Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2023

Supporting Information

Evaluating the roughness of structure-property relationships using pretrained molecular representations

David E. Graff[†,‡], Edward O. Pyzer-Knapp[¶], Kirk E. Jordan[#], Eugene I. Shakhnovich[†], and Connor W. Coley[*,‡,§]

[†]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, United States*
[‡]*Department of Chemical Engineering, MIT, Cambridge, MA 02139, United States*
[¶]*IBM Research Europe, Warrington WA4 4AD, United Kingdom*
[#]*IBM Thomas J. Watson Research Center, Cambridge, MA 02142, United States*
[§]*Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, United States*

E-mail: ccoley@mit.edu

Additional Methods

Differences between ROGI-XD and k-nearest neighbors

There are thematic similarities between the calculation of the ROGI-XD and the RMSE of a k-nearest neighbors (KNN) model. Indeed, the generally strong correlation between these two in most tasks suggests that the two are measuring similar values. Both methods effectively coarse-grain the dataset using some notion of a local neighborhood's average. Whereas a KNN model assumes a fixed neighborhood of size k, ROGI-XD uses a flexible definition of "neighborhood" depending on the "fractional coarse-graining." This allows the ROGI-XD to measure the impact of gradual coarse-graining on a dataset while KNN measures only a single snapshot at a given value of k. This matters especially when the *chosen* value of k is inappropriate for a given dataset (e.g., depending on the existence of outliers in representation space) and will lead to a false impression of a dataset's roughness. Consider two extremes: (1) $k = 0$ will result in average of RMSE of 0 and (2) $k = N$ for which the average RMSE will be the dataset's variance σ^2. Neither of these two quantities are particularly informative of overall dataset roughness. Moreover, while there exists some k that maximizes correlation between the RMSE of a KNN and alternative models, this is an empirical result specific to each dataset (Figure S3) and it is impossible to know this value *a priori*.

Table S1: Dimensionality of all representations tested

representation	dimensionality
descriptor	14
FP	512
VAE	128
GIN	300
ChemBERTa	384
ChemGPT	2048
random	128

Additional Results

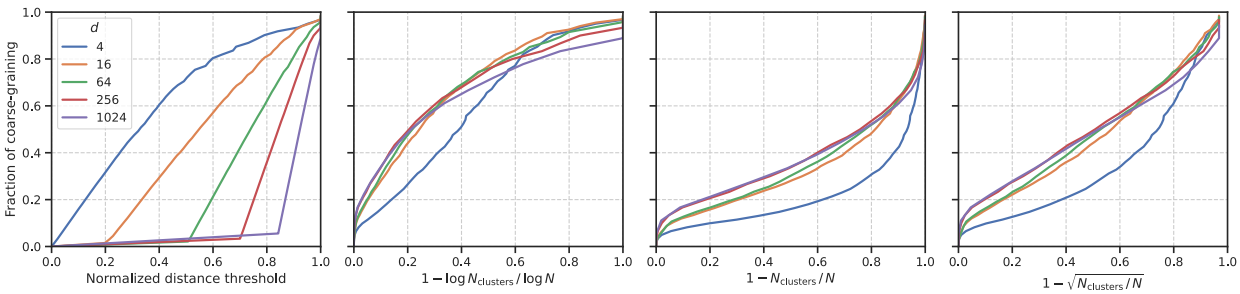

Figure S1: Examples of possible different formulations of the integration variable in the ROGI-XD formulation for 1000 points sampled from the domain $[0, 1]^d$. "Fraction of coarse-graining" is the fractional number of steps in the stepwise dendrogram of the clustering routine after subsampling. In all formulations aside from normalized distance threshold t, the curves of all datasets approximately overlap, indicating a domain of integration that is independent of representation dimensionality. However, only $1 - \log N_{\text{clusters}} / \log N$ is defined over the constant domain $[0, 1]$. Note that the approximate overlap is caused by the dendrogram subsampling, but this step is performed for computational efficiency.

