

SHIELD reduces information leakage in data splits through joint structural optimization

Ayman Karaki^{1,*}, Marwan Khraisheh¹

¹College of Science and Engineering, Hamad Bin Khalifa University, P.O. Box 34110, Doha, Qatar

*Corresponding author: ayka89500@hbku.edu.qa

Contents

S1	Overview of baseline methods	2
S1.1	Random split	
S1.2	Stratified split	
S1.3	k -means split	
S1.4	Agglomerative split	
S1.5	Kennard–Stone split	
S1.6	Leave-chemistry-system-out split	
S1.7	LP-PDBBind quality-based splits	
S1.8	DeepChem’s Scaffold split	
S1.9	DeepChem’s Butina clustering split	
S1.10	DeepChem’s Fingerprint split	
S1.11	DeepChem’s MaxMin split	
S1.12	DeepChem’s Molecular-weight split	
S1.13	DataSAIL split	
S2	Evaluating Extrapolative Splits	4
S2.1	Visual separation in the embedded feature space	
S2.2	Scaled cross-fold similarity	
S2.3	k -nearest-neighbor purity	
S2.4	α -shape intersection-over-union	
S2.5	Downstream regressor performance	
S2.6	Downstream classifier performance	
S3	Benchmark datasets	6
S4	Benchmark 1: LPBF dataset	6
S5	Benchmark 2: matbench_glass	9
S6	Benchmark 3: PDBbind	11
S7	Benchmark 4: tox21	13
S8	Benchmark 5: lipophilicity	16
S9	Benchmark 6: qm8	18
S10	Benchmark 7: qm9	22
S11	Benchmark 8: HIV	25
S12	Supplementary methods	27
S12.1	Cut identity for term M	
S12.2	Entropic transport for term D	
S12.3	Kernel discrepancy for term D	
S12.4	Stratification forms	
S12.5	Calibration scales	
S12.6	Computational complexity	

S1. Overview of baseline methods

SHIELD is compared against a broad set of existing data-splitting strategies across the eight benchmark datasets described in the following sections. These competitors span computationally simple statistical baselines, unsupervised clustering methods, deterministic diversity-based selection algorithms, chemistry- and materials-specific heuristics, and DataSAIL. Each is briefly described in the following subsections, together with its mechanism and its main strengths and weaknesses as a tool for out-of-distribution (OOD) evaluation. The specific subset applied to each benchmark and its exact configuration are given in the corresponding benchmark section.

S1.1 Random split

A random split¹ assigns each observation to a fold independently at random, subject only to the proportions of the target fold without any reference to similarity, hierarchy, or target value. It is computationally trivial and does not impose assumptions on the criterion that governs deployment relevance, making it a natural and universal point of reference. By construction, however, it offers no protection against related observations falling on opposite sides of the split, since a related pair is exactly as likely to be separated as to remain together, so it routinely leaks information and typically yields the most interpolative benchmark estimate of the splits evaluated.

S1.2 Stratified split

A stratified split¹ extends random assignment by additionally constraining each fold to preserve the marginal distribution of the target, drawn per class for a discrete label or per bin for a continuous target. It keeps target proportions representative across folds, which is useful for imbalanced classification tasks and remains computationally inexpensive. Because it constrains only the target distribution and not feature space similarity or hierarchies, it inherits the same vulnerability to information leakage as a random split, while adding no additional extrapolative difficulty.

S1.3 k -means split

A k -means split¹ partitions the observation feature vectors into k clusters by minimizing within-cluster squared distance to iteratively updated centroids, after which whole clusters, or balanced combinations of clusters, are assigned to folds. The procedure is entirely unsupervised and applicable to any feature representation and because it groups observations by generic proximity, it can incidentally achieve strong feature space separation when a dataset is naturally organized into discrete clusters. Since it does not preserve target balance or impose any constraints relative to the deployment, a fold can absorb entire classes, and the number of clusters is an additional dataset-dependent choice with no principled default.

S1.4 Agglomerative split

An agglomerative split¹ builds a hierarchical clustering by iteratively merging the closest pair of clusters under a chosen linkage criterion until the desired number of folds remains and then assigns each resulting cluster to a fold. Like k -means, it is unsupervised and generic and can produce visually distinct and well-separated clusters, particularly when a dataset is naturally organized into discrete clusters. The resulting fold sizes are not directly controllable and can be highly imbalanced, and as with k -means, cluster boundaries may not represent scientifically meaningful or deployment-relevant groupings.

S1.5 Kennard–Stone split

The Kennard–Stone algorithm² is a deterministic procedure that begins by selecting the two observations with the greatest mutual distance in feature space, then iteratively adds to the growing training fold the remaining observation that has the greatest minimum distance to the observations already selected, continuing until a target fraction of the data has been assigned; the unselected remainder becomes the held-out fold. Because it was originally designed to build a maximally representative calibration set rather than to withhold a genuinely novel region, the observations left for testing tend to lie within the span the training fold already covers, which can understate extrapolative difficulty relative to a split that deliberately targets a specific deployment setup.

S1.6 Leave-chemistry-system-out split

A leave-chemistry-system-out (LCSO) split is a standard leave-one-group-out (LOGO) split¹ where the group identifier is the chemical system. An LCSO withholds every composition that belongs to one or more entire chemical systems, a system being a specific combination of constituent elements, so that the held-out fold contains only systems absent from training. It is a domain-native and easily interpretable cold-split criterion for materials science, directly targeting a genuinely unseen element combination. A naive LOGO implementation gives no control over how similar an excluded system’s compositions are to the retained ones in a continuous compositional sense, so it can still permit near-duplication between folds even though no exact chemical system is shared.

S1.7 LP-PDBBind quality-based splits

Three further binary splits are constructed from the predefined LP-PDBBind cleaning level indicators CL1, CL2, and CL3,³ which denote membership in progressively stricter quality-controlled subsets of the dataset. Each indicator is treated independently and gives rise to a separate train-test partition rather than a single split combining all three. For a given indicator, complexes are divided according to whether they belong to the corresponding cleaning level, where the larger group is assigned to training and the smaller group is assigned to the held-out fold. Therefore, CL1, CL2, and CL3 splits are based on data quality criteria rather than molecular similarity, biological relatedness, or any clustering procedure. Because each split is built from a single cleaning level indicator without any constraints for similarity, hierarchy, or target balance, the resulting train-test partition carries no guaranteed relationship to any of the structural criteria evaluated elsewhere in this study, and because assignment follows the indicator’s own majority and minority group sizes rather than a fixed target proportion, the resulting fold sizes are not directly controllable. This construction simulates deployment across data quality regimes, evaluating generalization between complexes that satisfy different levels of the LP-PDBBind quality-control criteria.

S1.8 DeepChem’s Scaffold split

The scaffold splitter⁴ computes the Bemis-Murcko molecular scaffold, the ring systems and linking substructure of a molecule stripped of substituents, groups molecules that share an identical scaffold, and assigns scaffold groups to folds ordered by group size. It directly targets a chemically meaningful and widely used definition of structural relatedness and keeps molecules built on the same core intact within a fold, thereby reducing analog series leakage. It treats scaffold identity as a hard binary criterion and so ignores similarity between distinct but structurally close scaffolds, and the resulting fold composition is dictated entirely by the scaffold frequency distribution of the dataset, offering no control over target balance.

S1.9 DeepChem’s Butina clustering split

The Butina splitter⁴ applies the Butina clustering algorithm to pairwise Tanimoto similarity based on fingerprints, grouping molecules into clusters defined by a similarity threshold and assigning whole clusters to folds. Unlike methods that require the number of clusters to be specified in advance, the number and composition of clusters emerge from the selected similarity threshold. Like scaffold splitting, it keeps highly similar molecules together but can also capture relationships between close analogs that do not share an identical scaffold.

S1.10 DeepChem’s Fingerprint split

The fingerprint splitter⁴ separates molecules by fingerprint dissimilarity evaluated over the whole dataset, comparing every pair of molecules to build the fold assignment. This pairwise comparison captures dissimilarity structure that a coarser clustering or scaffold-based criterion can miss. The limitation of this pairwise comparison is the computational quadratic scaling with the number of molecules, which limits its practicality on larger benchmarks.

S1.11 DeepChem’s MaxMin split

The MaxMin splitter⁴ applies a greedy maximum-minimum diversity-picking procedure, iteratively adding to a subset whichever remaining molecule has the greatest minimum fingerprint distance to molecules already in that subset, the same general principle the Kennard–Stone algorithm applies to a continuous feature space, here applied to a discrete fingerprint representation. It produces a maximally diverse training subset with guaranteed coverage of the chemical space represented in the dataset. It shares the same underlying limitation as Kennard–Stone which can understate the difficulty of genuine extrapolation.

S1.12 DeepChem’s Molecular-weight split

The weight-based splitter⁴ orders molecules by molecular weight and assigns folds along this single ordering. It is computationally trivial and gives an interpretable extrapolation axis whenever molecular weight strongly correlates with the property of interest. Molecular weight is only one, often weak, proxy for the structural or mechanistic similarity that actually governs most property predictions. Consequently, this split may not separate the folds along the axis of similarity relevant to the deployment settings and can leave substantial structural leakage between the folds.

S1.13 DataSAIL split

DataSAIL⁵ treats fold assignment as a combinatorial optimization problem. It works over one precomputed similarity matrix, or two for paired data, and searches for an assignment that minimizes similarity crossing the fold boundary. The search is subject to target fold sizes and optional stratification constraints. While its single-entity modes optimize along one axis, such as protein identity or ligand identity, its two-entity cold mode optimizes both axes jointly. DataSAIL is the closest existing method to SHIELD. Both frame splitting as an explicit optimization problem rather than a heuristic clustering or ordering rule, and DataSAIL is the only baseline here that couples similarity-aware splitting with stratification for paired biological entities. However, it optimizes over a single relational input type, a similarity matrix, so it cannot jointly account for hierarchical group containment or an explicit target-distribution

objective the way SHIELD does. Under two-entity cold splitting, it resolves overlap between observations that are similar along both axes simultaneously by discarding those observations outright. This can remove a substantial fraction of the dataset as in the case of PDBbind benchmark.

S2. Evaluating Extrapolative Splits

Assessing whether a data partition provides a genuine test of OOD deployment or extrapolative performance requires more than a single metric. This section presents a set of quantitative and qualitative diagnostics used to evaluate the quality of a split, independently of which splitting method produced the partition. The rationale for a multi-metric assessment is that no single diagnostic is reliable on its own, which the results across the benchmarks support. Each metric encodes a distinct structural definition of information leakage, whether through a qualitative view of the embedded feature space, a global graph cut, a local neighborhood property, or a downstream empirical performance. Since a split can satisfy one of these diagnostics while failing another, an assessment based on any one metric in isolation can be misled. A data partition that fragments a class or a natural grouping, for example, can register a favorable cut-based score while producing degenerate folds, and a partition that looks well separated in an embedding can leave the specific structural subgroup responsible for leakage untouched. The approach taken throughout this work is therefore to evaluate every split along multiple complementary axes (Table S1) and to treat agreement or disagreement between them as itself informative, rather than to select any single metric as definitive.

Table S1 | Comparison of the split-quality diagnostics. Level distinguishes whether metrics diagnose the whole partition (global), or observation’s local neighborhood (local). Channel probed identifies which leakage channel the diagnostic is most directly informative about; holistic indicates that all three channels contribute jointly and cannot be disentangled from the result alone. Link to SHIELD’s objective indicates whether a diagnostic is confirmatory (it reproduces a quantity SHIELD directly minimizes), evidentiary (it is related to, but distinct from, an optimized term), or independent (no SHIELD term acts on the quantity it measures).

Diagnostic	Type	Level	ML fit required	Channel probed	Link to SHIELD’s objective
t-SNE visualization	Qualitative	Global	No	Holistic (H, M, D)	Independent
Scaled $L(\pi)$	Quantitative	Global	No	Primarily M	Confirmatory for M ; evidentiary for H, D
kNN purity	Quantitative	Local	No	Primarily M	Independent
α -shape IoU	Quantitative	Global	No	Primarily D	Evidentiary for D
Downstream performance	Quantitative	Global	Yes	Holistic (H, M, D)	Independent

S2.1 Visual separation in the embedded feature space

As a qualitative complement to the quantitative diagnostics below, a shared t-SNE embedding of the input feature space is fitted once per dataset, so that every evaluated split can be plotted on the identical coordinate system for direct visual comparison. This provides a check that can catch failure modes that a scalar metric can mask, such as a split that scores favorably in aggregate but corresponds to a visually implausible or degenerate fold. Because embeddings of this kind are stochastic and do not preserve global distances, apparent separation on the embedding is suggestive rather than quantitative, and is interpreted only alongside the metrics that follow.

S2.2 Scaled cross-fold similarity

Adapted from DataSAIL,⁵ the scaled cross-fold similarity $L(\pi)$ measures the total pairwise similarity weight cut by a partition π on a graph with edge set E and similarity weights w_{uv} :

$$L(\pi) = \frac{\sum_{(u,v) \in E} w_{uv} \mathbf{1}[\pi(u) \neq \pi(v)]}{\sum_{(u,v) \in E} w_{uv}} \quad (\text{S1})$$

so that $L(\pi) \in [0, 1]$. Lower values indicate that highly similar samples remain co-located within a fold and higher values indicate that similarity is broken more extensively across the partition boundary. DataSAIL’s original formulation weights each pair by its similarity and by the cardinality of the clusters the two members represent; the graph-based form used here folds both factors into a single edge weight w_{uv} , consistent with the affinity matrices SHIELD’s own M term operates on. Because $L(\pi)$ collapses a partition to a single aggregate scalar, it is sensitive to the similarity graph used for evaluation. In particular, partitions that fragment folds can reduce the total cut weight, and hence $L(\pi)$, without corresponding to physically meaningful separation.

S2.3 k -nearest-neighbor purity

Let $\text{Te} = \{i : \text{fold}_i = 1\}$ denote the held-out test observations, with cardinality $|\text{Te}|$, and let $\text{Tr} = \{j : \text{fold}_j = 0\}$ denote the training and validation observations. For $i \in \text{Te}$, let $\mathcal{N}_k(i)$ denote its k -nearest-neighbors (kNN) over the full dataset, excluding i itself. Adapting the neighborhood-purity concept used to assess single-cell multi-omics

integration⁶ to the present setting, kNN purity is defined as:

$$\text{kNN Purity} = \frac{1}{|\text{Te}|} \sum_{i \in \text{Te}} \frac{|\mathcal{N}_k(i) \cap \text{Tr}|}{k} \quad (\text{S2})$$

where kNN purity is evaluated over the held-out observations only, each queried against neighbors drawn from the full dataset. Lower purity indicates local isolation of the held-out region, with the neighborhood of a held-out point dominated by other held-out observations rather than training observations. Higher purity indicates that training points intrude into the immediate neighborhood of held-out points, reflecting greater local overlap between the two folds. Purity is an intuitive local diagnostic that complements $L(\pi)$'s global cut-based view with a per-observation neighborhood property, and it is sensitive to failure modes that a single aggregate cut sum can miss.

S2.4 α -shape intersection-over-union

Let $A_\alpha(\text{Tr})$ and $A_\alpha(\text{Te})$ denote the α -shapes enclosing the training and held-out point clouds computed on a shared two-dimensional PCA projection of the feature space. The concave-hull construction underlying an α -shape is computationally tractable and visually interpretable only in low dimensions. The parameter α controls the shape's concavity, interpolating from the convex hull at $\alpha = 0$ toward increasingly detailed, and eventually fragmented, boundaries as α grows. The overlap between the two folds is then quantified as the standard intersection-over-union (IoU) of the two enclosed areas:

$$\text{IoU}_\alpha = \frac{|A_\alpha(\text{Tr}) \cap A_\alpha(\text{Te})|}{|A_\alpha(\text{Tr}) \cup A_\alpha(\text{Te})|} \quad (\text{S3})$$

with $|\cdot|$ denoting area. A lower IoU_α indicates weaker geometric overlap between the train and test regions, which is generally desirable for OOD evaluation. We report the metric over a range α values to ensure that the reported IoU is not an artifact of an arbitrarily chosen α .

Since $L(\pi)$ and kNN purity are both computed from discrete relationships between individual observations, both are sensitive to how densely and how evenly a fold happens to be sampled. Instead, IoU_α collapses each fold to the region of feature space it occupies and asks whether those two regions overlap as wholes, independently of how many points populate them or how those points pairwise relate. This makes IoU_α the diagnostic most directly aligned with the distributional term D , whether folds occupy the same region of covariate space, complementing $L(\pi)$ and kNN purity's alignment with the graph-cut term M . A split can leave a held-out fold with low average similarity to training under $L(\pi)$, or with a locally isolated neighborhood under kNN purity, while its overall geometric footprint still sits inside the same broad region of feature space the training fold already spans, the same concern raised for diversity-picking splitters in Section S1.11. IoU_α is directly sensitive to that failure mode because it is defined on the occupied regions themselves rather than on pairwise or nearest-neighbor structure. The IoU_α diagnostic is reported only for the LPBF and matbench_glass benchmarks, both of which are naturally organized into discrete and densely sampled clusters in feature space.

S2.5 Downstream regressor performance

The diagnostics above characterize the structure of a partition directly, but the quantity that ultimately matters is whether that structure translates into a more realistic, often harder extrapolation task. For regression benchmarks, this is assessed with the coefficient of determination on the held-out fold:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (\text{S4})$$

its generalization gap $\Delta R^2 = R^2_{\text{train}} - R^2_{\text{test}}$,⁷ the root-mean-square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2} \quad (\text{S5})$$

and the proposed Discovery Normalized Squared Error (DNSE) metric. DNSE quantifies how strongly the test error grows relative to the training error for a given split, and for regression, it is defined as:

$$\text{DNSE} = \frac{\text{RMSE}_{\text{test}}}{\text{RMSE}_{\text{train}}} \times \text{RMSE}_{\text{test}} = \frac{\text{RMSE}_{\text{test}}^2}{\text{RMSE}_{\text{train}}} \quad (\text{S6})$$

where a smaller training error inflates DNSE for the same test error, unlike the R^2 gap, which is the difference between two values of the same metric rather than a ratio-weighted composite. Because the denominator is determined by the training fold and the trained model, rather than by the leakage between folds, two splits with the same test error can score different DNSE values. DNSE therefore characterizes how a given model behaves under a given split rather than measuring the split's leakage independently. To avoid an artificially inflated value from a near-zero denominator,

we did not report DNSE when $\text{RMSE}_{\text{train}} < 10^{-3}$.

S2.6 Downstream classifier performance

For classification benchmarks, a similar methodology is applied using the area under the receiver operating characteristic curve (ROC AUC), its generalization gap $\Delta \text{AUC} = \text{AUC}_{\text{train}} - \text{AUC}_{\text{test}}$, the Brier score:

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - o_i)^2 \quad (\text{S7})$$

with $o_i \in \{0, 1\}$ the true label and $\hat{p}_i$ the predicted probability, and the standard confusion-matrix quantities:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Sensitivity} = \frac{TP}{TP + FN} \quad (\text{S8})$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad F_1 = \frac{2 \text{ Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (\text{S9})$$

ROC AUC gives a threshold-independent measure of ranking quality, equal to the probability that a randomly chosen positive observation is scored above a randomly chosen negative one; the Brier score is a proper scoring rule that captures probability calibration in addition to discrimination; and precision, sensitivity, accuracy, and F1 score add the standard confusion-matrix-based view, which is particularly informative given the class imbalance present in several of the classification benchmarks in this study. Since the Brier score is already a mean-squared quantity, unlike the square-rooted RMSE, the classification form of DNSE places the training Brier score under a square root to preserve the same mathematical form as the regression variant:

$$\text{DNSE} = \frac{\text{Brier}_{\text{test}}}{\sqrt{\text{Brier}_{\text{train}}}} \quad (\text{S10})$$

As with the regression form, DNSE is omitted when $\text{Brier}_{\text{train}} < 10^{-3}$. This often indicates severe overfitting, only observed for the kNN model on the matbench_glass benchmark in this study.

S3. Benchmark datasets

SHIELD was applied to eight benchmark datasets that span quantum chemistry (qm8, qm9), toxicology (tox21), physicochemistry (lipophilicity), bioactivity (HIV), materials science (matbench_glass), additive manufacturing (LPBF), and structural biology (PDBbind). The datasets were configured to show the distinct capabilities of SHIELD, as listed in Table S2. In the following sections, we provide a brief background for each dataset, the preprocessing step to build SHIELD relational structure, and present related results.

Table S2 | SHIELD configurations applied to each benchmark.

