Supplementary Information

Scalable Molecular Representations Enabled by Multimodal Fusion and Sequence Distillation

Suvendu Kumar[1,†], Saveena Solanki[1,†,*], Mudit Gupta[1,2,†], Sonam Chauhan[1,†], Sanjay Kumar Mohanty[1,†], Shiva Satija[1], Abhinav Kumar Sharma[1], Subhadeep Duari[1], Arushi Sharma[1], Raidhani Shome[1], Vishakha Gautam[1], Sakshi Arora[1], Syed Yasser Ali[1], Adnan Raza[1], Sourav Sinha[1], Aayushi Mittal[1], Debarka Sengupta[1,2], Natarajan Arul Murugan[1], and Gaurav Ahuja[1,2,*]

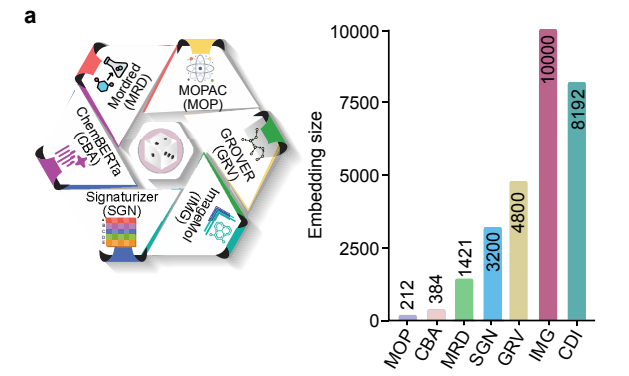
[1]Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi)

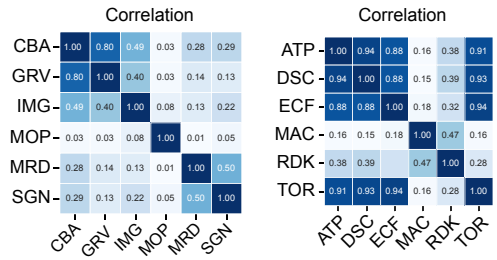
[2]Infosys Center for AI, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India

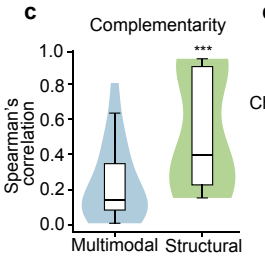
† Co-first authors * Correspondence: gaurav.ahuja@iiitd.ac.in; saveenas@iiitd.ac.in

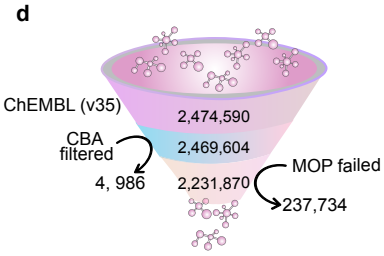
Data repository

The archived CDI data repository is available through Zenodo using the persistent DOI: https://doi.org/10.5281/zenodo.17582661.









e

Chemical Dice Integrator

1) Encoder Function generating latent represntations

$$\mathbf{h} = g(\boldsymbol{x})$$

2) Decoder reconstruct space from latent space

$$\tilde{\boldsymbol{x}} = f(\mathbf{h}) = f(g(\boldsymbol{x}))$$

3) Optimization objective minimizing reconstruction difference

$$\arg\min_{f,g} \quad \{(\Delta(\boldsymbol{x}), f(g(\boldsymbol{x}))\}$$

4) Dataset definition - each input has a embedding

$$\mathrm{D} = \{\boldsymbol{x}^i \mid i = 1, \ldots, M\}, \quad \boldsymbol{x}^i = \{\boldsymbol{d}^i_j \mid j = 1, \ldots, 6\}$$

where M is total number of molecules

5) Mean Squared Error(MSE) loss function

$$\mathrm{L}_{MSE}(\mathrm{x}, \mathrm{y}) = 1/\mathrm{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{y}_i|^2$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{R}^n$

6) Reconstruction Error (RE) to measure input-output differences

$$RE(\mathrm{x}, \tilde{\mathrm{x}}) = 1/\mathrm{n}\sum_{i=1}^{n} |\boldsymbol{x} - \tilde{\boldsymbol{x}}_i|^2$$

where $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathrm{R}^n$

7) ReLU activation function

$$ReLU(\boldsymbol{x}) = \max(\mathbf{0}, \boldsymbol{x})$$

8) Input and output dimension for each Semantic Commonality Autoencoder

$$\dim = \left[\sum_{k=1}^{6} |\mathbf{e}^i_k|\right] - |\mathbf{e}^i_j|$$

where|x| dimension of x

9) Concatenate all embeddings except jth one input to jth SCA block

$$\boldsymbol{a}^i_j = CONCAT(\boldsymbol{d}^i_{\cdot j})$$

10) Latent embedding representation of jth for SCA encoder

$$\boldsymbol{p}^i_j = g(\boldsymbol{a}^i_j)$$

11) Combined SCA loss in MSE and RE terms

$$L_{SCAj} = L_{RE}(\boldsymbol{a}^i_j, \tilde{\boldsymbol{a}}^i_j) + L_{MSE}(\boldsymbol{p}^i_j, \boldsymbol{d}^i_j)$$

12) Super Embedding Autoencoder loss

$$L_{SEA} = L_{RE}(\boldsymbol{r}^i, \tilde{\boldsymbol{r}}^i)$$

13) Total Loss

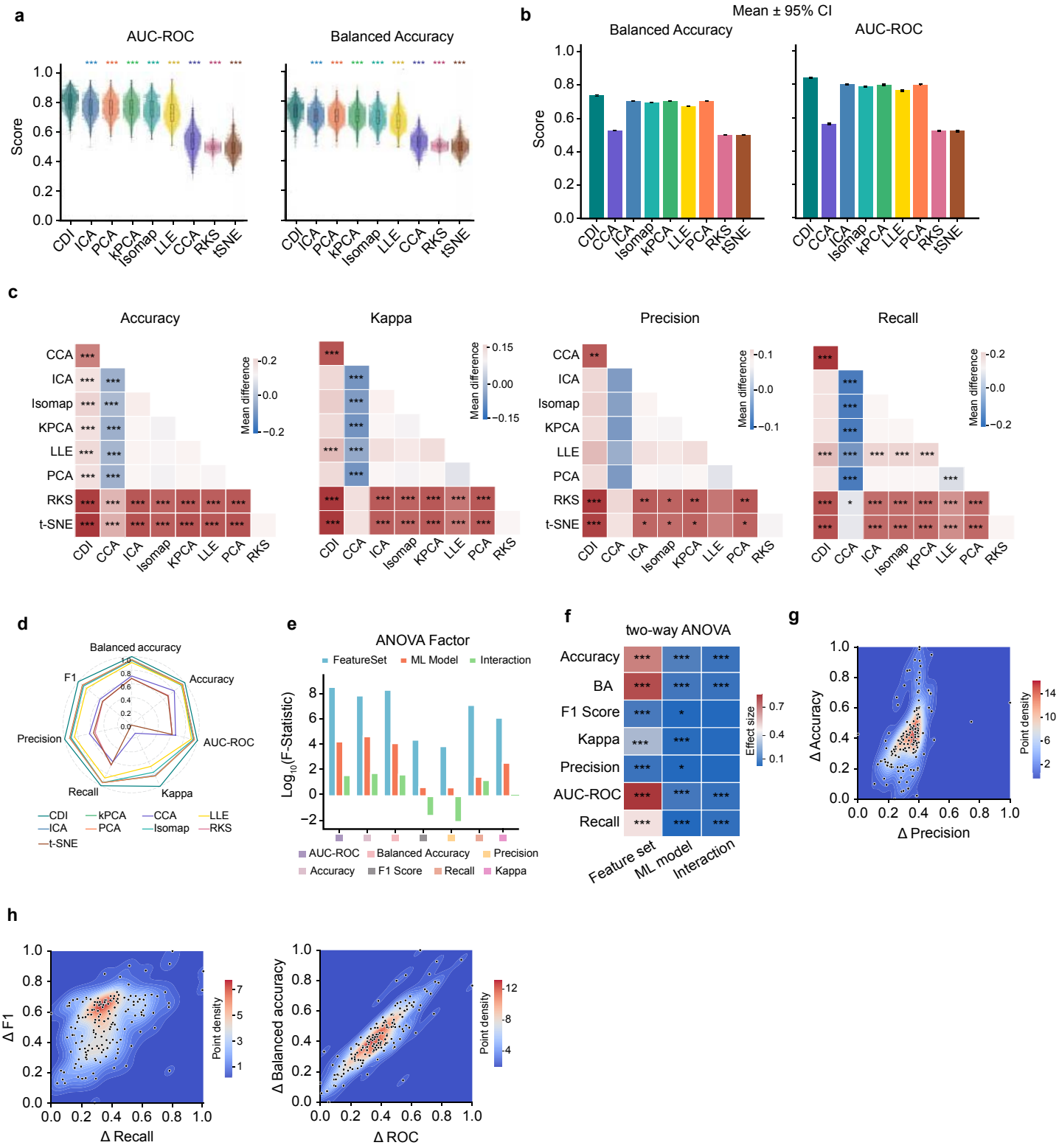
$$L^i_{\text{total}} = \frac{1}{6}\sum_{j=1}^{6} L^i_{\text{SCA}_j} + L^i_{\text{SEA}}$$