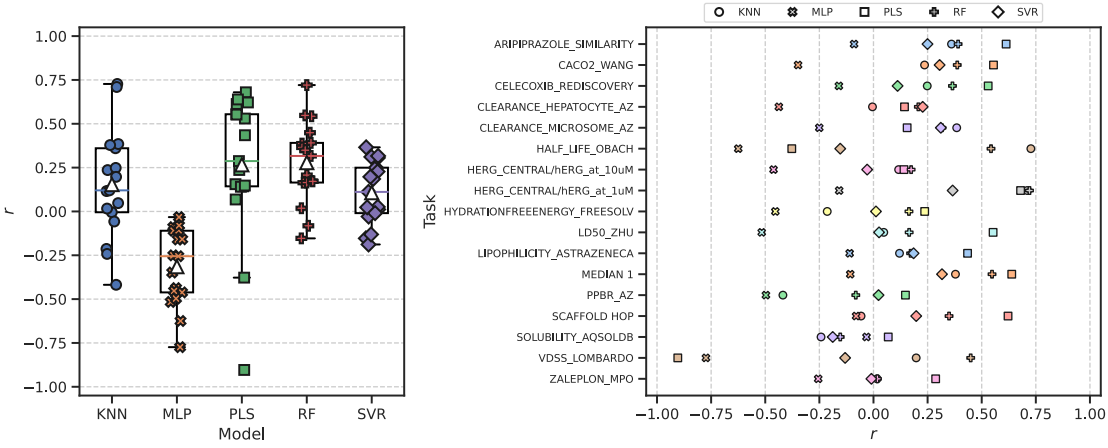

Figure S2: Pearson correlation coefficient r between ROGI and cross-validated RMSE across all representations evaluated for a given pair of ML model and task. *Left*: Box plot of correlations grouped by ML model architecture with individual data points plotted above. The median is depicted via the solid, colored line, and the mean by the white triangle (△). *Right*: Correlations grouped by task. *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression.

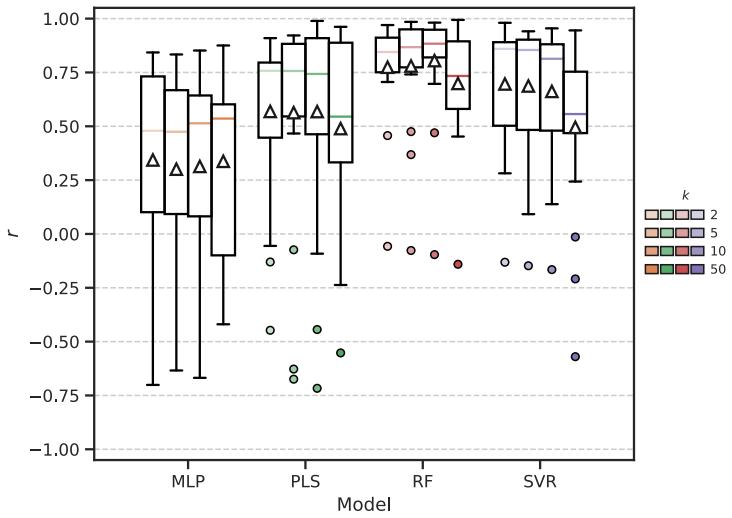

Figure S3: Boxplot of Pearson correlation coefficients r between KNN model RMSE over the entire dataset and cross-validated RMSE across all representations for a given combination of k, ML model architecture, and task. The median is depicted via the solid, colored line, and the mean by the white triangle (△). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression.

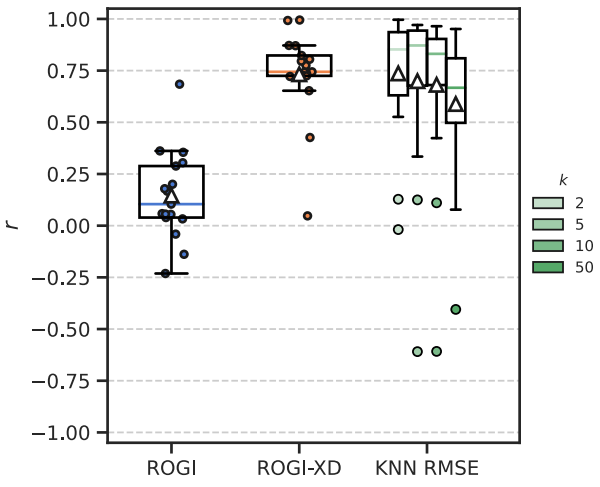

Figure S4: Boxplot of Pearson correlation coefficients r between minimum RMSE of a KNN model over the entire dataset and minimum cross-validated RMSE for a given task across all representations. The median is depicted via the solid, colored line, and the mean by the white triangle (△). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression.