Dataset	N	Task	Target	(α, β, γ)	D mode	Stratification (μ)	Coarsened?
LPBF	1,579	Regression	relative density (%)	1, 1, 1	shift	none	—
matbench_glass	5,680	Classification	glass-forming (binary)	.5, .2, .7	match	quantile (0.5)	—
PDBbind	19,087	Regression	binding affinity (pKd)	1, 1, 1	match	quantile (0.5)	—
tox21	7,823	Classification	12 assays (binary)	1, 1, 1	shift	none	—
lipophilicity	4,200	Regression	Solubility (logD)	1, .2, 3	match	quantile (0.5)	—
qm8	21,786	Regression	16 spectral properties	1, 1, 1	shift	none	Yes
qm9	133,885	Regression	12 quantum properties	1, 1, 1	shift	none	Yes
HIV	41,127	Classification	activity (binary)	1, .2, 3	shift	none	Yes

S4. Benchmark 1: LPBF dataset

Laser powder bed fusion (LPBF) is an additive manufacturing technology that builds a metal part by melting successive powder layers with a laser beam. The LPBF dataset,⁸ retrieved from Harvard Dataverse,⁹ maps material and process parameters to printed part density across 1,579 builds drawn from 6 alloys and 45 printer models. The target variable is the relative density (RD (%)), defined as the fraction of the nominal nonporous solid density achieved. In this setting, an extrapolative model should predict RD (%) for an alloy it has not seen, built on a machine it has not seen. Alloy family and machine identity are therefore two independent generalization axes, which is why the LPBF benchmark supports a two-dimensional split setting suitable for cold splits.

The relational inputs are constructed to be complementary and reflect the physics governing alloy densification. The hierarchy term H is defined at two levels: the lower level groups observations by alloy identity, with family similarity set to the cosine between the alloys’ elemental composition vectors; the upper level groups alloys by

solidification class (face-centered cubic, body-centered tetragonal, or a two-phase hexagonal/body-centered structure). This hierarchy is physically meaningful because alloys within the same solidification class share similar solidification behavior, while similarities in composition give rise to related thermophysical properties. Together, these factors influence melt pool stability and porosity formation. Respecting this hierarchy during partitioning limits the overlap of shared material-laser interaction across folds, thereby producing a more faithful evaluation of generalization to unseen alloy systems. The manifold affinity M is a cosine kNN graph built on 14 process parameters and physical descriptors, including the volumetric, linear, and areal energy densities, a normalized energy feature, the Péclet number, and a Marangoni proxy, all established predictors of melt pool dimensions and porosity. Alloy-constant thermophysical and composition columns are excluded from M so that it encodes LPBF process neighborhood rather than reencoding the alloy identity already carried by term H . The distributional term D uses a radial-basis-function (RBF) kernel over the full scaled feature space in shift mode, driving the per-fold feature distributions apart; since the process parameter distribution is the operative covariate for RD (%), separating it constitutes a direct test of process extrapolation.

SHIELD was configured with two folds at an 80/20 ratio, unit weights on all active terms, D in shift mode, and no stratification, since the intended structure is the domain-defined cold split rather than target balance. The alloy is designated the left entity and the printer the right entity. Four two-entity regimes were run under the full $H/M/D$ objective: warm, cold-alloy, cold-machine, and cold-both. Three single-term ablations, restricted to the warm regime and denoted H -SHIELD, M -SHIELD, and D -SHIELD, isolate the individual contribution of each term; the full joint objective is denoted HMD -SHIELD. Four baseline splits, random, stratified, k -means, and agglomerative, serve as points of reference.

The shared t-SNE embedding of the input features (Fig. S1 a-k) places every split on the same manifold, allowing direct visual comparison. Under the random and stratified baselines the held-out points interleave with the training points throughout the embedding, and kNN purity stays close to its baseline level, meaning most of a held-out point's neighbors remain in the training fold. The unsupervised clustering baselines, namely k -means and agglomerative, improve the separation of held-out points because they partition on generic feature proximity without reference to the physical structure described above. Among single-term ablations, H -SHIELD achieves markedly stronger separation than D -SHIELD and M -SHIELD, consistent with the six alloy families offering few degrees of freedom for the hierarchy term (Fig. S1 l) relative to the continuous process-parameter neighborhoods that M and D act on (Fig. S1 m). The α -shape overlap between the train and test point clouds (Fig. S2 c) corroborates this ranking across the tested range of α , falling from a near-complete overlap for the interpolative baselines to a near-zero overlap for the SHIELD-derived splits. Disagreement among diagnostics indicates that no single metric is sufficient to establish

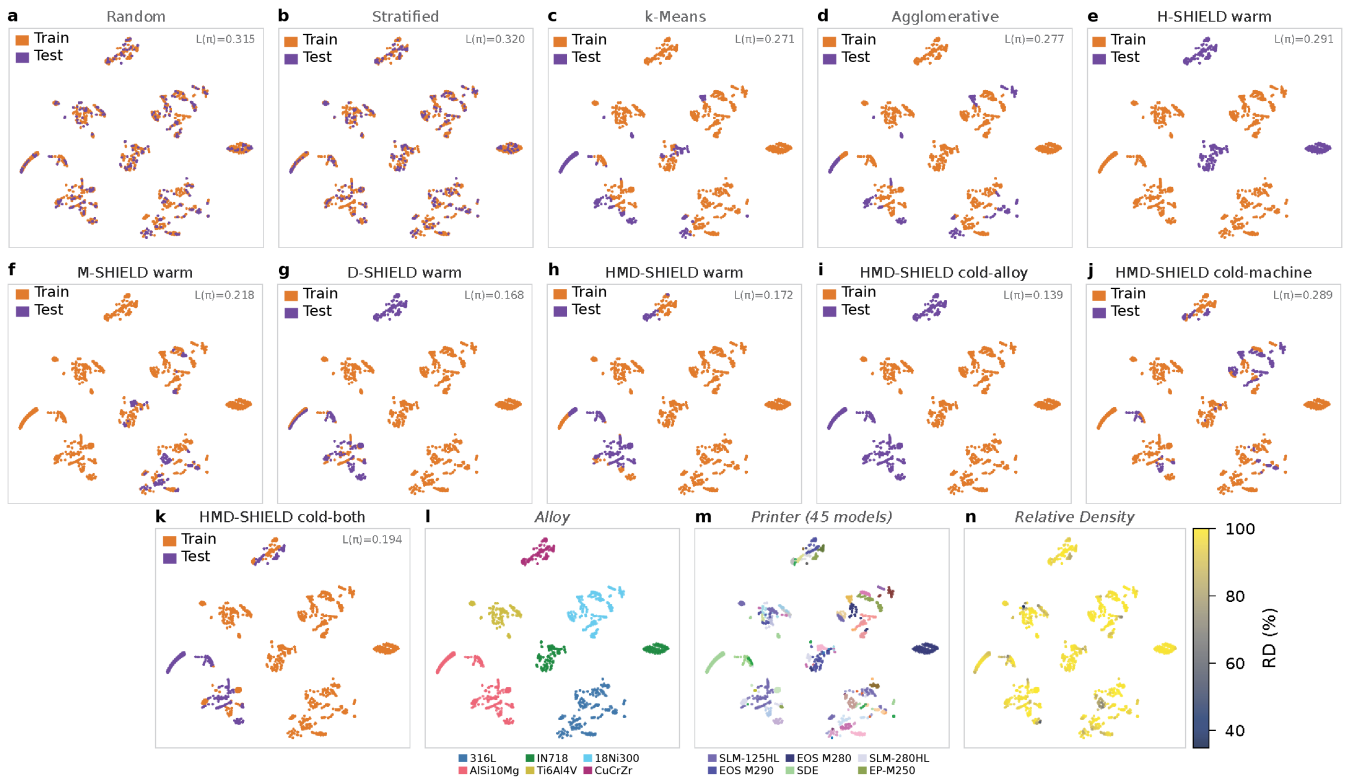


Fig. S1 | LPBF split atlas. Shared t-SNE embedding of the LPBF input features under every split evaluated. Baseline splits include random, stratified, k -means, and agglomerative (a-d). Single-term SHIELD ablations restricted to the warm regime (e-g). Full joint objective HMD -SHIELD under the warm, cold-alloy, cold-machine, and cold-both regimes (h-k). The same t-SNE embedding colored by alloy identity (6 alloys), printer identity (45 models, listing top 6 models labeled in the legend), and relative part density RD (%), respectively (l-n).

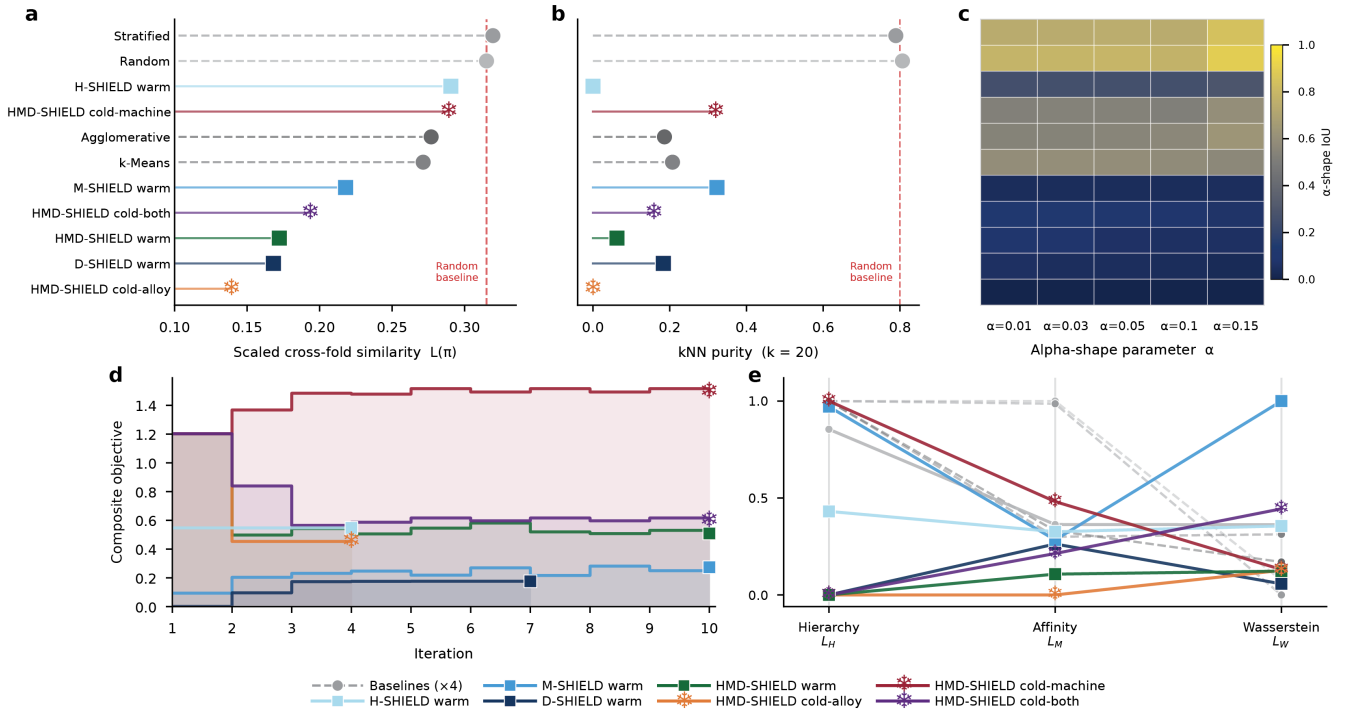


Fig. S2 | LPBF split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 20$) for every split, with the random-split as a reference (a,b). Overlap between the α -shapes of the training and held-out point clouds, across five values of the shape parameter α (c). SHIELD’s composite objective value across optimization iterations for every SHIELD configuration (d). Contributions of the hierarchy (L_H), affinity (L_M), and distributional (L_D) terms to the composite objective at convergence, for every SHIELD configuration and calculated as channel leakage for the baselines (e).

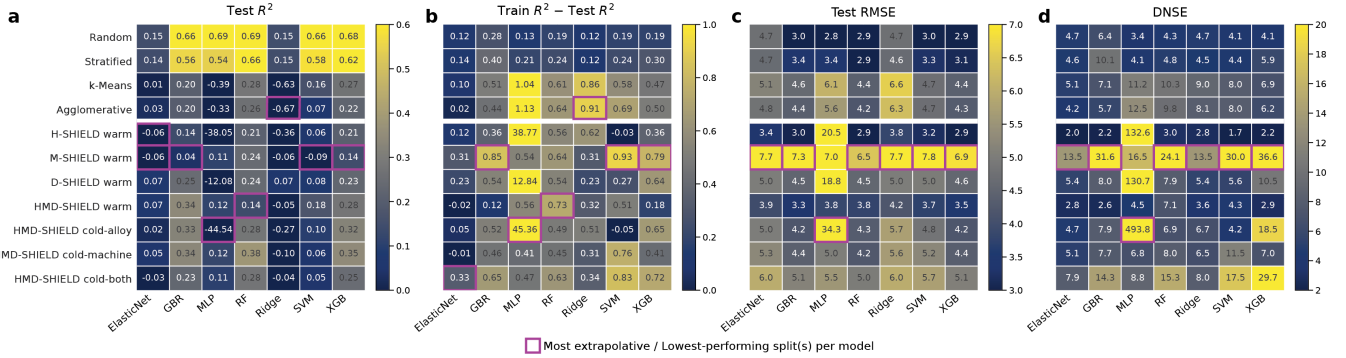


Fig. S3 | LPBF downstream model performance. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the eleven splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

geometric separation of the occupied feature space. For example, *H*-SHIELD achieves a high $L(\pi)$ despite exhibiting the lowest kNN purity and almost no overlap according to the α IoU metric. Likewise, kNN purity suggests a stronger fold separation than $L(\pi)$ and α IoU.

The two cold axes are not equally severe tests of extrapolation. Holding out an entire alloy (*HMD*-SHIELD cold-alloy, Fig. S1 i) achieves the lowest $L(\pi)$ of any split evaluated, where the terms M and D guide the selection of the held-out points into a further region on the embedding compared to *H*-SHIELD alone. Holding out a printer instead (*HMD*-SHIELD cold-machine, Fig. S1 j) is comparatively weak for OOD deployment purposes. This asymmetry is consistent with the physical rationale established above. Alloys in different families occupy genuinely distinct regions of the composition and thermophysical space, so removing one alloy removes a coherent process regime from training, whereas the 45 printers are not organized under any hierarchy and can share overlapping ranges of the process parameters that M and D act on, for example when two machines are operated at comparable laser power and beam diameter. Holding out one machine therefore need not remove the model from a process-parameter region it has already seen through other machines. The tuned regressors reflect this trend, with the MLP model achieving a test R^2 of approximately 0.69 under the random split, compared with 0.12 under the *HMD*-SHIELD warm- and cold-machine splits, and -44.5 under the cold-alloy split (Fig. S3a).

Beyond R^2 , the generalization gap (ΔR^2 , Fig. S3 b) and the error-based metrics (RMSE and DNSE, Fig. S3 c,d) point to the same conclusion from an independent angle. The tree-based, kernel, and linear models degrade gradually as the split moves from random to the SHIELD-derived conditions. The multilayer perceptron (MLP)

behaves qualitatively differently. Its test R^2 collapses to strongly negative values under H -SHIELD warm (-38.05), D -SHIELD warm (-12.08), and especially HMD -SHIELD cold-alloy (-44.54), accompanied by RMSE and DNSE values an order of magnitude above every other model-split combination. Under cold-alloy its test RMSE reaches 34.3 against a DNSE of 493.8, which by definition (Eq. S6) implies a training RMSE of order 2, roughly an order of magnitude below its test error. Because this collapse coincides precisely with the splits that most strongly separate the process manifold or the alloy hierarchy, rather than occurring uniformly, it indicates that the MLP is extrapolating far outside its training support in these regimes, a failure mode that tree ensembles are protected against since their predictions are bounded by the range of training-leaf values. Remarkably, despite producing only modest separation in the t-SNE embedding and scoring high on $L(\pi)$ and kNN purity, M -SHIELD warm yields the largest degradation in RMSE and DNSE across all regression models except the MLP. This discrepancy indicates that qualitative separation in the t-SNE embedding and local diagnostics did not reliably predict model degradation upon deployment, whereas the proposed α IoU metric correctly identified the split as challenging. By jointly encoding alloy thermophysical properties and process parameters, term M captures a rich axis of extrapolation than any other single SHIELD term, despite its weaker apparent separation under individual diagnostics.

S5. Benchmark 2: matbench_glass

Matbench_glass is a materials benchmark of 5,680 alloy compositions labeled for glass-forming ability,¹⁰ a binary indicator of whether the composition vitrifies rather than crystallizing upon rapid cooling. The only input is the composition, and the classes are imbalanced. Unlike the LPBF benchmark, where alloy and machine define two independent generalization axes, matbench_glass presents a single generalization axis, the composition itself, thus, an extrapolative model should predict glass-forming ability for compositions unlike those it has already seen, so the relevant leakage to control is compositional near-duplication and manifold proximity.

The hierarchy H is a single level of 60 k -means clusters on these vectors; compositions within the same cluster are compositionally similar, so keeping each cluster intact reduces the likelihood that alloys that share dominant elements are split across folds. The affinity M is a cosine kNN graph ($k = 20$) on the raw element-fraction vectors, with within-cluster edges set to zero so that M carries only cross-cluster similarity and stays complementary to the intra-cluster cohesion that H already enforces. Paired with quantile stratification, the distributional term D operates in match mode, where the transport cost is computed on Magpie composition descriptors and the optimization encourages each fold's composition distribution to match a uniform target marginal. Match mode is preferred over shift mode because the label imbalance makes representative folds more informative than maximally separated ones, while quantile stratification helps prevent degenerate folds.

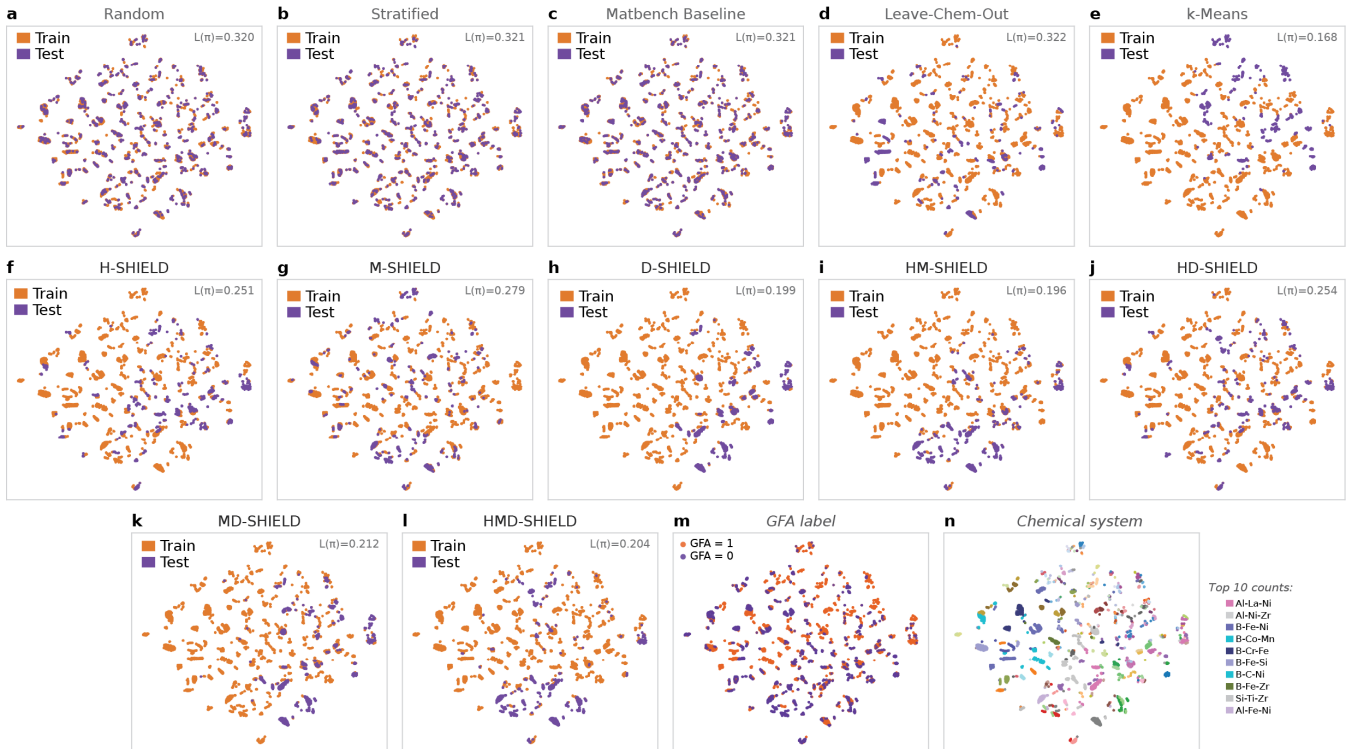


Fig. S4 | Matbench_glass split atlas. Shared t-SNE embedding of the composition features under every split evaluated. Baseline splits include random, stratified, the Matbench default, LCSO, and k -means (a-e). Single-term SHIELD ablations (f-h), pairwise SHIELD configurations (i-k), and the joint objective HMD -SHIELD (l). The same t-SNE embedding colored by glass-forming-ability label ($GFA = 1$, vitrifies; $GFA = 0$, crystallizes), and colored by the chemical system, with the ten most frequent systems labeled in the legend (m,n).

