# Supplementary

## JUMP-CP dataset

**JUMP-CP Preprocessing** The cellular phenomic dataset used in this study originates from the JUMP-Cell Painting (JUMP-CP) Consortium and consists of approximately 700,000 morphological profiles derived from ∼116000 small molecules tested in U2OS cells across 12 data-generating centers. Plate-specific raw data are publicly available via the Cell Painting Gallery on the Registry of Open Data on AWS (`https://registry.opendata.aws/cellpainting-gallery/`), specifically under the subset `cpg0016-jump`.

Each plate includes both positive and negative controls. Negative control wells, containing dimethyl sulfoxide (DMSO), are used to detect and correct for technical variation and serve as a reference for identifying morphological perturbations. Positive control wells contain eight compounds selected from the CPJUMP1 pilot study for their distinct and reproducible phenotypic signatures. While most compounds were assayed at a concentration of 10 $\mu$M, those from `source_7` were tested at 0.625 $\mu$M to assess dose-dependent effects. Morphological profiles are extracted from high-content fluorescence microscopy images using CellProfiler, yielding quantitative descriptors of cellular structure and organization.

**Centering on DMSO Controls** DMSO-based centering is carried out *per plate*. Let $X \in \mathbb{R}^{n \times d}$ be the raw feature matrix and $X_{\text{DMSO}} \subset X$ the rows corresponding to the 128 negative-control (DMSO) wells present on every plate.

For each feature $j$, we apply a **RobustScaler**:

$$\hat{x}_{ij} = \frac{x_{ij} - \text{Median}\big(X_{\text{DMSO},j}\big)}{\text{IQR}\big(X_{\text{DMSO},j}\big)} \qquad \text{with IQR} = Q_3 - Q_1,$$

so that every feature is expressed in units of its inter-quartile range (IQR) relative to the plate's DMSO baseline.

**Why Robust, not Standard, scaling?** Imaging conditions such as illumination, focus, and staining vary across plates, introducing systematic technical drift. Centering on the DMSO wells from the *same* plate corrects for this drift and places all perturbations in a consistent, plate-invariant reference frame. Moreover, the histograms in Fig. S1 reveal that many DMSO features exhibit substantial skew and heavy tails, often due to cell-level heterogeneity or segmentation artefacts. While the mean and standard deviation are sensitive to such outliers, the median and interquartile range (IQR) provide a more robust and reliable basis for scaling. For this reason, robust scaling has also been adopted in several prior studies working with Cell Painting features [1].