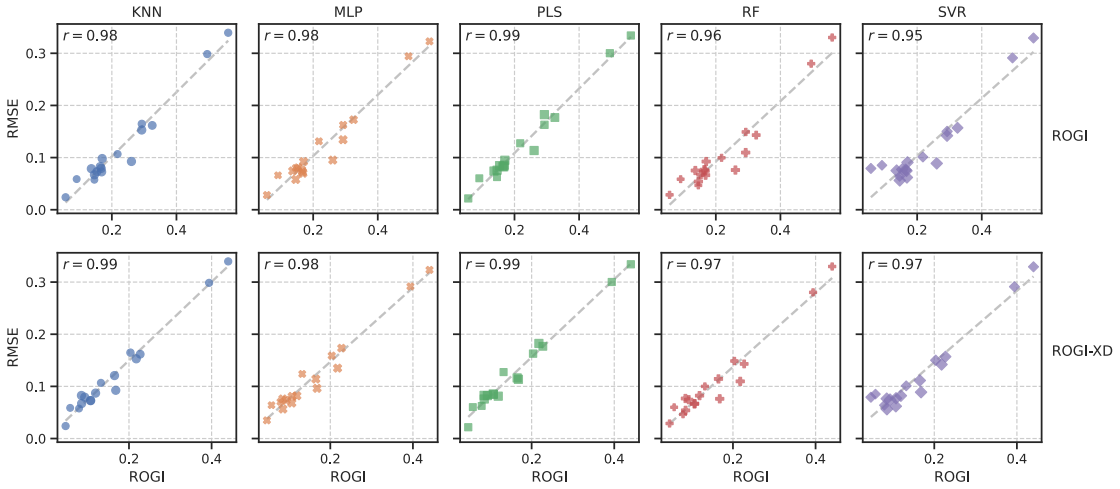

Figure S5: Roughness metric vs RMSE using descriptors across all tasks. Marker size is proportional to the natural logarithm of the dataset size. *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression.

