KTH ROYAL INSTITUTE
OF TECHNOLOGY

Topological regularization and relative latent representations

Alejandro García Castellanos

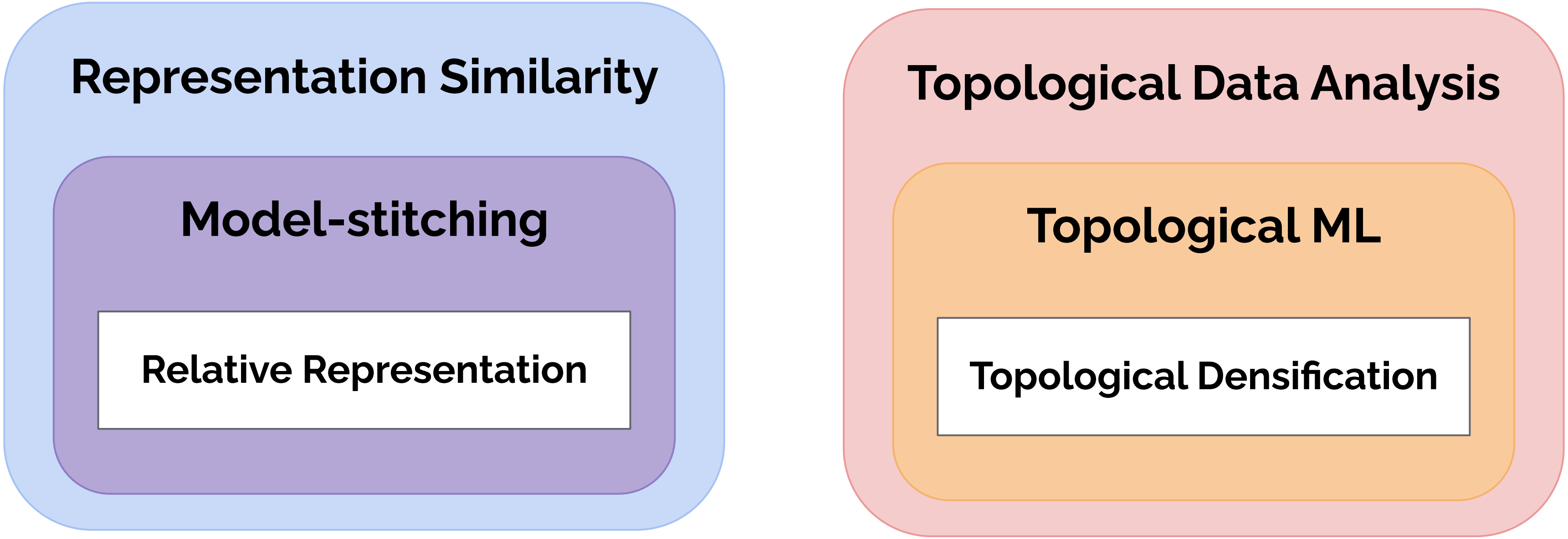
Supervisors: Martina Scolamiero, Giovanni Luca Marchetti

Examiner: Florian Pokorny

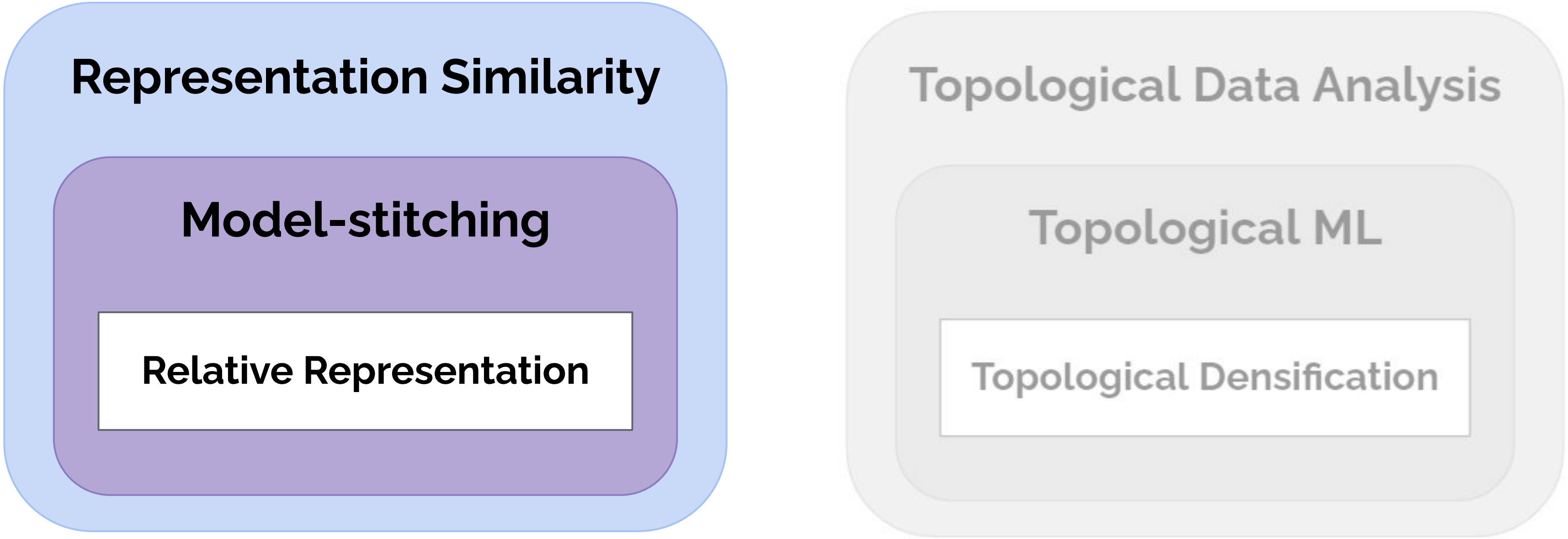
Opponent: Adhithyan Kalaivanan

Background

Overview


Representation Similarity
Model-stitching
Relative Representation
Topological Data Analysis
Topological ML
Topological Densification

Overview: Relative Latent Representations



Representation Similarity

How similar are the latent spaces between two random initializations?

Based on statistical similarity metrics:

- CCA
  - SVCCA
  - PWCCA
- **<u>CKA</u>**

“Well-performing” networks tend to have more similar representations

Wider networks with low-generalization error

ε-similar representations

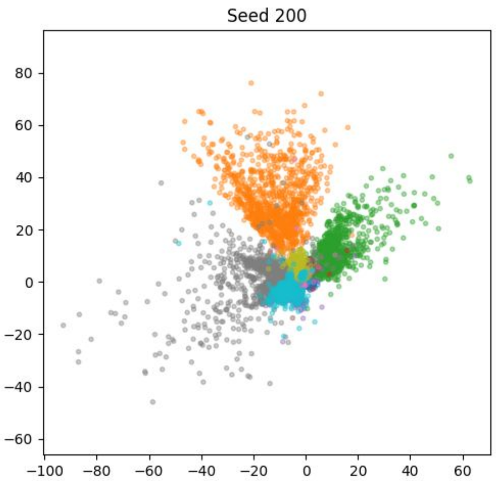
"Almost isometric up-to-scale"

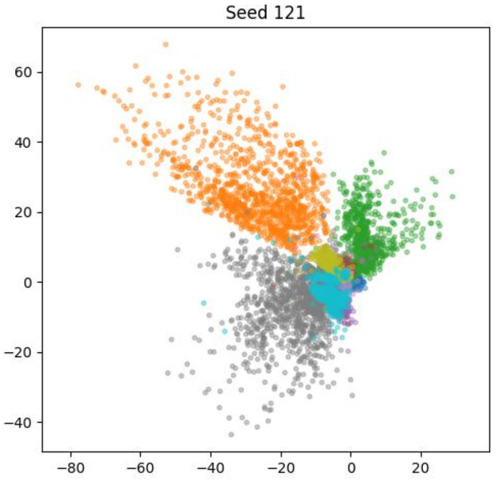
Two representations $X, Y \subseteq \mathbb{R}^n$ are ε-similar if there exist a bijection $T : X \to Y$ s.t. exists $\alpha \in \mathbb{R}^*$ for which $|d(T(x_1), T(x_2)) - \alpha \cdot d(x_1, x_2)| \leq \varepsilon$

This is mainly based on **empirical evidence**

$\Downarrow$

Need for theoretical foundation explaining the origin of the ε-similarities



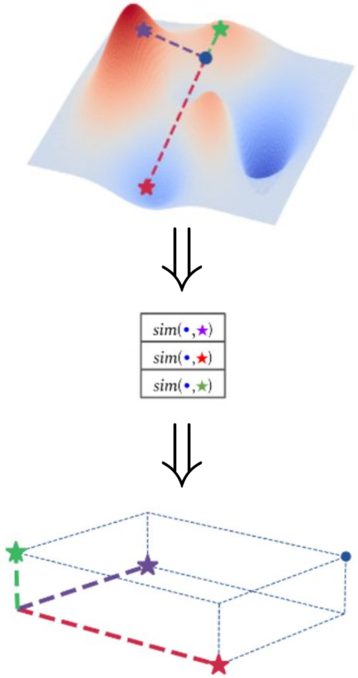


Relative representations

Let $\varphi : \mathcal{X} \to \mathcal{Z}$ the feature extractor component of your network, and $\mathcal{A} = \{a_1, ..., a_k\} \subset \mathcal{X}$ a set of points called *anchors*. Then for any similarity function sim we define the relative representation of a point $x \in S$ w.r.t. $\mathcal{A}$ as

$$(sim(\varphi(x), \varphi(a_1)), ..., sim(\varphi(x), \varphi(a_k)) \in \mathbb{R}^k.$$

