

---

# Supplementary material

---

Alessandro Benfenati<sup>1a</sup> Alfio Ferrara<sup>1b</sup> Alessio Marta<sup>1c</sup> Davide Riva<sup>2d</sup> Elisabetta Rocchetti<sup>1b\*</sup>

<sup>a</sup>Department of Environmental Science and Policy, <sup>b</sup>Department of Computer Science,

<sup>c</sup>Department of Mathematics, <sup>d</sup>Department of Control and Computer Engineering

<sup>1</sup>Università degli Studi di Milano, <sup>2</sup>Politecnico di Torino

{alessandro.benfenati, alfio.ferrara}@unimi.it

{alessio.marta, elisabetta.rocchetti}@unimi.it

{davide.riva}@polito.it

\* Corresponding author

## Outline

Here we provide the outline of the content included in this Supplementary Material. In Section 1 we provide the definitions necessary for the assumptions and proofs given respectively in Section 2 and 3. Section 4 gives the details of the adopted models, Section 5 presents additional results relative to the exploration analysis and Section 6 with respect to the interpretation analysis. Finally, Section 7 includes extended results related to explored volumes, and Section 8 gives the full picture on computation times.

## 1 Definitions

First we recall our geometric definitions of neural network and layer.

**Definition 1** (Neural Network). *A neural network is a sequence of  $\mathcal{C}^1$  maps  $\Lambda_i$  between manifolds of the form:*

$$M_0 \xrightarrow{\Lambda_1} M_1 \xrightarrow{\Lambda_2} M_2 \xrightarrow{\Lambda_4} \dots \xrightarrow{\Lambda_{n-1}} M_{n-1} \xrightarrow{\Lambda_n} M_n \quad (1)$$

*We call  $M_0$  the input manifold and  $M_n$  the output manifold. All the other manifolds of the sequence are called representation manifolds. The maps  $\Lambda_i$  are the layers of the neural network. We denote with  $\mathcal{N}_{(i)} = \Lambda_n \circ \dots \circ \Lambda_i : M_i \rightarrow M_n$  the mapping from the  $i$ -th representation layer to the output layer.*

**Definition 2** (Smooth layer). *A map  $\Lambda_i : M_{i-1} \rightarrow M_i$  is called a smooth layer if it is the restriction to  $M_{i-1}$  of a function  $\bar{\Lambda}^{(i)}(x) : \mathbb{R}^{d_{i-1}} \rightarrow \mathbb{R}^{d_i}$  of the form*

$$\bar{\Lambda}_\alpha^{(i)}(x) = F_\alpha^{(i)} \left( \sum_{\beta} A_{\alpha\beta}^{(i)} x_\beta + b_\alpha^{(i)} \right) \quad (2)$$

*for  $i = 1, \dots, n$ ,  $x \in \mathbb{R}^{d_i}$ ,  $b^{(i)} \in \mathbb{R}^{d_i}$  and  $A^{(i)} \in \mathbb{R}^{d_i \times d_{i-1}}$ , with  $F^{(i)} : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$  a diffeomorphism.*

We also need some standard definitions in differential geometry [4].

**Definition 3** (Submersion). *Let  $f : M \rightarrow N$  be a smooth map between manifolds. Then  $f$  is a submersion if, in any chart, the Jacobian  $J_f$  has rank  $\dim(N)$ .*

**Definition 4** (Embedding). *Let  $f : M \rightarrow N$  be a smooth map between manifolds.  $f$  is an embedding if its differential is everywhere injective and if it is an homeomorphism with its image. In other words,  $f$  is a diffeomorphism with its image.*

**Definition 5** (Distribution). A distribution  $\mathcal{D}$  of dimension  $k$  over a  $m$ -dimensional manifold  $M$  is a collection of  $k$  smooth vector fields  $v_1, \dots, v_k$  such that  $(v_1)_p, \dots, (v_k)_p$  form a basis of a vector subspace of dimension  $k$  in  $T_p M$  for every  $p \in M$ .

**Definition 6** (Integrable distribution). A distribution  $\mathcal{D}$  of dimension  $k$  is an integrable distribution if there exist a manifold  $M$  of dimension  $m \geq k$  such that the collection of  $k$  smooth vector fields  $v_1, \dots, v_k$  are generating a vector space of dimension  $k$  over  $T_p M$  for all  $p \in M$ .