f

Model	Basic	Generalized
Total Parameters	1.09B	0.10B
Learning Rate	0.05	0.00001
Training Size	100%	90%
Validation Size	0%	10%
Epochs	600	30
Total Molecules	2,231,870	2,231,870
Loss Function	MSE	MSE
Optimizer	SGD	AdamW
Scheduler	ReduceLROn Plateau	None
Batch Size	64	16

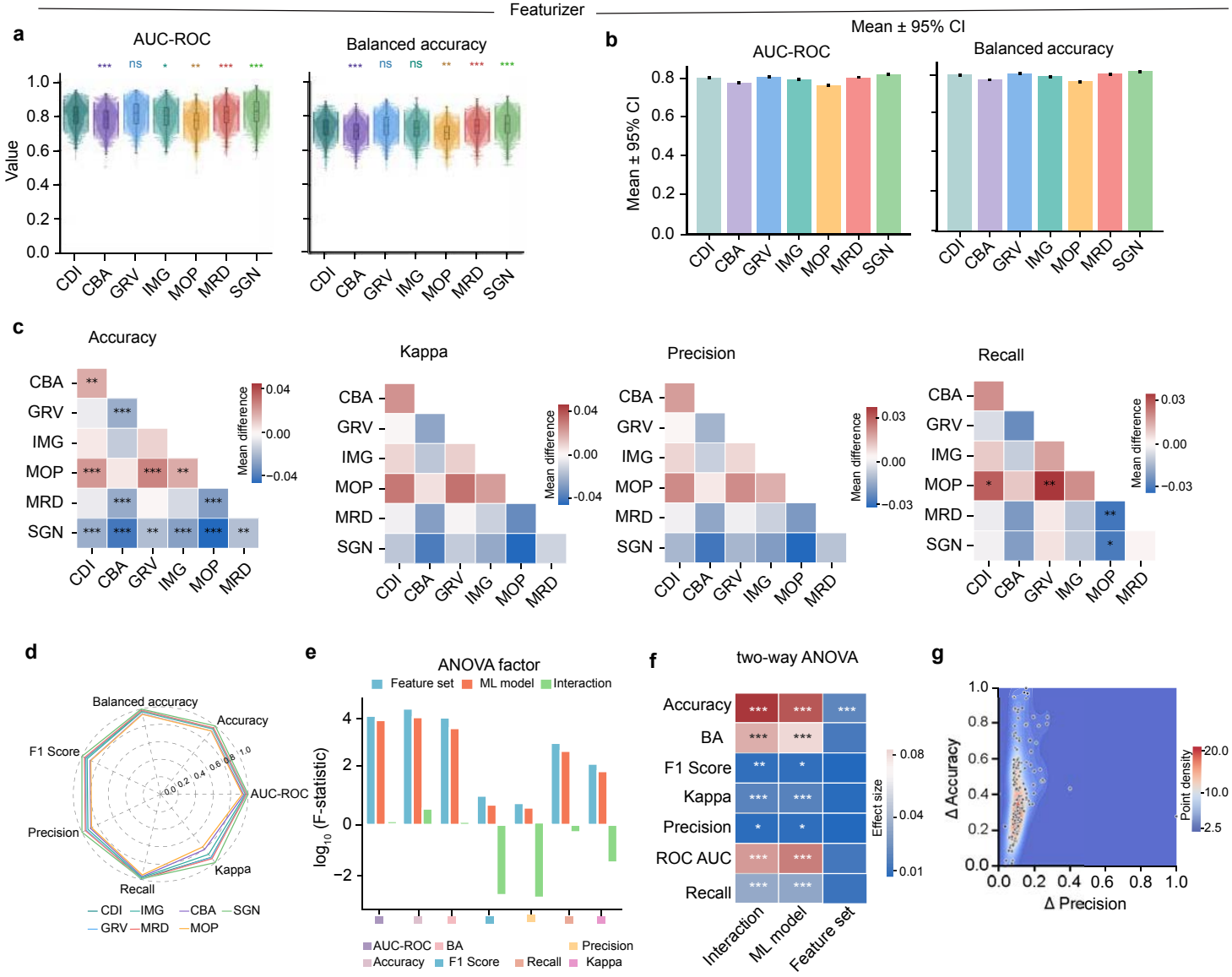
Supplementary Figure 1 | Mathematical formulation, dataset preparation and training configuration of the Chemical Dice Integrator.

(a) Overview of the six molecular representations integrated by CDI and their dimensionalities: MOPAC- derived quantum descriptors (MOP, 212), ChemBERTa (CBA, 384), Mordred descriptors (MRD, 1,421), Signaturizer (SGN, 3,200), GROVER (GRV, 4,800) and ImageMol (IMG, 10,000). The resulting CDI embedding has 8,192 dimensions. **(b)** Pairwise absolute Spearman-correlation matrices for the six multimodal representations and six conventional structural representations: AtomPair (ATP), descriptor set (DSC), extended-connectivity fingerprint (ECF), MACCS keys (MAC), RDKit fingerprint (RDK) and topological-torsion fingerprint (TOR). Lower off-diagonal correlations indicate reduced redundancy and greater informational complementarity. **(c)** Distribution of the 15 pairwise absolute Spearman correlations within the multimodal and structural representation groups. Violin plots show the distributions; internal box plots indicate the median and interquartile range, with whiskers extending to 1.5 times the interquartile range. Multimodal representations exhibited significantly lower pairwise correlations than structural representations (two-sided Mann-Whitney U test, n = 15 representation pairs per group, U = 49, P = 0.00897; Welch's two-sided t-test, t = –2.81, P = 0.00965). **(d)** Curation of the ChEMBL v35 training collection. Of 2,474,590 initial molecules, 4,986 were removed during CBA filtering, leaving 2,469,604 molecules; a further 237,734 molecules failed MOPAC processing, resulting in 2,231,870 molecules with aligned multimodal representations for CDI training. **(e)** Mathematical specification of the CDI framework, including the encoder and decoder functions, reconstruction-minimization objective, molecular-dataset definition, mean-squared-error and reconstruction-error terms, ReLU activation, SCA input dimensionality, leave-one-modality-out concatenation, latent representation, combined SCA loss, SEA reconstruction loss and total CDI-Basic objective. **(f)** Training configuration of CDI-Basic and CDI-Generalized. Source data are provided as a Source Data file.