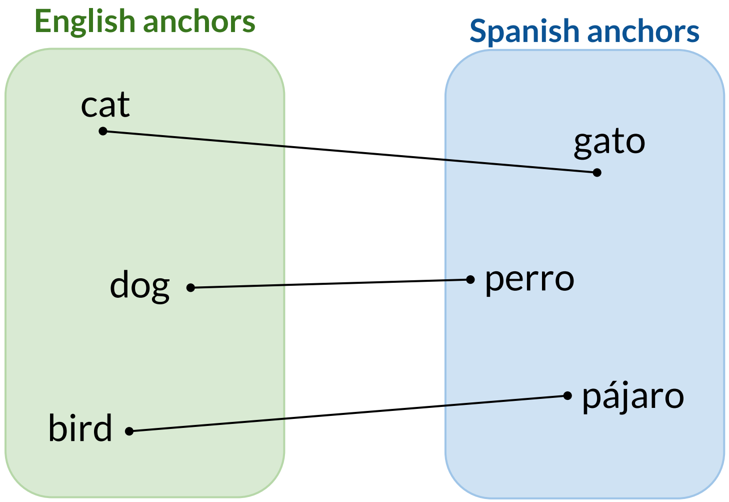
When we use the cosine similarity → we are **invariant to 0-similarities**



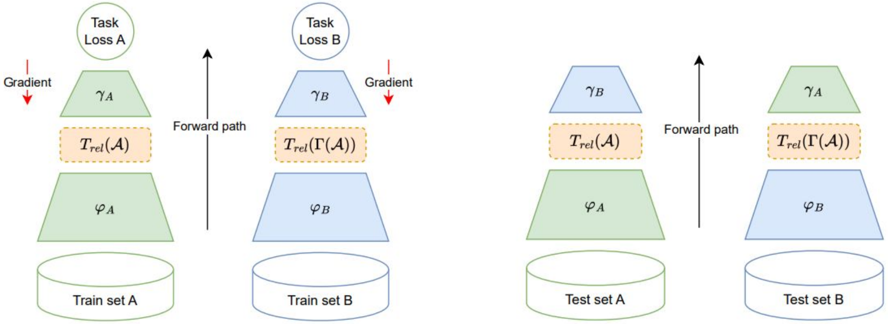
[1] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà, "Relative representations enable zero-shot latent space communication," Sep. 2022

Zero-shot cross-domain model stitching

Parallel anchors



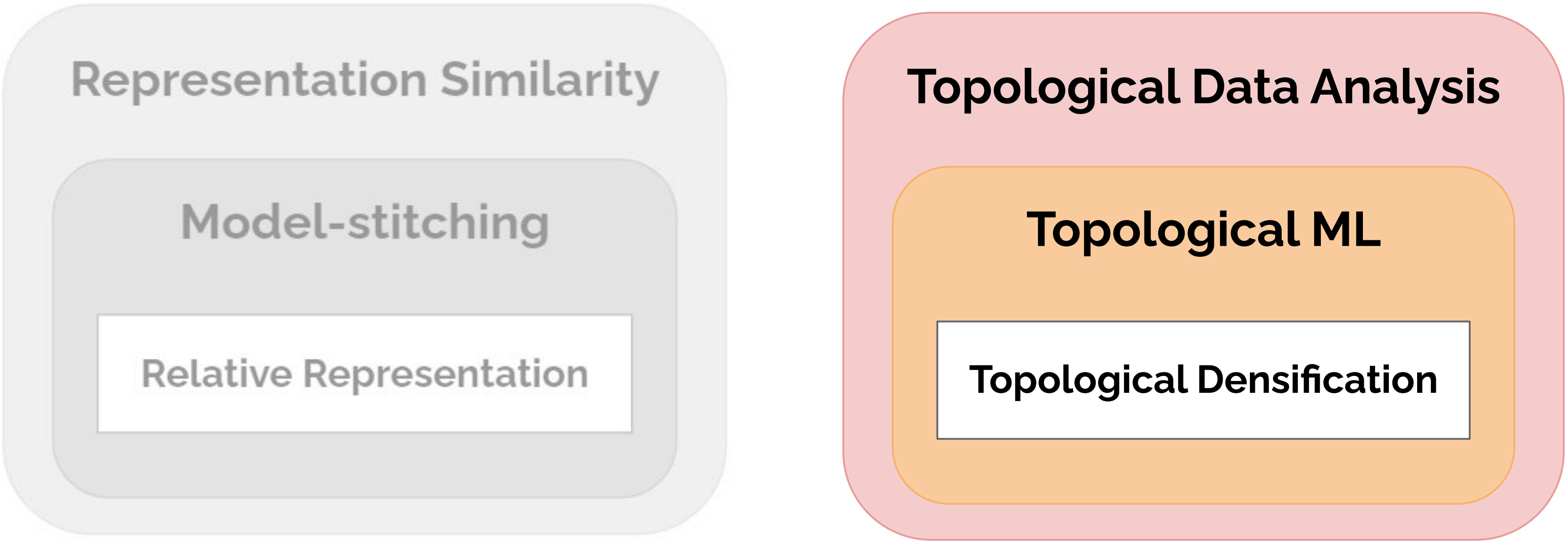
Original training and testing setup



(a) Train w/ relative transformations.

(b) Model stitching

Overview: Topological Densification



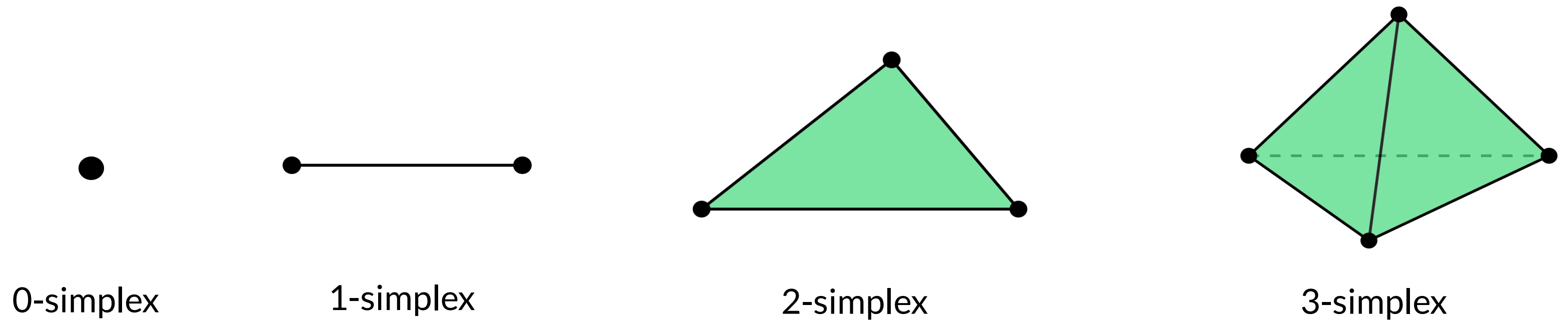
Topological data analysis

Topological data analysis (TDA) is an approach for the **analysis of the qualitative geometric properties** of datasets using topology techniques.

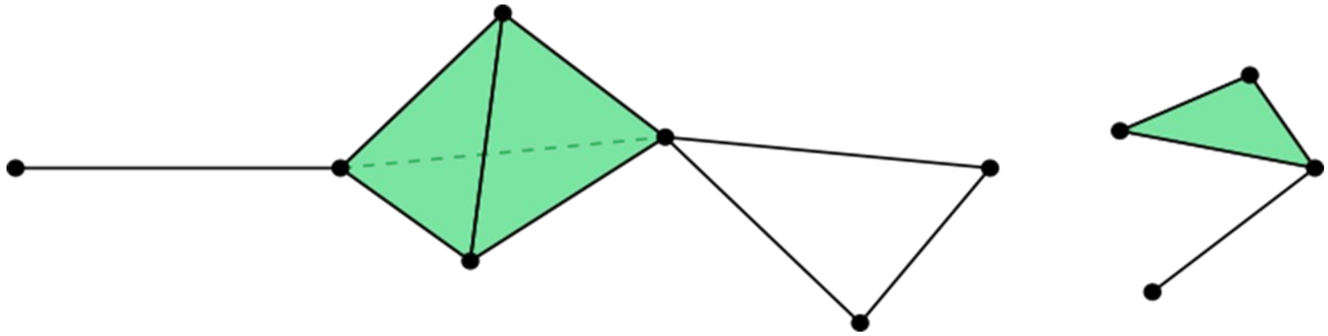
- Geometric qualitative properties: **connected components**, holes, cavities…
- **Advantages:**
  - Have a sense of the shape of higher-dimensional data that cannot be directly visualized.
  - Results are stable against noise.