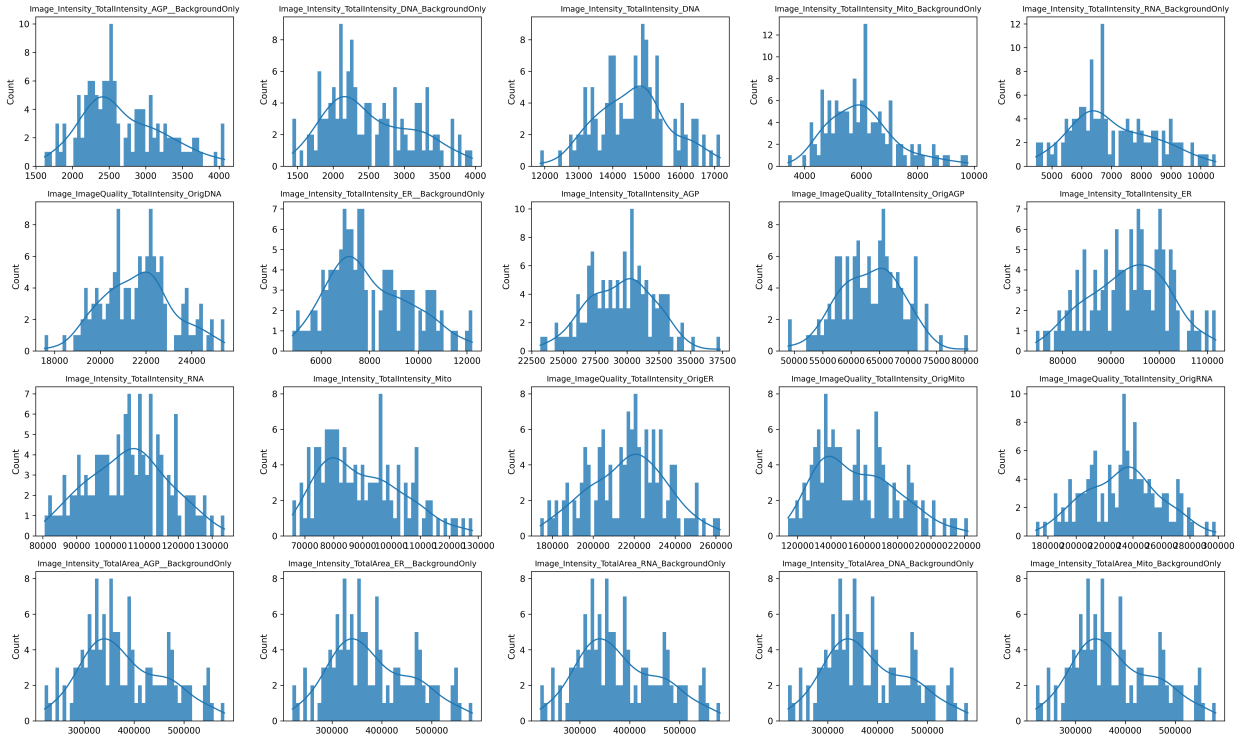



Figure S1: Distributions of the 20 most variable features in DMSO wells. The pronounced skew and long right tails motivate the use of a RobustScaler (median/IQR) rather than a mean/SD StandardScaler.

**Filtering Features** After centering, data from all plates are concatenated, and feature filtering is applied globally by removing features with low and those containing missing values. We first drop feature columns based on the following criteria:

- Features with standard deviation $\sigma_j$ below 0.1:

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

After preprocessing, each morphological profile contains 4709 features that describe diverse aspects of cellular morphology, including shape, texture, and intensity across cellular compartments. This rich and standardized dataset enables the development of transferable molecular representations by capturing biologically relevant variation induced by compound treatments.

**Effectiveness of Preprocessing Pipeline** To validate our preprocessing approach, we employ Uniform Manifold Approximation and Projection (UMAP) to visualize the morphological feature space before and after normalization. We perform two complementary analyses: (1) plate-based visualization of 20 randomly selected plates to assess plate-to-plate variability, and (2) compound-based visualization of 20 randomly selected compounds with all their replicates to evaluate replicate consistency.

The results demonstrate that our DMSO-based robust scaling effectively addresses technical confounders while preserving biological signal. In the original data, samples cluster primarily by plate identity rather than biological similarity, indicating strong batch effects. After normalization, plate-driven clustering is substantially reduced while compound replicates show improved consistency, validating that our preprocessing pipeline successfully enhances the signal-to-noise ratio by reducing technical variation while maintaining compound-specific morphological signatures.

**Handling Replicates during Training** Each compound has between 5 and 60 replicates, originating from the same plate, lab, or different labs. [ToDo: Include replicate distribution plot]

The dataloader pairs each SMILES with one randomly sampled morphological measurement per training instance. Across different epochs, the dataloader returns different replicates for the same SMILES, allowing the model to learn from the full range of biological variability.

## LINCS dataset

**LINCS Data Preprocessing** The LINCS L1000 dataset preprocessing involved several sequential steps to ensure data quality and experimental consistency. The preprocessing pipeline is illustrated through comprehensive visualizations that guided filtering decisions at each stage.

**Cell Line Selection** Cell lines were selected based on compound coverage to ensure sufficient data for robust analysis. We applied a minimum threshold of 1000 unique compounds per cell line, which balanced data availability with statistical power (Figure S3).

**Dose Standardization and Binning** Compounds were tested across multiple cell lines at various doses represented as absolute concentrations. The dose distribution analysis revealed substantial heterogeneity in testing concentrations across the dataset (Figure S4a). To convert these continuous dose values into categorical variables suitable for analysis, we implemented logarithmic binning spanning five orders of magnitude ($10^{-5}$ to $10^{5}$ $\mu$M). The dose level distribution analysis (Figure S4b) informed our selection of the 0.001-100 $\mu$M range, which captures the majority of experimental measurements while excluding extreme outliers.

**Time Point Standardization** Similarly, compounds were tested across different cell lines at various time points, creating temporal heterogeneity in the experimental design (Figure S5a). Analysis of time point prevalence (Figure S5b) revealed that 6-hour and 24-hour measurements represented the most frequently used experimental conditions. We selected these two time points to maximize data availability while maintaining temporal diversity in drug response profiles.

**Handling Genomic Conditions during Training** Each compound has genomic measurements across multiple cell lines, with each cell line tested at different doses and time points. For example, a compound might have data for cell line A at $2\mu$M/6h, $5\mu$M/6h, $2\mu$M/24h, and $5\mu$M/24h.

The dataloader randomly selects one dose-time combination per cell line per training instance. If a compound has data for 3 cell lines, the dataloader will return 3 genomic conditions (one per cell line) with randomly chosen dose-time combinations. Across different epochs, different dose-time combinations are selected, ensuring the model sees the full range of experimental conditions while keeping batch sizes manageable.

## Auto-Encoders Pretraining

**Hyperparameter Optimization** To identify optimal model configurations, we conducted comprehensive hyperparameter searches using grid search across multiple architectural and training parameters. The search was performed to systematically evaluate the impact of different model components on reconstruction performance.