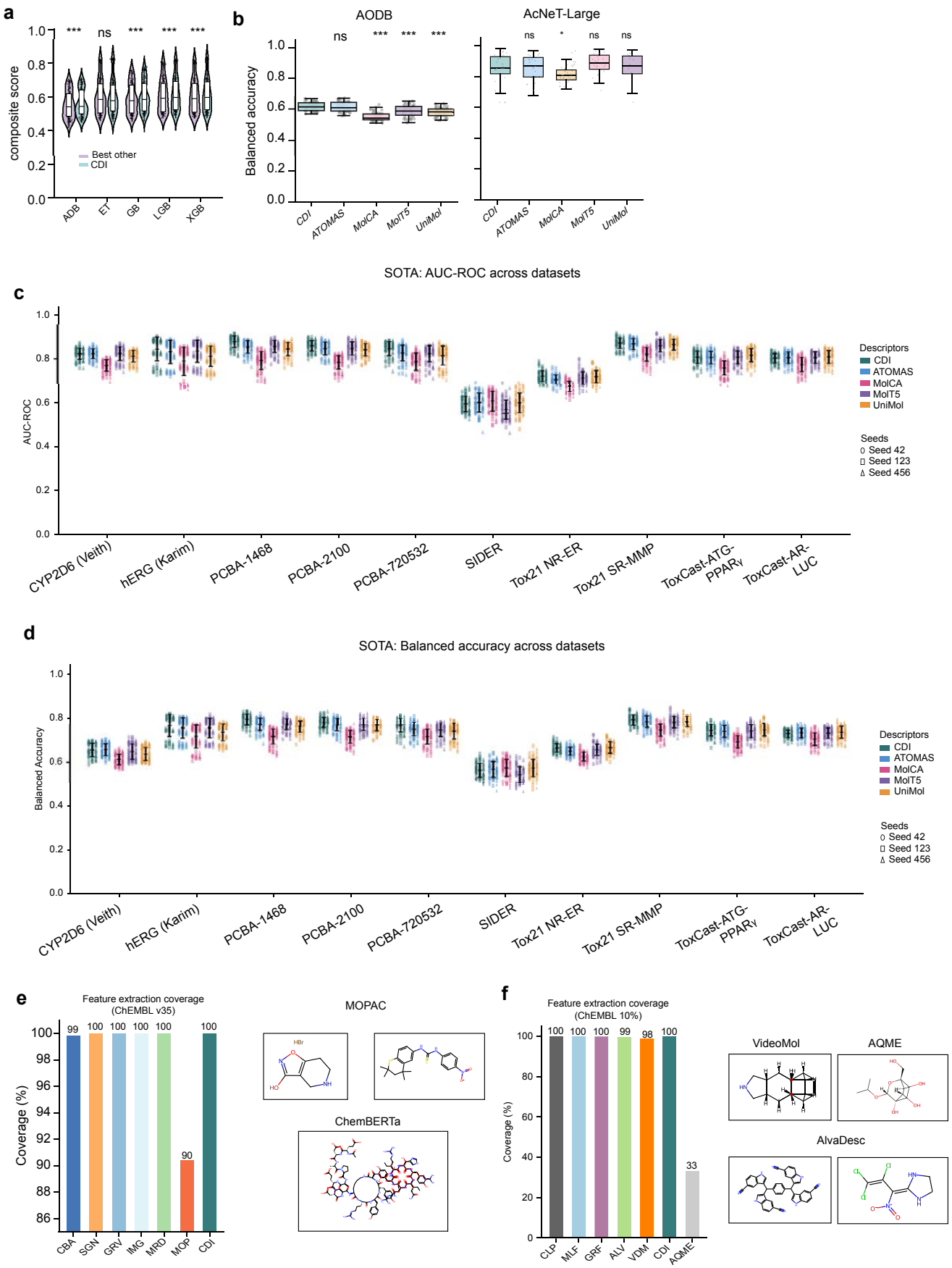
Supplementary Figure 2 | Aggregator-based benchmarking of feature representations across classification tasks.

(a) Raincloud distributions of ROC–AUC and balanced accuracy for CDI and eight aggregation baselines (CCA, ICA, Isomap, kPCA, LLE, PCA, RKS and t-SNE); n=1,539 model-dataset observations per feature set. CDI mean ± s.d.: ROC-AUC, 0.805 ± 0.069; balanced accuracy, 0.734 ± 0.064. CDI was compared with each baseline using two-sided Mann-Whitney U tests; all P < 0.001 **(b)** Mean ROC-AUC and balanced accuracy with 95% confidence intervals (mean ± 1.96 × s.d./n); n = 1,539 observations per feature set. No hypothesis test was applied. **(c)** Pairwise mean differences in accuracy, Cohen's κ, precision and recall after averaging across models, seeds and folds within dataset-task combinations; n = 171 observations per feature set (N = 1,539). One-way ANOVA followed by Tukey's HSD; dfbetween=8, dfresidual = 1,530. **(d)** Radar plot of normalized mean performance across seven classification metrics. **(e)** log10-transformed F-statistics from two-way ANOVA for feature set, ML model and their interaction. Feature-set effects were significant for all metrics (P < 0.001); AUC-ROC, F = 5,459.88; balanced accuracy, F = 4,360.98; accuracy, F = 2,781.06. **(f)** Partial ηp2 from the same ANOVA: AUC-ROC, 0.760; balanced accuracy, 0.717; accuracy, 0.618; recall, 0.434; Cohen's κ, 0.219. **(g)** Contour density of min-max-normalized ΔPrecision versus ΔAccuracy; n = 171. Pearson r = 0.474, P = 5.97×10−11; Spearman ρ = 0.636, P=9.05×10−21. **(h)** Contour densities for ΔRecall versus ΔF1 and ΔROC-AUC versus Δbalanced accuracy; n=171. Recall-F1: Pearson r = 0.470, P = 8.42×10−11; Spearman ρ=0.461, P = 2.33×10−10. ROC-AUC:balanced accuracy: Pearson r = 0.875, P = 4.19×10−55; Spearman ρ=0.876, P=2.33×10−55. Correlation tests were two-sided. Asterisks in a denote Mann-Whitney U tests and in c Tukey-adjusted comparisons: ns, Padj ≥ 0.05; *Padj < 0.05; **Padj < 0.01; ***Padj < 0.001. Source data are provided as a Source Data file.