**Definition 7** (Trivial fiber bundle). A trivial fiber bundle is a structure  $(E, B, \pi, F)$ , where  $E, B$  and  $F$  are topological spaces with  $E = B \times F$  and the map  $\pi : E \rightarrow B$  is the projection of  $B \times F$  on  $B$ . The space  $F$  is called typical fiber. In the case  $F$  is a vector space, then  $(E, B, \pi, F)$  is called a trivial vector bundle.

**Definition 8** (Vertical and horizontal spaces). Let  $(E, M, \pi, F)$  be a vector bundle over a manifold  $M$ . Then the vertical space  $\mathcal{V}_p E$  at  $p \in E$  is the vector space  $\mathcal{V}_p E = \text{Ker}(d_p \pi) \subset T_p E$ . The horizontal space  $\mathcal{H}_p E$  is a choice of a subspace of  $T_p E$  such that  $T_p E = \mathcal{V}_p E \oplus \mathcal{H}_p E$ . The spaces  $\mathcal{V}E := \sqcup_{p \in E} \mathcal{V}_p E$  and  $\mathcal{H}E := \sqcup_{p \in E} \mathcal{H}_p E$  are two bundles called vertical and horizontal bundles respectively.

## 2 Hypotheses

In the following lemmas and propositions we always assume the following hypotheses to hold true.

**Assumption 1.** The manifolds  $M_i$  are open and path-connected sets of dimension  $\dim M_i = d_i$ .

**Assumption 2.** The sequence of maps (1) satisfies the following properties:

- 1) If  $\dim(M_{i-1}) \leq \dim(M_i)$  the map  $\Lambda_i : M_{i-1} \rightarrow M_i$  is a smooth embedding.
- 2) If  $\dim(M_{i-1}) > \dim(M_i)$  the map  $\Lambda_i : M_{i-1} \rightarrow M_i$  is a smooth submersion.

**Assumption 3.** The manifold  $M_n$  is equipped with the structure of Riemannian manifold, with metric  $g^{(n)}$ .

**Assumption 4.** We assume that the manifolds  $M_i$  are diffeomorphic to  $\mathbb{R}^{d_i}$  for some  $d_1, \dots, d_n \in \mathbb{N}$ .

**Assumption 5.** The matrices of weights in the maps  $\Lambda_i$ ,  $i = 1, \dots, n$ , as per in Definition 2 are of full rank.

**Assumption 6.** The Riemannian manifold  $(M_n, g_n)$  is complete.

## 3 Proof of the propositions

**Proposition 1.** Let  $\gamma : [0, 1] \rightarrow M_i$  be a piecewise  $\mathcal{C}^1$  curve. Let  $k \in \{i, i+1, \dots, n\}$  and consider the curve  $\gamma_k = \Lambda_k \circ \dots \circ \Lambda_i \circ \gamma$  on  $M_k$ . Then  $Pl_i(\gamma) = Pl_k(\gamma_k)$

*Proof.* It is enough to notice that  $\gamma_k : (0, 1) \rightarrow M_k$  is still a piecewise  $\mathcal{C}^1$  curve and that

$$\begin{aligned} Pl_k(\gamma_k) &= \int_0^1 \sqrt{g_{\gamma_k(s)}^{(k)}(\dot{\gamma}_k(s), \dot{\gamma}_k(s))} ds \\ &= \int_0^1 \sqrt{((\Lambda_k \circ \dots \circ \Lambda_i)^* g^{(k)})_{\gamma(s)}(\dot{\gamma}(s), \dot{\gamma}(s))} ds \\ &= Pl_i(\gamma) \end{aligned}$$

where  $(\Lambda_k \circ \dots \circ \Lambda_i)^* g^{(k)}$  is the pullback of  $g^{(k)}$  via  $\Lambda_k \circ \dots \circ \Lambda_i$ . □

**Corollary 1.** Let  $\gamma : [0, 1] \rightarrow M_i$  be a piecewise  $\mathcal{C}^1$  curve. Consider the curve  $\Gamma = \mathcal{N}_i \circ \gamma$  on  $\mathcal{N}(M_0) \subseteq M_n$ . Then  $Pl_i(\gamma) = Pl_n(\Gamma)$ , with  $L_n$  the length of a curve defined using the Riemannian metric  $g^{(n)}$ .