**Search Space and Architecture Specifications** Table S1 summarizes the complete hyperparameter search space and architectural specifications explored in our optimization.

The hyperparameter search encompassed three architectural variants with increasing

Table S1: Hyperparameter search space and architectural specifications for autoencoder optimization

| Hyperparameter | Values/Range |
|---|---|
| **Model Configuration** | |
| Model Type | Autoencoder (AE) |
| Latent Dimension | 128 (fixed) |
| Normalization | BatchNorm / LayerNorm / None |
| Dropout Rate | 0.0 (fixed) |
| **Architecture Variants** | |
| Vanilla | Input $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512 $\rightarrow$ Output |
| Medium | Input $\rightarrow$ 1024 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512 $\rightarrow$ 1024 $\rightarrow$ Output |
| Large | Input $\rightarrow$ 2048 $\rightarrow$ 1024 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512 $\rightarrow$ 1024 $\rightarrow$ 2048 $\rightarrow$ Output |
| **Training Parameters** | |
| Learning Rate | $1\times10^{-5}$ to $1\times10^{-2}$ (6 values) |
| Weight Decay | 0.0 (fixed) |
| Batch Size | 64 (fixed) |

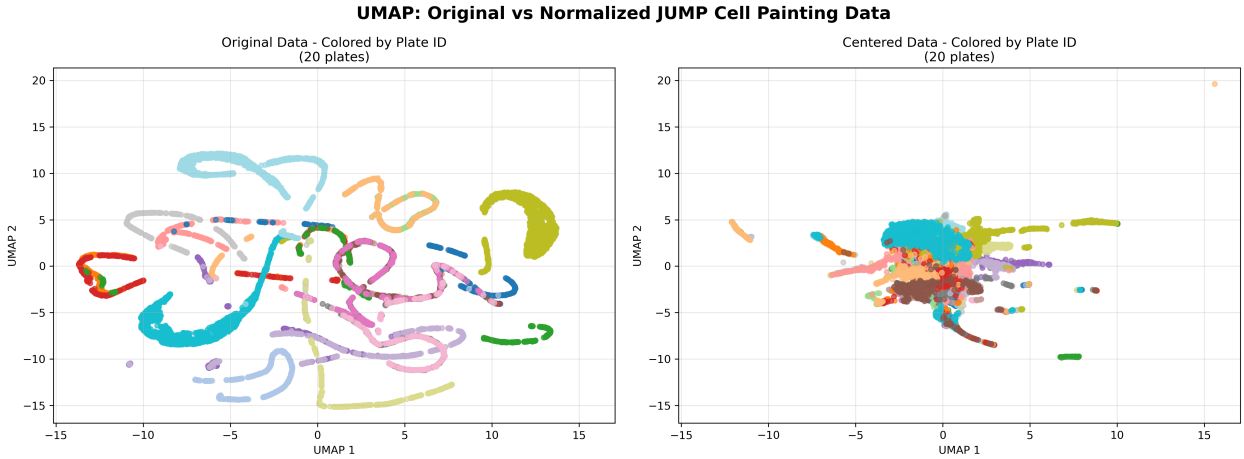
complexity: vanilla (2 hidden layers), medium (3 hidden layers), and large (4 hidden layers) configurations. Each architecture was tested with different normalization strategies to assess their impact on convergence and representation quality.

Learning rates spanned three orders of magnitude to capture both conservative and aggressive optimization regimes. The latent dimension was fixed at 128 to balance representational capacity with computational efficiency.

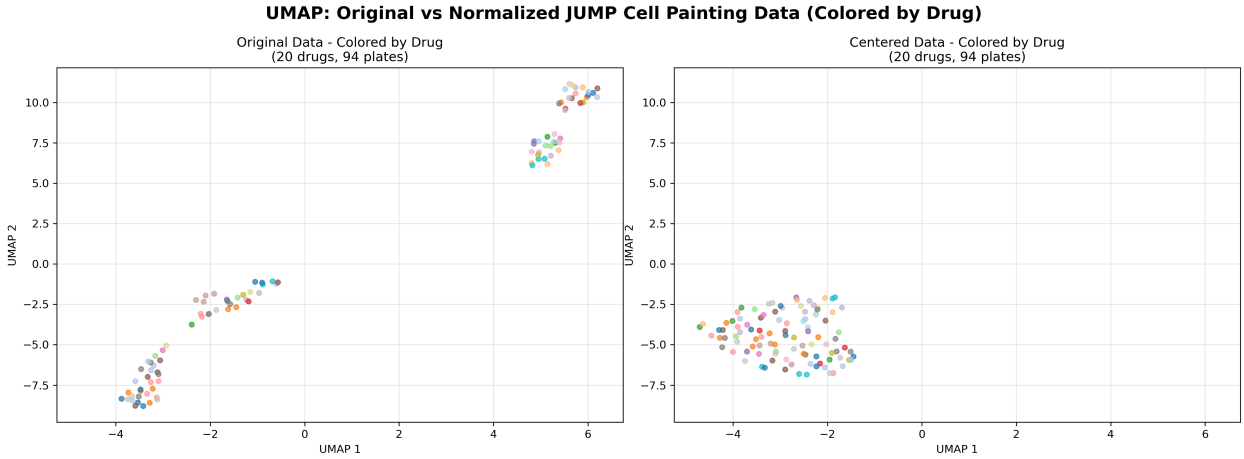
**Evaluation and Selection** Model performance was evaluated using reconstruction loss (mean squared error) on held-out validation sets. The optimal hyperparameter combination was selected based on the lowest validation loss as shown in Figure S6.