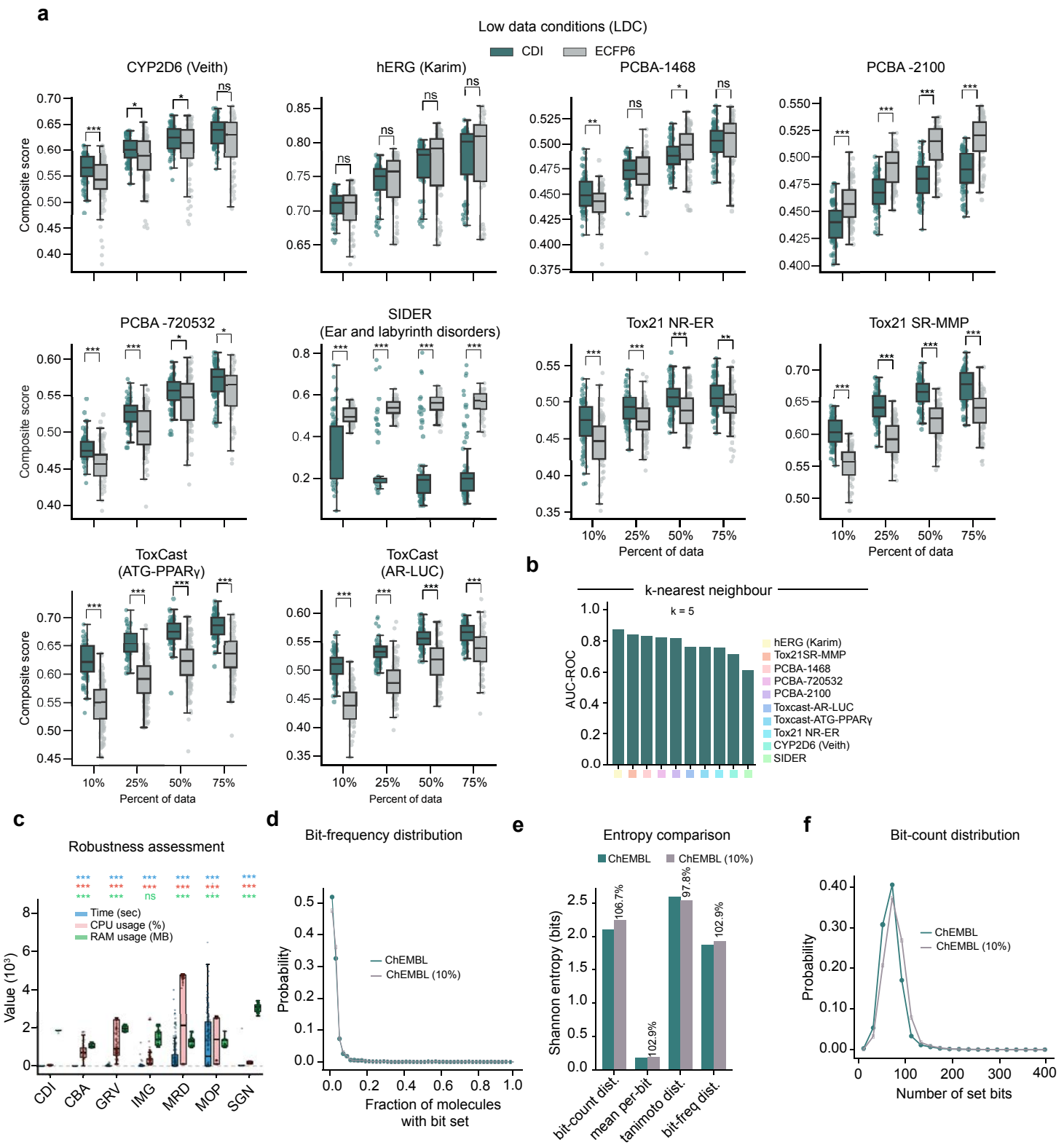
Supplementary Figure 3 | Featurizer-based benchmarking of CDI against individual molecular representations across classification tasks.

(a) Raincloud distributions of ROC-AUC and balanced accuracy for CDI and six constituent featurizers: ChemBERTa (CBA), GROVER (GRV), ImageMol (IMG), MOPAC (MOP), Mordred (MRD), and Signaturizer (SGN). Each representation contributed n = 1,539 model-dataset observations. CDI achieved mean ± s.d. values of 0.805 ± 0.069 for ROC-AUC and 0.734 ± 0.064 for balanced accuracy. Boxplot centre lines denote medians, box limits, first and third quartiles; whiskers, 1.5× the interquartile range, and points, individual observations. CDI was compared independently with each featurizer using two-sided Mann-Whitney U tests. Exact P values for ROC-AUC were CBA, 6.03 × 10-23; GRV, 8.07 × 10-3; IMG, 1.21 × 10-3; MOP, 1.54 × 10-51; MRD, 0.378; and SGN, 7.58 × 10-13. For balanced accuracy: CBA, 3.49 × 10-25; GRV, 2.12 × 10-2; IMG, 2.85 × 10-4; MOP, 5.02 × 10-47; MRD, 0.0678; and SGN, 6.42 × 10-12. **(b)** Mean ROC-AUC and balanced accuracy with 95% confidence intervals, calculated as mean ± 1.96 × s.d./√n; no additional test was applied. **(c)** Pairwise mean-difference heatmaps for accuracy, Cohen's κ, precision, and recall. Values were averaged within each dataset–task combination across models, seeds, and folds. Each representation contributed n = 171 observations; N = 1,197. Differences were assessed by one-way ANOVA followed by Tukey's HSD test (df = 6, dfresidual = 1,190). **(d)** Radar plot of normalized mean performance across seven metrics; no test was applied. (e, f) Two-way ANOVA F-statistics and partial η2 values for featurizer, model, and their interaction; n = 1,539 observations per representation, N = 10,773. **(g)** ΔPrecision versus ΔAccuracy across n = 171 tasks (Pearson r = 0.173, P = 0.0235; Spearman ρ = 0.416, P = 1.54 × 10-8). **(h)** ΔRecall versus ΔF1 (r = 0.342, P = 4.72 × 10-6; ρ = 0.416, P = 1.46 × 10-8) and ΔROC-AUC versus ΔBalanced Accuracy (r = 0.910, P = 1.48 × 10-66; ρ = 0.903, P = 8.69 × 10-64). Source data are provided as a Source Data file.



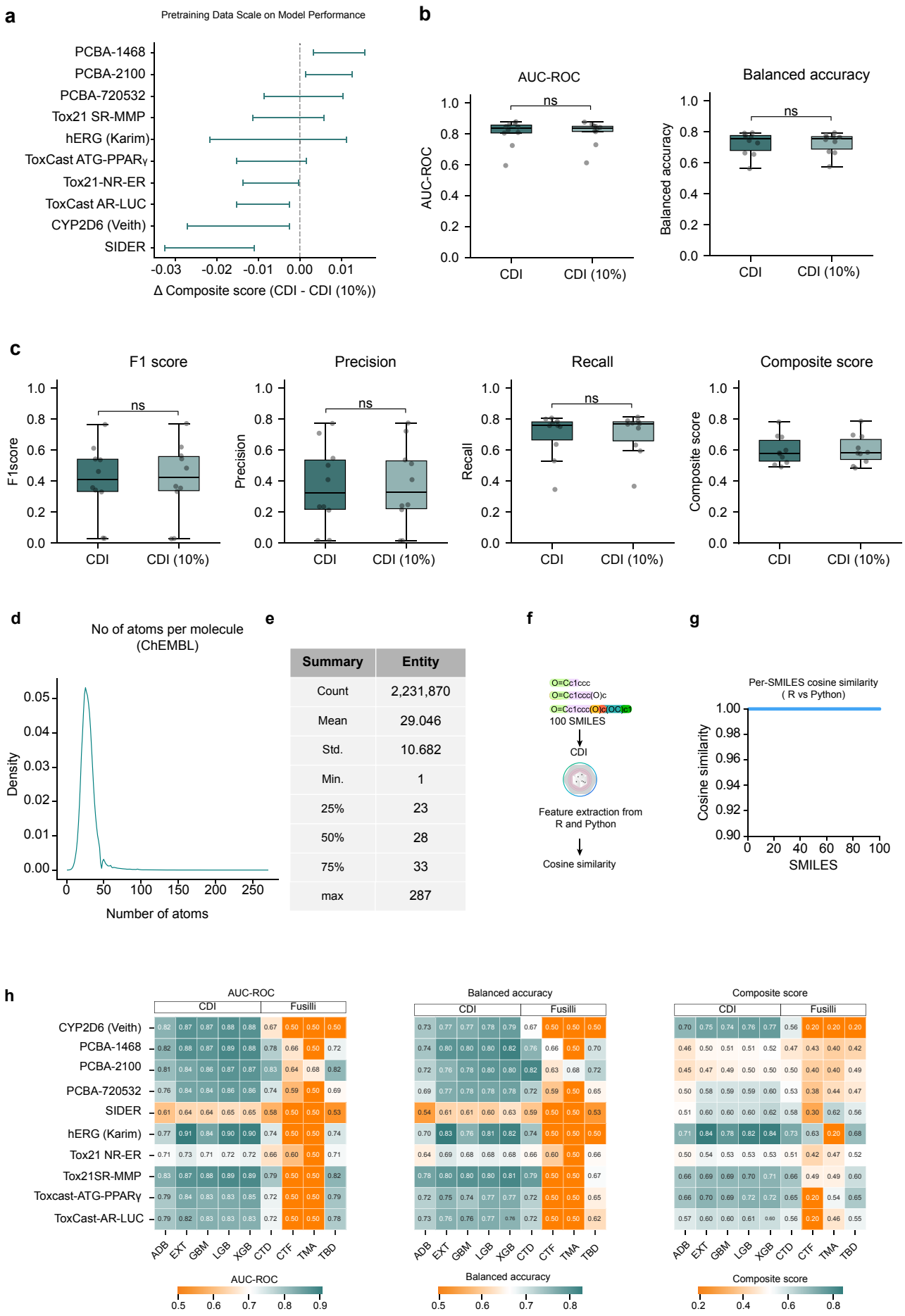
Supplementary Figure 4 | Additional SOTA benchmarking, external validation, and feature-extraction coverage.