*Proof.* The thesis immediately follows from Proposition 1 setting  $k = n$ . □

**Lemma 1.**  $M_i / \sim_i$  is an open, path-connected, Hausdorff, second-countable set.

*Proof.* An elementary property of quotient maps yields that  $M_i / \sim_i$  is still a path-connected space and by [3, Corollary 3.17] we also know that  $\pi_i$  is an open map, therefore the quotient set  $M_i / \sim_i$  is open. Since pseudometric spaces are completely regular [3, Section 7], we conclude that  $M_i / \sim_i$  is Tychonoff and therefore it is in particular  $T_2$ . At last we note that, since  $\pi_i$  is an open quotient,  $M_i / \sim_i$  is also second-countable.  $\square$

**Proposition 2.** *If two points  $p, q \in M_i$  are in the same class of equivalence, then  $\mathcal{N}_i(p) = \mathcal{N}_i(q)$ .*

*Proof.* Let  $p, q \in M_i$  two points in the same class of equivalence  $[p]$ . Then, since  $M_i$  is path connected by hypothesis, there is a piecewise  $\mathcal{C}^1$  null curve  $\gamma : [0, 1] \rightarrow M_0$  connecting  $q$  and  $p$ , with  $Pl_i(\gamma) = 0$ . Consider now the curve  $\Gamma = \mathcal{N}_i \circ \gamma$  on  $M_n$ . By Corollary 1 we conclude that also  $Pl_n(\Gamma) = 0$  and being  $g^{(n)}$  a Riemannian metric we have that  $\mathcal{N}_i(p) = \mathcal{N}_i(q)$ .  $\square$

We recall that a singular Riemannian manifold is geodesically connected if any pair of points can be connected by a pseudolength minimizing curve.

**Lemma 2.** *Let  $M_i = (a_i, b_i)^{d_i} \subseteq \mathbb{R}^{d_i}$  for every  $i = 0, \dots, n$ . Then every singular Riemannian manifold  $(M_i \subseteq \mathbb{R}^{d_i}, g_i)$  is geodesically connected.*

*Proof.* Let us consider the last layer  $\Lambda_n$  of the neural network. For simplicity, in the following we denote with  $M_{n-1}$  both the manifold and the image of its global chart in  $\mathbb{R}^{d_{n-1}}$ . According to Definition 15, we can see this layer as in the following diagram

$$M_{n-1} \subseteq \mathbb{R}^{d_{n-1}} \xrightarrow{A_n} Q_n \subseteq \mathbb{R}^{d_n} \xrightarrow{\sigma_n} M_n \subseteq \mathbb{R}^{d_n}$$

where  $Q_n = A_n(M_{n-1})$ . We can suppose without loss of generality that  $d_{n-1} > d_n$  – Otherwise by Assumption 2 the map realizing the layer is an embedding and there is nothing to prove, as we fall back in the case of a (non-singular) Riemannian scenario concerning an isometric embedding in an ambient space of higher dimension. A basic theorem in linear algebra then guarantees that we can find two linear isomorphisms  $P_n : \mathbb{R}^{d_n} \rightarrow \mathbb{R}^{d_n}$  and  $T_{n-1} : \mathbb{R}^{d_{n-1}} \rightarrow \mathbb{R}^{d_{n-1}}$  such that the linear map  $\overline{A}_n = T_{n-1}^{-1} A_n T_n$  is given by the projection on the first  $n$  coordinates, namely the application

$$(\overline{x}_1^{n-1}, \overline{x}_2^{n-1}, \dots, \overline{x}_{d_{n-1}}^{n-1}) \mapsto (\overline{x}_1^{n-1}, \overline{x}_2^{n-1}, \dots, \overline{x}_{d_n}^{n-1}).$$

Consider now the diagram

$$\begin{array}{ccc} M_{n-1} \subseteq \mathbb{R}^{d_{n-1}} & \xrightarrow{A_n} & Q_n \subseteq \mathbb{R}^{d_n} \xrightarrow{\sigma_n} M_n \subseteq \mathbb{R}^{d_n} \\ \downarrow T_{n-1} & & \downarrow P_n \\ \overline{M}_{n-1} \subseteq \mathbb{R}^{d_{n-1}} & \xrightarrow{\overline{A}_n} & \overline{Q}_n \subseteq \mathbb{R}^{d_n} \end{array}$$

The maps  $T_{n-1}$  and  $P_n \circ \sigma_n^{-1}$  can therefore be considered as a global change of charts of the manifolds  $M_{n-1}$  and  $M_n$  for which the map  $A_n$  reads as a projection on the first  $n$  coordinates. Computing the pullback of the Riemannian metric  $g_n$  in these charts yields

$$g_{n-1} = \begin{pmatrix} P_n^{-1T} J_{\sigma_n}^T g_n J_{\sigma_n} P_n^{-1} & \mathbb{O}_{d_n \times (d_{n-1} - d_n)} \\ \mathbb{O}_{(d_{n-1} - d_n) \times d_n} & \mathbb{O}_{(d_{n-1} - d_n) \times (d_{n-1} - d_n)} \end{pmatrix} \quad (3)$$