Simplicial complex

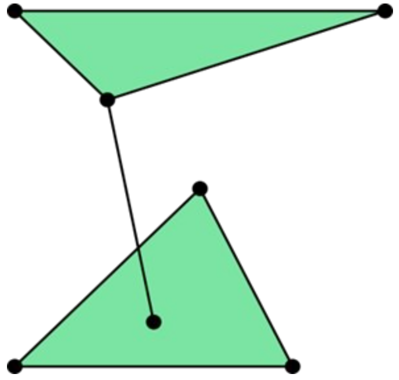
- Def: An k-simplex σ in $\mathbb{R}^d$ with $d \geq k$ is a k-dimensional triangle.



- Def: A *simplicial complex* is a finite collection of simplices K that satisfies that the (non-empty) intersections between the simplices are simplicies of lesser dimension, belonging to the simplicial complex K.



Is a simplicial complex

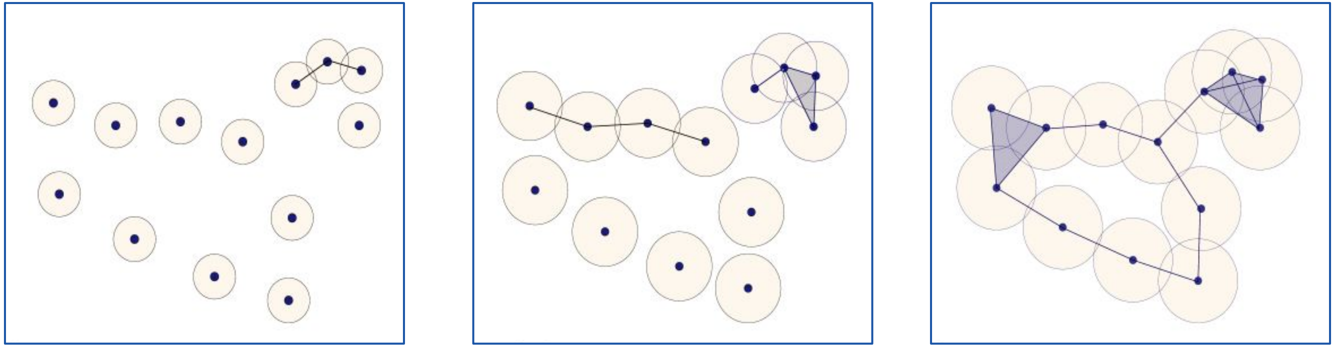


Not a simplicial complex

Vietoris-Rips complex

Definition Let $X \subset \mathbb{R}^d$ be a finite set of points. We call *Vietoris-Rips* complex of X of radius r to the abstract simplicial complex

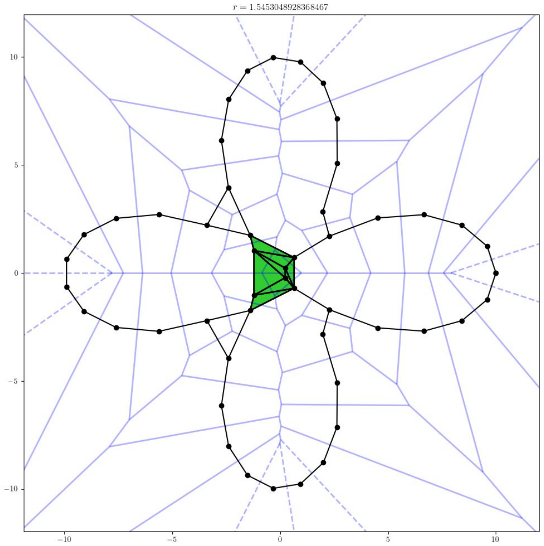
$$\begin{aligned} \mathrm{VR}(X, r) &= \{\sigma \subseteq X \mid \operatorname{diam} \sigma \leq r\} \\ &= \{\{x_0, \ldots, x_n\} \subseteq X \mid d(x_i, x_j) \leq r \ \forall i, j\} . \end{aligned}$$



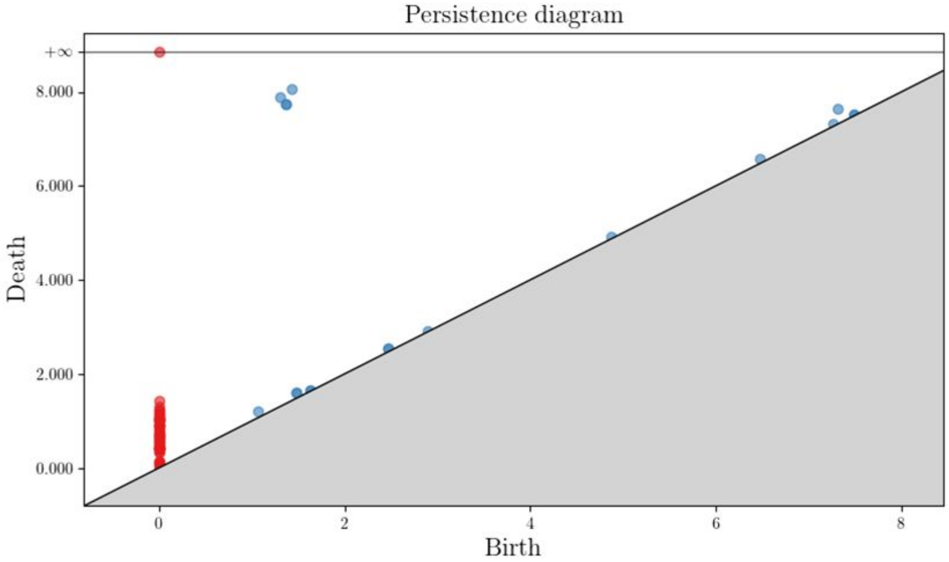
[2] Choudhary, Aruni - https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/26911, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=130411727

Persistent Homology

Each point (a_i, a_j) of the *Persistence Diagram* represents an l-dimensional hole that is born at "instant" a_i and dies at a_j

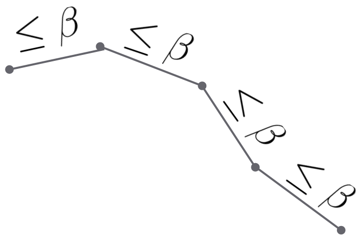


$\Rightarrow$



Topological Densification

High likelihood of β-connected → **Mass attract mass**

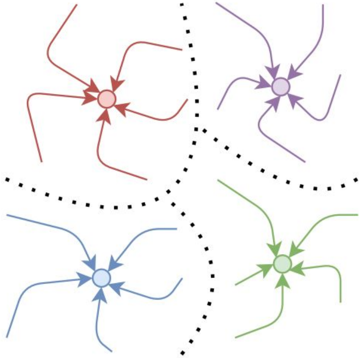


- Equal to having all H_0(VR) homology death-times in (0, β)
- Can be enforced with **regularization**:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{\beta}, \ \lambda > 0$$

where,

$$\mathcal{L}_{\beta} = \sum_{i=1}^{n} \sum_{d \in \dagger(\mathcal{B}_i)} |d - \beta|$$



- Condensate, for each class, its push-forward distributions inside their decision boundary
- **Reduce generalization error**

Latent space similarity study

Theoretical: Intertwiner Groups

Let $G_{\sigma_{n_i}}$ denote the set of invertible linear transformations that exhibit equivalent transformations before and after the nonlinear layer σ_{n_i}, i.e.,

$$G_{\sigma_{n_i}} \equiv \{A \in GL_{n_i}(\mathbb{R}) \mid \exists B \in GL_{n_i}(\mathbb{R}) \text{ s.t. } \sigma_{n_i} \circ A = B \circ \sigma_{n_i}\}$$

- All elements are of the form PD where $P \in \Sigma_n$ and D is diagonal
- Symmetries in weight space