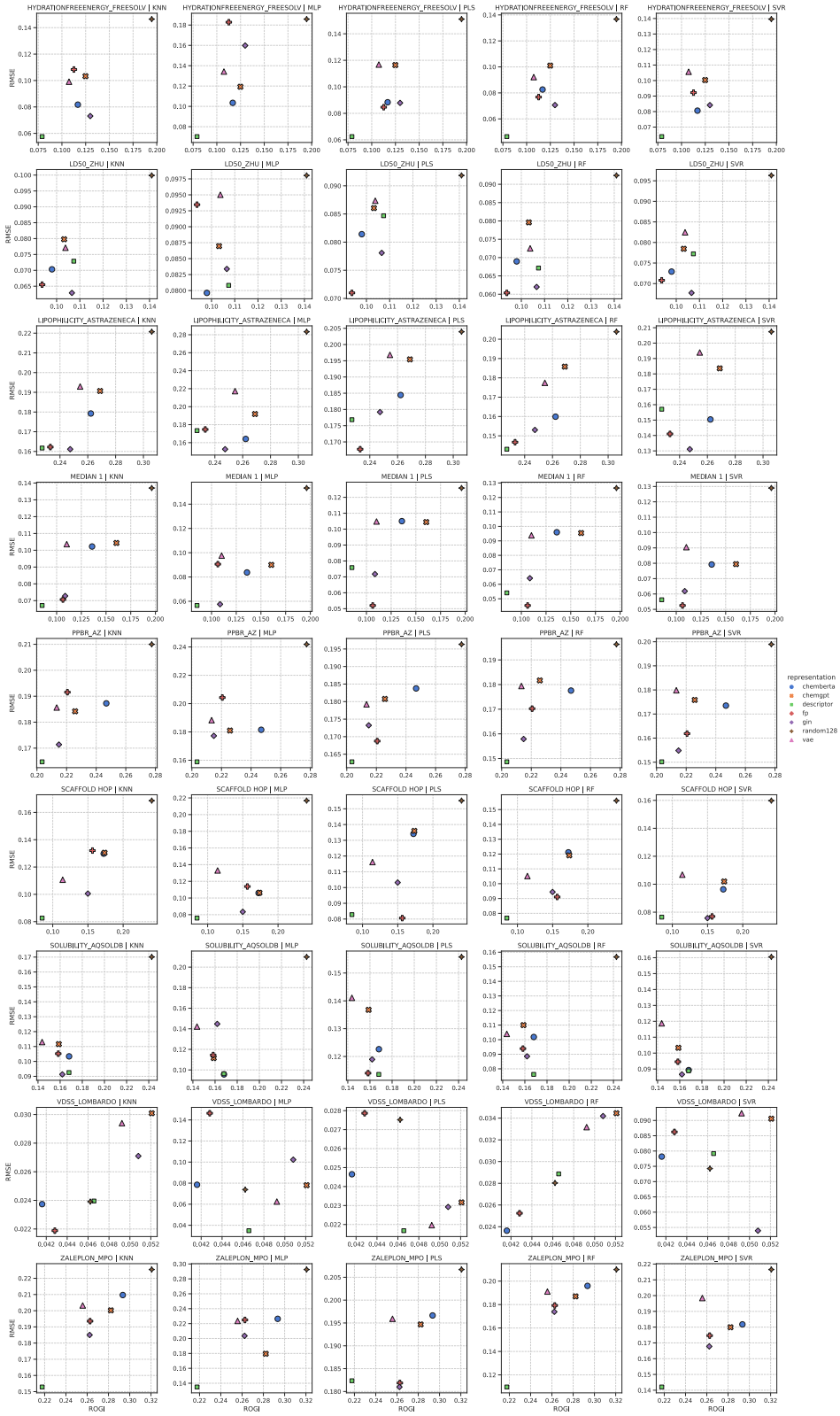

Figure S6: ROGI-XD vs. RMSE for each combination of task and ML model (1/2). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression; *FP*: Morgan fingerprint; *VAE*: variational autoencoder; *GIN*: graph isomorphism network.