Making use of the assumption that  $(M_n, g_n)$  is complete we can conclude that the singular Riemannian manifold  $(M_{n-1}, g_{n-1})$  is such that for any two points  $p, q \in M_{n-1}$  there exists a pseudolength minimizing curve connecting these two points due to the following reasoning. By the Hopf-Rinow theorem for Riemannian manifolds, the completeness of  $(M_n, g_n)$  is equivalent to its geodesic completeness and connectedness. In the global coordinates considered above, we can associate to the Riemannian geodesic

$$\gamma : [0, 1] \rightarrow M_n, \quad t \mapsto (\gamma_1(t), \dots, \gamma_{d_n}(t))$$

connecting  $\overline{A}_n p$  and  $\overline{A}_n q$ , the curve on  $(M_{n-1}, g_{n-1})$  given by

$$\Gamma : [0, 1] \rightarrow M_{n-1}, \quad t \mapsto (\gamma_1(t), \dots, \gamma_{d_n}(t), 0, \dots, 0).$$

By construction  $\Gamma$  is a pseudolenght minimizing curve for the degenerate Riemannian manifold  $(M_{n-1}, g_{n-1})$  connecting  $p$  with  $q$ . Indeed, another curve obtained modifying the first  $n$  components of  $\Gamma$  cannot be shorter, since we have  $Pl(\Gamma) = \ell(\gamma)$ ; On the other hand, the length of any curve

$$Z : [0, 1] \rightarrow M_{k-1}, t \mapsto (\gamma_1(t), \dots, \gamma_{d_n}(t), \zeta_1(t), \dots, \zeta_{d_{n-1}-d_n}(t))$$

with  $\zeta_1, \dots, \zeta_{d_{n-1}-d_n}$  piecewise differentiable functions cannot be shorter than the original curve  $\gamma$  since  $g_{n-1}(\dot{Z}, \dot{Z}) = g_{n-1}(\dot{\Gamma}, \dot{\Gamma})$  and therefore  $Pl(\Gamma) = Pl(Z)$ . Since this fact is true for an arbitrary pairs of points  $p, q \in M_{n-1}$ , we get that the geodesic connectedness of  $(M_{n-1}, g_{n-1})$ .

Next, we consider the pullback of the singular Riemannian metric  $g_{n-1}$  through the previous layer  $\Lambda_{n-1}$ . If  $d_{n-1} = d_{n-2}$ , the map  $\Lambda_{n-1}$  is a diffeomorphism and the previous result trivially holds true. Now we focus on the case  $d_{n-1} < d_{n-2}$ . Following same reasoning above we can find two global coordinates straightening the action of the linear map  $A_{n-1}$  as per the diagram

$$\begin{array}{ccccc} M_{n-2} \subseteq \mathbb{R}^{d_{n-2}} & \xrightarrow{A_{n-1}} & Q_{n-1} \subseteq \mathbb{R}^{d_{n-1}} & \xrightarrow{\sigma_{n-1}} & M_{n-1} \subseteq \mathbb{R}^{d_{n-1}} \\ \downarrow T_{n-2} & & \downarrow P_{n-1} & & \\ \overline{M_{n-2}} \subseteq \mathbb{R}^{d_{n-2}} & \xrightarrow{\overline{A_{n-1}}} & \overline{Q_{n-1}} \subseteq \mathbb{R}^{d_{n-1}} & & \end{array} \quad (4)$$

and such that  $\overline{A_{n-1}}$  is the projection over the first  $d_{n-1}$  coordinates. Computing the pullback of  $g_{n-1}$  with respect to the layer map in these global charts yields

$$g_{n-2} = \begin{pmatrix} P_{n-1}^{-1T} J_{\sigma_{n-1}}^T g_{n-1} J_{\sigma_n} P_{n-1}^{-1} & \mathbb{O}_{d_n \times (d_{n-2}-d_{n-1})} \\ \mathbb{O}_{(d_{n-2}-d_{n-1}) \times d_{n-1}} & \mathbb{O}_{(d_{n-2}-d_{n-1}) \times (d_{n-2}-d_{n-1})} \end{pmatrix} \quad (5)$$