To DO! include JUMP-CP AE HP comparison

## Downstream Evaluation




(a) Plate-based UMAP analysis




(b) Compound-based UMAP analysis

Figure S2: UMAP visualizations comparing original and normalized morphological profiles. **(a)** Strong plate effects in original data (left) are substantially reduced after DMSO centering (right). **(b)** Biological replicates of the same compound cluster more tightly after normalization, demonstrating improved signal-to-noise ratio while preserving biological diversity.

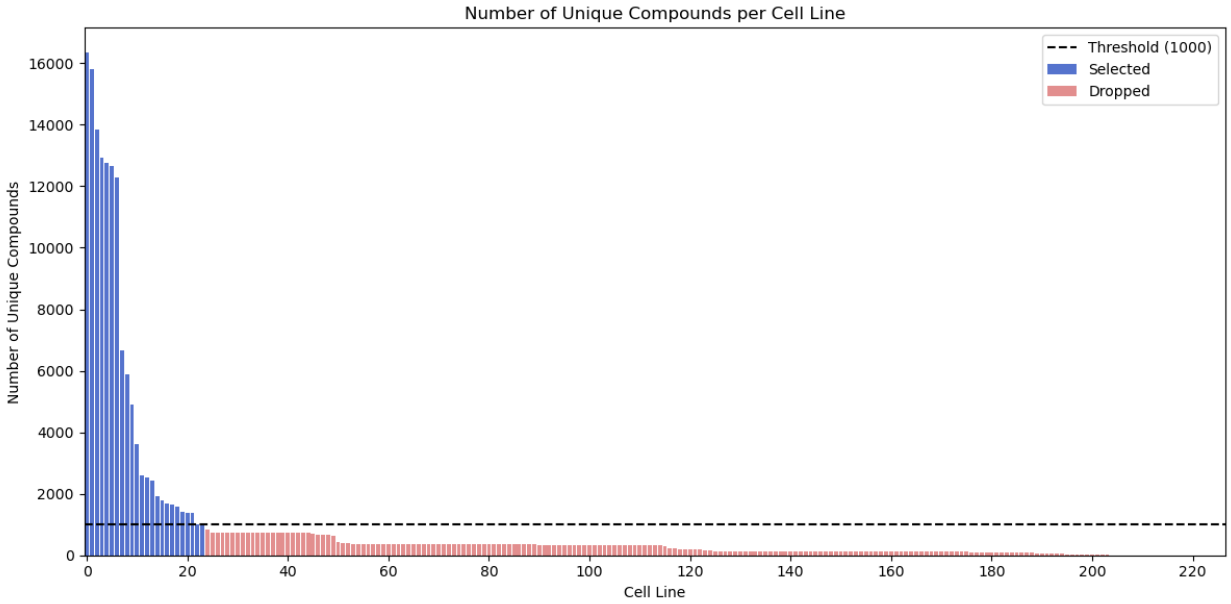
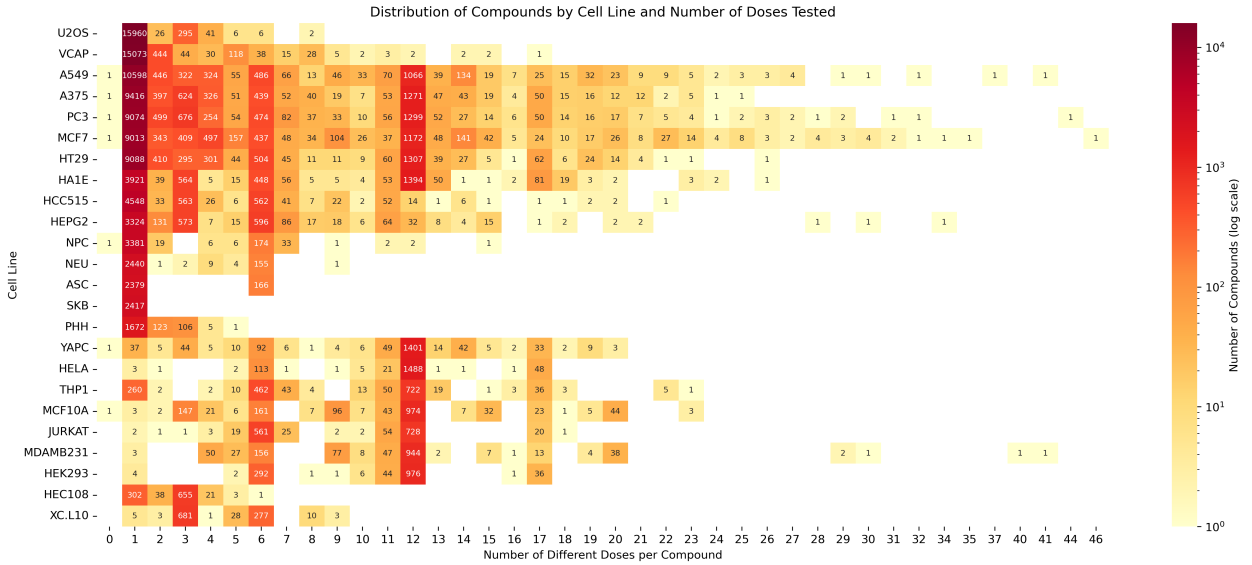
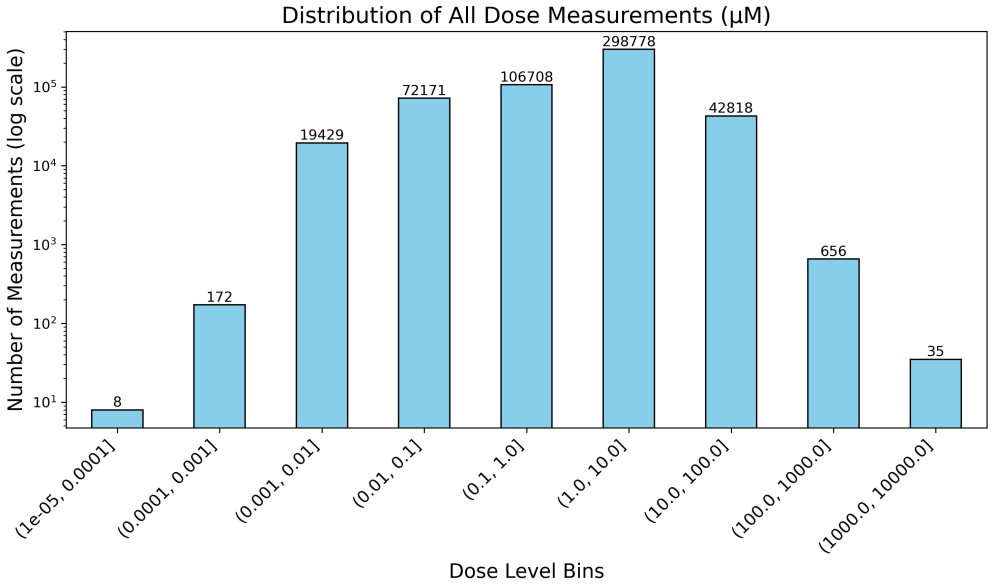