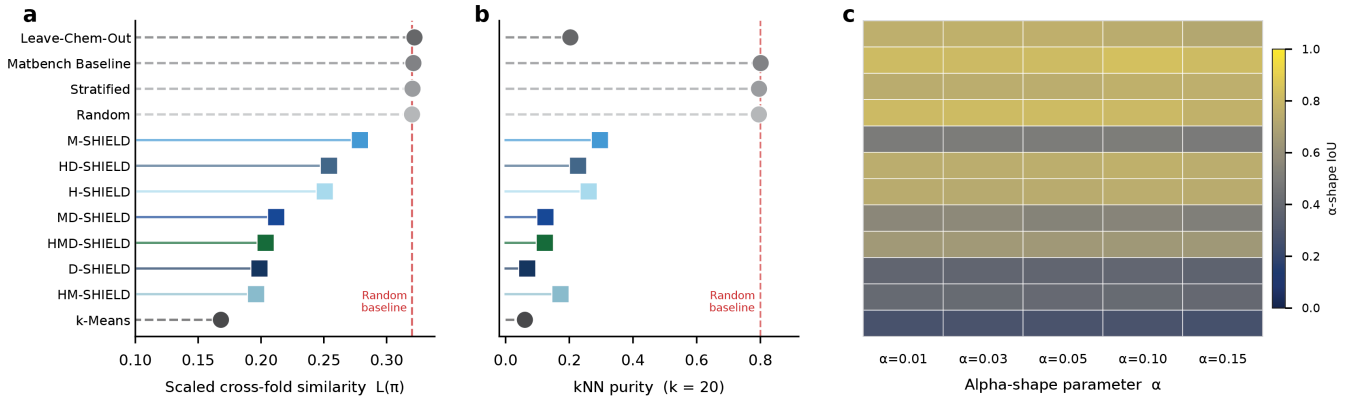


Fig. S5 | Matbench_glass split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 20$) for every split, with the random-split as a reference (a,b). Overlap between the α -shapes of the training and held-out point clouds, across five values of the shape parameter α (c).

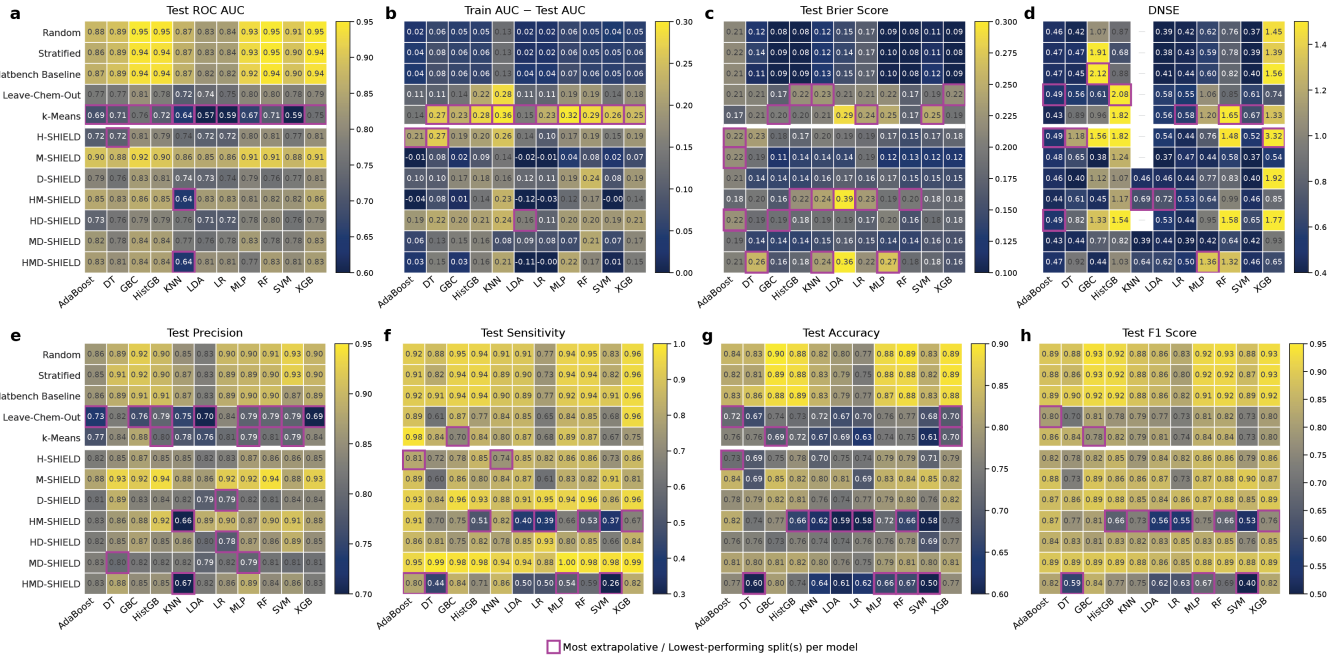


Fig. S6 | Matbench_glass downstream model performance. ML performance of eleven classification models, each independently hyperparameter-tuned per split, across the twelve splits evaluated. The reported metrics include test ROC AUC, generalization gap, Brier score, DNSE, precision, sensitivity, accuracy, and F1 score (a-h). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

SHIELD ran with two folds at an 80/20 ratio and weights $(\alpha, \beta, \gamma, \mu) = (0.5, 0.2, 0.7, 0.5)$, informed by the relational-term reasoning above. Because the input space here is a single composition vector rather than a cross-product of entity types, the meaningful ablation is over which of the three relational terms are active rather than over warm and cold regimes. The three single-term configurations (H -SHIELD, M -SHIELD, D -SHIELD), the three pairwise configurations (HM -SHIELD, HD -SHIELD, MD -SHIELD), and the joint objective (HMD -SHIELD) were therefore all evaluated. The five baselines used as reference splits include random, class-stratified, Matbench default split, a leave-chemistry-system-out split, and k -means.

Every baseline split, except for k -means, lands within a narrow similarity leakage band around ($L(\pi) = 0.320$) as in Fig. S5 a-d), confirming that neither label stratification nor a naive LCSO meaningfully reduces information leakage. The one exception is the k -means baseline as it reaches the single lowest $L(\pi)$ of any split evaluated (0.168), which may be attributed to the size of the dataset and the naturally clustered/discrete structure of alloy chemical systems in composition space (Fig. S5 e,n). Among the single-term SHIELD runs, D -SHIELD is strongest ($L(\pi) = 0.199$), H -SHIELD intermediate (0.251), and M -SHIELD weakest (0.279). This ordering follows directly from how M was constructed with within-cluster edges zeroed to keep it complementary to H ; M alone carries no information about the intra-cluster near-duplicates that term H is built to contain. This is further supported by the reduction in $L(\pi)$ in the cases of HM -SHIELD, MD -SHIELD, and HMD -SHIELD. Upon tuning classifiers, k -means, HM -SHIELD, and HMD -SHIELD splits reflect the worst performance across a range of metrics. While k -means scores the lowest on test ROC AUC (Fig. S6 a), HM -SHIELD and HMD -SHIELD have comparable or worse performance across other

metrics, such as Brier score, DNSE, sensitivity, accuracy, and F1 scores (Fig. S6 c,d,f-h). This apparent discrepancy is expected because the metrics probe different aspects of model behavior. ROC AUC measures the ranking of predicted scores independently of the decision threshold, whereas Accuracy, F1 score, DNSE, and Brier score also depend on the confidence and thresholding of those scores. Different extrapolation splits therefore need not produce the same ordering across these metrics. Unlike SHIELD, which enforces a per-class balance band, unconstrained k -means clustering provides no mechanism to preserve the imbalanced glass-forming label across folds. Instead, it exploits the same discrete structure to fragment that class arbitrarily, yielding a deceptively low $L(\pi)$ that reflects fold fragmentation rather than a controlled and physically meaningful separation. By contrast, SHIELD produces the intended controlled separation, and the term-by-term trends above indicate a structured removal of bias rather than a simple degradation in ML model performance.

S6. Benchmark 3: PDBbind

PDBbind pairs a protein with a bound ligand and records their binding affinity (pKd) as the target for 19,087 complexes spanning 12,452 unique proteins and 15,446 unique ligands.^{3,11} Since pKd affinity is governed jointly by the protein pocket and the ligand pharmacophore, an extrapolative model should be tested along both axes at once, here unseen protein families and unseen ligand chemotypes, rather than only unseen exact sequences or scaffolds. PDBbind is therefore used to apply cold-entity splitting along these two paired axes.

Proteins are represented by hashed 3-mer sequence features and ligands by ECFP4 fingerprints. The hierarchy H has four levels along two axes: a protein axis (protein type, then sequence sub-cluster) and a ligand axis (Murcko scaffold, then ECFP sub-cluster), with the fine levels weighted above the coarse levels. The affinity M is a sparse bilinear graph whose edge weight is the product of protein similarity and ligand similarity, kNN sparsified; two complexes are neighbors only when both their proteins and their ligands are independently similar, a stricter and sparser criterion than either axis alone would give. The distributional term D runs in match mode with quantile stratification, ensuring that the per-fold pKd ranges are representative.

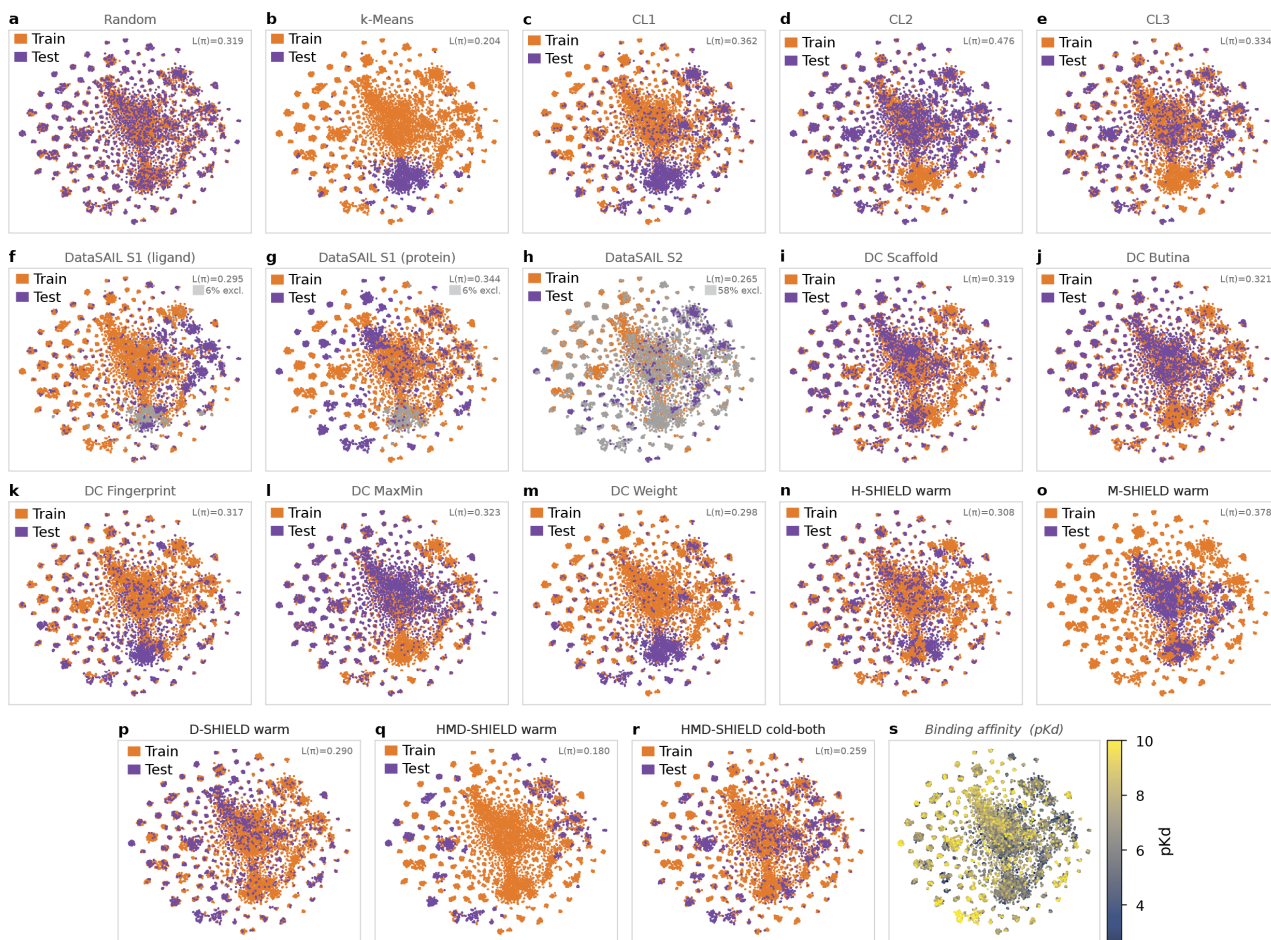


Fig. S7 | PDBbind split atlas. Shared t-SNE embedding of the complex features under every split evaluated. Baseline splits include random, k -means, quality-based splits from the LP-PDBBind CL1-CL3 flags, DataSAIL cold splits, and five DeepChem splitters (a-m). Single-term SHIELD ablations (n-p). Full joint objective HMD -SHIELD under the warm and cold-both regimes (q,r). Complexes excluded under DataSAIL's own cold-split rule are shown in gray and reported as a percentage of the dataset (f-h). The same t-SNE embedding colored by pKd binding affinity (s).

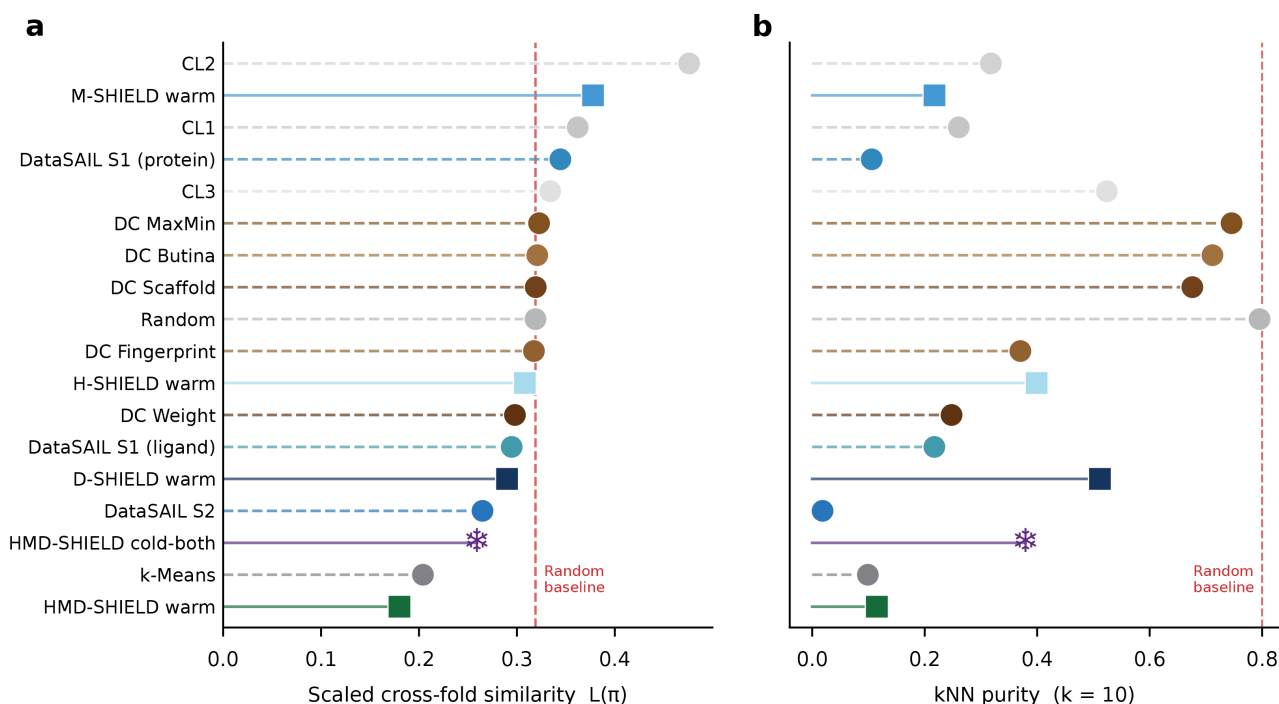


Fig. S8 | PDBbind split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 10$) for every split, with the random-split as a reference (a,b).

SHIELD ran with two folds at 80/20 ratio, the protein sequence sub-family as the left entity and the ligand ECFP chemotype as the right entity. SHIELD was run in warm and cold-both regimes, alongside the three single-term warm ablations. The baseline splitters include k -means, three predefined binary splits derived from the LP-PDBBind CL1, CL2, and CL3 flags, five DeepChem splitters (scaffold, Butina, fingerprint, MaxMin, and weight-based), and DataSAIL in its single-entity (protein, ligand) and two-entity cold split.

Across all diagnostics, baseline splitters separate the folds less aggressively than *HMD-SHIELD* warm split, with the exception of DataSAIL. For instance, k -means baseline is the most notable baseline with the lowest $L(\pi)$ of 0.204 and well separated cluster on t-SNE embedding (Fig. S8 b); this is plausibly explained by the discrete structure of the embedding, in which protein families and ligand chemotypes form well-separated clusters rather than a continuous manifold. Consequently, a generic clustering-based split recovers much of the same structural organization that SHIELD targets explicitly. DataSAIL’s single-entity cold modes (S1, Fig. S8 f,g) cluster well compared to other baselines, consistent with each optimizing similarity along only one axis of a problem whose leakage is jointly governed by the protein pocket and the ligand pharmacophore. DataSAIL’s two-entity cold split (S2, Fig. S8 h), which combines both axes explicitly, generates a similar $L(\pi)$ to that of *HMD-SHIELD* cold-both split, as well as a near-zero kNN purity. However, the reported kNN purity and downstream ML performance for DataSAIL’s S2 split are biased since 58% of the dataset is discarded as inter-fold overlap, far more than the 6% discarded under either of its S1 splits. Discarding over half of the available data is a significant practical limitation of DataSAIL and other two-entity splitters, particularly because constructing reliable scientific datasets is often expensive and time-intensive. *HMD-SHIELD* cold-both instead resolves inter-fold overlaps through an optimization-guided repair step performed after each iteration, allowing all samples to be retained. Despite repairing inter-fold overlaps, SHIELD preserves nearly the same structural separation ($L(\pi)$) and comparable downstream predictive performance as DataSAIL’s S2 split without excluding any observations.

Among the SHIELD single terms, *H-SHIELD* ($L(\pi) = 0.308$) outperforms *M-SHIELD* (0.378), consistent with *H*’s containment levels spanning both the protein and ligand axes, against *M*’s comparatively sparse bilinear graph, which requires similarity on both axes simultaneously. *D-SHIELD*, by contrast, gives the worst downstream performance of any split despite a moderate $L(\pi)$ of 0.290; test R^2 collapses to between -9.7 and -6.4 across every regressor, with a ΔR^2 of roughly 7–10 (Fig. S9 a,b), indicating complete failure on the held-out fold. This behavior is unexpected from the geometric diagnostics alone and indicates that moderate structural separation, as quantified by $L(\pi)$, does not necessarily translate into moderate extrapolation difficulty. A plausible explanation is that *D-match* constrains only the target pKd distribution, without explicitly preserving protein-family or ligand-chemotype structure. Consequently, folds may remain representative in the target variable while differing substantially along the structural axes on which the regressors ultimately rely. More broadly, this result illustrates that geometric diagnostics alone are insufficient to anticipate downstream extrapolation difficulty, highlighting the need for new split-quality diagnostics that better characterize extrapolation regimes.