(a) Composite-score comparison between CDI and the best-performing non-CDI descriptor for AdaBoost (ADB), Extra Trees (ET), Gradient Boosting (GB), LightGBM (LGB) and XGBoost (XGB). Composite score was the mean of ROC-AUC, balanced accuracy, F1 score, precision and recall. Violin plots show paired replicate-level distributions with embedded boxplots; centre line, median; box limits, first and third quartiles; whiskers, 1.5× the interquartile range. Each classifier comparison contained n = 150 matched dataset-target-fold-seed replicates. CDI differed significantly from the best alternative for ADB, GB, LGB and XGB, but not ET. **(b)** Balanced-accuracy validation on AODB and ACNet-Large for CDI, ATOMAS, MolCA, MolT5 and UniMol. AODB included n = 50 replicates per representation (N = 250); ACNet-Large included n = 15 per representation (N = 75). CDI differed significantly from MolCA, MolT5 and UniMol on AODB, but not ATOMAS; no CDI-versus-comparator difference was significant on ACNet-Large. **(c)** ROC- AUC benchmarking across ten classification tasks. Each task–representation combination contributed up to n = 75 valid observations, yielding N = 3,750 across the panel. Points show individual runs and error bars denote mean ± s.d. **(d)** Balanced-accuracy benchmarking across the same tasks and representations; sample sizes and plotting conventions are as in panel d. CDI-versus-comparator differences in panels d and e were assessed using two-sided Mann-Whitney U tests with Benjamini-Hochberg correction. **(e)** Feature-extraction coverage across ChEMBL (v35): CBA, 99.8%; SGN, GRV, IMG and MRD, 100%; MOP, 90.39%; and CDI, 100%. Representative CBA and MOP failures are shown. **(f)** Coverage of replacement encoders on the representative 10% ChEMBL subset: CLAMP, MolFormer and CDI, 100%; Graphormer, 99.997%; AlvaDesc, 99.81%; VideoMol, 98.95%; and AQME, 33.25%. Representative failures are shown for VideoMol, AQME and AlvaDesc. Significance notation: ns, adjusted P ≥ 0.05; *adjusted P < 0.05; **adjusted P < 0.01; ***adjusted P < 0.001. Source data are provided as a Source Data file.



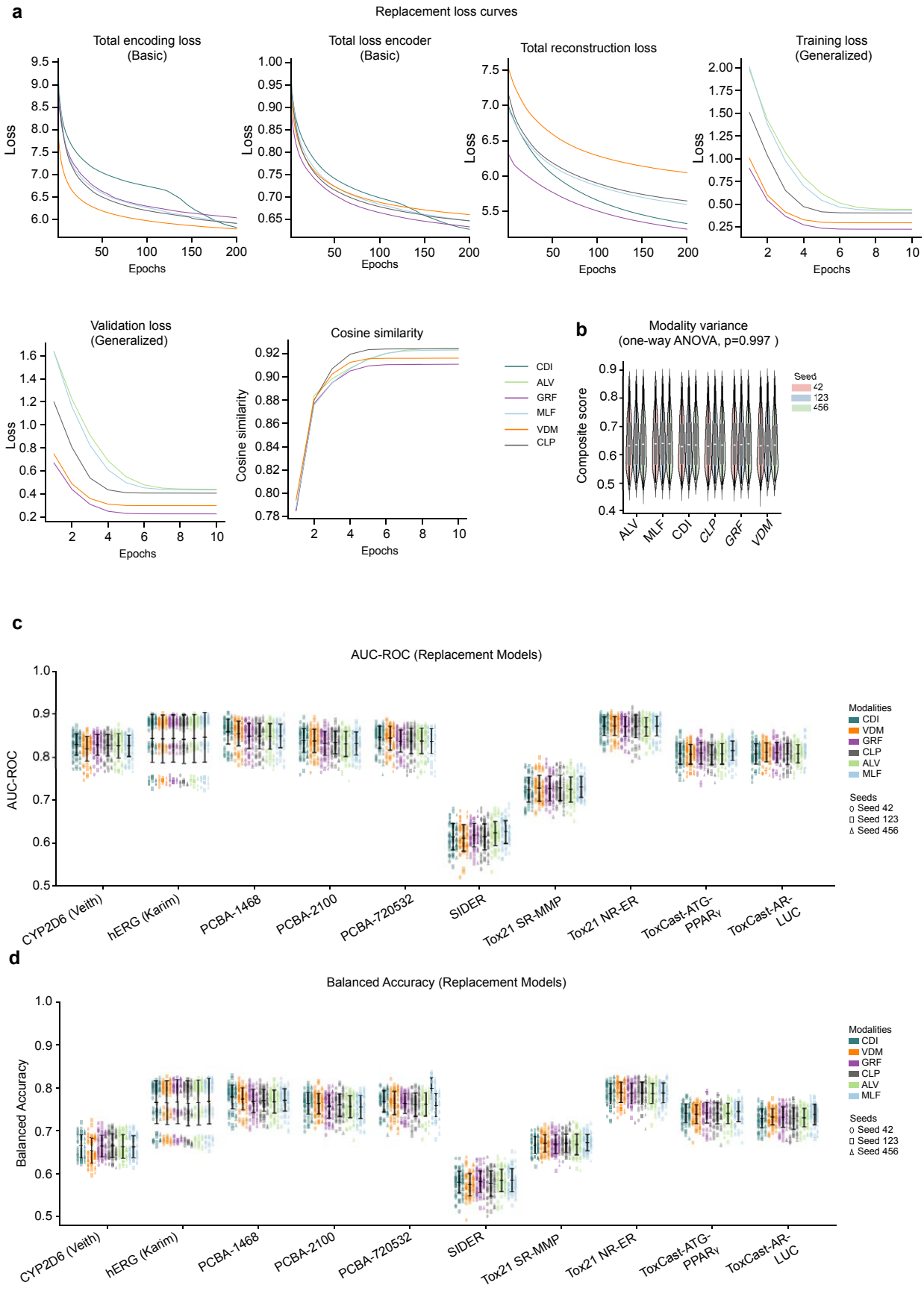
Supplementary Figure 5 | Low-data performance, neighbourhood preservation, computational robustness and representativeness of the 10% ChEMBL subset.

(a) Composite-score distributions for CDI and ECFP6 under low-data conditions using 10%, 25%, 50% and 75% of each training fold across ten classification benchmark tasks. The held-out test fold was unchanged across fractions. Boxplots show replicate-level performance, with individual observations overlaid; significance annotations are shown above the corresponding comparisons. **(b)** Leave-one-out k-nearest-neighbour classification using CDI embeddings with k = 5 and cosine distance across the ten benchmark datasets. Bars show dataset-level ROC-AUC values, demonstrating preservation of local predictive neighbourhoods. **(c)** Computational robustness profiling of CDI and the six source-featurizer pipelines (ChemBERTa, GROVER, ImageMol, Mordred, MOPAC and Signaturizer), including runtime, CPU usage and RAM usage. **(d)** ECFP6 bit-frequency distributions for the full ChEMBL training collection and the chemically representative 10% ChEMBL subset. **(e)** Shannon-entropy comparison of the fingerprint distributions in the full and 10% ChEMBL collections across the categories shown. **(f)** Distribution of ECFP6 bit counts in the full and 10% ChEMBL collections. Panels d-f assess whether reduced-scale sampling preserves the structural and information-content characteristics of the full pretraining collection. Exact numerical values and statistical outputs are provided in the accompanying Source Data. Source data are provided as a Source Data file.