Figure S3: Cell line selection based on compound coverage. Selected cell lines (blue) meet the minimum threshold of 1000 unique compounds, while excluded cell lines (red) fall below this criterion. The dashed line indicates the selection threshold.

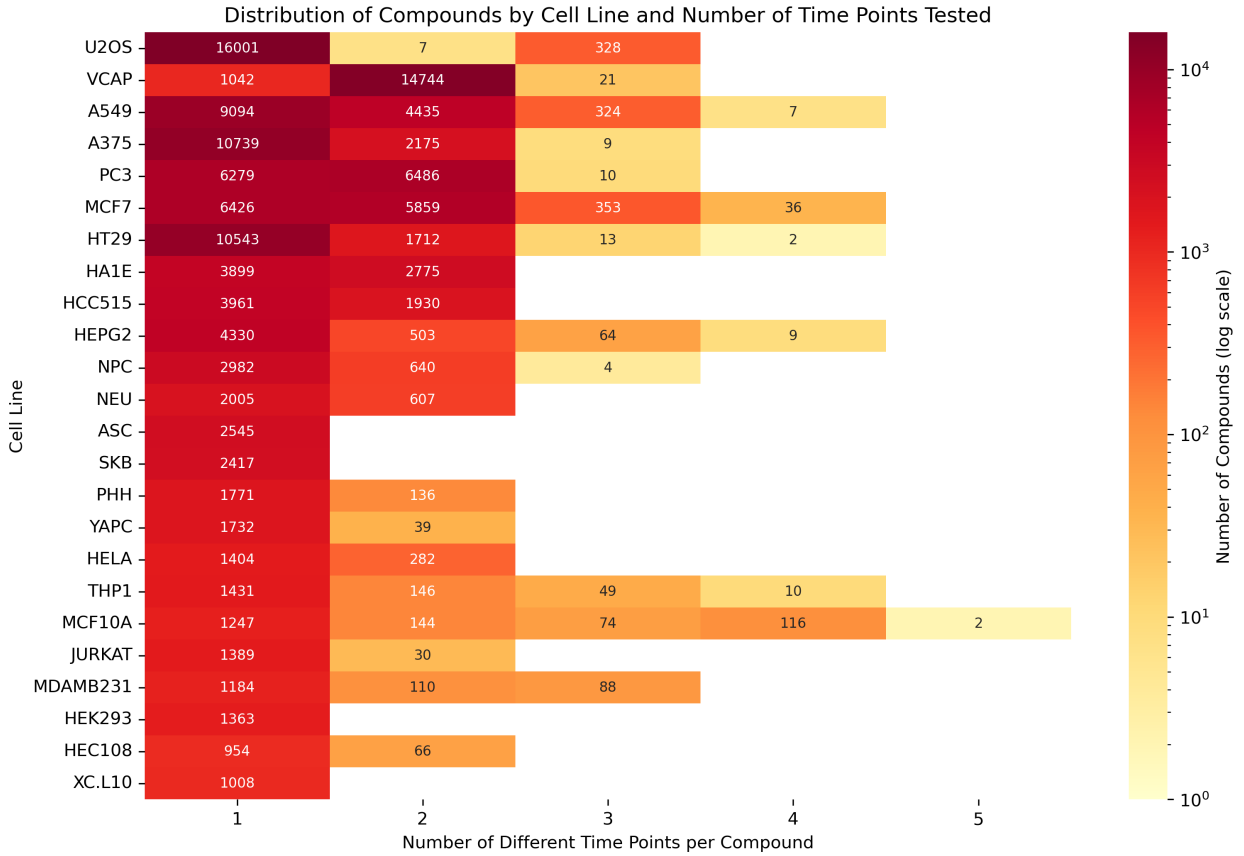



(a) Dose-response coverage across cell lines

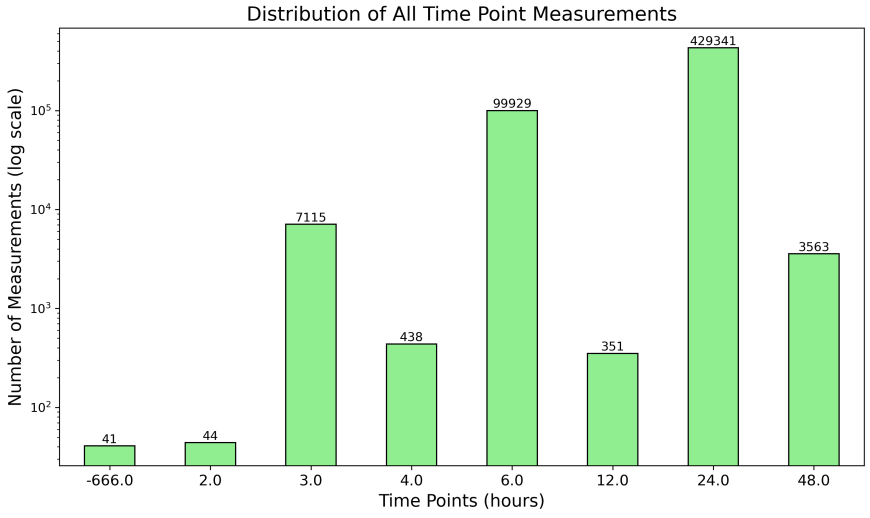



(b) Dose level distribution after logarithmic binning

Figure S4: Dose standardization and binning analysis. **(a)** Log-scale heatmap showing the number of compounds tested with varying numbers of dose levels across different cell lines, revealing the heterogeneity in dose-response experimental design. **(b)** Distribution of dose measurements across logarithmic bins ($10^{-5}$ to $10^{4}$ $\mu$M), demonstrating the concentration ranges selected for analysis (0.001-100 $\mu$M covers the majority of measurements).




(a) Temporal coverage across cell lines




(b) Time point distribution

Figure S5: Time point standardization analysis. **(a)** Log-scale heatmap displaying the number of compounds tested with varying numbers of time points across different cell lines, showing experimental design variability. **(b)** Distribution of measurements across different time points, demonstrating the prevalence of 6-hour and 24-hour measurements that were selected for the final dataset.