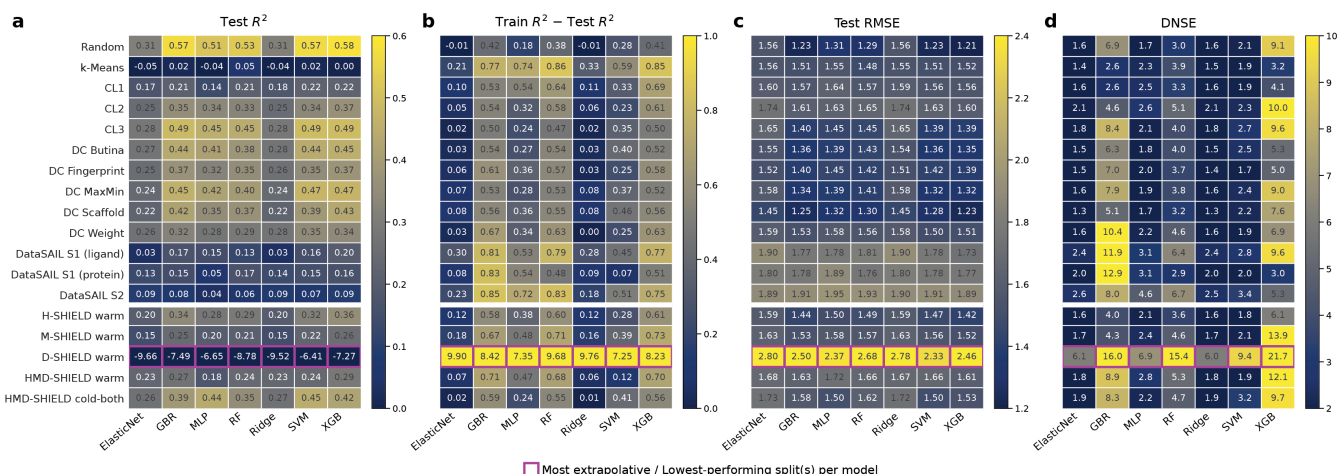


Fig. S9 | PDBbind downstream model performance. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the eighteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

Warm *HMD-SHIELD* reaches the lowest $L(\pi)$ of any split evaluated (0.180), while cold-both reaches a higher value (0.259), still below every baseline except *k*-means. This difference arises because warm optimizes assignments at the resolution of individual protein-ligand complexes, matching the granularity of both the manifold graph and the distributional objective. By contrast, cold-both requires a repair step to restore the simultaneous holdout constraints on protein sub-families and ligand chemotypes while retaining all observations. These repairs inevitably deviate from the unconstrained optimum, increasing the final cut cost. The resulting increase in $L(\pi)$ therefore reflects a limitation of the current repair strategy rather than the cold-both objective itself, suggesting that more effective repair algorithms could further improve partition quality while preserving the desired extrapolation guarantees.

S7. Benchmark 4: tox21

Tox21 screens 7,823 molecules against 12 toxicity assays spanning nuclear receptor and stress response pathways.¹¹ The labels are binary and sparse, with many molecules untested in individual assays and positive rates below 5% for several assays. Unlike *matbench_glass*'s single binary target, tox21 carries twelve simultaneous assays with no single covariate governing all of them at once, so the leakage-relevant structure here is chemical; molecules sharing a

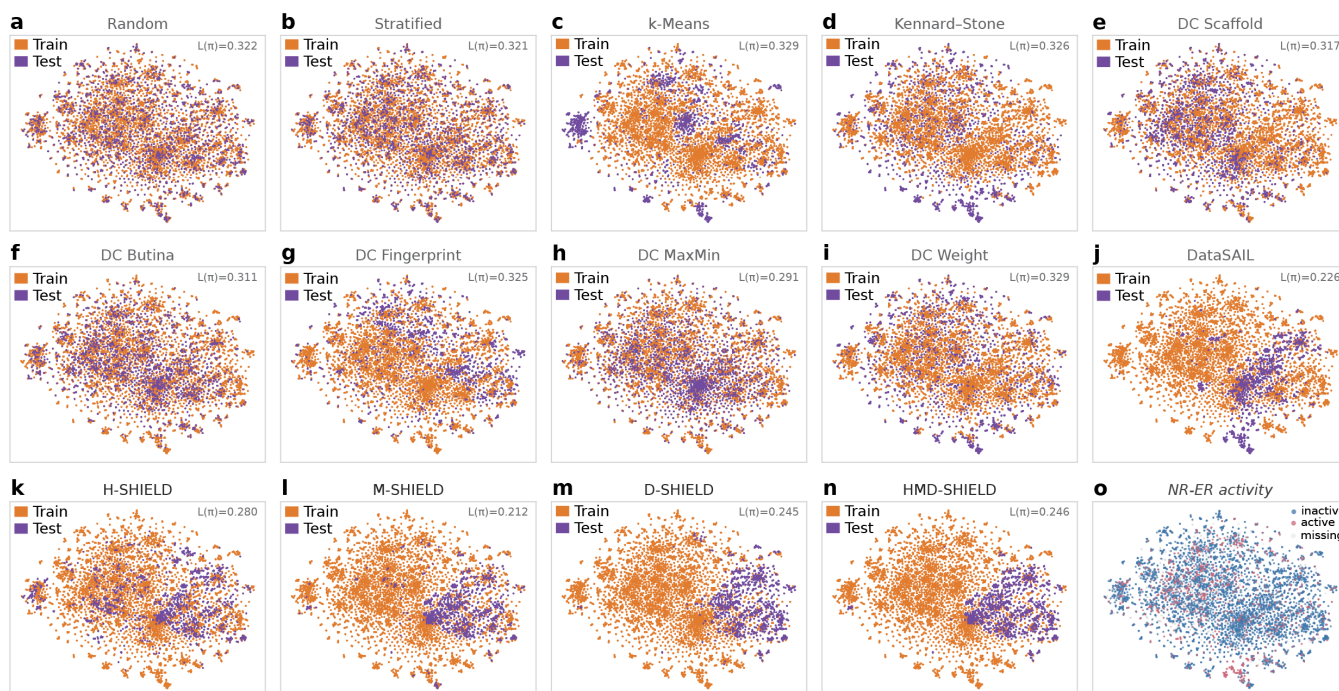


Fig. S10 | tox21 split atlas. Shared t-SNE embedding of the ECFP4 fingerprint space under every split evaluated. Baseline splits include random, stratified, *k*-means, Kennard-Stone, DataSAIL, and five DeepChem splitters (a-j). Single-term SHIELD ablations and the joint objective *HMD-SHIELD* (k-n). The same t-SNE embedding colored by activity on the *NR-ER* label (o).

scaffold or mechanism of action are expected to behave similarly across similar assays, independent of which specific assays they were tested on.

Molecules are represented by ECFP4 Morgan fingerprints. The hierarchy H has two levels drawn from a curated chemical taxonomy: a scaffold and functional-group class above a mechanism and target class, so molecules with the same pharmacophore stay together. The affinity M is a scaffold-gated kNN graph on Morgan-cosine similarity, where an edge is kept when two molecules exceed the similarity threshold and share a Bemis-Murcko scaffold, with a second tier of ring-topology-class edges connecting structurally unique molecules. The distributional term D runs in shift

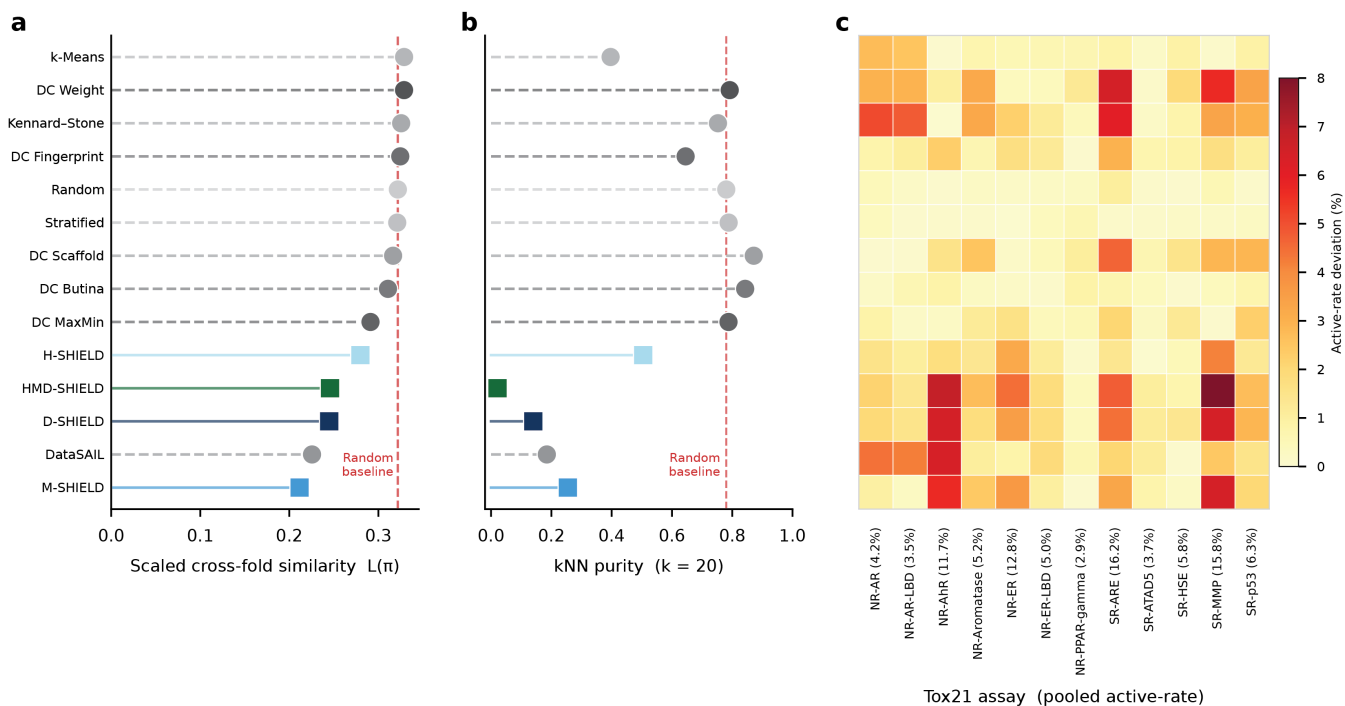


Fig. S11 | tox21 split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 20$) for every split, with the random-split as a reference (a,b). Deviation between each fold's active rate and the global active rate, in percentage points, for each of the twelve tox21 assays (c).

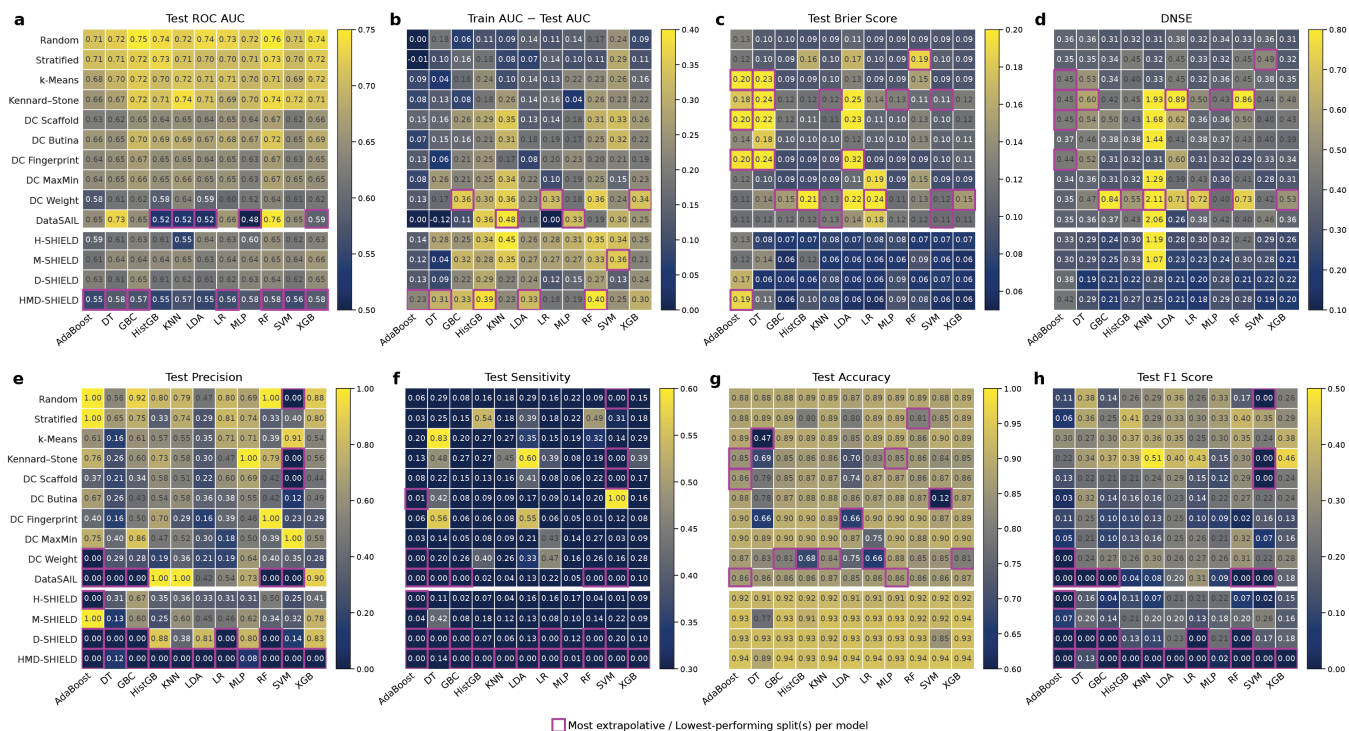


Fig. S12 | tox21 downstream model performance for NR-ER activity label. ML performance of eleven classification models, each independently hyperparameter-tuned per split, across the fourteen splits evaluated. The reported metrics include test ROC AUC, generalization gap, Brier score, DNSE, precision, sensitivity, accuracy, and F1 score (a-h). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

mode in the fingerprint-kernel space, pulling each fold’s overall chemical-space occupancy apart.

SHIELD ran with two folds at 80/20 ratio, unit term weights, and no stratification term. The three single-term SHIELD ablations and the full joint objective were evaluated against random and class-stratified reference splits and a competitor set of k -means, Kennard–Stone, five DeepChem splitters (scaffold, Butina, fingerprint, MaxMin, and weight-based), and DataSAIL.

SHIELD splits separate the folds well below every baseline, with only k -means and DataSAIL approaching comparable separation, according to the $L(\pi)$ and kNN purity results (Fig. S11 a,b). M -SHIELD and D -SHIELD produce similar hold-out folds in the embedded space since both terms operate in the same ECFP4 space (Fig. S10 l,m). H -SHIELD alone provides weaker structural separation (Fig. S10 k), but when combined with the other objectives in HMD -SHIELD (Fig. S10 n), it produces the lowest near-zero kNN purity (Fig. S11 b) and the poorest predictive performance across all classifiers and toxicity assays (Fig. S12 and Fig. S13). Among baseline splits, only DataSAIL and the DeepChem weight-based splitter approach SHIELD’s downstream performance. Despite optimizing a single structural criterion, the DeepChem weight-based splitter is nearly indistinguishable from a random split under the leakage assessment metrics, yet consistently achieves lower predictive performance across classifiers. This discrepancy highlights the limitations of existing leakage metrics and highlights the need for assessment criteria that better predict downstream generalization before model training and deployment. Although $L(\pi)$ and kNN purity generally agree on the relative ordering of the splitting strategies, notable deviations occur for the k -means split under kNN purity and the DeepChem weight-based splitter under both metrics, illustrating that different leakage diagnostics emphasize distinct aspects of fold separation. Unlike most of the other benchmarks, tox21 exhibits weaker agreement between $L(\pi)$ and kNN purity, with discrepancies for the k -means split under kNN purity and the DeepChem weight-based splitter under both metrics. On both the single primary assay (Fig. S12) and the averages across all twelve assays (Fig. S13), the HMD -SHIELD split consistently degrades minority-class performance across all tuned models, while overall accuracy remains comparatively high due to class imbalance. This indicates that HMD -SHIELD most effectively limits structural overlap between the folds, thereby requiring greater extrapolation to unseen compounds.

Tox21 additionally serves as the benchmark on which the two rounding routines available for converting a relaxed

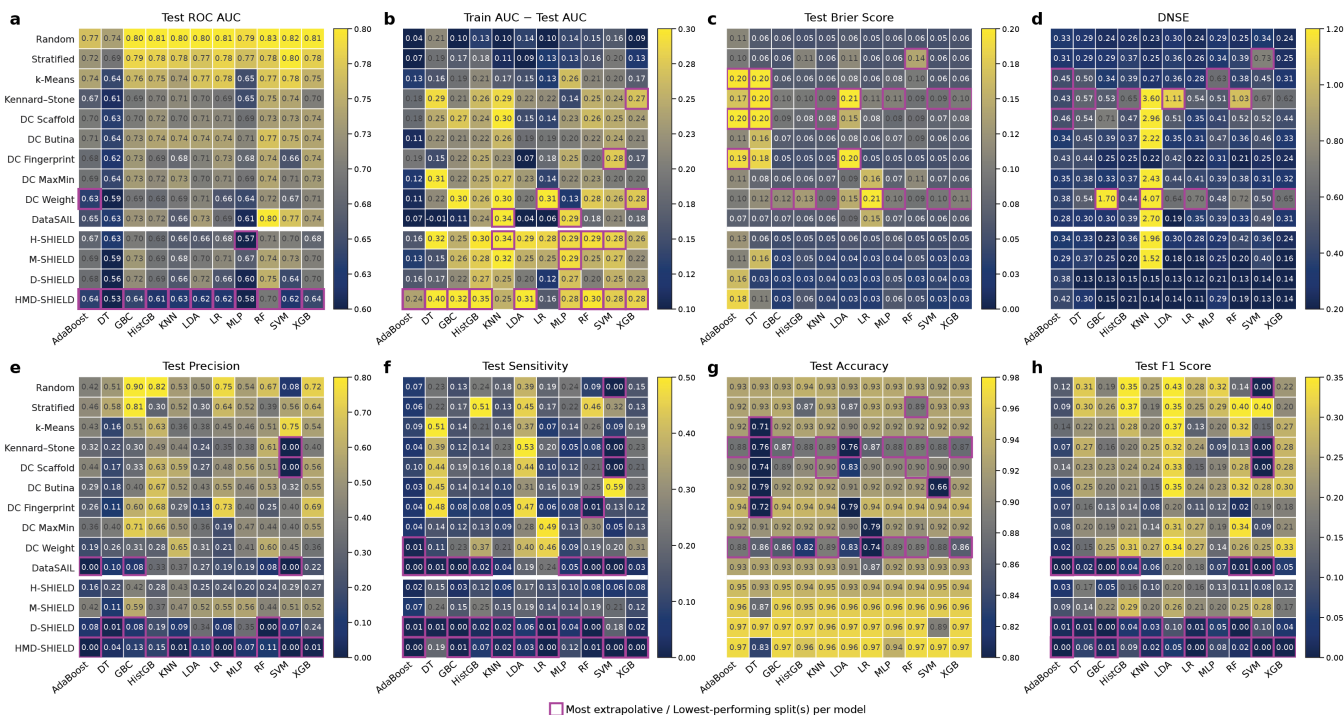


Fig. S13 | tox21 downstream model performance averaged across all 12 assays. ML performance of eleven classification models, each independently hyperparameter-tuned per split, across the fourteen splits evaluated. The reported metrics include test ROC AUC, generalization gap, Brier score, DNSE, precision, sensitivity, accuracy, and F1 score (a-h). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

Table S3 | tox21 full HMD-SHIELD split under the two rounding routines. The pipage and confidence-gap routines return the same partition, so the scaled cross-fold similarity and the three channel leakages coincide.

Rounding routine	$L(\pi)$	L_H	L_M	L_D
Pipage (dependent sampling)	0.24572	85.681	147.10	0.05302
Confidence-gap (deterministic)	0.24572	85.681	147.10	0.05302

assignment into a final partition, dependent-sampling pipage and deterministic confidence-gap, are directly compared. Table S3 reports the full *HMD*-SHIELD split results under both routines; the two routines return an identical partition, agreeing on the scaled cross-fold similarity and all three leakage channels to the reported precision.

S8. Benchmark 5: lipophilicity

The lipophilicity benchmark holds 4,200 drug-like molecules with a measured octanol-water distribution coefficient, $\log D$ at pH 7.4, spanning 2,408 Bemis-Murcko scaffolds.¹¹ $\log D$ is a continuous property that governs a compound’s solubility and membrane permeation. Unlike *tox21*’s twelve binary targets, lipophilicity offers a single continuous target, the clean setting in which to isolate the effect of the stratification form and the distributional mode independently of any other confound, which is why this benchmark carries a dedicated comparison of both alongside the standard term ablations.

