Mixing Configurations for Downstream Prediction

Juntang Wang* Hao Wu* Yihan Wang* Dongmian Zou Shixin Xu

Zu Chongzhi Center, Duke Kunshan University (* equal contribution)

ICML International Conference On Machine Learning

Motivation

- **Clustering**-based features widely used.
- Most methods choose ONE resolution γ — that is global, fixed and ad hoc.
- Real data exhibits multi-scale structure.
- Sweeping $\gamma \in [0, \infty)$ is intractable.
- Recent work introduces *configurations* — a finite set of stable partitions (see below)

Q: adaptively weight all stable partitions (configurations) for downstream prediction?

Preliminaries: Configurations

Def. 3.1: *A partition of n items is represented by an assignment vector* $\boldsymbol{\omega} \in \{1, \ldots, |\boldsymbol{\omega}|\}^n$.

Hamiltonian energy (LeCun et al., 2006):

$$H(\boldsymbol{\omega}) = h_a + \gamma h_r \tag{1}$$

$$h_a = -\sum_{k=1}^{|\boldsymbol{\omega}|} \sum_{i<j} w_{ij} \mathbf{1}_{\omega_i=\omega_j=k} \quad \text{(attraction)}$$

$$h_r = \sum_{k=1}^{|\boldsymbol{\omega}|} \left(\sum_{i<j} w_{ij} \mathbf{1}_{\omega_i=k} \right)^2 \quad \text{(repulsion)}$$

where $w_{ij} \geq 0$ are pairwise geodesic distance.

Def. 3.2: *Let* $\boldsymbol{\omega}^*(\gamma) \coloneqq \arg\min H(\boldsymbol{\omega})$. *A configuration* $\boldsymbol{\omega} = \boldsymbol{\omega}^*(\gamma)$ *for all* $\gamma \in [\gamma^-, \gamma^+]$. *Sweeping* $\gamma \in [0, \infty)$ *yields a finite set of distinct configurations* $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m\}$.

Parallel-DT (Pitsianis et al., 2023) extracts all such coarse→fine configurations in $O(n \log n)$.

Configuration-Mixed Prediction

Jointly learn predictor f_ψ and selector $\boldsymbol{w}_\theta(\boldsymbol{x})$:

$$\mathcal{L}_{\text{CMP}} \coloneqq \frac{1}{n} \sum_{t=1}^{n} \ell\big(f_\psi\big([\boldsymbol{x}_t; \boldsymbol{z}_\theta(\boldsymbol{x}_t)]\big), y_t\big) \,,$$

$$\boldsymbol{z}_\theta(\boldsymbol{x}) \coloneqq \sum_{i=1}^{m} w_{\theta,i}(\boldsymbol{x})\, \phi(\boldsymbol{\omega}_i, \boldsymbol{x}), \tag{2}$$

$$(\theta^\star, \psi^\star) \in \arg\min_{\theta,\psi} \; \mathcal{L}_{\text{CMP}}(\theta, \psi).$$

where $\boldsymbol{w}_\theta(\boldsymbol{x}) = (w_{\theta,1}(\boldsymbol{x}), \ldots, w_{\theta,m}(\boldsymbol{x}))$; ϕ encodes $\boldsymbol{x}$ under $\boldsymbol{\omega}_i$; and ℓ is a prediction loss.

We group five predictor classes by how the mixing is done:

1. $\mathcal{F}_{\text{raw}}$: $f_\psi(\boldsymbol{X})$ — uses raw features only. = Base
2. $\mathcal{F}_{\text{single}}$: $f_\psi(\boldsymbol{X}, \boldsymbol{\omega}_i)$ — augments with one configuration. = +Single
3. $\mathcal{F}_{\text{static}}$: $f_\psi(\boldsymbol{X}) + \bar{\boldsymbol{w}}^T \boldsymbol{\Omega}$ — with a global weight. = +Static (global $\bar{\boldsymbol{w}}$)
4. $\mathcal{F}_{\text{standard}}$: $f_\psi(\boldsymbol{X}, \boldsymbol{\Omega})$ — concatenates all as features. = +Config
5. $\mathcal{F}_{\text{adaptive}}$: $f_\psi(\boldsymbol{X}, \boldsymbol{w}_\theta(\boldsymbol{X})^T \boldsymbol{\Omega})$ — learns $\boldsymbol{w}_\theta(\boldsymbol{X})$. = +MixConfig