As in the previous case, on the last  $d_{n-2} - d_{n-1}$  coordinates  $g_{n-2}$  is degenerate. However, also  $P_{n-1}^{-1T} J_{\sigma_{n-1}}^T g_{n-1} J_{\sigma_n} P_{n-1}^{-1}$  is a degenerate metric, therefore it is convenient to look for a global coordinate system in which is evident that  $g_{n-2}$  is the direct sum of a Riemannian metric and a null one. Let  $\bar{x}_1^{n-2}, \dots, \bar{x}_{d_{n-2}}^{n-2}$  be the original coordinates of  $\overline{M_{n-2}} \subseteq \mathbb{R}^{d_{n-2}}$ . To this end, we keep the last  $d_{n-2} - d_{n-1}$  coordinates untouched while we build a transformation of  $\bar{x}_1^{n-2}, \dots, \bar{x}_{d_{n-1}}^{n-2}$  as follows. Since  $\overline{A_{n-1}}$  is the identity on  $\bar{x}_1^{n-2}, \dots, \bar{x}_{d_{n-1}}^{n-2}$ , we can transform  $\bar{x}_1^{n-2}, \dots, \bar{x}_{d_{n-1}}^{n-2}$  via the diffeomorphism  $T_{n-1} \circ \sigma_{n-1} \circ P_{n-1}^{-1}$  to obtain a new set of coordinates

$$\bar{x}_1^{n-1}, \dots, \bar{x}_{d_{n-1}}^{n-1}, \bar{x}_{d_{n-1}+1}^{n-2}, \dots, \bar{x}_{d_{n-2}}^{n-2}$$

in which  $g_{n-2}$  reads

$$g_{n-2} = \begin{pmatrix} P_n^{-1T} J_{\sigma_n}^T g_n J_{\sigma_n} P_n^{-1} & \mathbb{O}_{d_n \times (d_{n-2}-d_n)} \\ \mathbb{O}_{(d_{n-2}-d_n) \times d_n} & \mathbb{O}_{(d_{n-2}-d_n) \times (d_{n-2}-d_n)} \end{pmatrix} \quad (6)$$

Using this global chart we can repeat the argument above to conclude that  $(M_{n-2}, g_{n-2})$  is geodesically connected. When  $d_{n-1} > d_{n-2}$ , the map  $\Lambda_{n-1}$  is an embedding. Proceeding as above, we can build a diagram as in Equation (4) such that  $\overline{A_{n-1}}$  is the map given by

$$(\bar{x}_1^{n-2}, \bar{x}_2^{n-2}, \dots, \bar{x}_{d_{n-2}}^{n-2}) \mapsto (\bar{x}_1^{n-2}, \bar{x}_2^{n-2}, \dots, \bar{x}_{d_{n-2}}^{n-2}, 0, \dots, 0).$$

Let

$$\gamma : [0, 1] \rightarrow M_{d-1}, t \mapsto (\gamma_1(t), \dots, \gamma_{d_{d-1}}(t))$$

be a geodesic connecting  $\overline{A_{n-1}}p$  with  $\overline{A_{n-1}}q$  and consider the curve on  $(M_{n-2}, g_{n-2})$  built taking only the first  $d_{n-2}$  components of  $\gamma$ , namely

$$\Gamma : [0, 1] \rightarrow M_{n-2}, t \mapsto (\gamma_1(t), \dots, \gamma_{d_{n-2}}(t))$$

connecting  $p$  with  $q$ . We claim that  $\Gamma$  is a geodesic. Suppose, by absurd, that there exists a curve  $\Delta : [0, 1] \rightarrow M_{n-2}, t \mapsto (\delta_1(t), \dots, \delta_{d_{n-2}}(t))$  connecting  $p$  with  $q$  and such that  $Pl(\Delta) < Pl(\Gamma)$ . Then, by construction, we obtain that the curve  $\delta = \overline{A_{n-1}}\Delta$  satisfies  $Pl(\delta) < Pl(\gamma)$ , in contrast with the hypothesis that  $\gamma$  is a geodesic. This proves that  $\gamma$  must be a geodesic connecting  $p$  with  $q$ . Since the same reasoning applies to any pairs of points of  $M_{n-2}$ , we proved that  $(M_{n-2}, g_{n-2})$  is geodesically connected.

Proceeding as above layer by layer up to  $\Lambda_1$ , we get the thesis.  $\square$

Thanks to this lemma, we can prove the following result.