The hierarchy *H* comprises two levels defined by chemical function rather than substructure counts, with a polarity and functional class above an ionization class. These levels represent the primary physicochemical determinants of $\log D$, and preserving them within a fold groups molecules whose lipophilicity arises from the same underlying mechanisms. The affinity *M* is a congeneric-series graph in which two analogs within a shared scaffold are linked when their Tanimoto similarity exceeds 0.4, capturing local structure–property smoothness where closely related molecules exhibit incremental changes in $\log D$ that can otherwise transfer across folds. The distributional term *D* operates in match mode using an RBF similarity kernel κ defined on standardized physicochemical descriptors,

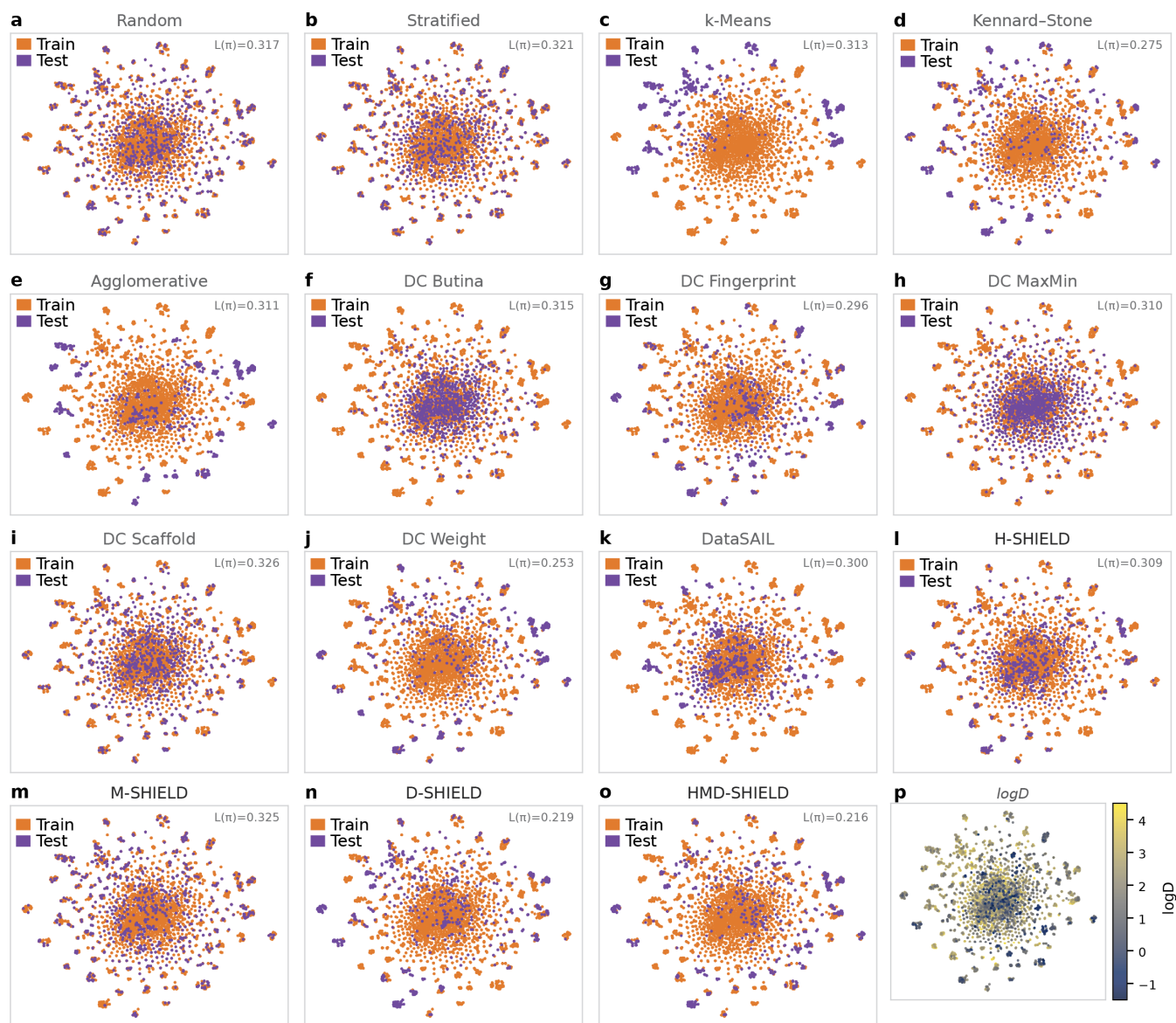


Fig. S14 | Lipophilicity split atlas. Shared t-SNE embedding of the molecular feature space under every split evaluated. Baseline splits include random, stratified, *k*-means, Kennard–Stone, agglomerative, five DeepChem splitters, and DataSAIL (a–k). Single-term SHIELD ablations and the joint objective *HMD*-SHIELD (l–o). The same t-SNE embedding colored by $\log D$ (p).

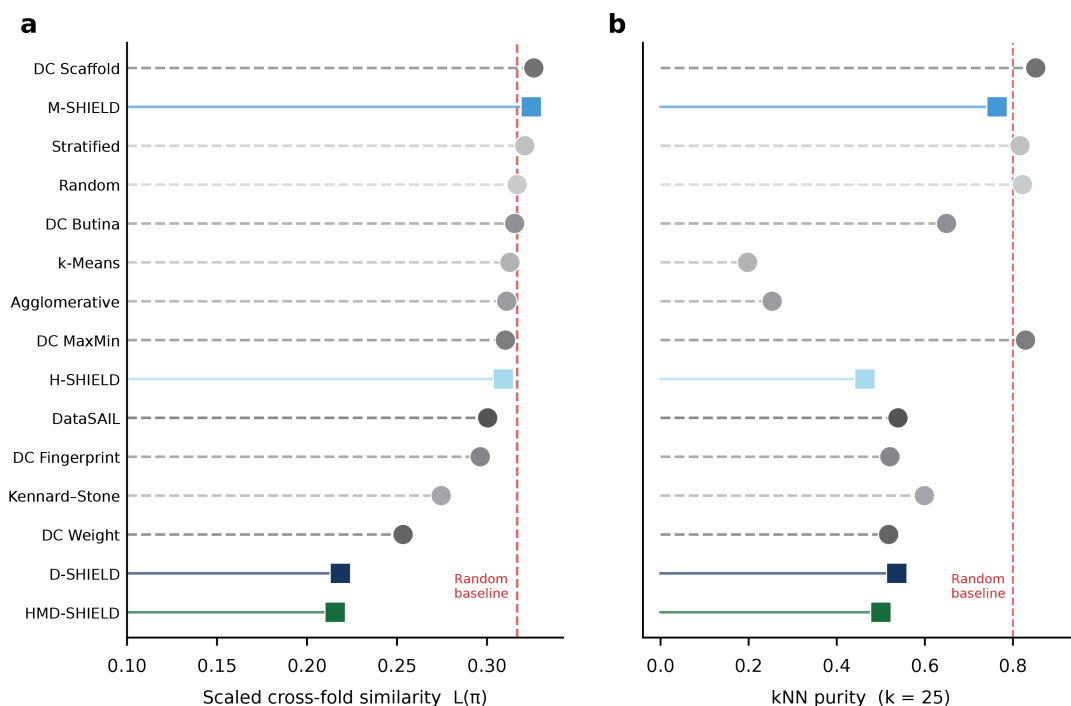


Fig. S15 | Lipophilicity split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 25$) for every split, with the random-split as a reference (a,b).

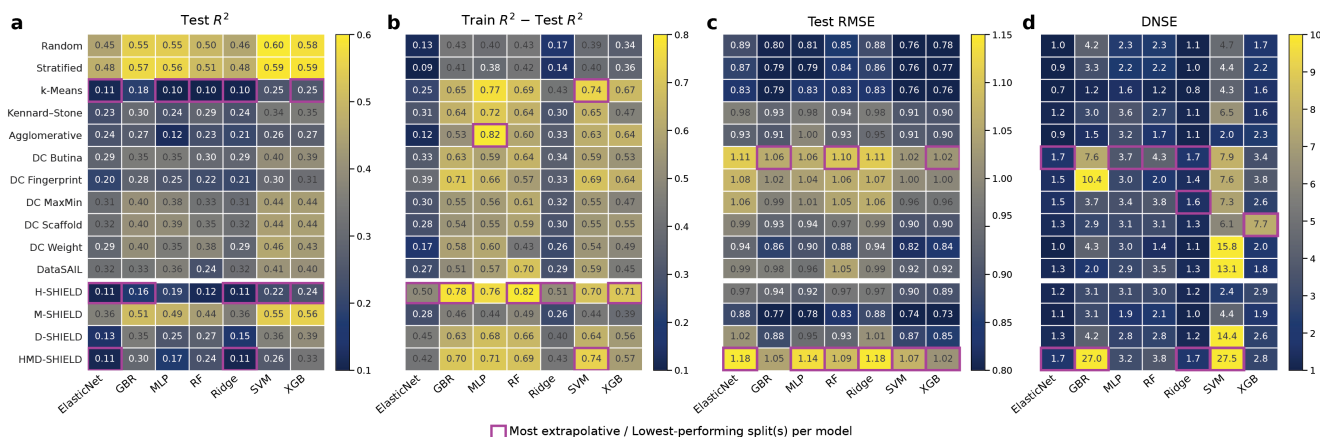


Fig. S16 | Lipophilicity downstream model performance. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the fifteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

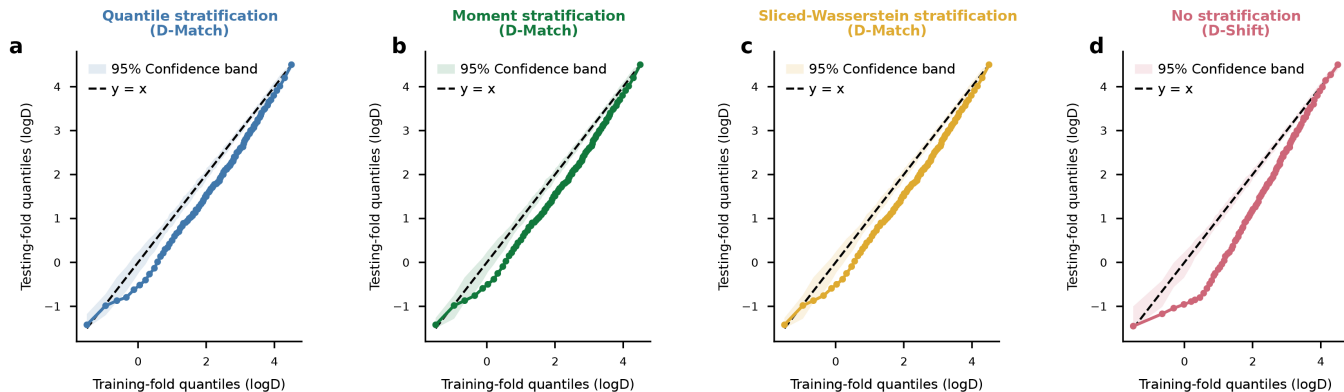
aligning each fold’s descriptor distribution toward a shared target through quantile stratification. The same kernel defines the cross-fold similarity in Eq. (S1) and the shift-mode alternative examined below, in which D instead separates fold-level descriptor distributions without stratification.

SHIELD ran with two folds at 80/20 ratio, weights $(\alpha, \beta, \gamma) = (1, 0.2, 3)$, D in match mode with quantile stratification ($\mu = 0.5$), for the primary configuration shown in the main figures. The three single-term ablations and the full joint objective were run under this same configuration. A dedicated ablation study then varied one design choice at a time relative to this primary run, comparing quantile, moment, and sliced-Wasserstein stratification under match mode, as well as replacing distributional matching with distributional shifting and removing stratification. Baselines and competitors include random, stratified, k -means, Kennard–Stone, agglomerative, DataSAIL, and five DeepChem splitters.

Fold separation is not straightforward to read from any single diagnostic. Multiple baselines achieve good visual separation on the shared t-SNE embedding (Fig. S14), yet the four diagnostics used throughout this study, visual separation on the embedding, $L(\pi)$, kNN purity, and downstream regressor performance, do not agree with one another. The agglomerative split illustrates this directly as it produces sparse and visually distinct clusters on the embedding, but its $L(\pi)$ is nonetheless close to random’s (0.311 versus 0.317), it records the second-lowest kNN purity of any split evaluated, and its downstream R^2 remains near the average across all models; four diagnostics, four

Table S4 | Lipophilicity stratification forms and distributional mode design choice studies.

D mode	Stratification form	$L(\pi)$	Train-test W_1 (logD)
Match	Quantile (primary)	0.216	0.418
Match	Moment	0.220	0.406
Match	Sliced-Wasserstein	0.220	0.406
Shift	None	0.201	0.650

**Fig. S17** | Lipophilicity stratification forms and distributional mode design choice studies. Held-out versus training logD quantiles for the *HMD*-SHIELD split match mode paired with quantile stratification (a), moment stratification (b), and sliced-wasserstein stratification (c), and for shift mode with no stratification (d). The dashed line marks perfect train-test agreement ($y = x$); the shaded band is the 95% confidence interval.

different rankings. DeepChem Butina shows the reverse pattern where substantial train-test overlap is visible on the embedding and both $L(\pi)$ (0.315) and kNN purity are high, yet its best-model test RMSE (1.02) and DNSE (1.7) are numerically identical, to the reported precision, to *HMD*-SHIELD’s own best-model values, the strongest SHIELD configuration by every leakage diagnostic. As a comparison between splitters, this is a null result; both partitions produce indistinguishable downstream difficulty despite markedly different structural leakage. As evidence about the diagnostics, however, it provides a second instance of the same conclusion that structural leakage alone does not determine downstream predictive difficulty. Taken together, these observations show that the diagnostics rank the same partitions differently, reinforcing our conclusion that no single metric provides a complete assessment of information leakage. This reflects the fact that leakage can occur through different channels, motivating multi-metric adoption for better evaluation of split quality.

Holding the H and M relational terms fixed, we performed a dedicated ablation study to isolate two design decisions. First, we compared the three stratification forms available under the distributional matching objective. Second, we replaced the matching objective with distributional shifting, thereby disabling stratification. Table S4 reports the resulting train-test Wasserstein-1 distance on logD for each configuration. The primary configuration, comprising distributional matching with quantile stratification, achieves $L(\pi) = 0.216$ while reducing the train-test distance to 0.418. Moment and sliced-Wasserstein stratification forms produce nearly identical distances of approximately 0.41. By contrast, switching term D to shift mode without stratification increases the train-test distance to its largest value (0.650) while reducing $L(\pi)$ to 0.201, consistent with the pronounced departure from the $y = x$ line in Fig. S17 d. This behavior arises because shift mode operates on the physicochemical-descriptor kernel rather than directly on logD, so the induced mismatch corresponds to only a moderate separation in the target space. Together, these results highlight the trade-off between the two objectives. Distributional shifting promotes stronger OOD separation at the cost of greater target-space drift, whereas distributional matching with stratification controls target-space drift while accepting greater cross-fold similarity. Neither objective is intrinsically preferable; the appropriate choice depends on the intended deployment scenario.

S9. Benchmark 6: qm8

qm8 reports 16 electronic-spectrum properties, excitation energies and oscillator strengths for 21,786 small molecules.¹¹ The 16-dimensional target makes joint per-fold representativeness harder to achieve than for lipophilicity’s single target. Qm8 additionally introduces a coarsening study, investigating the proposed H-first coarsening for datasets with more than 20,000 observations, and testing whether the leakage control achieved at full resolution survives when the partition is instead computed on a much smaller set of super-nodes.

The hierarchy H has two levels, a chromophore type above a transition type, because the chromophore governs the excited-state response and is therefore a meaningful containment unit. The affinity M is an ECFP4 Tanimoto kNN graph that links structural analogs. The distributional term D runs in shift mode in the same Tanimoto-kernel

space as M , separating the folds in chemical space rather than along any one spectral coordinate.

The primary SHIELD run used two folds at 80/20 ratio, unit term weights, and no stratification, so the primary split is driven entirely by structure. The three single-term ablations and the joint objective were run under this same configuration. Two dedicated studies then hold this relational configuration fixed while varying one further design choice, a stratification study on the primary spectral coordinate (E1-CC2) comparing two stratification forms, and a coarsening study comparing the full-resolution partition against the H-first super-node partition at five coarser resolutions. Baselines and competitors include random, stratified, k -means, Kennard-Stone, agglomerative, DataSAIL, five DeepChem splitters.

Among baseline splitters (Fig. S18 a-k), k -means, Kennard-Stone, DC Scaffold, and DataSAIL are the strongest in terms of t-SNE embedding separation and $L(\pi)$. Among single-term SHIELD configurations (Fig. S18 l-n), M -SHIELD attains the lowest $L(\pi) = 0.210$ because its affinity graph is constructed from the same Tanimoto similarity on which $L(\pi)$ is evaluated, so the optimization objective L_M differs from $L(\pi)$ only by normalization on this benchmark. By contrast, H -SHIELD and D -SHIELD produce well-separated clusters in the embedding while yielding substantially higher $L(\pi)$ values, comparable to that of the random baseline. This independence is expected for H , which operates on the chromophore hierarchy rather than the affinity graph, and only indirectly extends to D . HMD -SHIELD achieves the strongest overall performance across $L(\pi)$, kNN purity, and per-target distribution shift simultaneously (Fig. S19 a-c), as well as downstream performance (Figs. S20 and S21). The latter holds true for the

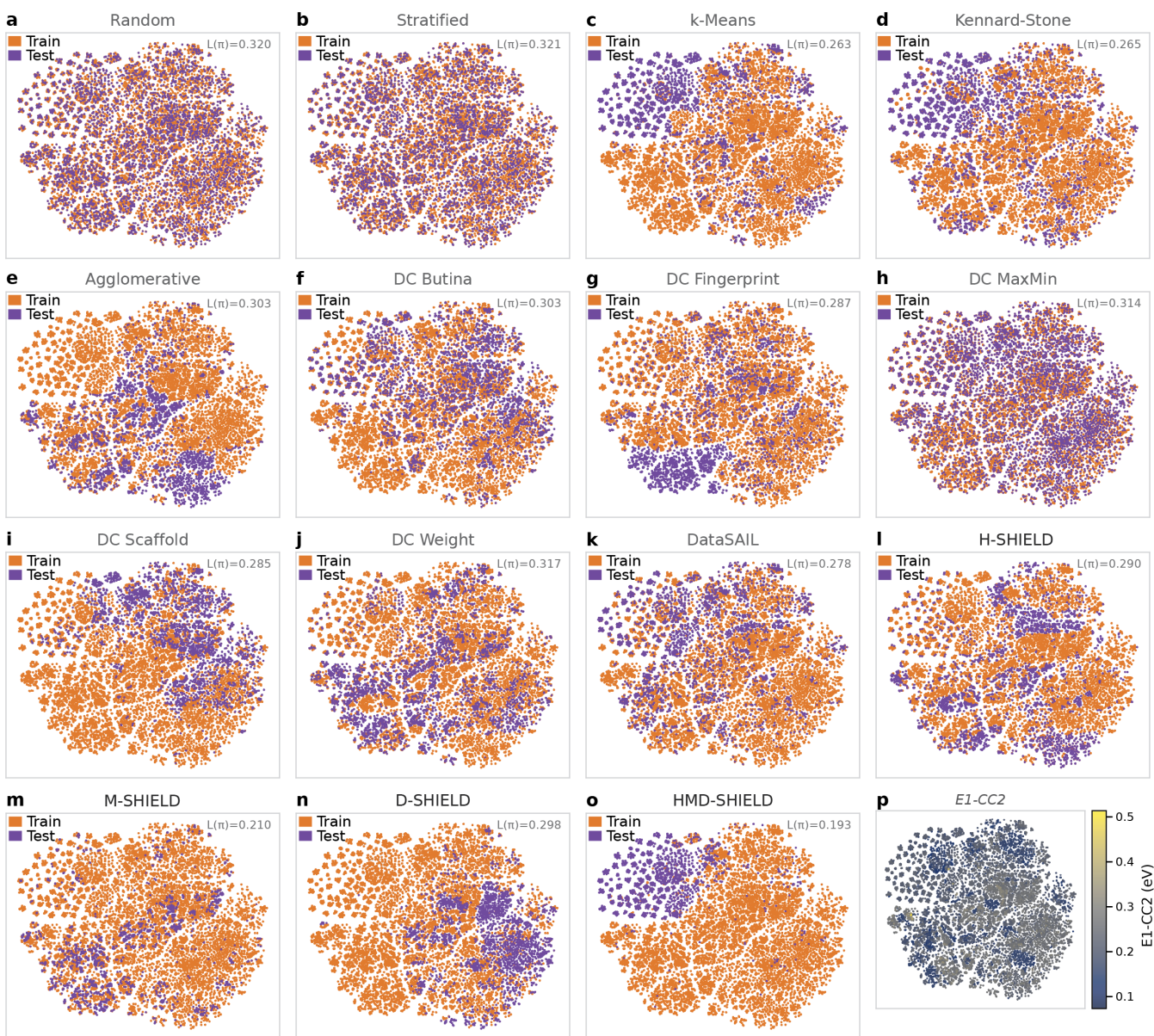


Fig. S18 | Qm8 split atlas. Shared t-SNE embedding of the ECFP4 fingerprint space under every split evaluated. Baseline splits include random, stratified, k -means, Kennard-Stone, agglomerative, five DeepChem splitters, and DataSAIL (a-k). Single-term SHIELD ablations and the joint objective HMD -SHIELD (l-o). The same t-SNE embedding colored by the E1-CC2 energy (p).

