Supplementary Materials

A Dataset Analysis

A.1 Dataset Statistics

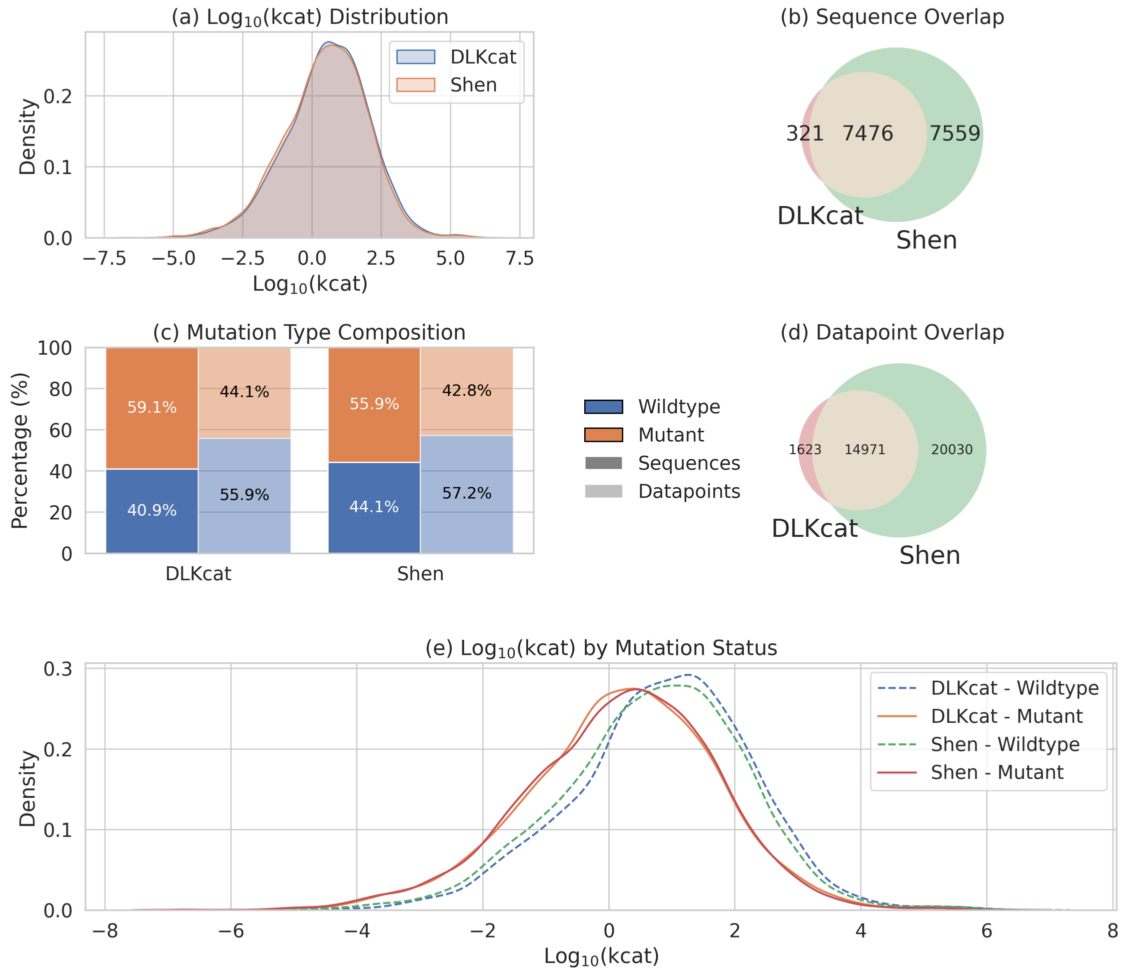

Figure 1: Overall, the Shen dataset is larger than DLKcat while maintaining similar k_{cat} distributions and wild type to mutant ratios, with most DLKcat data contained within Shen. **(a)** Distribution of $\log_{10} k_{cat}$ values in the DLKcat and Shen datasets. **(b)** Overlap of unique protein sequences between DLKcat and Shen. **(c)** Proportion of wild type versus mutant sequences and datapoints in each dataset. **(d)** Overlap in complete datapoints (defined as identical sequence, SMILES, and k_{cat} value) between DLKcat and Shen. **(e)** Distribution of $\log_{10} k_{cat}$ values in wild type and mutant subsets of DLKcat and Shen.