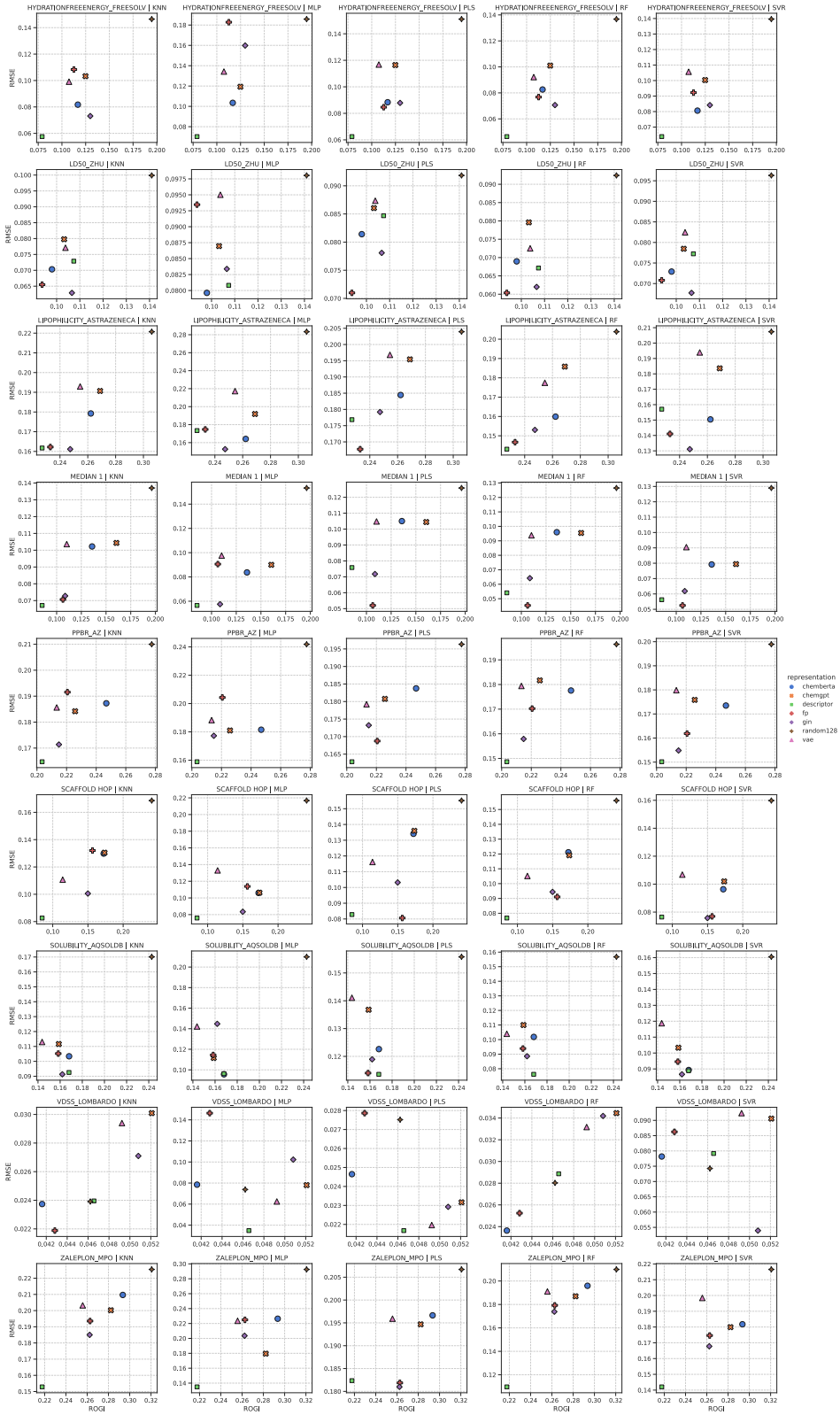

Figure S7: ROGI-XD vs. RMSE for each combination of task and ML model (2/2). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression; *FP*: Morgan fingerprint; *VAE*: variational autoencoder; *GIN*: graph isomorphism network.