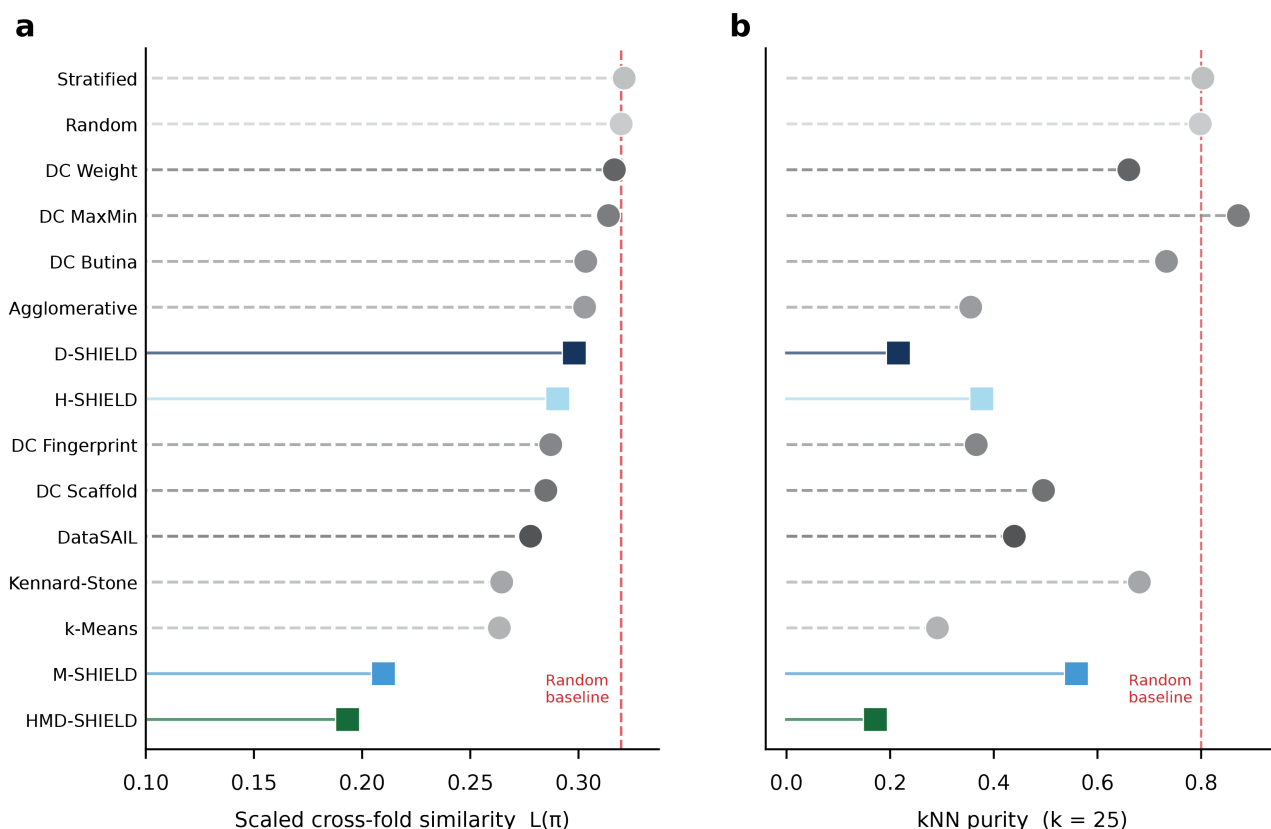


Fig. S19 | Qm8 split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 25$) for every split, with the random-split as a reference (a,b). Train-test distribution shift, in units of each target's pooled standard deviation, across all 16 Qm8 target variables with pooled σ given in parentheses for each target(c).

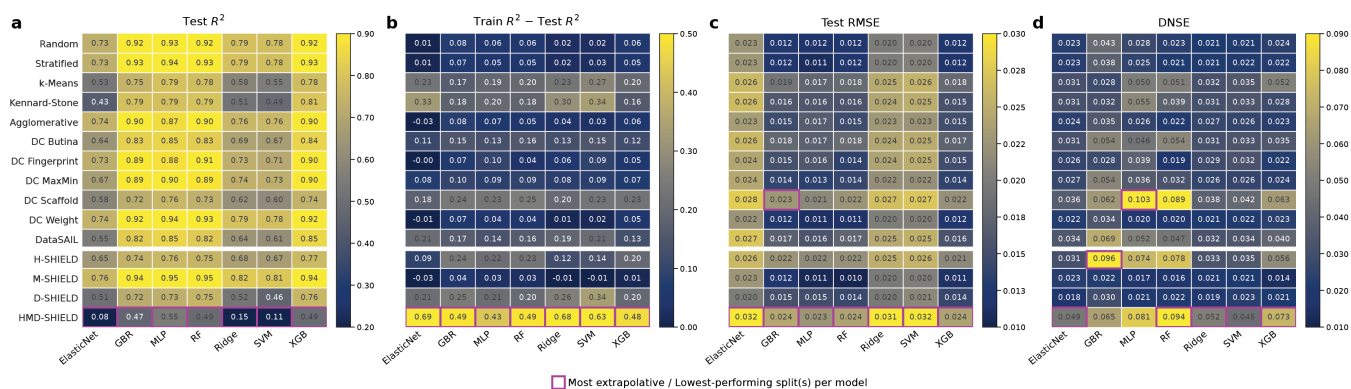


Fig. S20 | Qm8 downstream model performance for E1-CC2 target. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the fifteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest performance (most extrapolative) split(s) for each model.

primary target (E1-CC2) and for the average performance across all 16 targets, where *HMD-SHIELD* sets a harder ceiling across all regressors.

Fig. S22 and Table S5 report the stratification study on the primary target (E1-CC2). Unlike the analogous study for lipophilicity, every configuration here, including both stratification forms, keeps *D* in shift mode. This choice puts stratification and *D* in direct tension, even when defined on different space. Stratification constrains each fold's target marginal to resemble the others, the same goal *D*'s match mode pursues by construction, pulling each fold's measure toward a shared target distribution; *D*'s shift mode instead maximizes the distance between the folds' distributions in that same kernel space, so enabling stratification alongside shift mode asks the optimizer to satisfy two objectives pulling in opposite directions over the same coordinate, rather than the reinforcing pairing stratification has with match mode. Fig. S22 shows this directly on quantile-quantile plots where the non-stratified split departs from the $y = x$ line smoothly and monotonically (Fig. S22 c), the clean signature of shift mode acting alone. Meanwhile, both stratified forms (Fig. S22 a,b) wobble back toward the diagonal over parts of the range

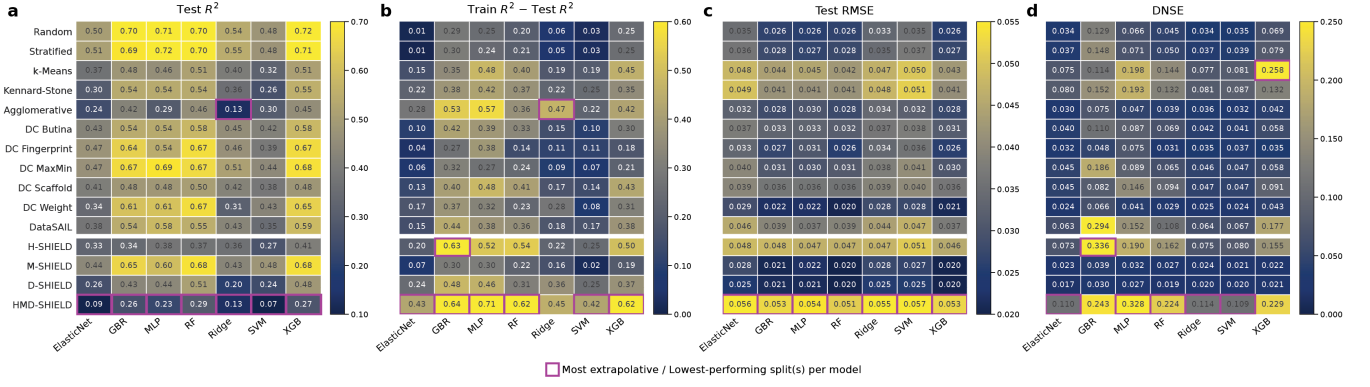


Fig. S21 | Qm8 downstream model performance averaged across all 16 targets. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the fifteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

Table S5 | Qm8 stratification study on the primary spectral coordinate (E1-CC2).

D mode	Stratification form	$L(\pi)$	Train-test W_1
Shift	None	0.193	0.0276
Shift	Quantile	0.198	0.0265
Shift	Sliced-Wasserstein	0.214	0.0217

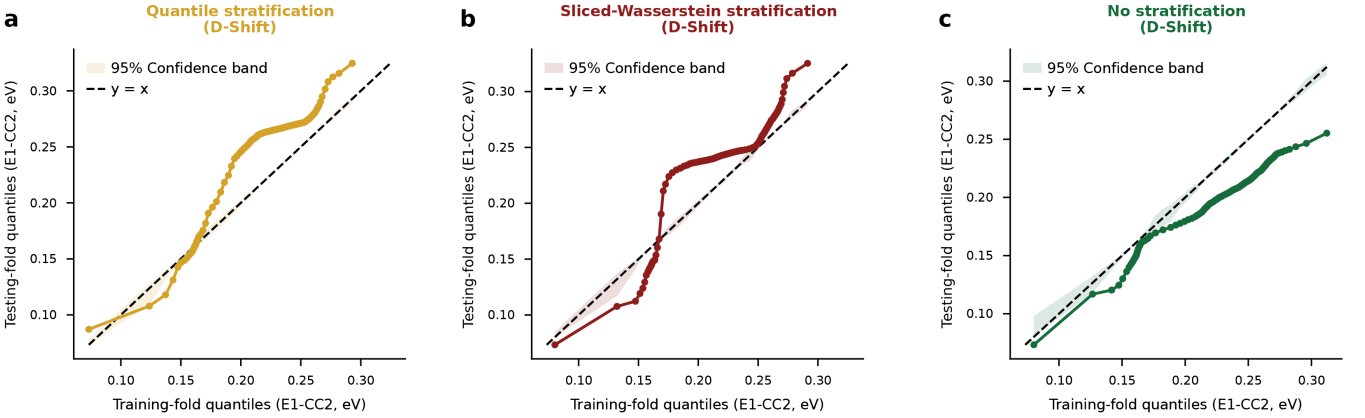


Fig. S22 | Qm8 stratification study. Held-out versus training quantiles of E1-CC2 for the *HMD-SHIELD* split with D in shift mode throughout. Investigated stratification forms are quantile (a), sliced-wasserstein (b), and no stratification (c). The dashed line marks perfect train-test agreement ($y = x$); the shaded band is the 95% confidence interval.

before departing from it again elsewhere, consistent with stratification repeatedly pulling the quantiles back into alignment even as shift mode continues to pull them apart. Stratification still narrows the aggregate Wasserstein distance to some extent, as Table S5 shows, but it does not produce the uniformly close fit seen when stratification is instead paired with match mode, as illustrated for the lipophilicity benchmark. On this evidence, stratification is not recommended alongside D in shift mode, since the two could pursue opposing objectives.

The coarsening study (Fig. S23) places the full-resolution *HMD-SHIELD* split alongside five coarser H-first super-node partitions, from $n' = 13$, the resolution of the chromophore-type hierarchy groups themselves, up to $n' = 1,000$. $L(\pi)$ and separation on the t-SNE embedding both stay within a narrow band across every resolution tested, at or below the full-resolution value throughout rather than drifting upward as the super-node count shrinks. The achieved separation is therefore retained under coarsening rather than eroded by it. At the coarsest resolution tested, $n' = 13$, this represents a compression ratio of roughly 1,668:1 where each super-node is represented by its centroid and where each super-node is represented as a single indivisible point for terms M and D . H is the only term still able to act on structure at this resolution, since its containment levels are what generated the super-node boundaries under H-first coarsening. Because $n' = 13$ coincides exactly with the number of chromophore-type hierarchy groups, every super-node at this resolution corresponds to exactly one hierarchy group, so respecting super-node integrity and respecting H 's containment become the same constraint, and H therefore transitions from a weighted term in the optimization to a hard constraint at this compression ratio, with no freedom for the optimizer to trade against it. Sustaining a stable split at this level of compression follows directly from coarsening along the

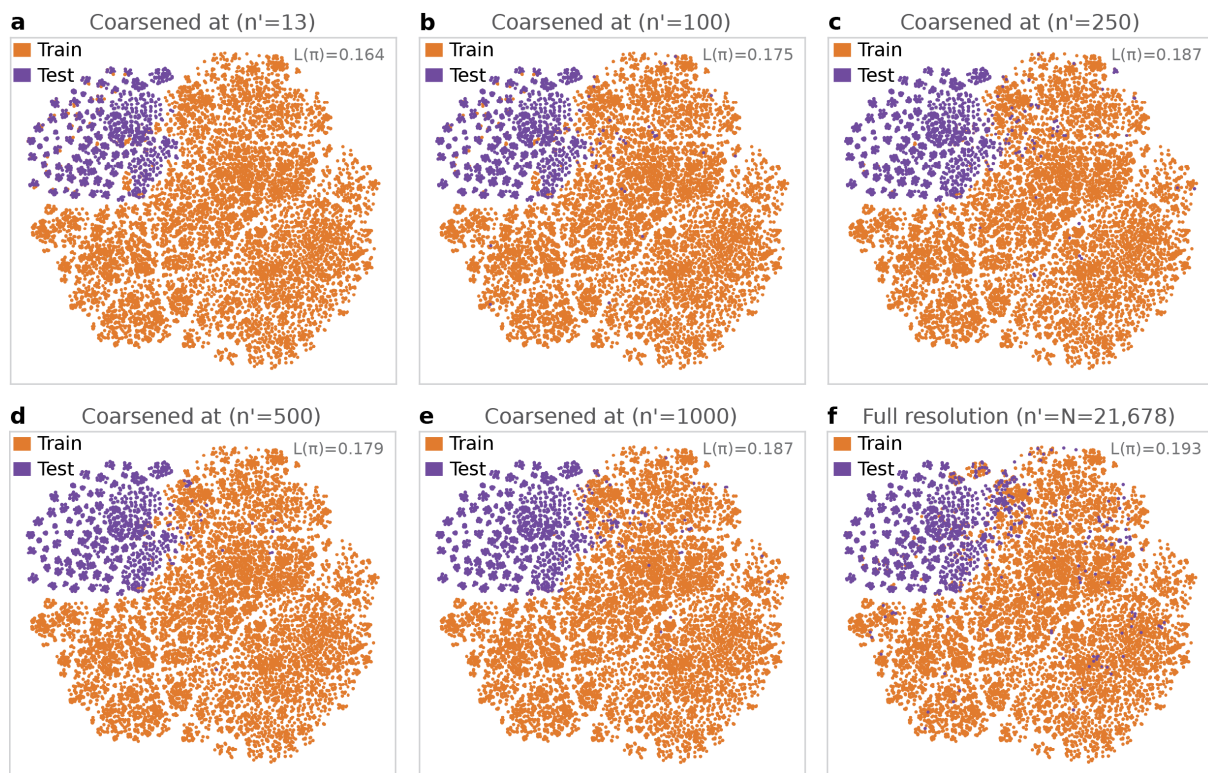


Fig. S23 | Qm8 coarsening study. t-SNE embedding of the *HMD-SHIELD* splits for qm8 dataset coarsened to $n' = 13$, 100, 250, 500, and 1,000 super-nodes (a-e), alongside the full-resolution split (f).

hierarchy rather than from an ad-hoc distance-based clustering. Because H-first super-nodes are anchored to a chemically meaningful grouping rather than to a clustering result that must itself be re-found at each resolution, the partition problem does not become less well-posed as the super-node count shrinks. A coarsening scheme whose super-nodes are instead defined by a distance-based clustering has no such anchor. The clustering must be re-derived at each resolution, and the grouping it returns need not correspond to any structure the split is intended to respect, so the relationship between the coarsened problem and the full-resolution problem is not fixed by construction.

S10. Benchmark 7: qm9

Qm9 holds 19 density-functional quantum properties for 133,885 small organic molecules with up to nine heavy atoms.^{11,12} At this size the observation-level affinity and kernel matrices are too large to form, so coarsening to super-nodes is not an optional refinement here, as it was for qm8, but the primary route. Therefore, we apply H-first coarsening such that SHIELD solves the partition problem on super-nodes before lifting the assignment back to molecules. The primary target was selected as the internal energy of a molecule at 0 Kelvin (U_0).

The hierarchy H has two levels, a property family above a heavy-atom-count size bin, with the size bin treated as the primary OOD axis, since extrapolating to larger molecules could form a hard generalization setup for quantum properties. The affinity M is an RBF kNN graph on standardized physicochemical descriptors, evaluated between super-node centroids. The distributional term D runs in shift mode with the same descriptor kernel raised to the power two, also at the super-node level.

The primary SHIELD runs with two folds at 80/20 ratio, unit term weights, and no stratification, solved on 512 super-nodes. Single-term ablations (H -SHIELD, M -SHIELD, D -SHIELD) and the joint objective (HMD -SHIELD) were evaluated. A dedicated coarsening-resolution sweep separately compares split stability from this primary 512-super-node resolution down to the coarsest resolution H 's own hierarchy structure permits. Reference baselines comprised random, stratified, k -means, Kennard–Stone, agglomerative, five DeepChem splitters, and DataSAIL, all run without coarsening.

HMD -SHIELD reaches by far the lowest cross-fold similarity of any split evaluated ($L(\pi) = 0.190$) and a kNN purity below 0.2, well below every baseline and competitor except for k -means (Fig. S25 a,b). Among single-term ablations, no term comes close to this on its own. As in earlier benchmarks, it is the combination of all three relational terms, not any one of them, that produces the large drop across every leakage metric simultaneously, since the optimizer can only exploit their complementary structure when H , M , and D are all active together.