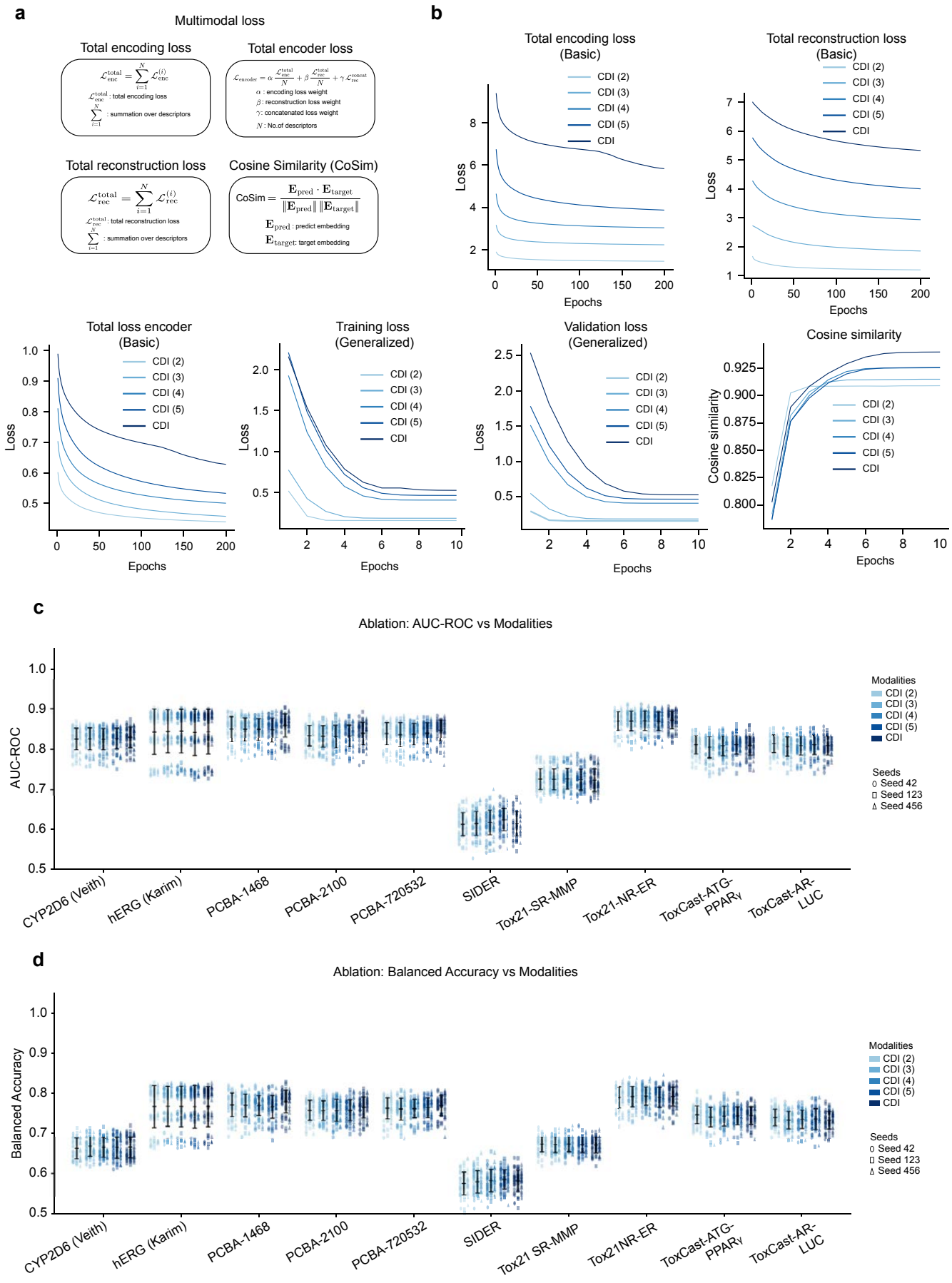
Supplementary Figure 6 | Effect of representative pretraining subset (10% ChEMBL) on CDI performance and dataset characterization.

(a) Paired composite-score differences between CDI pretrained on full ChEMBL and the representative 10% subset, calculated as CDI − CDI (10%). Points show mean differences and horizontal bars indicate 95% confidence intervals. **(b)** AUC–ROC and balanced accuracy for CDI and CDI (10%) across n=10 benchmark datasets. Centre line, median; box limits, interquartile range (IQR); whiskers, 1.5× IQR; points, individual datasets. Two-sided paired Wilcoxon signed-rank tests; ns, not significant. **(c)** F1 score, precision, recall and composite score for CDI and CDI (10%) across n = 10 datasets. Boxplot definitions and statistical testing are as in b; ns, not significant. **(d)** Kernel-density distribution of atom counts in the ChEMBL pretraining dataset. **(e)** Atom-count statistics for n = 2,231,870 molecules: mean, 29.046; s.d., 10.682; median, 28; IQR, 23-33; range, 1-287 atoms. **(f)** Workflow comparing CDI implementations in R and Python using 100 SMILES and cosine similarity between corresponding embeddings. **(g)** Per-molecule cosine similarity between R- and Python-derived CDI embeddings (n = 100); values were approximately 1.0 throughout. **(h)** Heatmaps of AUC-ROC, balanced accuracy and composite score for CDI and Fusilli across benchmark dataset-classifier combinations; cells show mean performance. Source data are provided as a Source Data file.



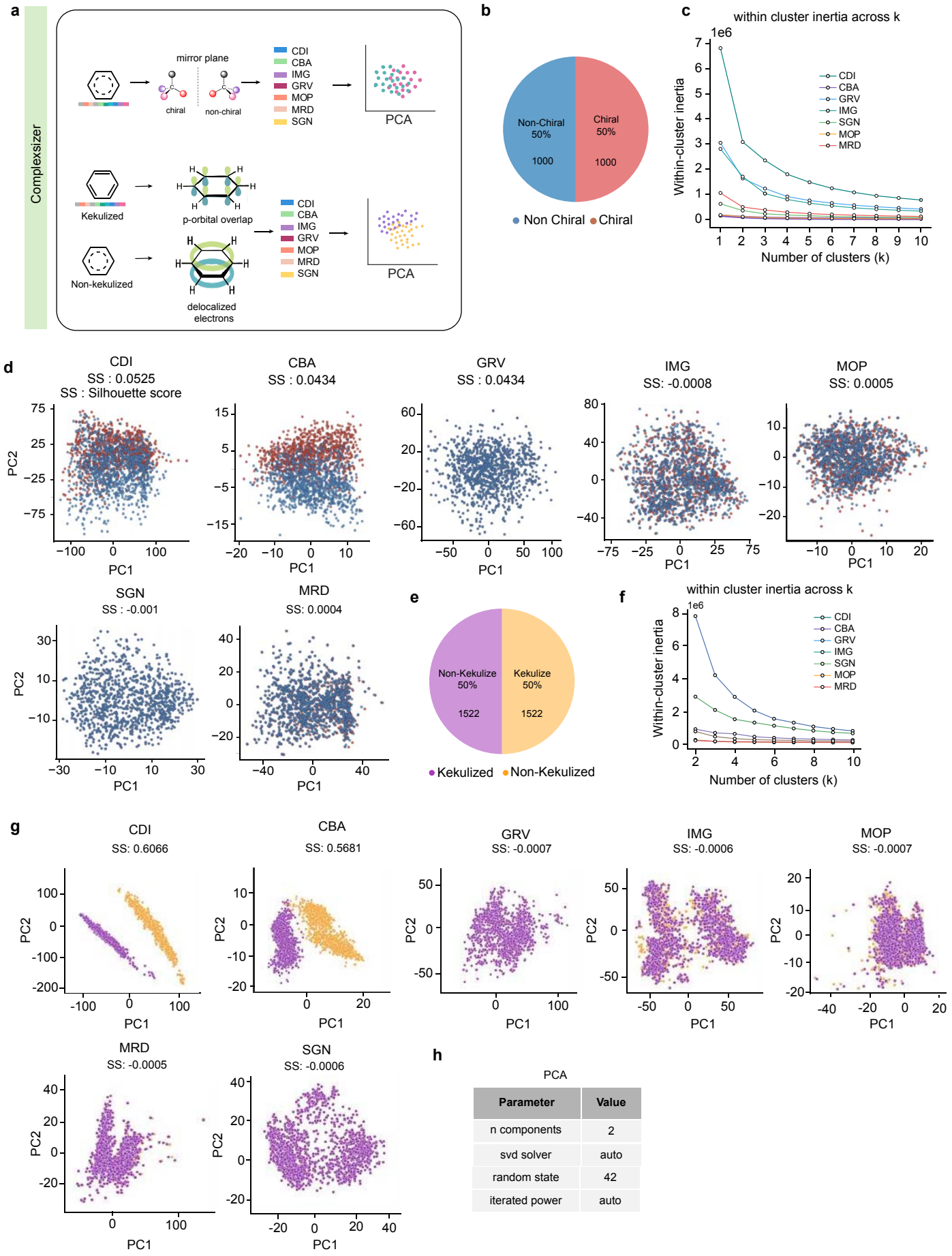
Supplementary Figure 7 | Training dynamics and modality-replacement analysis of CDI representations.