$\Downarrow$

Symmetries in latent representations

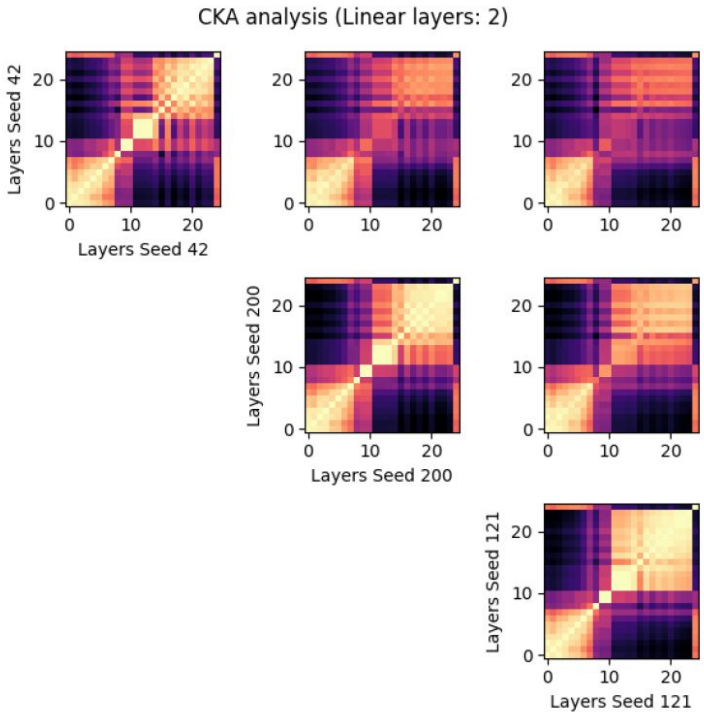
Robust relative transformation

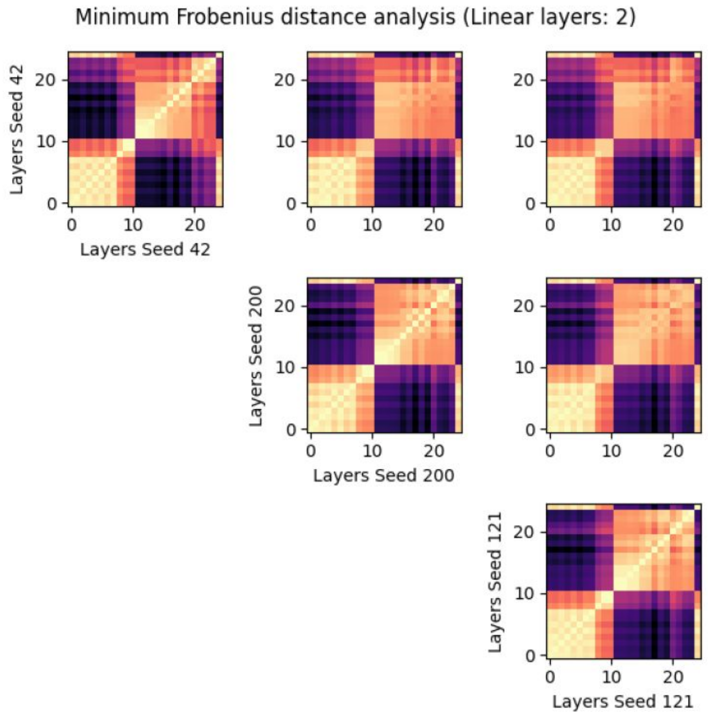
We apply BatchNorm without the learnable affine transformation before computing the cosine sim.

$\Downarrow$

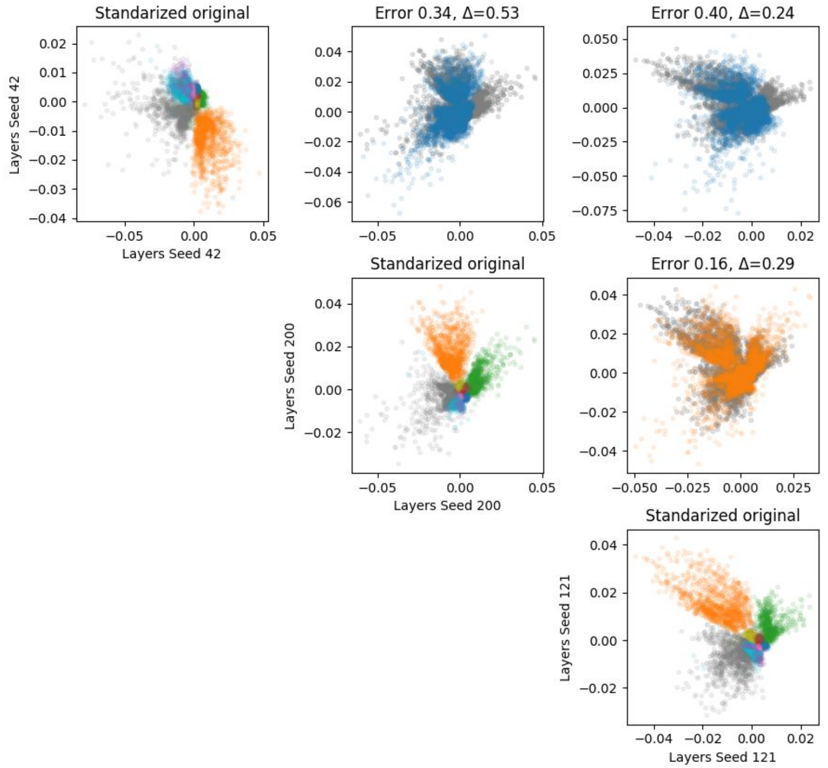
Invariant to intertwiner group actions and 0-similarities