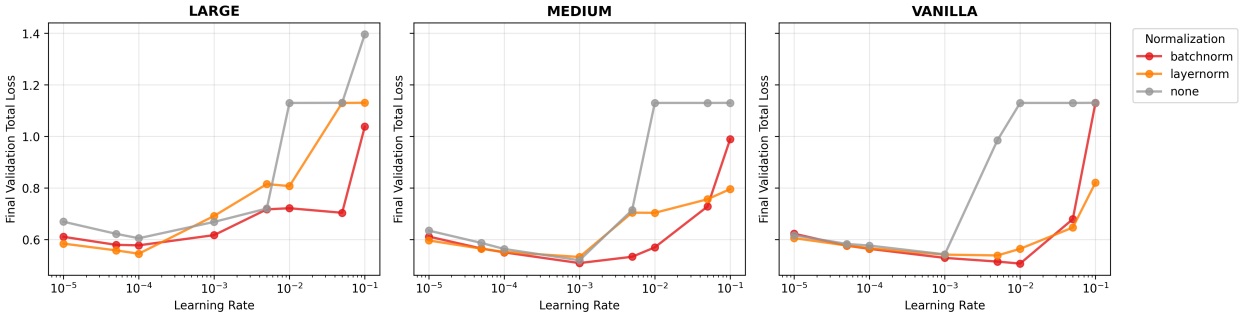



Figure S6: Hyperparameter tuning of the LINCS autoencoder: The vanilla architecture with batch normalization, trained at a learning rate of $1 \times 10^{-5}$, achieved the lowest validation reconstruction loss.

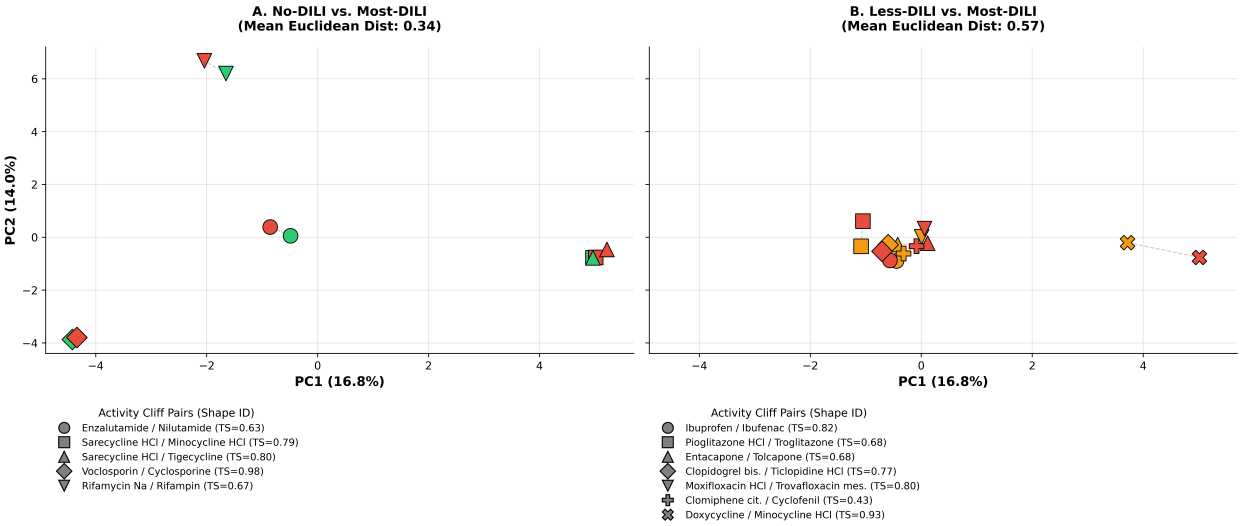



Figure S7: Principal component analysis of ECFP fingerprint representations for activity cliff compound pairs. Each panel shows the two-dimensional PCA projection of 1,024-bit ECFP4 fingerprints for held-out activity cliff compounds. Panel A shows No-DILI vs. Most-DILI pairs (ordinal distance $d = 2$; mean Euclidean separation: 0.34) and Panel B shows Less-DILI vs. Most-DILI pairs (ordinal distance $d = 1$; mean Euclidean separation: 0.57). Point color indicates DILI severity class (green: vNo-DILI-concern; orange: vLess-DILI-concern; red: vMost-DILI-concern) and point shape identifies each structural analog pair. Dashed lines connect members of the same pair. Tanimoto similarity (TS) between each pair is reported in the legend. The compact clustering and low inter-pair Euclidean distances reflect the structural similarity that defines activity cliffs and the limited capacity of topology-based fingerprints to encode divergent biological outcomes.