A.2 Sequence Clusters in Each dataset

To characterise sequence redundancy in the DLKcat and Shen datasets, protein sequences in each dataset were clustered at 80% identity and we analysed cluster size distributions. We calculate and plot diversity metrics seen in the figure below. The Gini coefficient is defined as $G = 1 - 2\int_0^1 L(F)\,dF$, where $L(F)$ is the Lorenz curve representing the cumulative fraction of sequences as a function of the cumulative fraction of clusters F. This quantifies inequality in cluster sizes, where $G = 0$ indicates perfect equality (all clusters contain the same number of sequences), while $G = 1$ indicates maximal inequality (all sequences belong to a single cluster). The Simpson effective number represents the number of equally sized clusters required to match the probability that two randomly chosen sequences belong to the same cluster; and the Shannon effective number gives the number of equally sized clusters needed to achieve the observed Shannon entropy of the cluster size distribution.

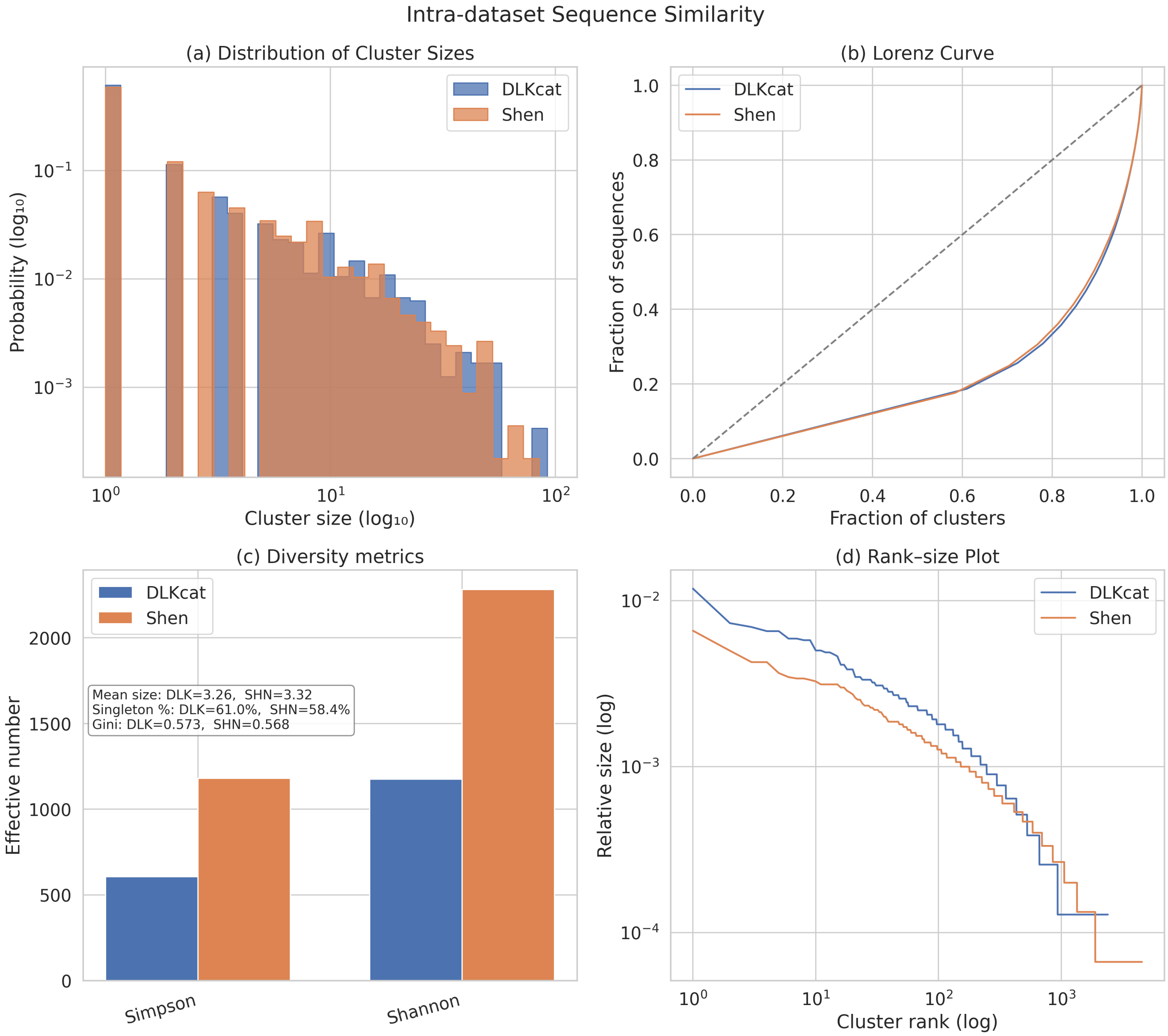

Figure 2: Shen and DLKcat datasets exhibit comparable sequence redundancy patterns, but the Shen dataset spans a broader sequence space, capturing roughly twice the effective diversity of DLKcat while maintaining similar levels of redundancy. **(a)** The distribution of cluster sizes on a log–log scale, where the x-axis represents the number of sequences per cluster and the y-axis indicates the probability of observing a cluster of that size. **(b)** Lorenz curve, which plots the cumulative fraction of clusters (x-axis), sorted from smallest to largest, against the cumulative fraction of sequences they contain (y-axis). A curve closer to the diagonal indicates a more even distribution of sequences across clusters, while curvature away from the diagonal reflects inequality. **(c)** Compares diversity metrics across the two datasets. The bar plots show the Simpson effective number ($\frac{1}{\sum p_i^2}$) and the Shannon effective number ($e^{-\sum p_i \ln(p_i)}$), where p_i is the proportion of sequences in cluster i. The inset text reports the mean cluster size, the percentage of clusters containing only one sequence, and the Gini coefficient of the sequences in each dataset (DLK and SHN). **(d)** Presents the rank–size distribution of clusters on a log scale. The x-axis indicates the rank of each cluster (with rank 1 being the largest), and the y-axis shows the relative size of the cluster, defined as the number of sequences it contains as a fraction of all sequences.