Numerical analysis: 2-dimensional autoencoder

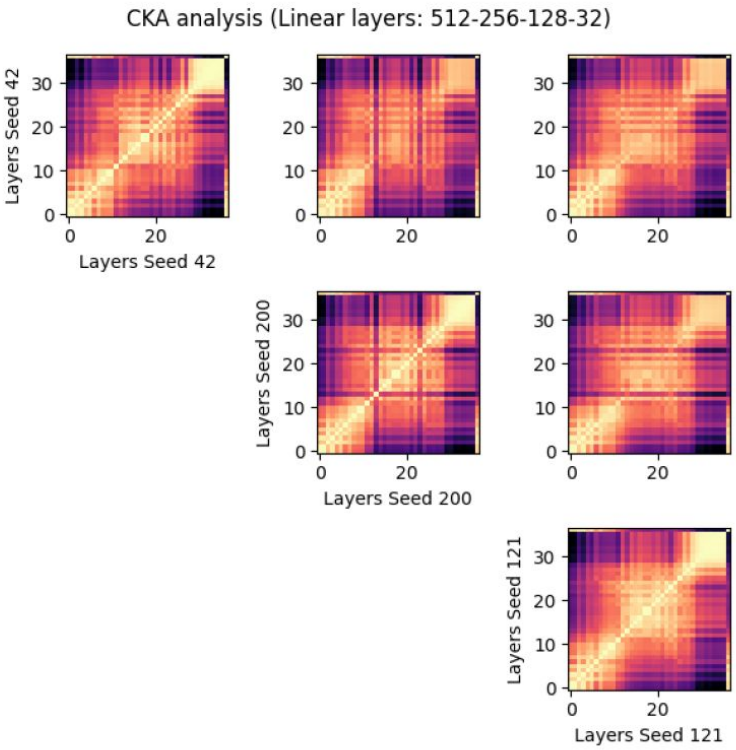


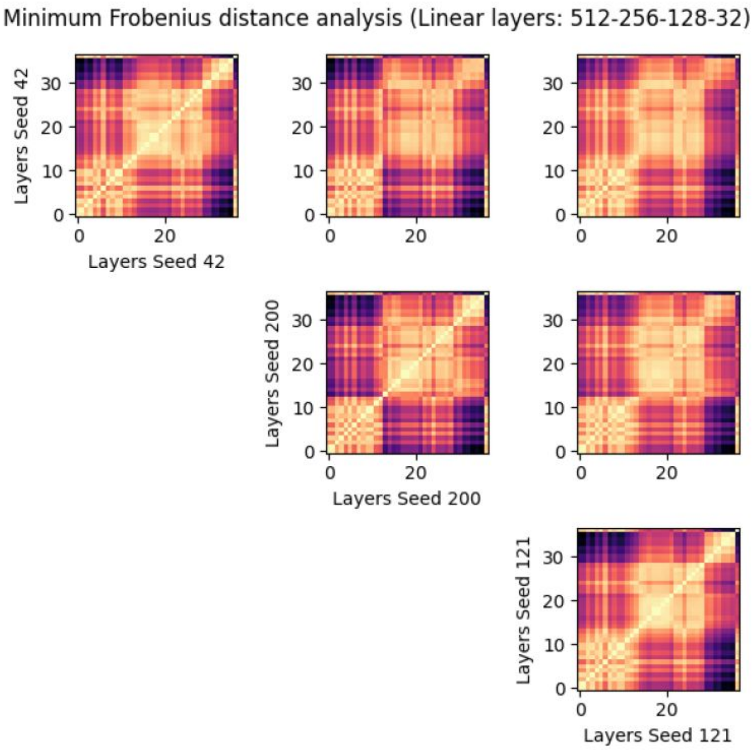


Procrustes analysis: 2-dimensional autoencoder

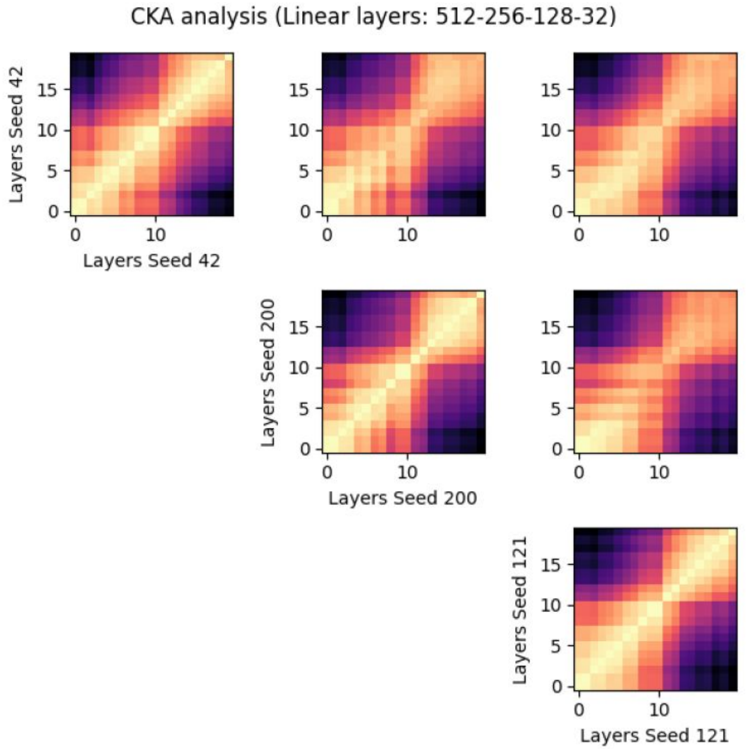


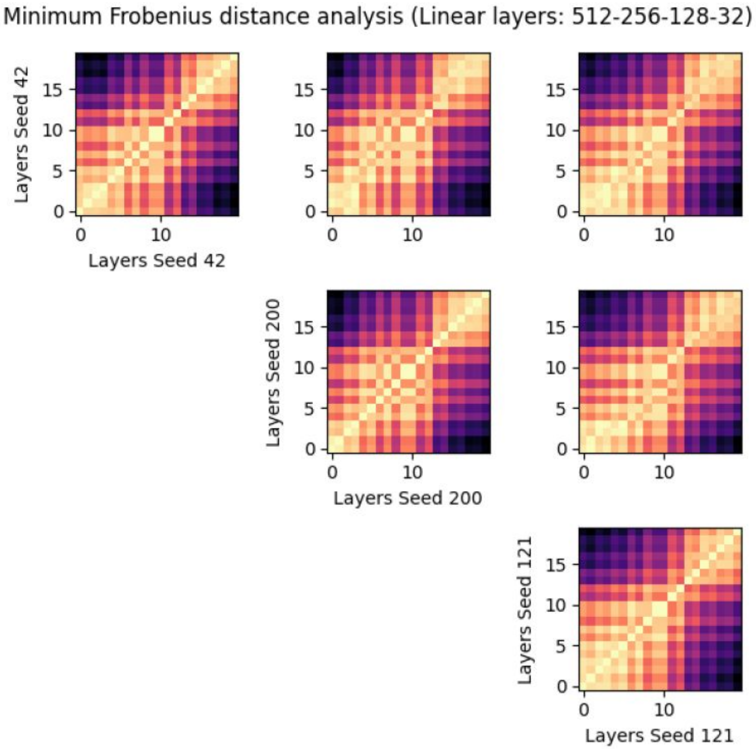
Numerical analysis: 32-dimensional autoencoder





Numerical analysis: classifier



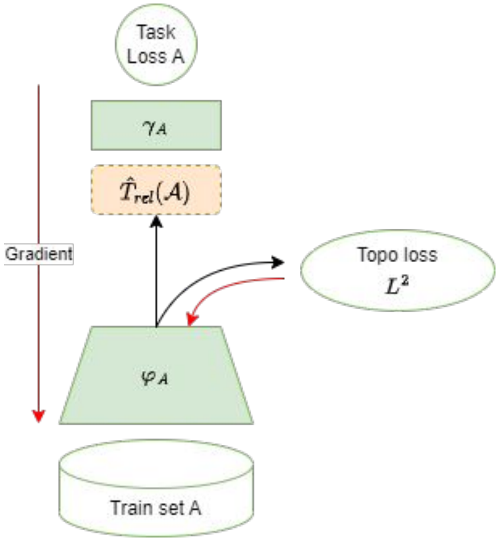


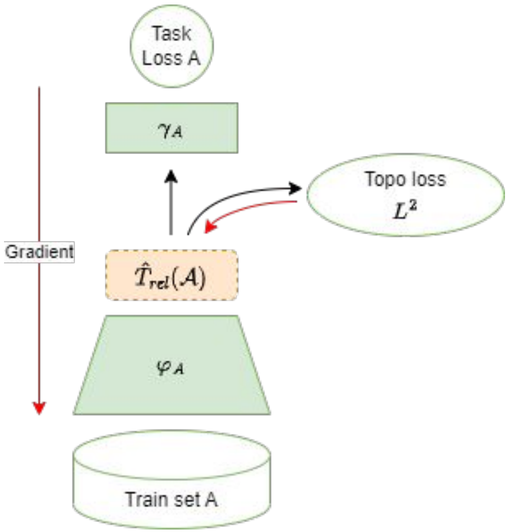
Cross-domain model-stitching analysis

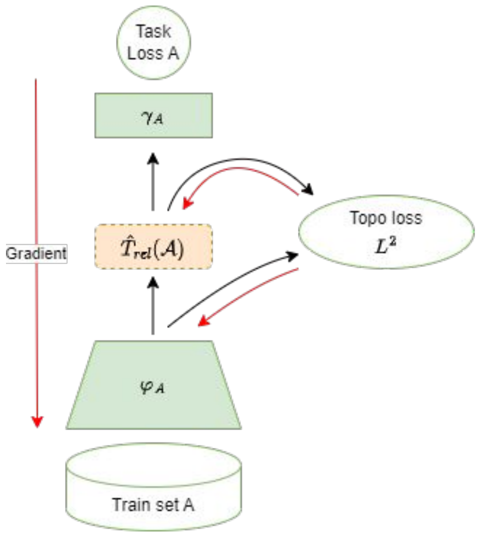
Multilingual model-stitching setup

Investigate the impact of topological densification on zero-shot stitching performance while using relative representations

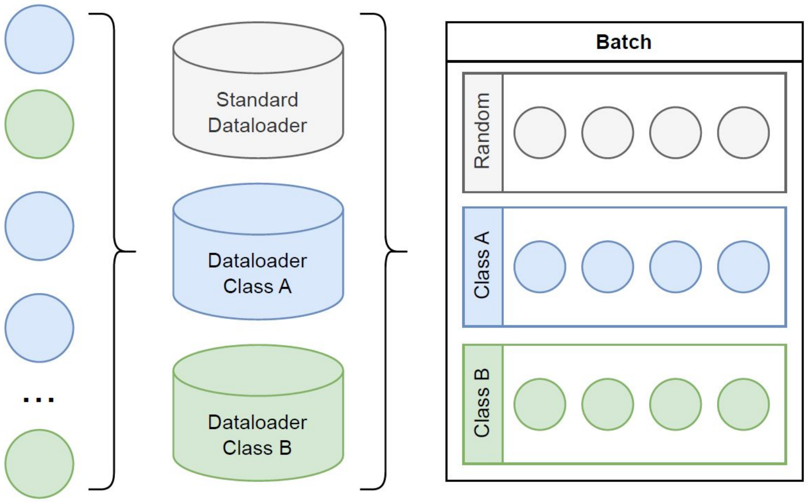
Pre-relative	Post-relative	Both







Topological densification dataloader



Debiasing trick:

1. **Freeze Linear and LayerNorm** modules and set **BatchNorm1d and LayerNorm to training** mode
2. Pass the “random” mini-batch
3. **Unfreeze Linear and LayerNorm** modules and set **BatchNorm1d and LayerNorm to eval** mode
4. Pass the remaining mini-batches

Baselines

Full-finetune

- **Relative:** better overall
- **Absolute:**
  - Better non-stitching
  - Worse stitching