(a) Training dynamics of CDI and five modality-replacement configurations: AlvaDesc (ALV), CLAMP (CLP), Graphormer (GRF), MolFormer (MLF) and VideoMol (VDM), replacing Mordred(MRD), Signaturizer (SGN), GROVER (GRV), ChemBERTa (CBA) and ImageMol (IMG), respectively. CDI-Basic curves show total encoding, encoder and reconstruction losses over 200 epochs; CDI-Generalized curves show training loss, validation loss and cosine similarity over 10 epochs. Each line represents one training trajectory; no error bars or confidence bands are shown. **(b)** Composite-score distributions across CDI and the five replacement configurations, where composite score was the mean of ROC-AUC, balanced accuracy and F1 score. Each representation contributed n = 750 observations, comprising 250 runs per seed across three seeds (42, 123 and 456), for N = 4,500 observations. Violin plots show replicate-level distributions. A two-sided one-way ANOVA found no overall difference among representations ($F(5,4494) = 0.1244$, $P = 0.997$). Pairwise CDI-versus-replacement comparisons used two-sided paired Wilcoxon signed-rank tests with Benjamini-Hochberg correction across five comparisons: VDM, adjusted $P = 5.42 \times 10\text{-}4$, $W = 118{,}616$; GRF, $P = 0.0220$, $W = 126{,}726$; CLP, $P = 5.42 \times 10\text{-}4$, $W = 118{,}862$; ALV, $P = 0.00192$, $W = 121{,}518$; and MLF, $P = 0.480$, $W = 136{,}616$. **(c)** Dataset-wise ROC-AUC comparison across ten classification tasks. Each representation and dataset-task combination contributed n = 75 observations from five classifiers × three seeds × five folds, yielding N = 4,500 observations. Points denote individual runs, marker shapes denote seeds and vertical error bars show mean ± s.d. **(d)** Balanced-accuracy comparison using the same design and sample sizes. For panels c and d, CDI was compared with each replacement using two-sided paired Wilcoxon signed-rank tests with Benjamini-Hochberg correction. Exact statistics and adjusted P values are provided in the Source Data. CDI-versus-replacement differences were assessed separately for each dataset-task combination using two-sided paired Wilcoxon signed-rank tests, followed by Benjamini-Hochberg correction across all comparisons within the panel. Exact test statistics and raw and adjusted P values are provided in the accompanying Source Data. Source data are provided as a Source Data file.



Supplementary Figure 8 | Training dynamics and systematic modality-ablation analysis of CDI representations.

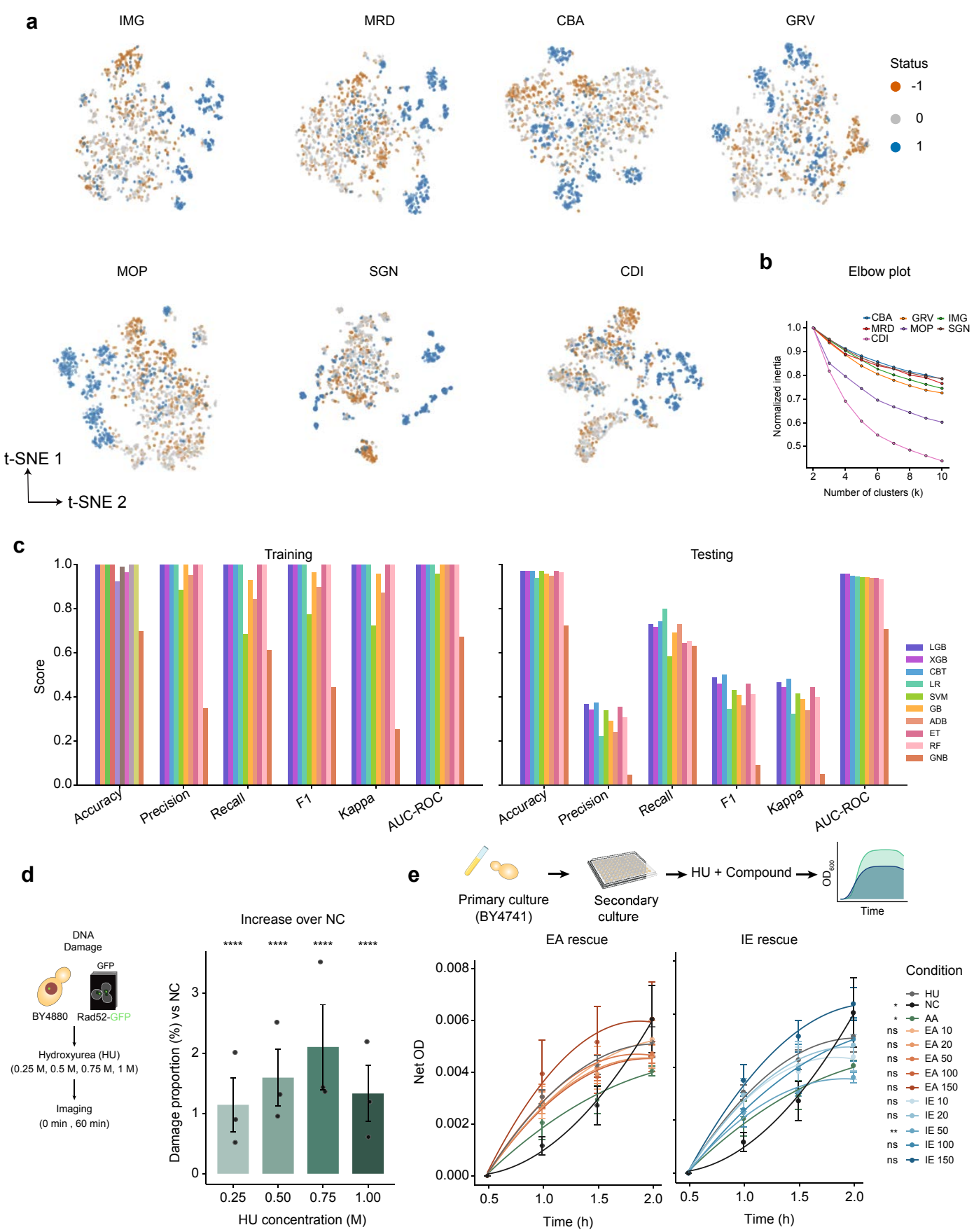
(a) Mathematical formulation of the multimodal CDI objective, including modality-specific encoding and reconstruction losses, the combined encoder objective and cosine similarity between predicted and target embeddings. **(b)** Training dynamics for full CDI and sequentially ablated configurations. CDI(5) excludes MOPAC; CDI(4) excludes MOPAC and Mordred; CDI(3) additionally excludes GROVER; and CDI(2) additionally excludes Signaturizer, retaining ImageMol and ChemBERTa. Modalities were removed according to their ΔAUL-based contribution ranking. CDI-Basic curves show total encoding, reconstruction and encoder losses over 200 epochs; CDI-Generalized curves show training loss, validation loss and cosine similarity over 10 epochs. Each line represents one trajectory; no error bars or confidence bands are shown. **(c)** Dataset-wise ROC-AUC comparison of CDI, CDI(5), CDI(4), CDI(3) and CDI(2) across ten classification tasks. Each representation and dataset-task combination contributed n = 75 observations from five classifiers × three seeds × five folds, yielding N = 3,750 observations. Points denote individual model runs, marker shapes denote seeds and vertical error bars show mean ± s.d. **(d)** Balanced-accuracy comparison for the same representations, datasets and sample sizes. For panels c and d, full CDI was compared with each ablated representation using two-sided paired Wilcoxon signed-rank tests with Benjamini-Hochberg correction across comparisons. Exact WWW statistics and raw and adjusted P values are provided in the Source Data. No inferential test was applied to the mathematical formulation in panel a or the training trajectories in panel b. Significance notation: ns, BH-adjusted P ≥ 0.05; *0.01 ≤ adjusted P < 0.05; **0.001 ≤ adjusted P < 0.01; and ***adjusted P < 0.001. Source data are provided as a Source Data file.