**Proposition 3.** *Let  $x, y \in M_i$ , then  $x \sim_i y$  if and only if  $x \sim_{\mathcal{N}_i} y$ .*

*Proof.* If  $x \sim_i y$ , then there is a piecewise  $\mathcal{C}^1$  null geodesic curve  $\gamma$  with  $\gamma(0) = x$  and  $\gamma(1) = y$  and we have that  $Pl_i(\gamma) = Pl_n(\mathcal{N}_i \circ \gamma) = 0$ . Since  $g^{(n)}$  is a non-degenerate Riemannian metric,  $Pl_n(\mathcal{N}_i \circ \gamma) = 0$  entails that the tangent vector to  $\mathcal{N}_i \circ \gamma(s)$  is the zero vector for every  $s \in (0, 1)$  and therefore  $\mathcal{N}_i \circ \gamma$  is the constant curve  $\mathcal{N}_i \circ \gamma(s) = \mathcal{N}_i(x)$ . This proves  $x \sim_i y \Rightarrow x \sim_{\mathcal{N}_i} y$ . Let us now assume that  $x \sim_{\mathcal{N}_i} y$ . By definition we know that there is a piecewise  $\mathcal{C}^1$  curve  $\gamma : [0, 1] \rightarrow M_i$  such that  $\gamma(0) = x, \gamma(1) = y$  and  $\mathcal{N}_i \circ \gamma(s) = \mathcal{N}_i(x) \forall s \in [0, 1]$ . It remains to prove that  $\gamma$  is a null curve. This follows from the fact that, being  $\mathcal{N}_i \circ \gamma$  a constant curve, then  $Pl_i(\gamma) = l(\mathcal{N} \circ \gamma) = 0$ .  $\square$

**Corollary 2.** *Under the hypothesis of Proposition 3, one has that  $M_i / \sim_i = M_i / \sim_{\mathcal{N}_{i+1}}$ . Moreover, if two points  $p, q \in M_i$  are connected by a  $\mathcal{C}^1$  curve  $\gamma : [0, 1] \rightarrow M_i$  satisfying  $\mathcal{N}_i(p) = \mathcal{N}_i \circ \gamma(s)$  for every  $s \in [0, 1]$ , then they lie in the same class of equivalence.*

*Proof.* The statement follows immediately from Propositions 2 and 3 making use of the definitions of the quotients  $\sim_i$  and  $\sim_{\mathcal{N}_{i+1}}$ .  $\square$

**Theorem 1** (Godement’s criterion, [2, 1]). *Let  $X$  be a smooth manifold and  $R \subset X \times X$  be an equivalence relation. The quotient  $X/R$  is a smooth manifold if and only if*

- 1)  *$R$  is a submanifold of  $X \times X$*
- 2) *The projection map on the second component  $pr_2 : R \subset X \times X \rightarrow X$  is a submersion.*

Now we can prove that  $M_i / \sim_i$  is a smooth manifold.

**Proposition 4.**  $\frac{M_i}{\sim_i}$  *is a smooth manifold of dimension  $\dim(\mathcal{N}(M_0))$ .*

*Proof.* We prove that the quotient  $M_i / \sim_i$  is a smooth manifold using Godement’s criterion (Theorem 1). The graph  $\mathcal{G}_{i+1}$  of  $\sim_{\mathcal{N}_i}$  is the union of  $C_p \times C_p$ , with  $C_p$  a connected component of  $\mathcal{N}_i^{-1}(p)$ , with  $p \in \mathcal{N}_i(M_{i-1}) \subseteq M_n$  and therefore  $\mathcal{G}_{i+1}$  is a submanifold of  $M_i \times M_i$ . Furthermore, the restriction of the projection  $pr_2$  to  $R$  is the restriction of the identity map to  $C_p$  for some  $p \in M_i$ , which is a diffeomorphism with its image and therefore a submersion. The statement about the dimension follows from the proof of 2)  $\Rightarrow$  1) of Theorem 1, see [2, Lemma 9.4], taking in account that  $T_p \mathcal{N}_i = \dim(\text{Ker}(g^{(i)}))$  is constant.  $\square$

This proposition, along with [2, Lemma 9.4 and Lemma 9.9], yields that the classes of equivalence  $[p]$  are the leaves of a simple foliation of  $M_i$  and that  $\pi_i$  is a smooth submersion.