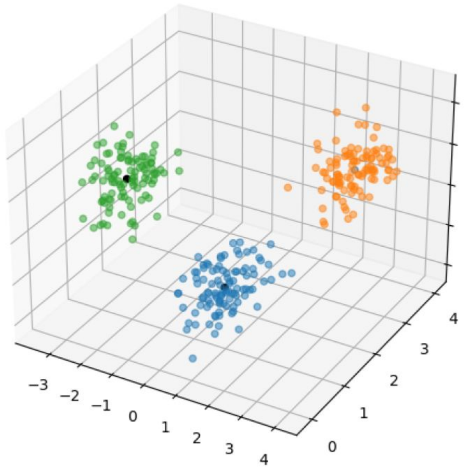
Biased dataloader + Debiasing trick

- Slightly worse results
- Enables topological regularization

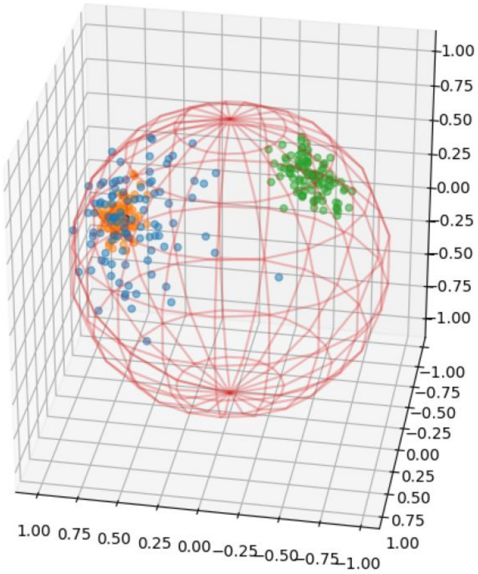
		Absolute			Relative		
Decoder	Encoder	Acc $\times$ 100	FScore $\times$ 100	MAE $\times$ 100	Acc $\times$ 100	FScore $\times$ 100	MAE $\times$ 100
en	en	59.08 ± 0.20	59.08 ± 0.85	48.47 ± 0.64	61.30 ± 0.28	60.84 ± 0.77	44.87 ± 0.92
	fr	35.06 ± 4.36	31.39 ± 4.62	101.75 ± 4.26	48.48 ± 0.08	48.74 ± 0.20	59.26 ± 0.37
fr	en	27.04 ± 6.14	25.86 ± 5.75	115.04 ± 9.79	60.87 ± 1.15	60.25 ± 1.63	45.08 ± 1.87
	fr	48.74 ± 0.62	48.99 ± 0.06	62.53 ± 0.92	49.37 ± 0.30	50.07 ± 0.19	58.24 ± 0.79

Pre-relative topological densification

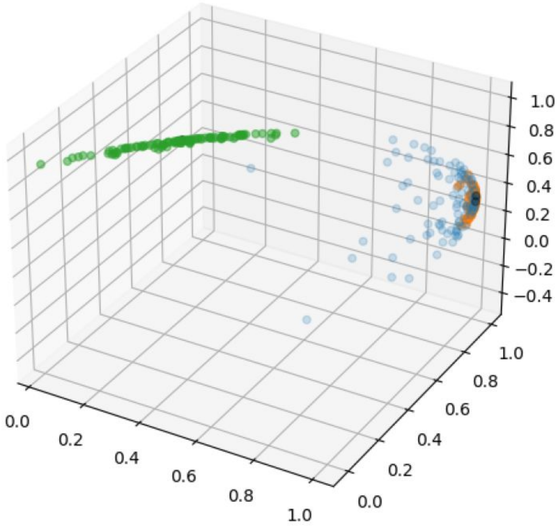
The relative transformation is not always cluster-preserving



$\Rightarrow$

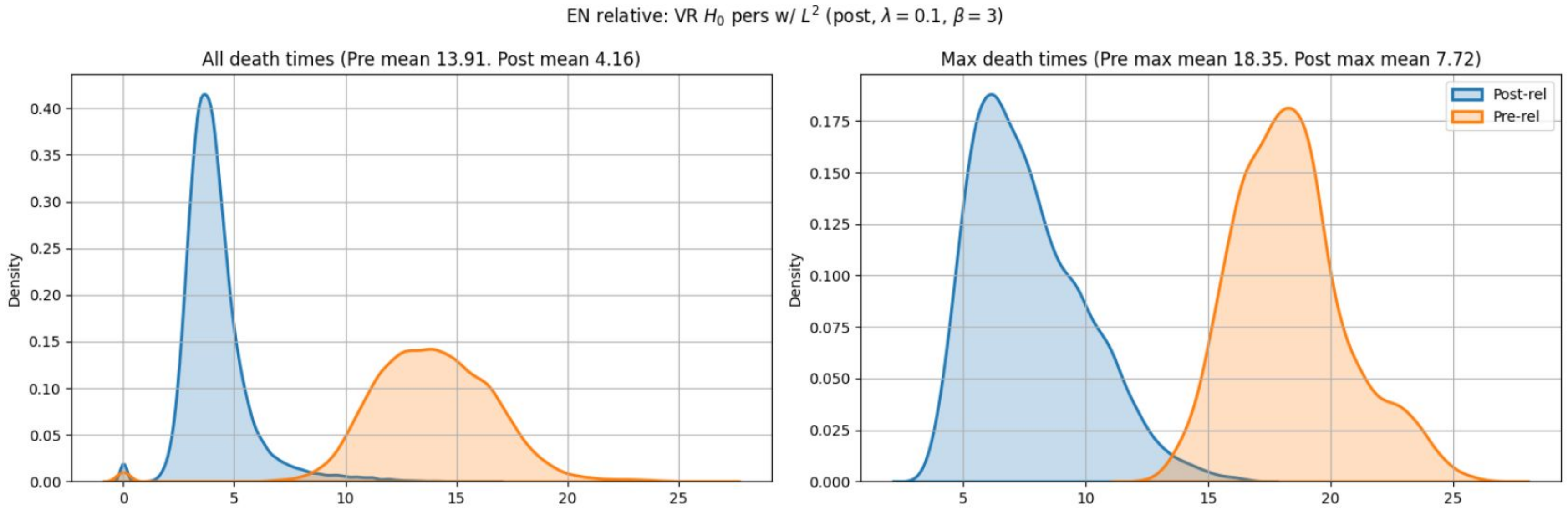


$\Rightarrow$



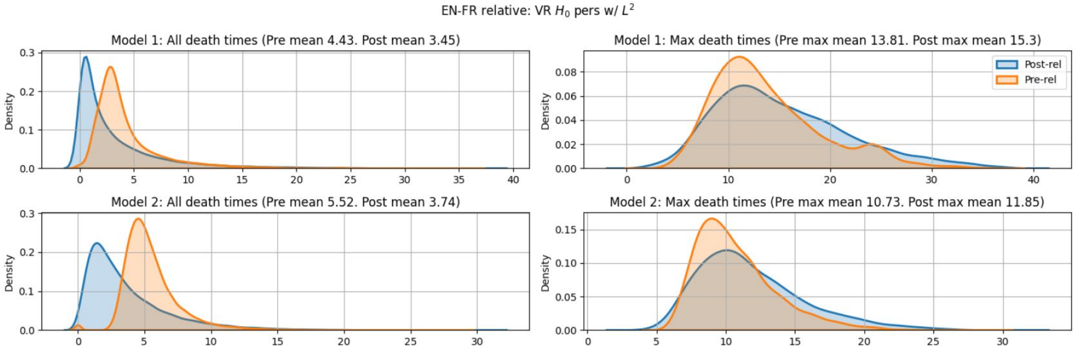
Post-relative topological densification

High mismatch of H_0 homology → Potential information bottleneck

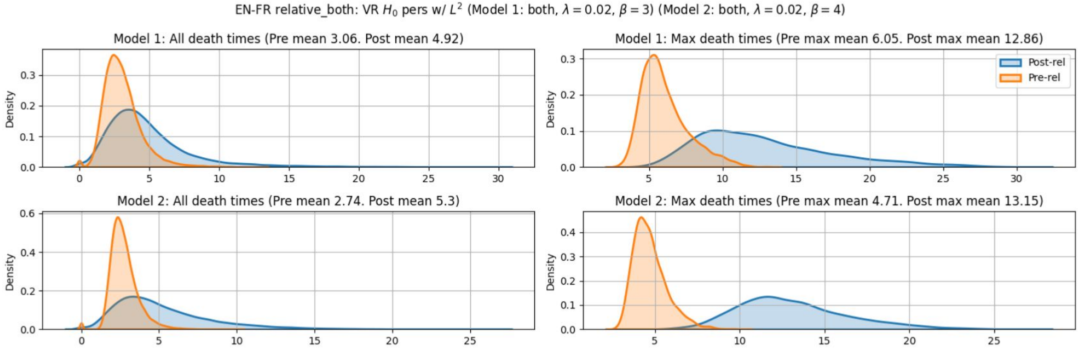


Both pre and post-relative topological densification

Vanilla



Topological densified



Topological densification: results

Vanilla

Decoder	Encoder	Absolute Acc × 100	Absolute FScore × 100	Absolute MAE × 100	Relative Acc × 100	Relative FScore × 100	Relative MAE × 100
en	en	59.08 ± 0.20	59.08 ± 0.85	48.47 ± 0.64	61.30 ± 0.28	60.84 ± 0.77	44.87 ± 0.92
	fr	35.06 ± 4.36	31.39 ± 4.62	101.75 ± 4.26	48.48 ± 0.08	48.74 ± 0.20	59.26 ± 0.37
fr	en	27.04 ± 6.14	25.86 ± 5.75	115.04 ± 9.79	60.87 ± 1.15	60.25 ± 1.63	45.08 ± 1.87
	fr	48.74 ± 0.62	48.99 ± 0.06	62.53 ± 0.92	49.37 ± 0.30	50.07 ± 0.19	58.24 ± 0.79