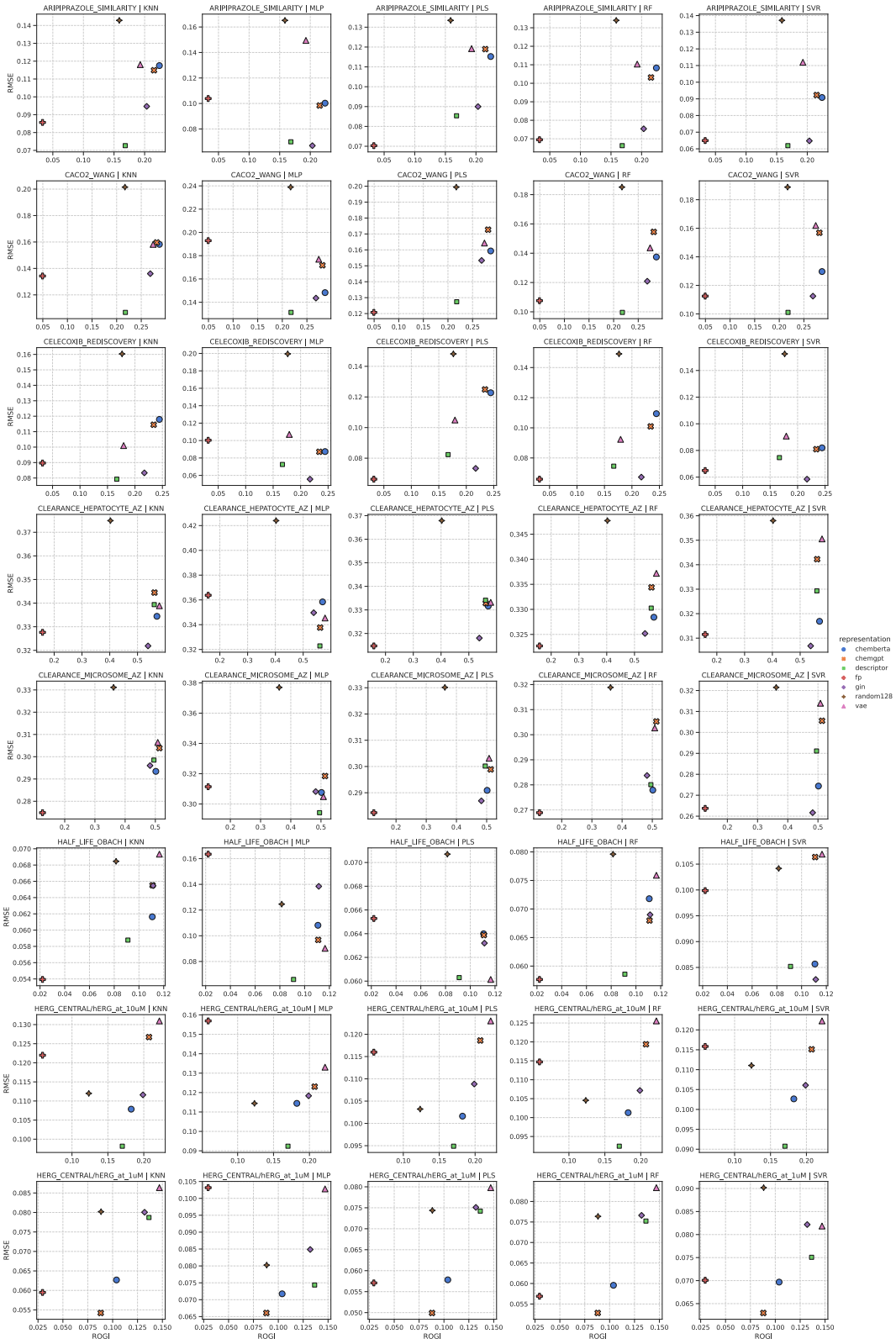

Figure S8: ROGI vs. RMSE for each combination of task and ML model (1/2). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression; *FP*: Morgan fingerprint; *VAE*: variational autoencoder; *GIN*: graph isomorphism network.