**Proposition 5.**  $\pi_i : M_i \rightarrow M_i / \sim_i$  *is a smooth fiber bundle, with  $\text{Ker}(d\pi_i) = \mathcal{V}M_i$ , which is therefore an integrable distribution. Every class of equivalence  $[p]$  is a path-connected submanifold of  $M_i$  and coincide with the fiber of the bundle over  $p$ .*

*Proof.* The first part of the statement follows applying Proposition 4 together with [2, Lemma 9.4 and Lemma 9.9]. The second part of the statement is then a consequence of the definitions of equivalence class and vertical bundle.  $\square$

## 4 Model implementation details

We trained one ViT model for each image dataset, always considering  $2 \times 2$  patches and using GELU as differentiable activation function. The model achieving the highest accuracy on MNIST has 4 attention heads and 4 layers with hidden size 48, while the best performing model on CIFAR10 has 8 heads, 6 layers and hidden size 384. These models have been trained with AdamW optimizer and a learning rate of 0.01, for 20 and 100 epochs respectively.

BERT-based models, instead, are pre-trained models having 12 heads, 12 layers and hidden size 768 in the case of text classification<sup>1</sup>, and 8 heads, 4 layers and hidden size 512 for MLM<sup>2</sup>.

## 5 Additional exploration examples

In this section, we propose other notable examples that have not been discussed in the main paper.

Figure 1 shows the exploration results from both SiMEC and SiMExp (lower middle and right panels, respectively) on an example from MNIST dataset. SiMExp updates eventually lead to a prediction change, causing the original input to be classified as “5”. Interestingly, while both algorithms modify the original image, it appears that changes to background pixels play a crucial role in driving this prediction shift. This effect is clearly visible in the bottom right plot of Figure 1, where the probability assigned to class “5” increases over iterations, eventually surpassing the probability of class “8”. Furthermore, it is noteworthy that the points predicted as “5” (represented by square markers in the top right part of the figure) diverge in direction from those explored by SiMEC. This illustrates SiMEC’s tendency to remain within regions of the embedding space that do not alter the output probability distribution, whereas SiMExp actively explores directions that lead to different classifications. However, these observations should be taken with a grain of salt, as the visualizations are based on two-dimensional projections of high-dimensional embeddings. As such, spatial relations and directions in the plots may not fully capture the underlying geometric properties of the original embedding space.

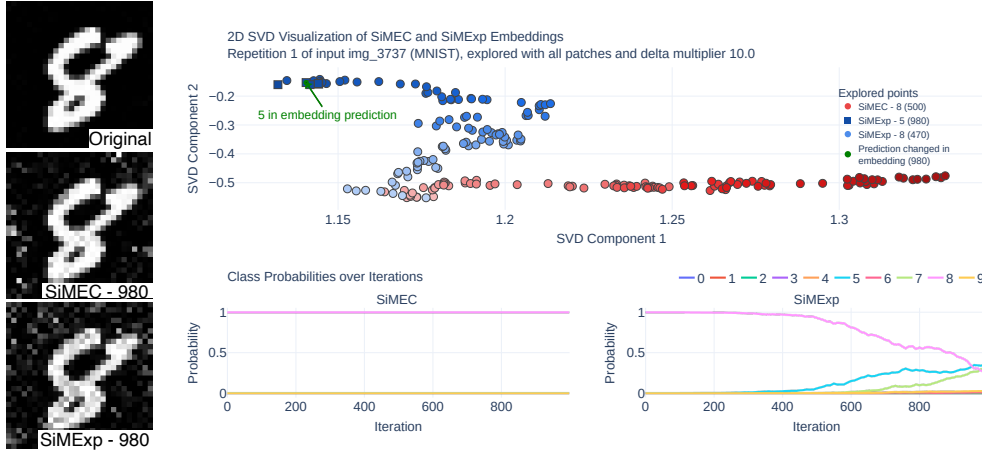


Figure 1: Example of exploration on a MNIST image using SiMEC and SiMExp. *Left*: Original image, followed by interpretation outputs of  $x_{980}$  from SiMEC (middle) and SiMExp (bottom). *Top right*: SVD projection of the explored points  $x^{(1)}, \dots, x^{(K)}$  for SiMEC (red) and SiMExp (blue), where color intensity encodes iteration progress (darker colors correspond to later iterations), and point shapes indicate predicted class labels. *Bottom right*: Evolution of class probabilities over iterations, for SiMEC (middle) and SiMExp (right).