B SMILES Representation Results

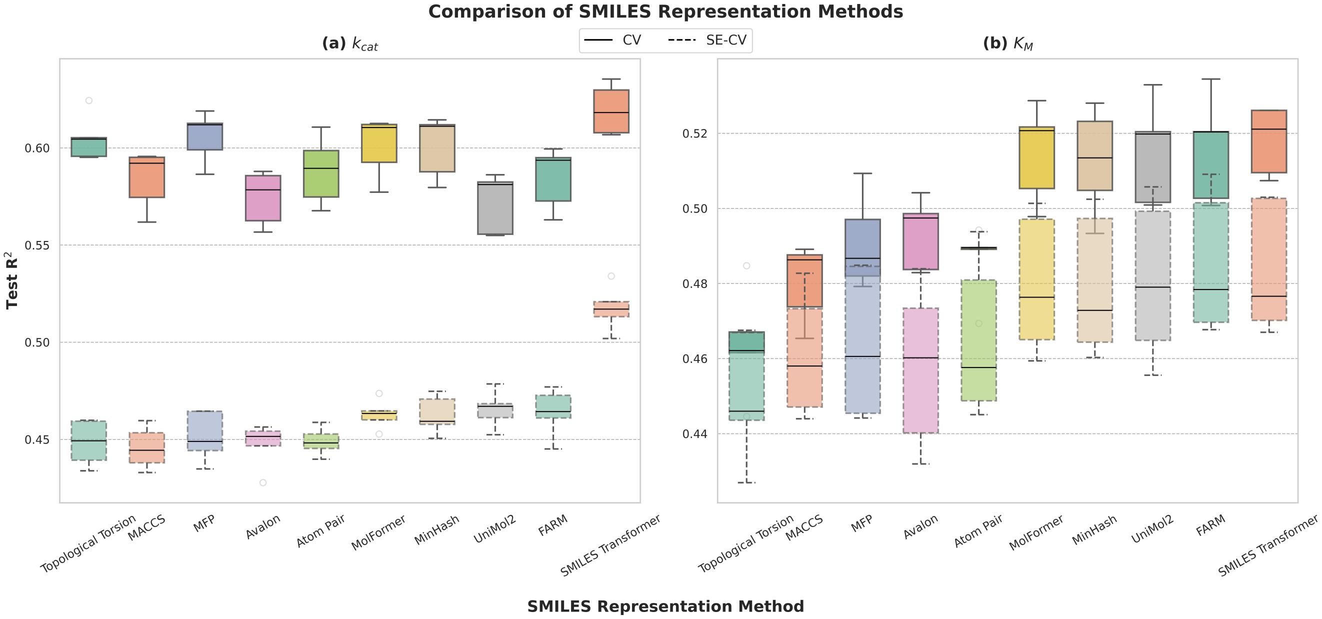

Figure 3: The SMILES Transformer performs best across both tasks and cross-validation methods, with the performance gap between SE-CV and CV notably smaller for $\mathbf{K_M}$ than for $\mathbf{k_{cat}}$. R^2 scores across five folds for different SMILES representations. Extra Trees is trained with PCA-reduced ESMC+ESM-2+T5 (300 components) protein representation. K_M experiments are conducted with ESMC + ESM2 + T5 (binding-weighted pooling concatenated with global pooling) protein representation. Low-opacity, dashed lines represent SE-CV, while high-opacity, solid lines represent standard CV.

C PCA Variance

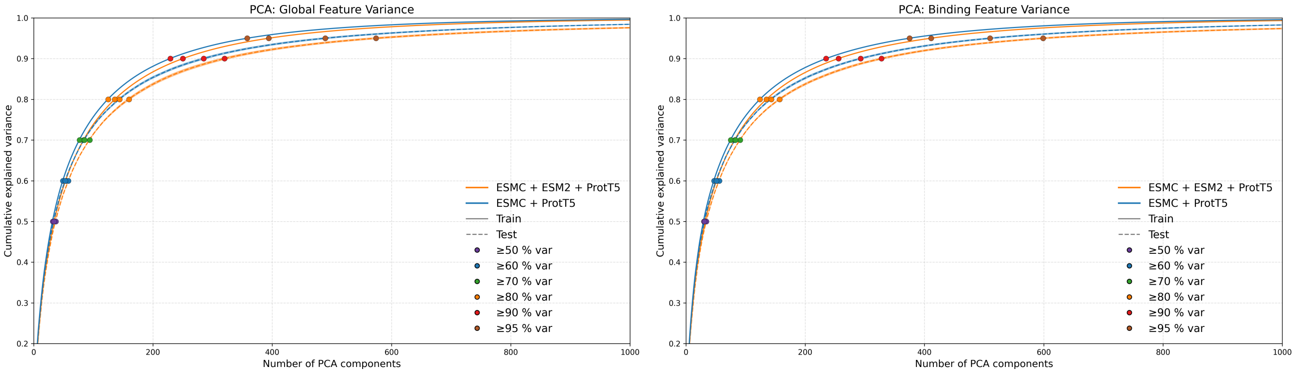

Figure 4: Explained variance captured by PCA applied to protein representations. Each plot shows the cumulative explained variance as a function of the number of PCA components, computed separately for the global vector (left) and the binding-weighted vector (right). PCA is fit on the training set of each SE-CV fold, and mean explained variance across folds is plotted for both the training set (solid lines) and test set (dashed lines). Curves are shown for two protein representation configurations: ESMC+ProtT5 and ESMC+ProtT5+ESM-2 (indicated by colour). Shaded regions show the standard deviation across folds, though they are small and largely obscured. Coloured dots indicate the number of components required to reach 50%, 60%, 70%, 80%, 90%, and 95% explained variance. Line and dot styles are explained in the legend.
The original dimension of the representations is 3456; capturing 95% of variance on the test data requires approximately ~ 600 components for both global and binding vectors and ~ 400 components are needed to capture 95% of the training data variance.