Parameter	Value
n components	2
svd solver	auto
random state	42
iterated power	auto

Supplementary Figure 9 | Complexizer analysis of structural sensitivity across molecular representations.

(a) Overview of the Complexizer evaluation framework. Seven molecular representations CDI, ChemBERTa (CBA), GROVER (GRV), ImageMol (IMG), MOPAC (MOP), Mordred (MRD) and Signaturizer (SGN) were evaluated for sensitivity to stereochemistry and bond-order changes. PCA was used for two-dimensional visualization. **(b)** Composition of the stereochemistry dataset comprising n = 1,000 chiral and n = 1,000 non-chiral molecules (N = 2,000). **(c)** Elbow curves showing within-cluster inertia for k = 1–10 clusters using PCA-reduced embeddings. Each curve was computed from the N = 2,000 molecules in panel b. **(d)** PCA projections of chiral and non-chiral molecules for each representation. Each point represents one molecule (n = 1,000 per class). Class separation was quantified descriptively using the silhouette score (SS). **(e)** Composition of the bond-order dataset comprising n = 1,522 kekulized and n = 1,522 non-kekulized molecules (N = 3,044). **(f)** Elbow curves showing within-cluster inertia for k = 1-10 clusters using PCA-reduced embeddings from the bond-order dataset. **(g)** PCA projections of kekulized and non-kekulized molecules. Each point represents one molecule (n = 1,522 per class; N = 3,044 per representation). Silhouette scores quantify separation between the two classes. **(h)** PCA configuration used for dimensionality reduction: two principal components, svd_solver='auto', iterated_power='auto', and random_state=42. This figure is descriptive; no hypothesis testing, multiple-testing correction or significance notation was applied. Source data are provided as a Source Data file.



Supplementary Figure 10 | Embedding separability, predictive performance and experimental validation of CDI-prioritized genomic-stability candidates.

(a) Two-dimensional t-SNE projections of embeddings generated by ImageMol (IMG), Mordred (MRD), ChemBERTa (CBA), GROVER (GRV), MOPAC (MOP), Signaturizer (SGN) and CDI, coloured by genomic-instability class: genotoxic (–1), non-genotoxic/neutral (0) and pro-genomic stability (1). IMG, MRD, CBA, GRV, SGN, and CDI each contained n = 1,803 molecules (601 per class); MOP contained n = 1,784 molecules owing to descriptor-generation failures. **(b)** Elbow curves of normalized within-cluster inertia for k = 2-10 clusters using the embeddings shown in panel a. **(c)** Training and independent testing performance of ten machine-learning classifiers (LightGBM, XGBoost, CatBoost, logistic regression, support vector machine, gradient boosting, AdaBoost, Extra Trees, Random Forest and Gaussian naïve Bayes) evaluated using accuracy, precision, recall, F1 score, Cohen's κ and ROC-AUC. Each bar represents one classifier-level estimate (n = 10 classifiers per metric). **(d)** Validation of hydroxyurea-induced DNA damage in Saccharomyces cerevisiae BY4880 expressing the Rad52-GFP reporter. Cells were treated with 0.25, 0.50, 0.75 or 1.00 M hydroxyurea and imaged at 0 and 60 min. Bars show mean normalized DNA damage from n = 3 independent biological experiments; error bars denote s.e.m. Hydroxyurea concentrations were compared with the negative control using two-sided Mann- Whitney U tests. **(e)** Yeast outgrowth assay used to select eugenyl acetate (EA) and isoeugenol (IE) concentrations. Cultures were treated with hydroxyurea alone or with EA or IE (10-150 µg mL^{-1}) and monitored by OD600 or approximately 2 h. Points show mean net OD; error bars denote s.e.m. across n = 9 well-level measurements per condition and time point (n = 6 for water and DMSO controls). Three independent biological experiments were performed. Comparisons with hydroxyurea used two-sided Mann-Whitney U tests; exact unadjusted P values are reported in the Source Data. No multiple-comparison correction was applied. Source data are provided as a Source Data file.

Supplementary Data captions

Six Supplementary Data workbooks are submitted separately. Each workbook contains a README worksheet describing its included worksheets and contents.

Supplementary Data 1. Benchmark datasets and software versions.

Summary of the classification and regression benchmark datasets used to evaluate the Chemical Dice Integrator (CDI), including dataset characteristics, prediction endpoints, task information, molecular counts, data sources, and software or package versions used in the analyses.

Supplementary Data 2. Aggregator, featurizer, and state-of-the-art benchmarking results.

Results from the internal benchmarking of CDI against dimensionality-reduction aggregators and individual molecular featurizers, together with comparisons against state-of-the-art molecular representations. The file includes classification and regression performance metrics, model-level and task-level results, win-rate calculations, effect-size summaries, and associated statistical comparisons.

Supplementary Data 3. External validation, generalization, neighborhood, and computational benchmarking.

Results from independent-dataset validation, scaffold-based out-of-distribution evaluation, low-data-condition analyses, k-nearest-neighbor and leave-one-out evaluations, embedding-coverage assessments, and computational-efficiency benchmarking.

Supplementary Data 4. Modality replacement, sequential ablation, multimodal fusion, and Complexizer analyses.

Results from controlled modality-replacement experiments, sequential modality-ablation analyses, comparison with generic multimodal tabular-fusion approaches, and Complexizer evaluations of chirality and kekulization sensitivity. The file includes performance metrics, win rates, effect sizes, and relevant statistical analyses.

Supplementary Data 5. Genomic-instability prediction and experimental validation.

Curated genomic-instability training data, class-balanced model-development data, predictions for the independent in-house screening library, candidate-prioritization outputs, yeast wet-lab measurements, and the corresponding statistical results.

Supplementary Data 6. CDI-Generator outputs.

Outputs generated using CDI-Generator, including the molecular-generation and optimization results reported in the manuscript.