Topological densified

Decoder	Encoder	Absolute Acc × 100	Absolute FScore × 100	Absolute MAE × 100	Relative Acc × 100	Relative FScore × 100	Relative MAE × 100
en	en	60.20 ± 0.88	59.69 ± 0.37	46.33 ± 0.47	61.25 ± 0.24	61.37 ± 0.07	44.50 ± 0.17
	fr	30.04 ± 0.93	18.56 ± 1.73	121.52 ± 16.07	50.14 ± 0.76	50.55 ± 0.50	58.81 ± 0.16
fr	en	41.01 ± 5.53	29.78 ± 11.70	87.95 ± 7.62	60.49 ± 0.78	60.90 ± 0.54	44.96 ± 0.34
	fr	51.06 ± 0.00	51.81 ± 0.04	56.63 ± 0.01	51.27 ± 0.01	51.71 ± 0.19	57.94 ± 0.74

Topological densification: extra

Having the same densification parameter can benefit the stitching performance

Decoder	Encoder	Relative Acc × 100	Relative FScore × 100	Relative MAE × 100
en	en	61.25 ± 0.24	61.37 ± 0.07	44.50 ± 0.17
	fr	$\mathbf{50.90 \pm 0.65}$	$\mathbf{51.50 \pm 0.66}$	$\mathbf{57.27 \pm 0.07}$
fr	en	$\mathbf{60.87 \pm 0.95}$	$\mathbf{61.27 \pm 0.77}$	$\mathbf{44.56 \pm 0.71}$
	fr	50.11 ± 0.38	50.58 ± 0.79	57.78 ± 0.14

L^∞ metric for VR filtration $\Rightarrow$ β parameter relates to the optimal spread of the clusters in terms of angle $\Rightarrow$ Helps hyperparameter tuning

Future work

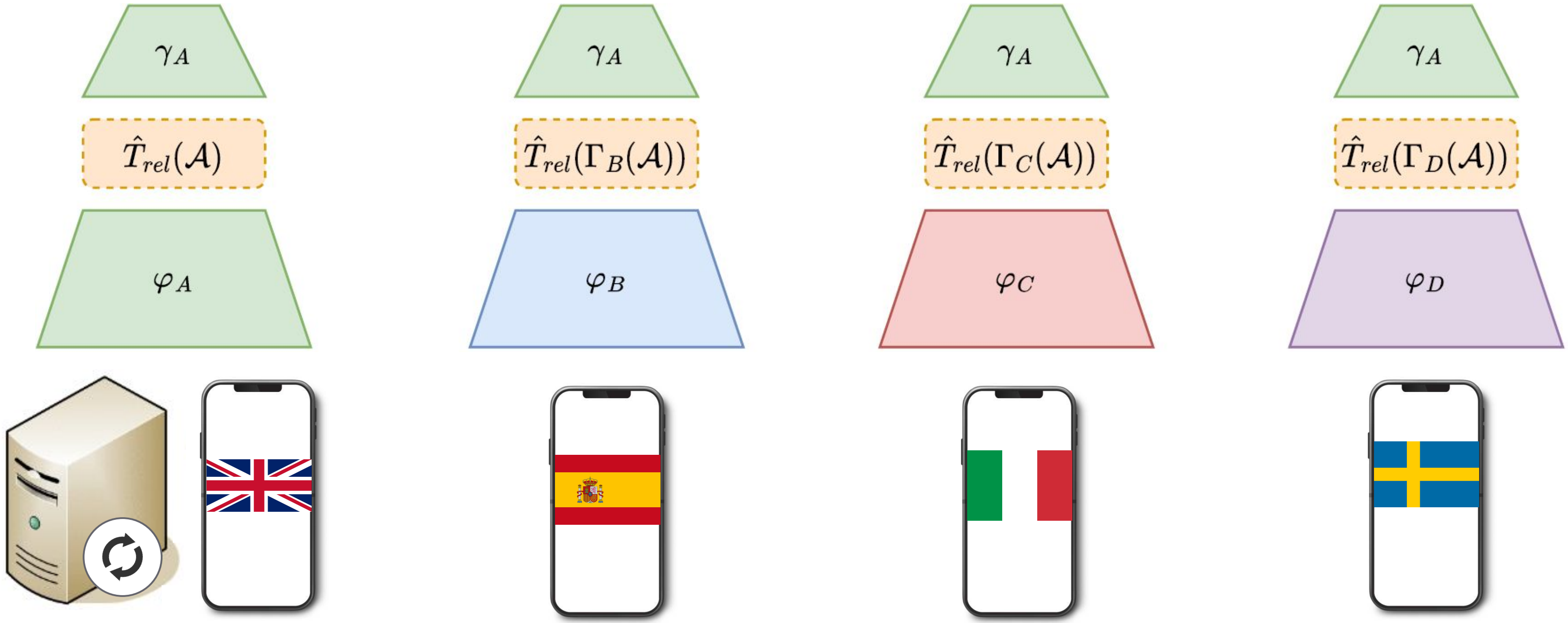
Future work

- Investigation of **alternative simplicial complex** constructions: *Lazy witness complex*
- Analysis of **representation similarity in multilingual model stitching**: *CKA analysis*
- Testing topological regularization on large models with **increased GPU VRAM**
- Exploring **other modalities:** *Image-Text*
- Exploring **higher dimensional homology**

Thank you!

Extra

Potential use case

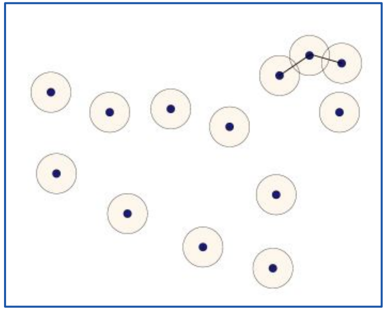


Simplicial homology

- Algebraic formalism that will allow us to count:
  - **Connected components.**
  - Holes.
  - Cavities.
  - Etc.

- Def: Let X be a geometric object, we define $\beta_i(X)$, the *i-th Betti number of X*, as the **number of i-dimensional holes of X**.

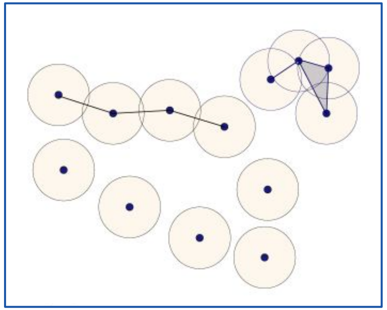
- It will allow us to calculate the Betti numbers of a simplicial complex using linear algebra.



$\beta_0(K) = 11$ *(Connected comp.)*

$\beta_1(K) = 0$ *(Holes)*

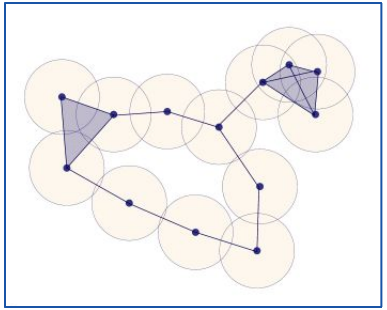
$\beta_2(K) = 0$ *(Cavities)*



$\beta_0(K) = 7$ *(Connected comp.)*

$\beta_1(K) = 0$ *(Holes)*

$\beta_2(K) = 0$ *(Cavities)*



$\beta_0(K) = 1$ *(Connected comp.)*

$\beta_1(K) = 1$ *(Holes)*

$\beta_2(K) = 0$ *(Cavities)*

Intertwiner Groups: properties

Symmetries in weight space → Symmetries in latent representations

Proposition *Suppose* $A_i \in G_{\sigma_{n_i}}$ *for* $1 \leq i \leq k-1$, *and let*

$$\widetilde{W} = (A_1 W_1, A_1 b_1, A_2 W_2 \phi_\sigma(A_1^{-1}), A_2 b_2, \ldots, W_k \phi_\sigma(A_{k-1}^{-1}), b_k)$$

Then, as functions, for each m

$$f_{\leq m}(x, \widetilde{W}) = \phi_\sigma(A_m) \circ f_{\leq m}(x, W),$$
$$f_{>m}(x, \widetilde{W}) = f_{>m}(x, W) \circ \phi_\sigma(A_m)^{-1},$$

where $f_{\leq m}$ *and* $f_{>m}$ *represent the truncations of the network before and after layer* m, *respectively. In particular,* $f(x, \widetilde{W}) = f(x, W)$ *for all* $x \in \mathbb{R}^{n_0}$.