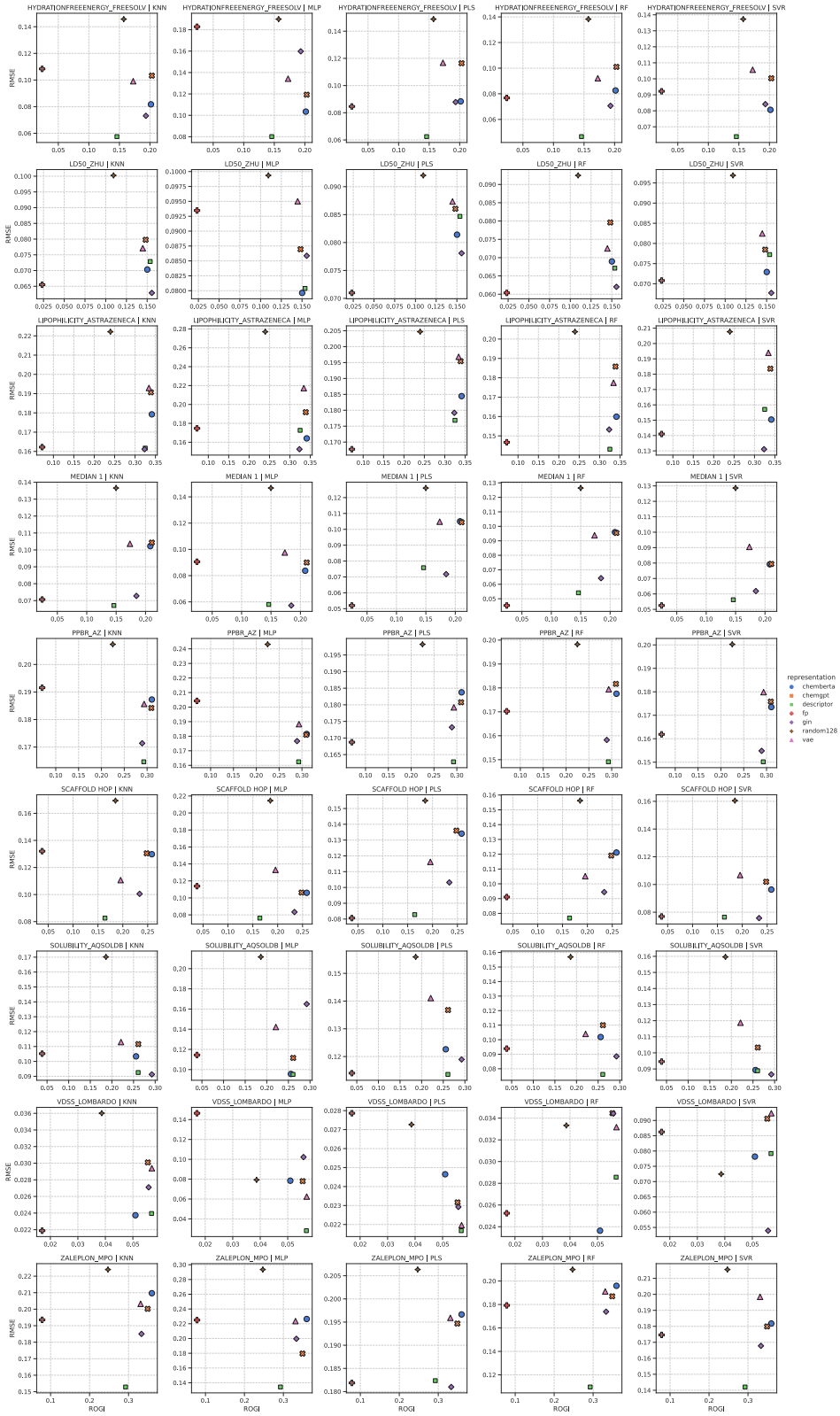

Figure S9: ROGI vs. RMSE for each combination of task and ML model (2/2). *KNN*: k-nearest neighbors; *MLP*: multilayer perceptron; *PLS*: partial least squares; *RF*: random forest; *SVR*: support vector regression; *FP*: Morgan fingerprint; *VAE*: variational autoencoder; *GIN*: graph isomorphism network.

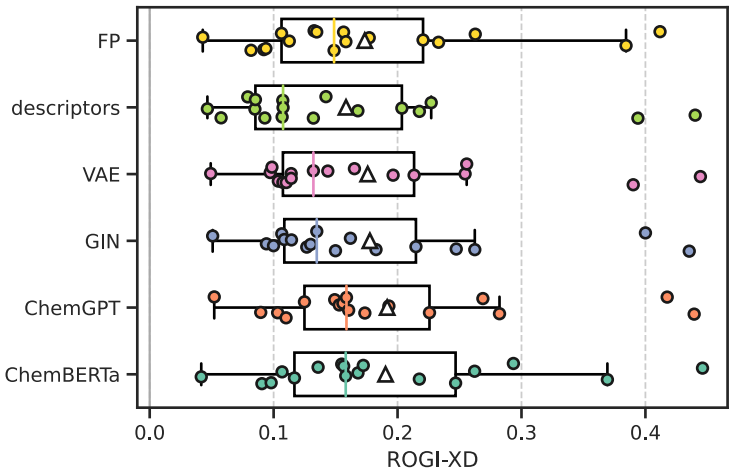

Figure S10: Distribution of ROGI-XD values on all tasks for the respective representation. The median is depicted via the solid, colored line, and the mean by the white triangle (△). *FP*: Morgan fingerprint; *VAE*: variational autoencoder; *GIN*: graph isomorphism network.