By construction, $\mathcal{F}_{\text{raw}} \subseteq \mathcal{F}_{\text{single/static}} \subseteq \mathcal{F}_{\text{standard}} \subseteq \mathcal{F}_{\text{adaptive}}$.

Prop. 3.4: *The inclusion* $\mathcal{F}_{static} \subset \mathcal{F}_{adaptive}$ *is strict. A deep* $\mathcal{F}_{standard}$ *may in principle match it.* Empirically, it does not (see Ablations).

MixConfig

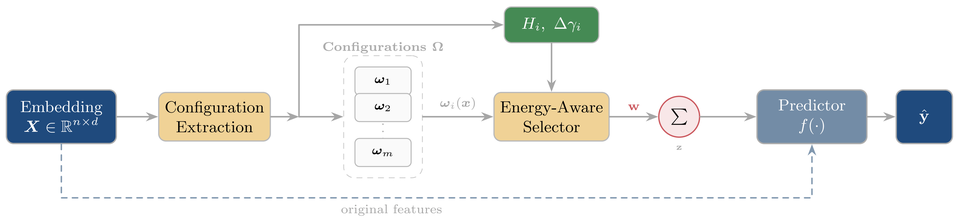

Figure 1. Frozen embedding → MixConfig (extract configs + energy-aware mix) → predictor.

The energy-aware selector learns *which resolutions matter for which samples*, leveraging stability stats as inductive bias:

$$\boldsymbol{h} = \text{MLP}_{\text{enc}}(\boldsymbol{x}) \quad \text{(sample context)} \tag{3}$$

$$\boldsymbol{c}_i = \text{Embed}_i(\boldsymbol{\omega}_i(\boldsymbol{x})) \quad \text{(cluster assignment)} \tag{4}$$

$$\boldsymbol{e}_i = [H_i, h_a^{(i)}, h_r^{(i)}, \Delta\gamma_i] \quad \text{(energy statistics)} \tag{5}$$

$$s_i = \text{MLP}_{\text{score}}([\boldsymbol{h}; \boldsymbol{c}_i; \boldsymbol{e}_i]) \quad \text{(compatibility score)} \tag{6}$$

$$w_i(\boldsymbol{x}) = \text{softmax}(s_i(\boldsymbol{x})) \qquad \boldsymbol{w}(\boldsymbol{x}) \in \Delta^{m-1} \tag{7}$$

We then form a mixed configuration representation $\boldsymbol{z}(\boldsymbol{x}) = \sum_{i=1}^{m} w_i(\boldsymbol{x})\, \boldsymbol{c}_i$ and pass the augmented features $[\boldsymbol{x}; \boldsymbol{z}(\boldsymbol{x})]$ to the downstream predictor.

Consistent Gains Across Four Domains

We evaluate MixConfig across four benchmark families. Our configuration-based methods (+Config, +MixConfig) consistently win.

Benchmark	Embedding	Base	+DeepCluster	+HDBSCAN	+Config	+MixConfig
Tabular (OpenML-CC18)						
Binary (AUC ↑)	TabPFN	.895±.004	.900**±.005	.898±.004	.903**±.004	**.907**±.003
Multi-class (Acc. ↑)	FT-Trans.	.782±.006	.779±.007	.785±.006	.789**±.005	**.793**±.004
Vision						
CIFAR-100 (Top-1 ↑)	CLIP	.742±.003	.740±.005	.746±.003	.751**±.003	**.758**±.002
ImageNet-1K (Top-1 ↑)	CLIP	.763±.002	.764±.003	.768±.003	.773**±.002	**.782**±.002
Molecular						
MolHIV (AUC ↑)	GIN	.804±.008	.801±.010	.809±.007	.814**±.007	**.819**±.005
QM9 (MAE ↓)	DimeNet++	.0112±.0002	.0114±.0004	.0109±.0002	**.0106**±.0002	**.0106**±.0001
Text						
SST-2 (Acc. ↑)	RoBERTa	.945±.003	.944±.004	.948±.003	.951**±.003	**.955**±.002
AG News (Acc. ↑)	BERT	.942±.004	.946**±.004	.944±.004	.948**±.003	**.951**±.003