D PCA Shen Dataset Results

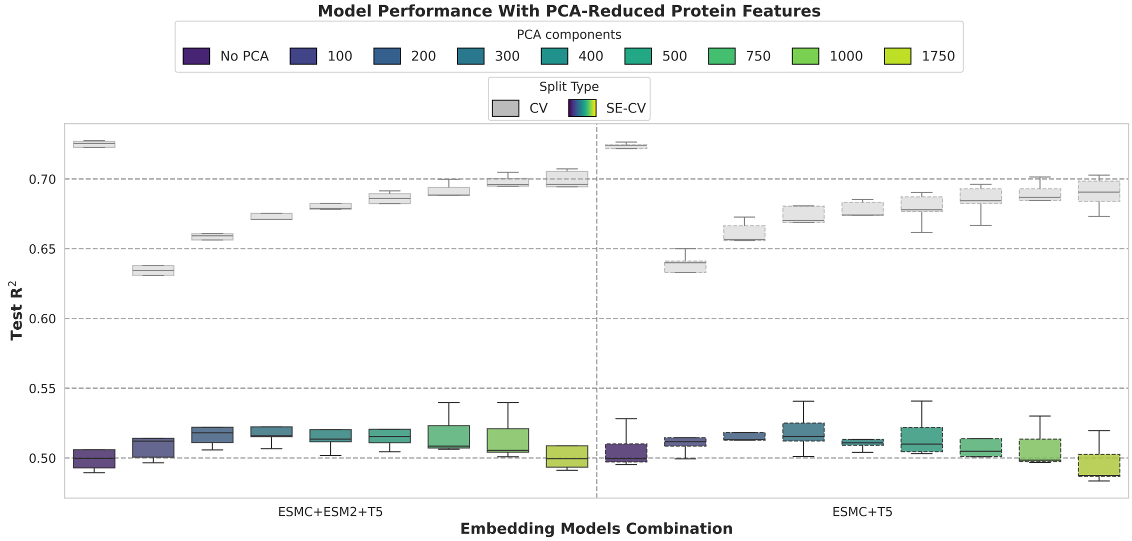

Figure 5: Effect of PCA-based dimensionality reduction on predictive performance on the Shen dataset. This figure mirrors the analysis shown in Figure 8 of the main text but applied to the Shen dataset. R^2 scores across five folds for Extra Trees models trained with and without PCA using two protein embedding combinations: ESMC+T5 and ESMC+ESM-2+T5. PCA was applied with varying numbers of components (100 to 1750). Coloured boxes correspond to SE-CV results; grey boxes correspond to standard CV.

E Protein Representation DLKcat Dataset Results

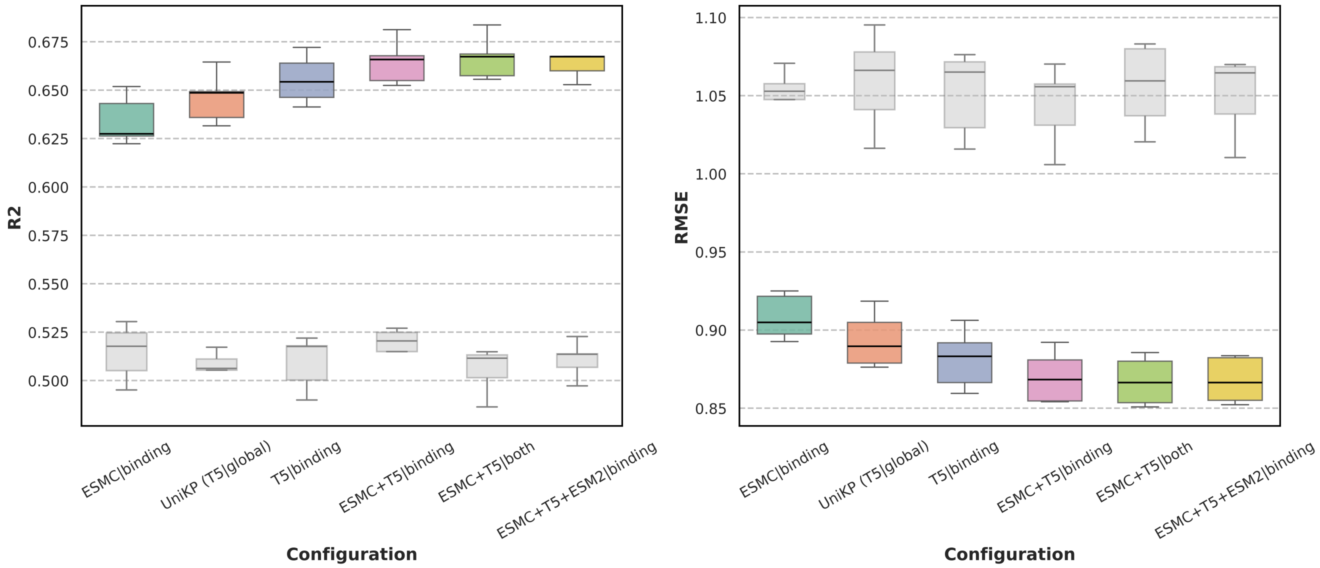

Figure 6: Effect of protein representation strategies on predictive performance across cross-validation settings on the DLKcat dataset. This mirrors the analysis shown in Figure 7 of the main text but applied to the DLKcat dataset. R^2 scores across five folds for Extra Trees models trained with different configuration. Top six configurations are shown. Coloured boxes correspond to standard CV results; grey boxes correspond to SE-CV.