Downstream ML performance metrics on the primary target (U_0), namely RMSE and DNSE, align with the ranking proposed by $L(\pi)$ and kNN purity (Fig. S26 c,d), indicating that HMD -SHIELD simulated the strongest OOD split over U_0 . However, the R^2 and its generalization gap do not align with this ranking, with DeepChem's

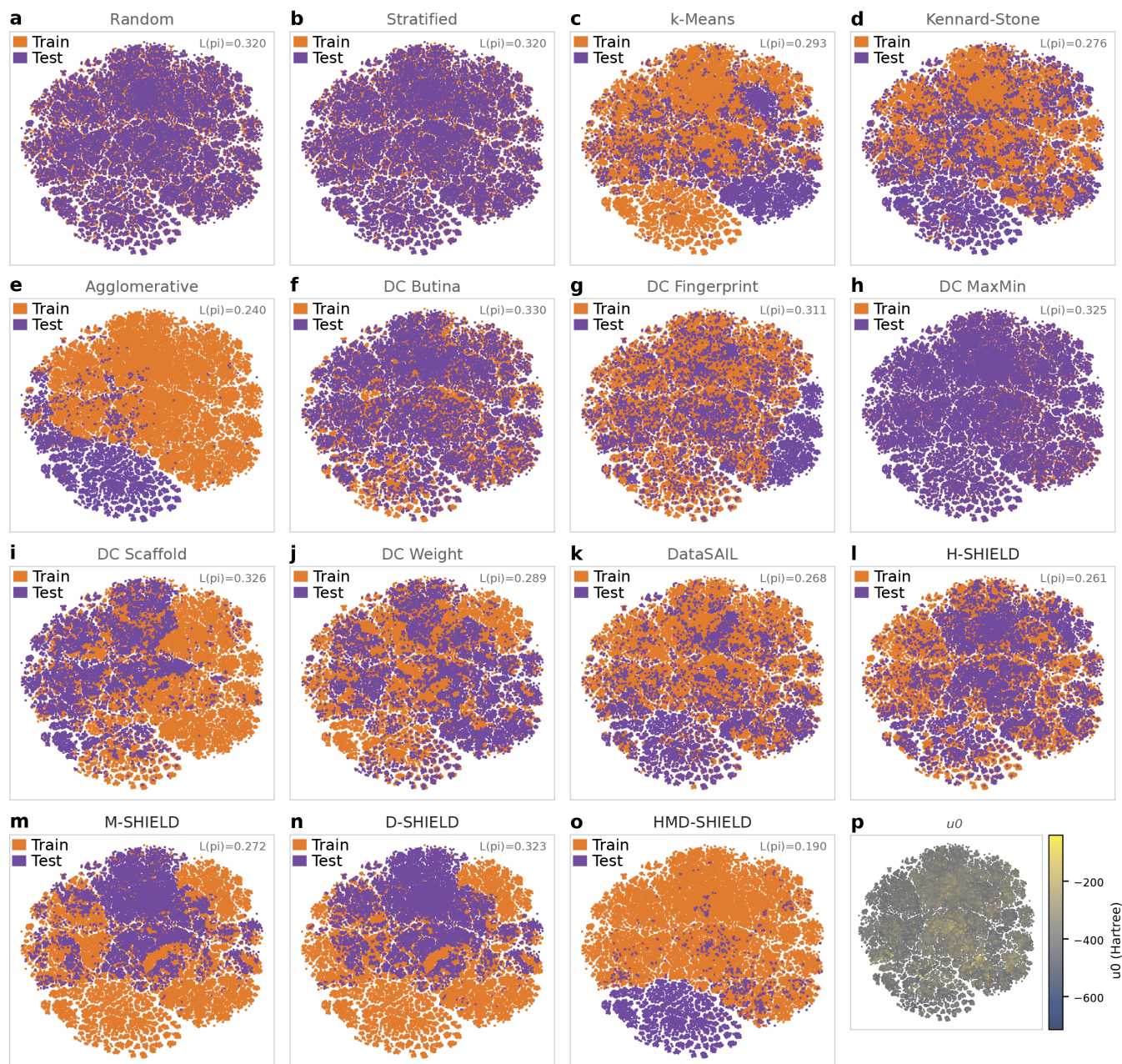


Fig. S24 | Qm9 split atlas. Shared t-SNE embedding of the molecular feature space under every split evaluated. Baseline splits include random, stratified, k -means, Kennard-Stone, agglomerative, five DeepChem splitters, and DataSAIL (a-k). Single-term SHIELD ablations and the joint objective HMD -SHIELD (l-o). The same t-SNE embedding colored by the U_0 target (p).

weight-based split and H -SHIELD ranking lowest (Fig. S26 a,b). This discrepancy mirrors that observed for tox21, as R^2 and RMSE capture complementary aspects of predictive performance. RMSE measures the absolute prediction error, whereas R^2 evaluates those errors relative to the variance of the held-out fold. Exploring the average downstream ML performance across all 19 targets, DeepChem’s weight-based split and H -SHIELD rank lowest for RMSE and DNSE, while D -SHIELD and HMD -SHIELD rank lowest for R^2 (Fig. S27 a-d). The relative ordering remains target- and metric-dependent, indicating that no split is universally optimal and that split selection should be guided by the deployment scenario of interest. Because the HMD -SHIELD configuration was optimized around U_0 as the primary target, it is expected to degrade downstream ML performance for this property rather than uniformly across all 19 targets. Consistent with this single-target formulation, HMD -SHIELD is the most challenging split for U_0 based on RMSE and DNSE, but its averaged performance across all properties remains competitive and is no longer the worst-performing split under these metrics. Nevertheless, several non-primary targets also exhibit substantial performance degradation under splits optimized for U_0 , suggesting that target shifts can propagate beyond the optimized property. This highlights a current limitation of the formulation, where H , M , D , and stratification objectives are optimized for a single target at a time. Future extensions could aggregate these terms across multiple properties and kernels, enabling splits tailored to deployment scenarios involving several target

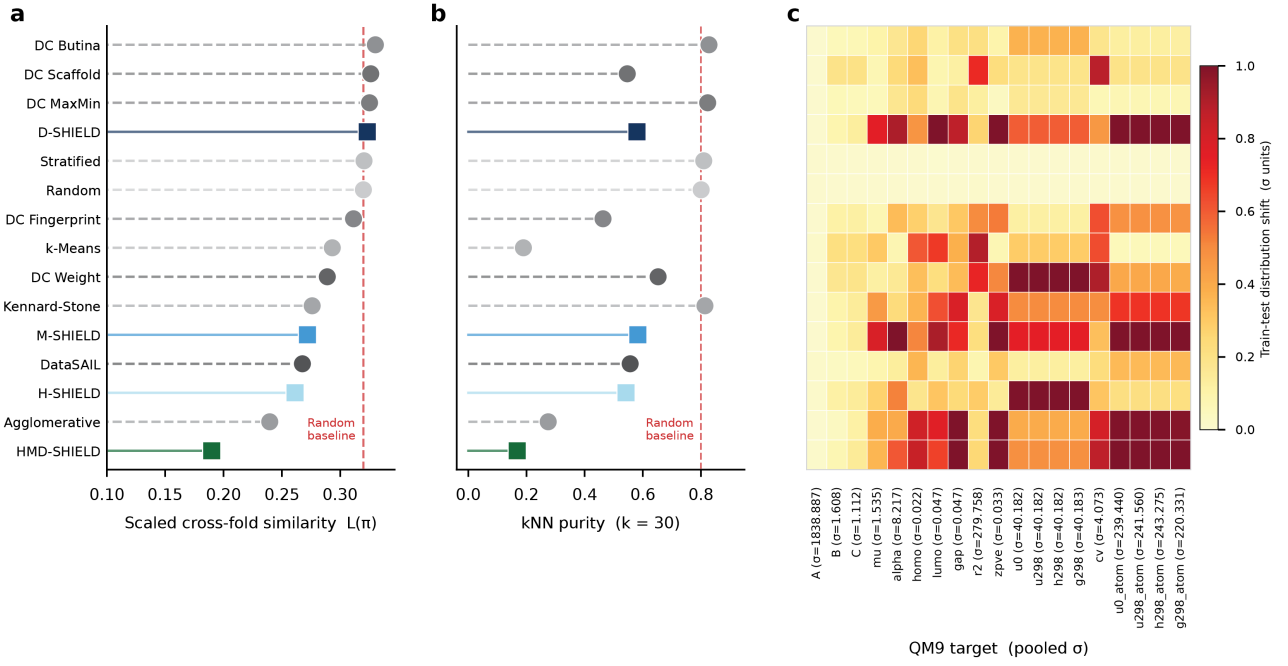


Fig. S25 | Qm9 split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 30$) for every split, with the random-split as a reference (a,b). Train-test distribution shift, in units of each target's pooled standard deviation, across all 19 Qm9 target variables with pooled σ given in parentheses for each target(c).

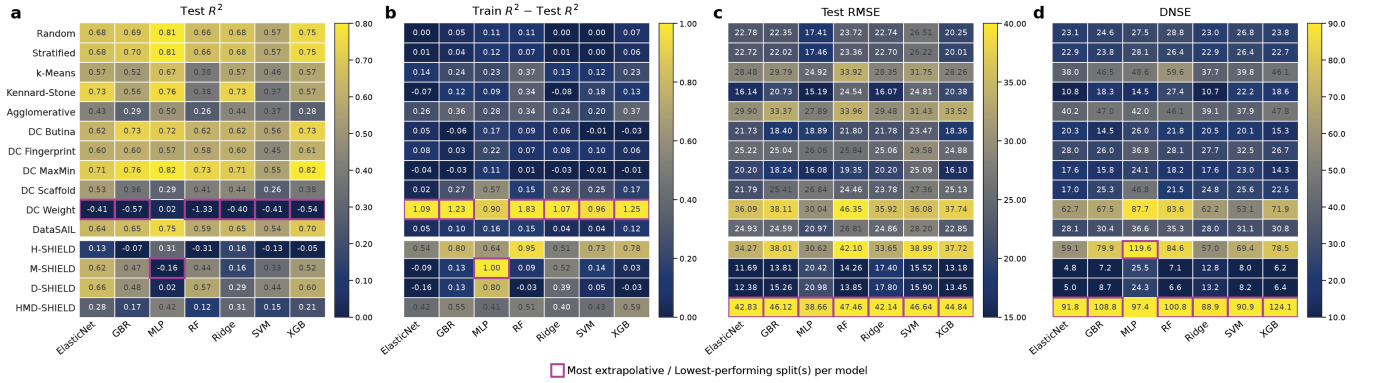


Fig. S26 | Qm9 downstream model performance for U_0 target. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the fifteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

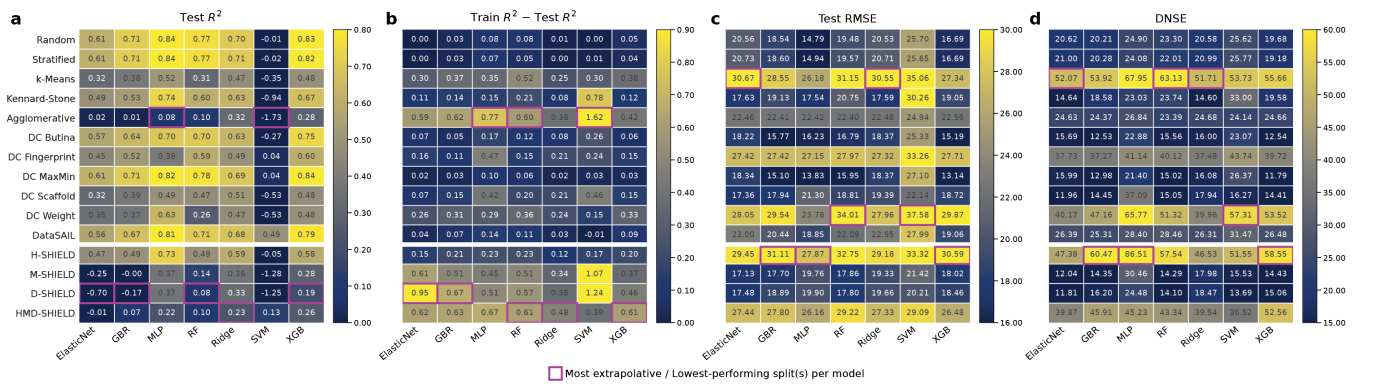


Fig. S27 | Qm9 downstream model performance averaged across all 19 targets. ML performance of seven regression models, each independently hyperparameter-tuned per split, across the fifteen splits evaluated. The reported metrics include Test R^2 , generalization gap, RMSE, and DNSE (a-d). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

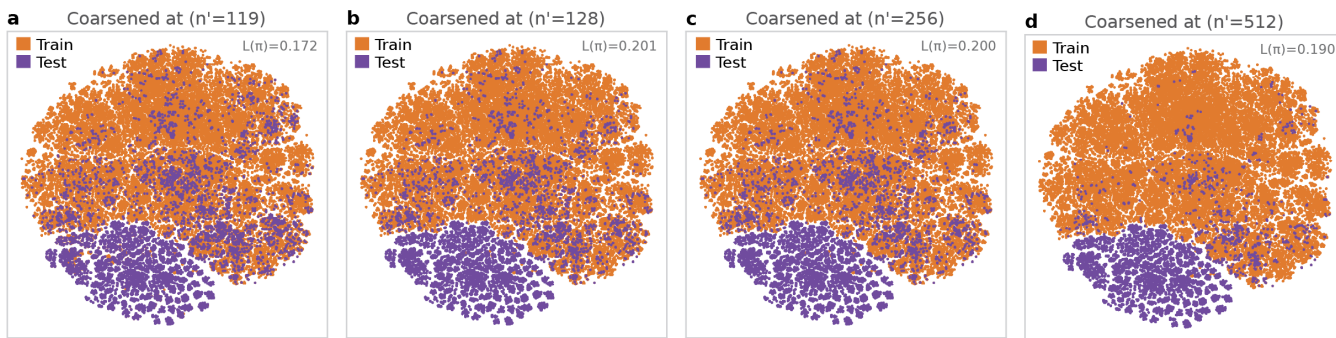


Fig. S28 | Qm9 coarsening study. t-SNE embedding of the *HMD-SHIELD* splits for qm9 dataset coarsened to $n' = 119$, 128, 256, and 512 super-nodes (a-e).

distributions simultaneously.

The coarsening-resolution sweep (Fig. S28) shows that the super-node approximation, even at the coarsest resolution tested, does not compromise split quality. The coarser resolutions shown in the figure ($n' = 119$, 128, and 256), all below the primary 512-super-node resolution, remain in the same 0.17-0.20 band for $L(\pi)$, with similarly structured held-out folds on t-SNE embedding. This confirms the same behavior established for qm8 where coarsening the full qm9 dataset to 119 super-nodes, the coarsest resolution *H*'s own hierarchy structure permits, a compression ratio of roughly 1,089:1, still yields a stable split within the same controlled $L(\pi)$ band the primary resolution achieves. This stability follows from the H-first coarsening scheme itself rather than from anything particular to this dataset. Because super-node boundaries are anchored to a physically meaningful grouping rather than to a distance-based clustering, high compression ratios are attainable without destabilizing the resulting split.

S11. Benchmark 8: HIV

The HIV benchmark labels 41,127 compounds for their ability to inhibit viral replication, a binary target with about 3.5% actives.¹¹ The actives are structurally diverse, their mean pairwise Tanimoto similarity is only about 0.10, so a split must respect the chemotype while still preserving the rare positive class across folds. Unlike tox21's twelve assays, HIV is a single binary target, but its extreme class imbalance combined with structurally scattered actives raises a different combination of constraints.

The hierarchy *H* has two levels, a chemotype defined by the Bemis-Murcko scaffold above a framework defined by the ring-system class (ring count, aromaticity, fusion, and heteroatom content). The affinity *M* is a scaffold-gated Tanimoto kNN graph evaluated between super-node centroids. Analogs are linked only within a shared scaffold, removing the spurious cross-scaffold edges that raw fingerprint overlap would otherwise create. The distributional term *D* runs in shift mode in the same Tanimoto-kernel space, separating the folds in chemical space, because the activity here is series-driven and structural separation could serve as the appropriate OOD lever. Unlike the matbench_glass benchmark, where term *D* runs in match mode, the HIV benchmark uses term *D* in shift mode with stratification intentionally disabled to assess whether optimization under extreme class imbalance, without an explicit class-balancing mechanism, produces degenerate folds.

SHIELD ran with two folds at 80/20 ratio and weights $(\alpha, \beta, \gamma) = (1, 0.2, 3)$ with no stratification. The dataset was coarsened moderately to 1,024 class-pure super-nodes by an H-first scheme before partitioning. The super-node count is fixed rather than swept, unlike the qm8 and qm9 coarsening studies, considering the light compression ratio of about 40:1. Because each super-node contains samples from only one class, class balance is enforced by assigning whole active- or inactive-class super-nodes to different folds, rather than by partitioning mixed-class super-nodes. This design removes mixed-class super-nodes as a confounding factor, ensuring that any degenerate fold arises from the optimization algorithm itself rather than from class mixing within super-nodes. Otherwise, mixed-class super-nodes would anchor active and inactive samples together, making active-free folds substantially harder to obtain without aggressively violating the hierarchy term *H*. The three single-term configurations (*H-SHIELD*, *M-SHIELD*, *D-SHIELD*), the three pairwise configurations (*HM-SHIELD*, *HD-SHIELD*, *MD-SHIELD*), and the joint objective (*HMD-SHIELD*) were evaluated. Reference baselines comprised random, stratified, *k*-means, Kennard-Stone, five DeepChem splitters, and DataSAIL, all run without coarsening.

None of the single-term SHIELD splits generates a well-separated held-out fold on its own, nor does *HM-SHIELD*, evident from high $L(\pi)$ and kNN purity values (Fig. S30 a,b) and substantial overlap with the training fold in the embedded space (Fig. S29). Only when the distributional term *D* is paired with another relational term does the held-out fold become well separated. *MD-SHIELD* achieves the lowest $L(\pi)$ of any configuration (0.228), followed by *HMD-SHIELD* (0.233) and *HD-SHIELD* (0.240), all substantially below every baseline and all exhibiting clear separation in the embedding. This behavior reflects the limitation of applying shift mode in isolation along an axis that is weakly coupled to the target distribution. Although the Tanimoto-kernel space captures chemical dissimilarity, global separation in this space does not necessarily translate into meaningful target separation or reduced

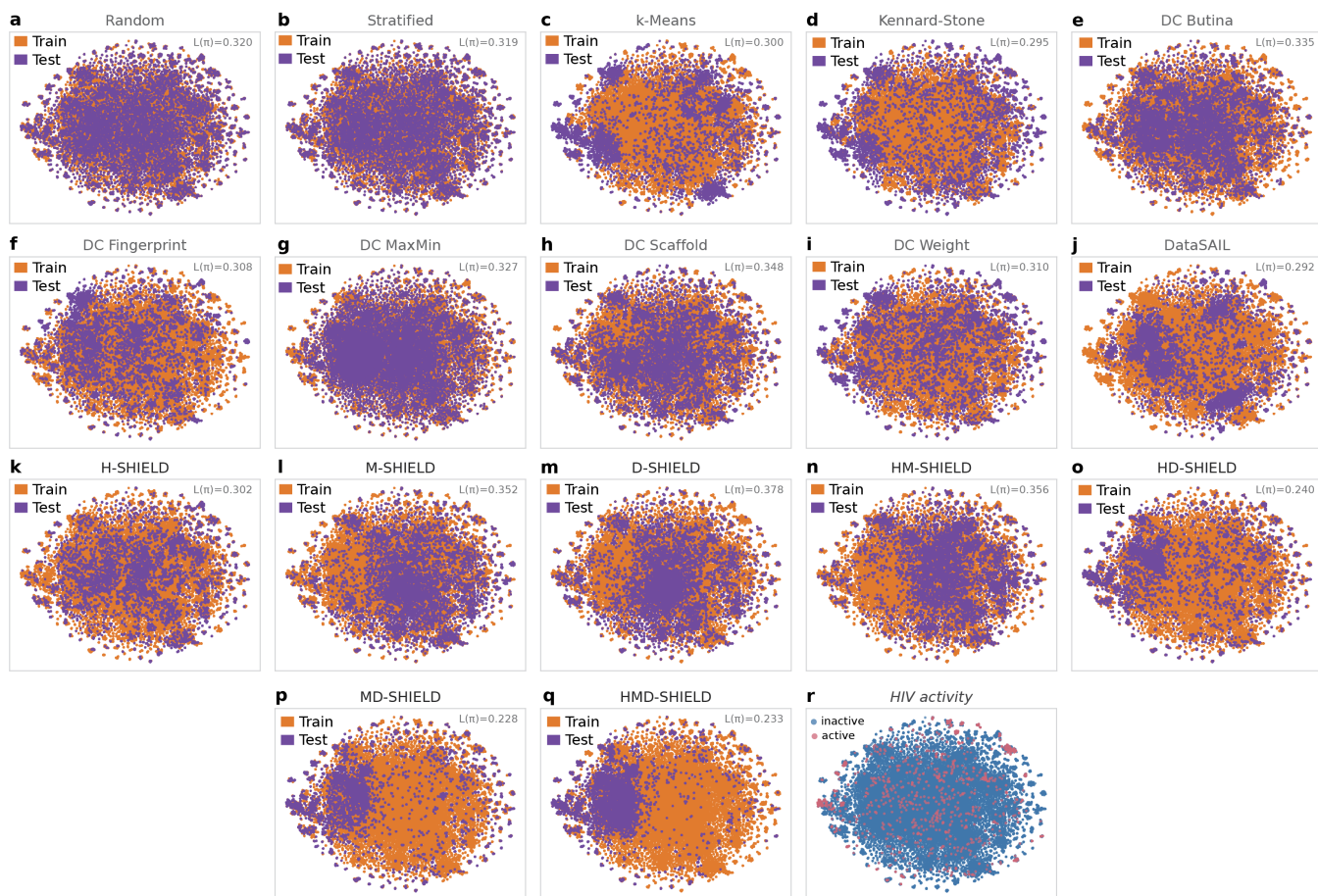


Fig. S29 | HIV split atlas. Shared t-SNE embedding of the ECFP4 fingerprint space under every split evaluated. Baseline splits include random, stratified, *k*-means, Kennard–Stone, five DeepChem splitters, and DataSAIL. (a–j). Single-term SHIELD ablations (k–m), pairwise SHIELD configurations (n–p), and the joint objective *HMD*-SHIELD (q). The same t-SNE embedding colored by HIV activity label (r).

