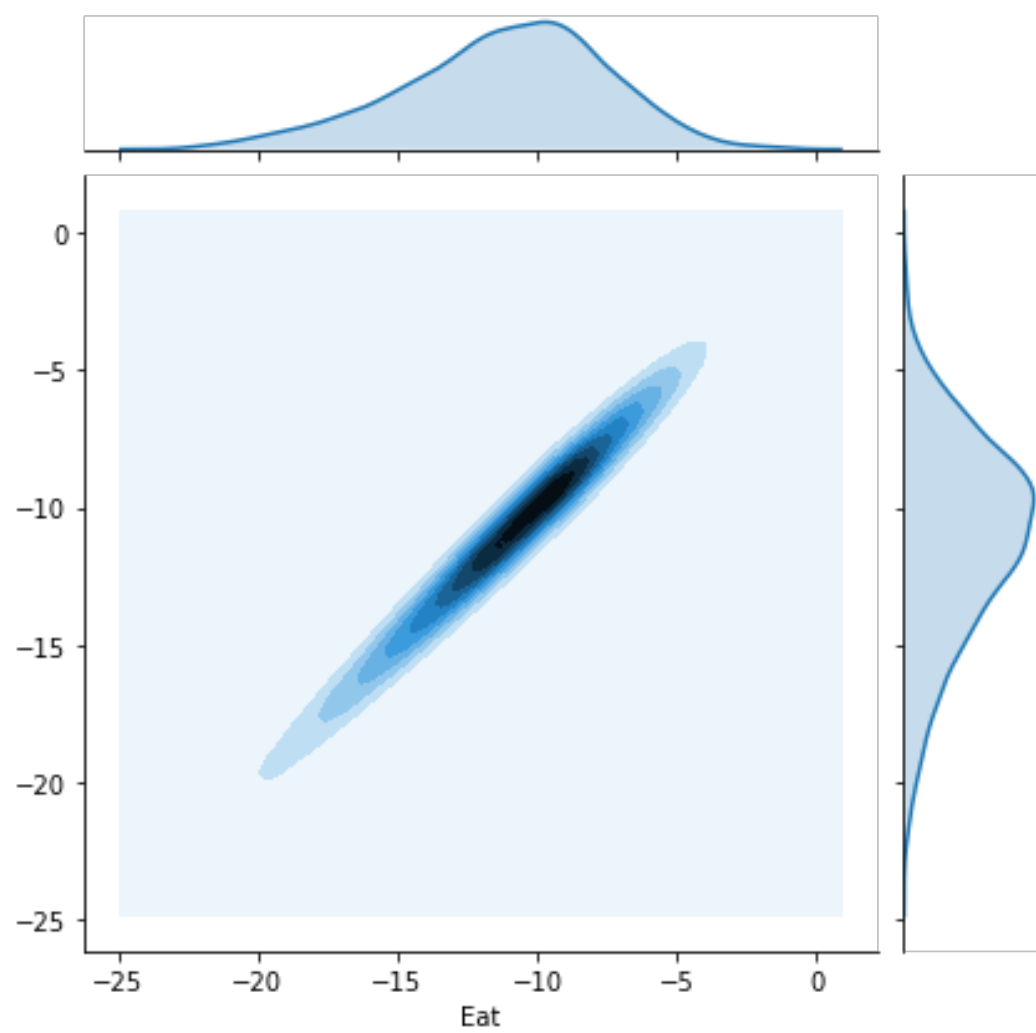
MODELS: RESULTS

MODEL TYPE	RMSE[Ry]	TEST SCORE	RMSE [kcal/mol]
LightGBM	0.0050735	0.99963	1.588
XGBoost	0.00522919	0.99957	1.636
Gradient Boosting	0.09485	0.9993	29.682
Extra Tree	0.12229	0.9989	38.268
Bagging	0.148732	0.9983	46.543
Random Forest	0.284132	0.994	88.914
K-Nearest Neighbors (k=3)	0.547451	0.9776	171.314
Linear Regression (PCA)	0.809693	0.9509	253.378
Elastic Net	0.882564	0.9417	276.182
Passive Aggressive	1.022587	0.922	319.999
Support Vector Machine Regression	1.698415	0.7848	531.487
Linear Regression (Kernel PCA)	3.3874	0.1439	1060.023
Linear Regression	116.05917	-1004.0646	36318.519

LIGHTGBM



XGBOOST



FURTHER