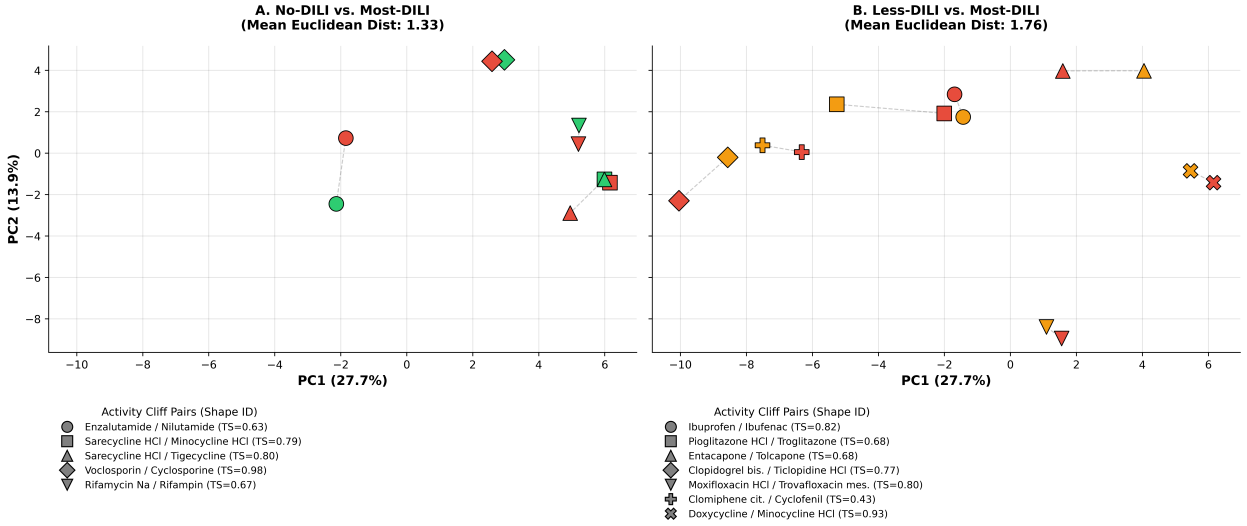



Figure S8: Principal component analysis of Bio$\mathcal{X}$Mol-Soft representations for activity cliff compound pairs. Panels are organized as in Figure S7. Panel A shows No-DILI vs. Most-DILI pairs ($d = 2$; mean Euclidean separation: 1.33) and Panel B shows Less-DILI vs. Most-DILI pairs ($d = 1$; mean Euclidean separation: 1.76). Compared to ECFP representations (Figure S7), Bio$\mathcal{X}$Mol-Soft embeddings exhibit substantially greater inter-pair separation despite equivalent structural similarity between pair members, indicating that soft phenotypic alignment during pretraining introduces toxicity-relevant variation not recoverable from molecular topology alone.

**Input:** Molecular dataset $\mathcal{D}$, hyperparameter search space $\mathcal{H}$, number of folds $K = 5$, number of repetitions $R = 5$

**Output:** Performance metrics $\{p_{r,k}^{\text{train}}, p_{r,k}^{\text{val}}, p_{r,k}^{\text{test}}\}$ for all $r, k$

**for** $r \leftarrow 1$ **to** $R$ **do**

  Generate scaffold-based folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ from $\mathcal{D}$;

  **for** $k \leftarrow 1$ **to** $K$ **do**

    $\mathcal{D}_{\text{test}} \leftarrow \mathcal{D}_k, \quad \mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \setminus \mathcal{D}_k$;

    Split $\mathcal{D}_{\text{train}}$ into $\mathcal{D}_{\text{train}}^{\text{inner}}$ and $\mathcal{D}_{\text{val}}$ for hyperparameter tuning;

    **for** *each* $h \in \mathcal{H}$ **do**

      Train model $M(h)$ on $\mathcal{D}_{\text{train}}^{\text{inner}}$ and evaluate on $\mathcal{D}_{\text{val}}$ to obtain score $s(h)$;

    **end**

    Select best hyperparameters $h^* = \arg\max_{h \in \mathcal{H}} s(h)$;

    Retrain $M(h^*)$ on full $\mathcal{D}_{\text{train}}$;

    Evaluate $M(h^*)$ on $\mathcal{D}_{\text{test}}$;

    Store results as $p_{r,k}^{\text{train}}$, $p_{r,k}^{\text{val}}$, and $p_{r,k}^{\text{test}}$;

  **end**

**end**

**return** $\{p_{r,k}^{train}, p_{r,k}^{val}, p_{r,k}^{test}\}$ *for all* $r, k$;

**Algorithm 1:** 5×5 Cross-Validation with Hyperparameter Optimization using Scaffold Splitting

# References

(1) Way, G. P.; Kost-Alimova, M.; Shibue, T.; Harrington, W. F.; Gill, S.; Piccioni, F.; Becker, T.; Shafqat-Abbasi, H.; Hahn, W. C.; Carpenter, A. E.; Vazquez, F.; Singh, S. Predicting cell health phenotypes using image-based morphology profiling. *Molecular Biology of the Cell* **2021**, *32*, 995–1005.