Robust relative transformation

Definition — Let $\varphi : \mathcal{X} \rightarrow \mathcal{Z} = \mathbb{R}^m$ be our encoder, and $\mathbb{A} \in \mathbb{R}^{d\times k}$, $\mathbb{B} \in \mathbb{R}^{d\times n}$ the matrix representation of $\mathcal{A}$ and $\mathcal{B}$. Then, the *robust relative representation* of $\mathcal{B} \subset \mathcal{X}$ w.r.t. $\mathcal{A}$ is

$$\hat{T}_\varphi(\mathcal{B}, \mathcal{A}) = \left(\widehat{\varphi(\mathbb{A})} D_{\mathbb{A}}\right)^T \left(\widehat{\varphi(\mathbb{B})} D_{\mathbb{B}}\right) \in \mathbb{R}^{k\times n},$$

where

$$D_{\mathbb{A}} = \mathrm{Diag}\left(\frac{1}{\sum_{i=1}^m \widehat{\varphi(\mathbb{A})}_{i,1}^2}, \ldots, \frac{1}{\sum_{i=1}^m \widehat{\varphi(\mathbb{A})}_{i,k}^2}\right),$$
$$D_{\mathbb{B}} = \mathrm{Diag}\left(\frac{1}{\sum_{i=1}^m \widehat{\varphi(\mathbb{B})}_{i,1}^2}, \ldots, \frac{1}{\sum_{i=1}^m \widehat{\varphi(\mathbb{B})}_{i,n}^2}\right),$$

and $\widehat{\varphi(\mathbb{A})}$ and $\widehat{\varphi(\mathbb{B})}$ represent the respective BatchNorm mean and variance standardizations of the anchor and batch images (without the learnable affine transformation). When the batch and the encoder are implied, we can denote this transformation by $\hat{T}_{rel}(\mathcal{A})$.

Numerical similarity metrics: formulas

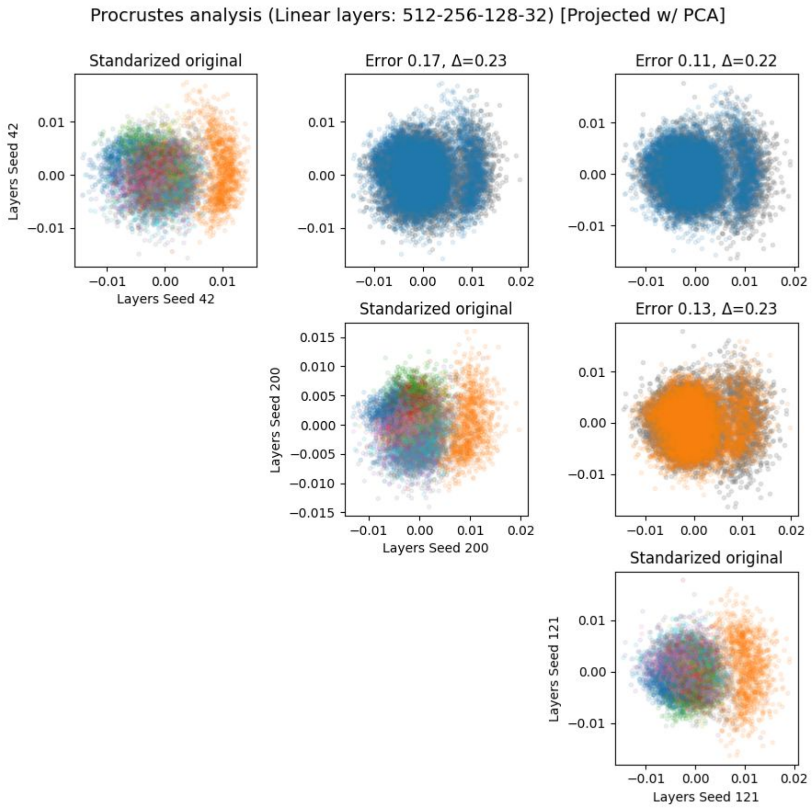
$$\mathrm{CKA}(X,Y) = \frac{\left\|\Sigma_{X,Y}\right\|_F^2}{\sqrt{\left\|\Sigma_{X,X}\right\|_F^2 \cdot \left\|\Sigma_{Y,Y}\right\|_F^2}}$$

where Σ represents the covariance matrix

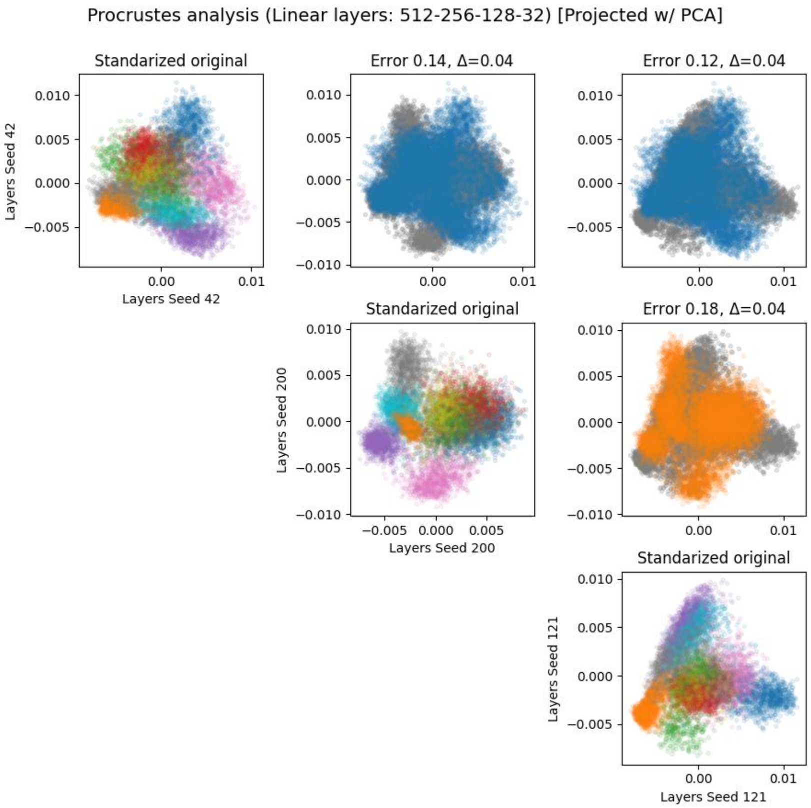
$$\mathrm{minFrob}(A,B) = \min_{P \in \Pi} \left\| \frac{A}{\|A\|_F} - P \frac{B}{\|B\|_F} \right\|_F$$

where A and B are the distance matrices of X and Y

Procrustes analysis: 32-dimensional autoencoder



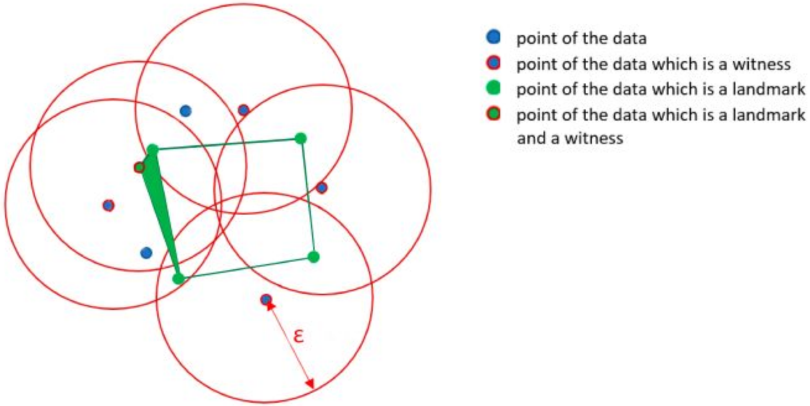
Procrustes analysis: classifier



Lazy Witness complex

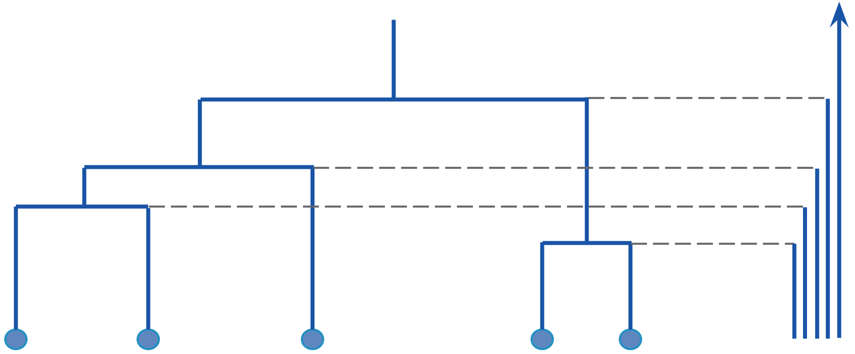
Definition —— (Nested family of witness complexes [10]). Let $(\mathcal{X}, d)$ be a metric space, $X \subset \mathcal{X}$ be a dataset, $L = \{l_0, \ldots, l_n\} \subseteq X$ be a set of landmark points, and $\varepsilon > 0$. Then the k-simplex $\sigma = \{u_1, \ldots, u_k\}$ with $u_i \in L$ belongs to the *Lazy Witness complex* $\mathrm{W}_\varepsilon(X, L)$ iff all its faces belong to $\mathrm{W}_\varepsilon(X, L)$ and there is a witness $x \in X$, such that:

$$\max\{d(u_i, x) \mid u_i \in \{u_1, \ldots, u_k\}\} \leq \varepsilon\,.$$



Exploring higher dimensional homology

Single Linkage Hierarchical Clustering ↔ $H_0(VR)$



Controlling $H_0(VR)$ → Topological densification

What beneficial properties for classification can we obtain by controlling $H_n(VR)$ for $n>0$?