Table 1. Bold = best; ** $p < 0.05$ vs Base (paired t-test, 5 seeds); all Cohen's $d > 0.8$.

Ablations

We isolate MixConfig's components and compare mixing strategies.

Variant	BBBP	CIFAR-100
MixConfig (full)	**.903**	**.758**
w/o energy features	.882	.738
w/o sample context $\boldsymbol{h}$	.890	.747
w/o cluster embeddings $\boldsymbol{c}_i$	.886	.743
Using only m' of m configs:		
$m' = 2$	.864	.716
$m' = 4$	.882	.737
$m' = 6$	.895	.751
$m' = m$ (all)	.903	.758

Table 2. Each component helps.

Mixing strategy	BBBP	CIFAR-100
Base (no configs)	.847	.742
Random partitions (same dim)	.856	.737
CSPA consensus	.862	.743
+Config (standard concat.)	.869	.751
Uniform mixing (w_i=1/m)	.873	.746
Global learnable weights ($\mathcal{F}_{\text{static}}$)	.878	.749
Stronger fusion ([$\boldsymbol{x}$; $\boldsymbol{\Omega}$]→2L-MLP)	.884	.753
+MixConfig	**.903**	**.758**

Table 3. All controls < MixConfig ⇒ energy-aware bias, not capacity.

Gains Grow as Labels Shrink

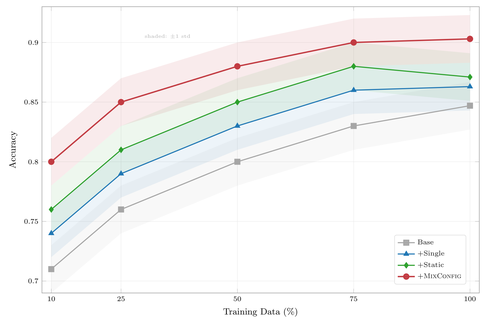

Figure 2. Performance vs. training set size.

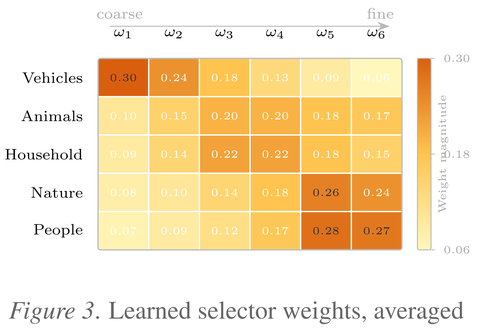

Figure 3. Learned selector weights, averaged over samples from each superclass.

Deployment

- Predictor-agnostic on frozen embeddings; d_c=32 selector MLP.
- Batch-transductive ($n \gtrsim 50$); inductive kNN fallback — no test-time extraction, retains ≥75% of gains.
- One-time extraction ∼ 150× cheaper than a resolution sweep.

Takeaways

- **Energy-aware mixing**: uniform / global $\bar{\boldsymbol{w}}$ / stronger-MLP controls all < MixConfig; adds 32 dims, beats +Config's 78.
- **Plug-and-play** on any frozen embedding & predictor.
- **Biggest gains when labels are scarce** (∼ 5× efficiency).
- **Best on** hierarchical, heterogeneous, low-data problems; **not** single-scale data or tiny batches.

Links & References

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. J. A Tutorial on Energy-Based Learning. 2006.

Pitsianis, N., Floros, D., Liu, T., and Sun, X. Parallel Clustering with Resolution Variation. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, 2023.

Code Paper