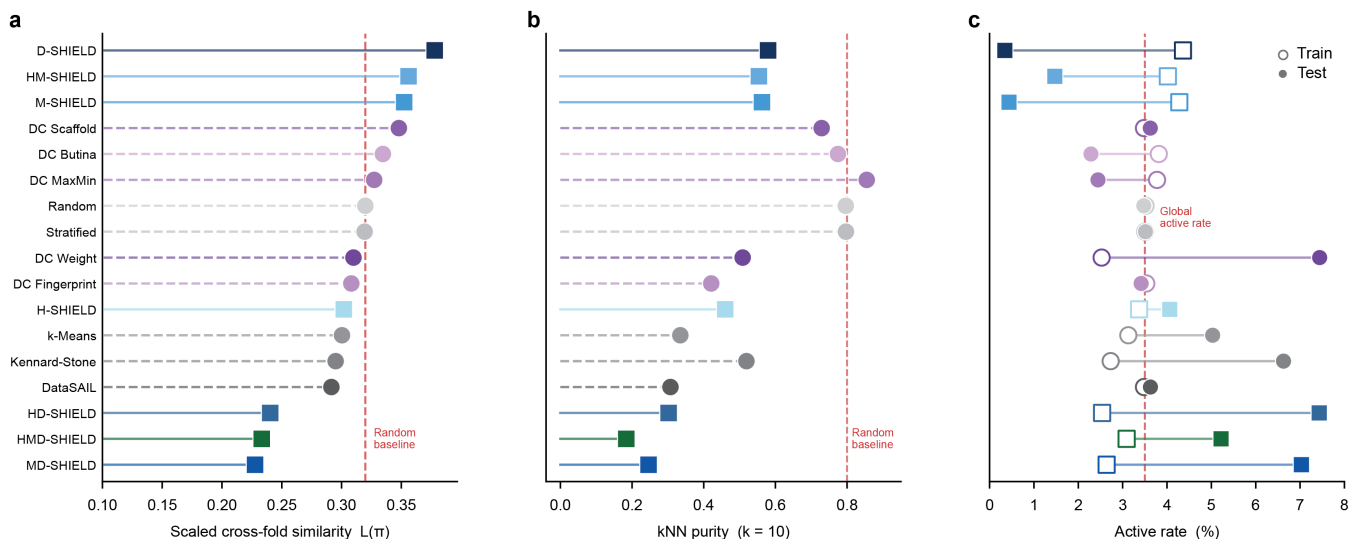


Fig. S30 | HIV split-quality diagnostics. Scaled cross-fold similarity $L(\pi)$ and kNN purity ($k = 10$) for every split, with the random-split as a reference (a,b). Active-rate distributions in the training and test partitions for each split (c). Open and filled markers denote the train and test sets, respectively, and the dashed red line indicates the global active rate of the full dataset.

information leakage without preserving scaffold-level and local neighborhood relationships. Pairing *D* with either *H*'s scaffold containment or *M*'s within-scaffold neighborhood graph supplies additional constraints, converting the aggregate distributional shift into a genuine reduction in information leakage. Downstream ML performance, however, follows a markedly different ranking. In particular, *M*-SHIELD, *D*-SHIELD, and *HM*-SHIELD achieve the highest

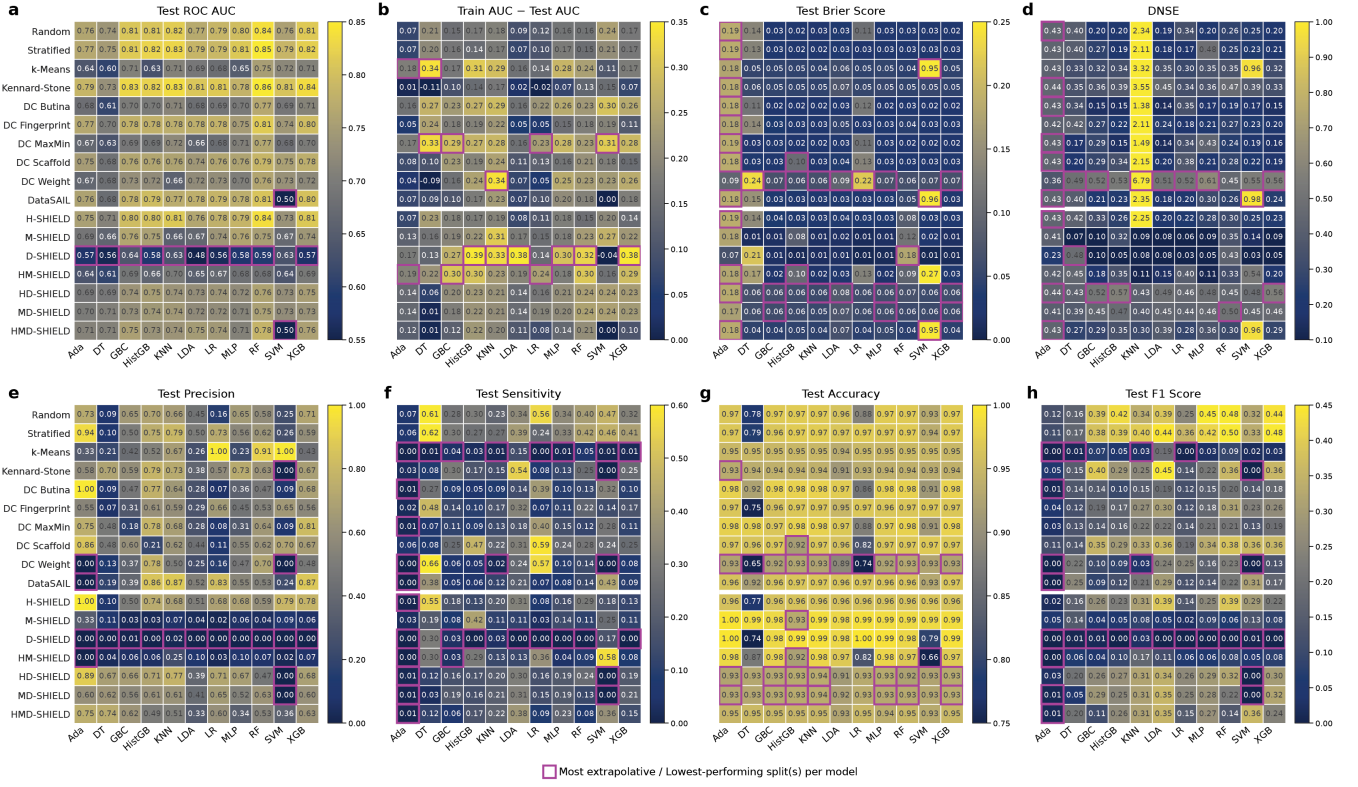


Fig. S31 | HIV downstream model performance. ML performance of eleven classification models, each independently hyperparameter-tuned per split, across the seventeen splits evaluated. The reported metrics include test ROC AUC, generalization gap, Brier score, DNSE, precision, sensitivity, accuracy, and F1 score (a-h). Boxed cells mark the lowest-performing (most extrapolative) split(s) for each model.

classification accuracy while simultaneously yielding the lowest precision. Inspection of the fold compositions reveals that these three configurations produce degenerate held-out folds containing only a handful of active compounds (Fig. S30 c), a consequence of intentionally optimizing in shift mode without stratification under extreme class imbalance. The resulting scarcity of positives inflates accuracy through the dominance of the inactive class while severely degrading precision, demonstrating that conventional downstream metrics can become misleading when the held-out class distribution collapses. These results therefore underscore the importance of incorporating explicit class-balancing safeguards, such as stratification or equivalent constraints, when partitioning highly imbalanced classification problems.

S12. Supplementary methods

S12.1 Cut identity for term M

The equivalence between the pairwise cut sum and the quadratic form $\text{tr}(z^\top L z)$ stated for term M in the main text follows by fixing a pair of observations i, j and inspecting the inner sum over folds in L_M . When i and j share a fold, $z_{ik} = z_{jk}$ for every k and the summand vanishes. When they do not, exactly two coordinates of $z_i - z_j$ equal ± 1 , so the squared sum over k equals two and the leading factor of one half restores a weight of A_{ij} for that crossing pair. Summing over all pairs recovers the cut sum. The same quadratic form expands as:

$$\frac{1}{2} \sum_{i,j} A_{ij} (z_{ik} - z_{jk})^2 = \sum_{i,j} z_{ik} (D_{ij} - A_{ij}) z_{jk} \quad (\text{S11})$$

with $D = \text{diag}(A\mathbf{1})$, which identifies the cut sum with $\text{tr}(z^\top L z)$ for the unnormalized Laplacian $L = D - A$. SHIELD uses this unnormalized Laplacian throughout the composite objective and its gradient, so that the value and the gradient of Eq. 11 remain mutually consistent. The symmetric normalized Laplacian is used only inside the spectral cascade of the initialization stage (main text, Algorithm 1), where it corresponds to a normalized-cut objective rather than to L_M itself. Both L_M and its gradient $2Lz$ are evaluated as a single sparse-dense product between L and the full $N \times K$ assignment matrix z , not as K separate per-column calls. A sparse-dense product of this shape is mathematically equivalent to K independent sparse matrix-vector products, each costing $O(\text{nnz}(A))$, since every nonzero of L must touch every one of the K columns, for a total of $O(\text{nnz}(A)K)$. All K columns are needed at once because the assignment linear program consumes the full $N \times K$ gradient simultaneously to build its linear coefficient θ , so the term cannot be evaluated one fold at a time, and this cost is paid once per alternation iteration, which makes explicit the fold-count scaling only implicit in the sparse-affinity cost given in the main text.

S12.2 Entropic transport for term D

SHIELD solves the entropic optimal-transport problem underlying match-mode D (Eq. 12) in the log domain rather than by exponentiating the cost matrix directly, which is unstable at small regularization ϵ . Writing $f \in \mathbb{R}^N$ and $g \in \mathbb{R}^m$ for the dual potentials of the source and target marginals, each Sinkhorn sweep updates f against the source marginal u_k through a log-sum-exp reduction over the target coordinates, then updates g symmetrically against the target measure ν_k^* , with every exponent evaluated through a stabilized log-sum-exp rather than a direct exponential. The iteration for a given fold stops once the largest coordinate-wise change in either potential falls below a tolerance, or once the dual objective proxy $\langle u_k, f \rangle + \langle \nu_k^*, g \rangle$ stabilizes in relative terms, with either condition required to hold for several consecutive iterations before the plan, reconstructed as $\Pi_{ij}^k = \exp((f_i + g_j - C_{ij})/\epsilon)$ and renormalized to unit mass, is accepted. The value entering calibration, the optimization history, and the composite objective is the transport-cost component $\sum_k \rho_k \langle C, \Pi^k \rangle$ alone. The entropic term is excluded from all three because it is not sign-stable across plans of differing diffuseness, though it remains inside the solver itself and inside the gradient of Eq. 13, so the quantity reported and the quantity differentiated are related but not identical.

S12.3 Kernel discrepancy for term D

Evaluating shift-mode D (Eq. 14) and its gradient (Eq. 15) over every fold pair reduces to a single dense matrix product rather than to $K(K-1)/2$ separate evaluations. Collecting the fold memberships into $U = [u_1, \dots, u_K]$, SHIELD forms the product κU once per evaluation, after which every pairwise MMD value and its gradient follow from inner products between the columns of U and the columns of κU . This costs $O(N^2 K)$ time, dominated by the single matrix product, consistent with the dense $O(N^2)$ memory cost already noted for shift-mode D in the main text.

S12.4 Stratification forms

The soft stratification forms admit closed-form gradients because each is differentiated fully through the fold-mass normalization rather than by treating fold membership sums as fixed, which is what makes them usable inside the same block-coordinate loop as the other relational terms. For the quantile form, each output coordinate j is discretized into B bins, with $\text{bin}_j(i)$ the bin of observation i and $W_{j,b} = \sum_{i:\text{bin}_j(i)=b} w_i$ the pooled mass of bin b . The residual and its gradient are:

$$L_{\text{strat}}^q = \sum_j \sum_k \sum_b \left(\sum_{i:\text{bin}_j(i)=b} w_i z_{ik} - r_k W_{j,b} \right)^2, \quad \frac{\partial L_{\text{strat}}^q}{\partial z_{ik}} = \sum_j 2w_i (\text{tally}_{k,\text{bin}_j(i)} - r_k W_{j,\text{bin}_j(i)}) \quad (\text{S12})$$

where $\text{tally}_{k,b}$ is the mass fold k currently places in bin b . The moment form compares each fold's central moments, orders one through four, against the pooled moments $M_{k,j}^{(m)}$, weighted per order and coordinate by $\omega_{m,j}$. Writing $\mu_{k,j}$ for the fold- k mean of coordinate j and $S_k = \sum_i w_i z_{ik}$ for the fold mass, differentiating the first and higher central moments exactly through their dependence on z gives:

$$\frac{\partial M_{k,j}^{(1)}}{\partial z_{l,k}} = \frac{w_l (y_{l,j} - \mu_{k,j})}{S_k}, \quad (\text{S13})$$

$$\frac{\partial M_{k,j}^{(m)}}{\partial z_{l,k}} = \frac{w_l}{S_k} \left[(y_{l,j} - \mu_{k,j})^m - m (y_{l,j} - \mu_{k,j}) M_{k,j}^{(m-1)} - M_{k,j}^{(m)} \right], \quad m = 2, 3, 4 \quad (\text{S14})$$

with the central first moment identically zero. The correction terms beyond the leading $(y_{l,j} - \mu_{k,j})^m$ contribution account for how the fold mean and fold mass themselves shift as $z_{l,k}$ changes, so the gradient remains exact rather than approximate even for folds with few members. The sliced-Wasserstein form draws a fixed number of random unit projections of the multi-dimensional target, reduces each projection to a one-dimensional weighted quantile function on a uniform grid, and integrates the squared difference between a fold's projected quantile function and the pooled one by the trapezoidal rule. Because this Monte-Carlo estimate carries its own standard error across projections, SHIELD widens the term's convergence tolerance by that error, so the alternation loop does not stop on projection noise rather than genuine convergence. A fold whose mass falls below a small numerical threshold has no well-defined moments or weighted quantile function. SHIELD detects this case and sets the affected fold's contribution to the stratification residual and its gradient to zero rather than letting a near-zero fold mass inflate either quantity, though the hard balance constraint keeps ordinary runs from exercising this guard.

S12.5 Calibration scales

Each active relational and stratification term is divided by a calibration scale before entering the composite objective, so that the weights $\alpha, \beta, \gamma, \mu$ act on comparable magnitudes as described in the main text. The soft balance penalty in the assignment program is calibrated the same way, SHIELD estimating S_{bal} from $\sum_k (\sum_i w_i z_{ik} - r_k W)^2$ on the same random balanced draws used for the other terms and dividing η by S_{bal} in Eq. 17, exactly as $\alpha, \beta, \gamma, \mu$ are divided by their own scales. The default scale for every term is a Monte-Carlo estimate obtained by drawing random

balanced assignments, evaluating the term on each, and taking the mean absolute value, the absolute value making the estimate robust to sign, which matters for the entropic match objective that can take a negative value on a diffuse random plan. Four terms depart from this default because a random balanced assignment sits close to the null distribution each of them measures against and would otherwise drive the Monte-Carlo estimate toward zero. For match-mode D , the scale is the transport-cost component alone, following the same exclusion of the entropic term used for calibration and history. For the quantile form, the per-bin residual is maximized over the feasibility simplex when an entire bin’s mass enters the fold with the smallest target proportion, giving the tight bound:

$$S_{\max}^q = \left(\sum_{j,b} W_{j,b}^2 \right) \left(1 - 2 \min_k r_k + \sum_k r_k^2 \right) \quad (S15)$$

so that the normalized residual lies in the unit interval. For the moment form, a single range-based bound would be dominated by the order-four term and a single pooled-relative bound collapses on the odd orders for symmetric data, so SHIELD instead uses a mixed per-order scale built from the per-coordinate range $R_j = \max_i y_{ij} - \min_i y_{ij}$ following:

$$S_{\max}^{\text{mom}} = K \sum_j \left(\omega_{1,j} R_j^2 + \omega_{2,j} \max(|M_{\star,j}^{(2)}|, \epsilon_2)^2 + \omega_{3,j} \frac{R_j^6}{64} + \omega_{4,j} \max(|M_{\star,j}^{(4)}|, \epsilon_4)^2 \right) \quad (S16)$$

with floors $\epsilon_2 = 10^{-6} R_j^2$ and $\epsilon_4 = 10^{-6} R_j^4$. The range bound $R_j^6/64$ for the third order is the worst-case squared third central moment attainable by data confined to range R_j , chosen because the pooled third moment itself collapses to zero on symmetric distributions and cannot serve as its own scale. For the sliced-Wasserstein form, the scale is the covariance-trace reference:

$$S_{\max}^{\text{SW}} = K \text{tr}(\text{Cov}_w(y)) = K \sum_j \text{Var}_w(y_j) \quad (S17)$$

which shares the squared- y units of the residual, is deterministic rather than sampled, and weights each observation’s contribution by its own weight rather than by the square of its projection, which keeps it resistant to outliers. For shift-mode D , any positive-semidefinite kernel satisfies $\text{MMD}^2 \leq 2 \max_i \kappa_{ii} - 2 \min_{ij} \kappa_{ij}$, which summed over fold pairs gives the worst-case scale:

$$S_{\max}^{\text{shift}} = \left(\sum_{k \neq k'} \rho_{kk'} \right) \max \left(2 \max_i \kappa_{ii} - 2 \min_{ij} \kappa_{ij}, 0 \right) \quad (S18)$$

When the kernel is close to constant and this bound collapses toward zero, SHIELD falls back to an average-pairwise scale that replaces the two extremes with the kernel’s diagonal mean and its overall mean.

S12.6 Computational complexity

Let N denote the number of observations, or the super-node count n' once coarsening has been applied, K the fold count, n_c the number of classes, d the feature dimension, q the outcome dimension, $\text{nnz}(A)$ the affinity edge count, m the optimal-transport support size, and I the number of alternation iterations, with n_c unrelated to the match-mode cost matrix C of Section S12.2.

The hierarchy LP has $O(\sum_{\ell} (n_{\text{nodes},\ell} + n_{\text{edges},\ell}) K)$ variables and is solved as a continuous linear program at every iteration rather than by branch-and-cut, since y and w remain relaxed throughout the loop. The value and gradient of L_M cost $O(\text{nnz}(A) K)$ for a sparse affinity and $O(N^2 K)$ for a dense one. Match-mode D costs $O(K N m I_{\text{sink}})$ across the K folds, with the number of Sinkhorn iterations I_{sink} cut sharply by warm-starting the duals from the previous outer iteration, refining the per-sweep cost given in the main text to the total across sweeps. Shift-mode D forms the matrix product κU at $O(N^2 K)$ time and $O(N^2)$ memory, this dense kernel the binding memory cost of the whole pipeline, unreduced by coarsening unless the kernel itself is rebuilt on the coarsened representation. The assignment program has NK continuous variables together with K balance constraints, $n_c K$ class constraints, and per-entity and per-family constraints, only its optional integer-polish stage solved by branch-and-cut.

The quantile and moment stratification forms cost $O(NqK)$ per iteration, each built from a fold-by-observation intermediate formed once per coordinate. The sliced-Wasserstein form costs $O(P K N \log N)$, dominated by sorting each fold’s weighted projections onto the quantile grid.

Pipage and confidence-gap rounding cost $O(NK + N \log N)$ over a constant number of balance passes. The leakage-aware swap costs $O((N + \text{nnz}(A))K)$ per pass, which for a kNN affinity with $\text{nnz}(A) = O(Nk)$ becomes $O(NK^2)$.

Total running time is $O(I \cdot (T_{\text{LP}} + T_D + T_M + T_{\text{strat}}))$ beyond the one-time coarsening and calibration cost, where T_{LP} is the assignment solve. Total space is $O(NK)$ for the assignment variables, plus $O(\text{nnz}(A))$ for a sparse affinity, $O(Nm)$ for the transport cost matrix, $O(Nd)$ for the input features, and $O(N^2)$ only when a dense shift-mode kernel is used. Coarsening substitutes n' for N in every relational term and is the recommended route to large N .

References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
2. Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137–148 (1969).
3. Li, J., Guan, X., Zhang, O., *et al.* Leak Proof PDBBind: A Reorganized Data Set of Protein–Ligand Complexes for More Generalizable Binding Affinity Prediction. *The Journal of Physical Chemistry B* **130**, 730–740 (Jan. 2026).
4. Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More* (O’Reilly Media, Sebastopol, CA, 2019).
5. Joeres, R., Blumenthal, D. B. & Kalinina, O. V. Data splitting to avoid information leakage with DataSAIL. *Nature Communications* **16**, 3337 (2025).
6. Huizing, G.-J., Deutschmann, I. M., Peyré, G. & Cantini, L. Paired single-cell multi-omics data integration with Mowgli. *Nature Communications* **14**, 7711 (2023).
7. Karaki, A., Hammoud, A., Alabtah, F. G., AbdelGawad, M. & Khraisheh, M. CorrosionOptiMap links additive manufacturing process parameters to corrosion of stainless steel. *npj Materials Degradation* **9** (2025).
8. Barrionuevo, G. O., La Fé-Perdomo, I. & Ramos-Grez, J. A. Laser powder bed fusion dataset for relative density prediction of commercial metallic alloys. *Scientific Data* **12**, 375 (2025).
9. La Fé-Perdomo, I., Barrionuevo, G. O. & Ramos-Grez, J. A. *A dataset for modeling and optimizing the relative density of metallic alloys in laser-based powder bed fusion* version V2. 2024.
10. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **6**, 138 (2020).
11. Wu, Z., Ramsundar, B., Feinberg, E. N., *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
12. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **1**, 140022 (2014).