An exploration example on a Winobias sentence is shown in Figure 2. In this case, the input sentence itself is not directly modified through the interpretation of the explored embeddings. However, while the original sentence requires predicting the token “they” (substituting the “[MASK]” token), SiMExp successfully explores the embedding space, eventually leading to a different prediction—“he”. This indicates a shift in the model’s behavior induced by the exploration. In the top panel of Figure 2, we highlight the embedding corresponding to this changed prediction within the exploration plot. As in previous cases, the figure illustrates how SiMExp explores more diverse directions in the embedding space, whereas SiMEC focuses its search in a localized region. Nonetheless, it is important to reiterate that these visualizations are based on two-dimensional projections of inherently high-dimensional embeddings. Therefore, spatial relations and apparent directions in the plot may not perfectly reflect the true geometry of the embedding space.

<sup>1</sup><https://huggingface.co/ctorean/hate-speech-bert>

<sup>2</sup><https://huggingface.co/gaunernst/bert-small-uncased>

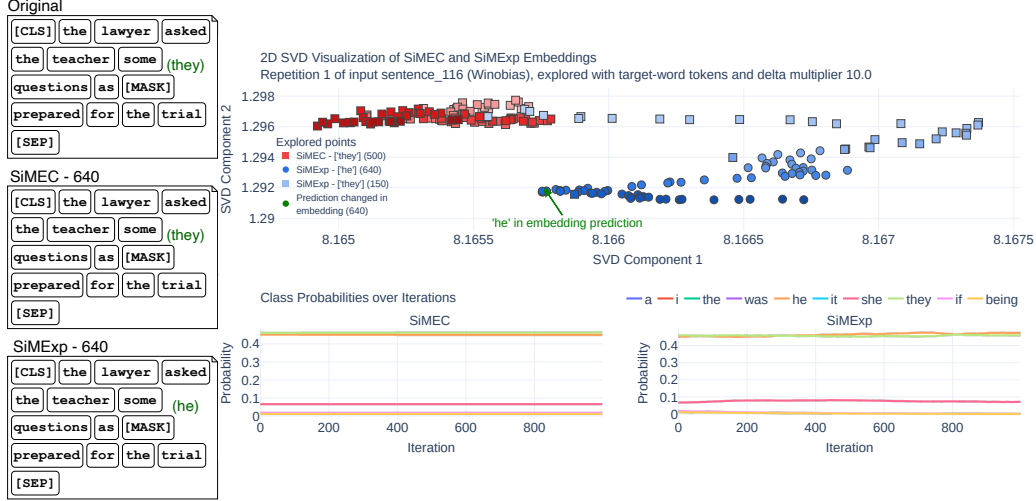


Figure 2: Example of exploration on a Winobias sentence using SiMEC and SiMExp. *Left*: Original sentence, followed by interpretation outputs of  $x_{640}$  from SiMEC (middle) and SiMExp (bottom). *Top right*: SVD projection of the explored points  $x^{(1)}, \dots, x^{(K)}$  for SiMEC (red) and SiMExp (blue), where color intensity encodes iteration progress (darker colors correspond to later iterations), and point shapes indicate predicted class labels. *Bottom right*: Evolution of class probabilities over iterations, for SiMEC (middle) and SiMExp (right).

Figure 3 shows an additional exploration example on an MHS input sentence, highlighting a different aspect of SiMEC’s behavior. Specifically, this example illustrates how SiMEC can discover alternative sentences by modifying individual tokens while preserving semantic consistency. On the left side of Figure 3, we see the token “murderers” being replaced by its synonym “killers” at the 640th SiMEC iteration, without affecting the predicted label. Unlike the other MHS experiments presented in the main text, this case focuses on exploring a single token. As a result, the probability variations observed in the SiMExp plot (bottom right) are more subtle, since the exploration is constrained to just one token embedding within the entire sentence.

## 6 Comparison of embedding and interpretation predictions during exploration

In the following, we provide detailed tables reporting on the results from exploration and interpretation analysis. For each dataset, we experimented with two values for hyperparameter  $\eta$ , namely 1 and 10. For the 3 classification datasets we adopted 3 different patch/token configurations: *all*, in which all patches/tokens are subject to SiMEC/SiMExp updates, *one*, in which only the patch/token with the highest attribution value is selected, and *q2*, in which half of the patches/tokens are selected according to their attribution scores. For WinoBias, instead, we apply the algorithms to the token given in the ground truth (i.e. *target-word*), since the dataset is itself targeted at association between tokens.

First we measure changes in the predictions given by re-applying the Transformer model to the pure and interpreted outputs of our algorithms. In particular, we rely on 3 metrics: (i) the absolute number of recorded prediction changes for a single input image/sentence in  $K = 1000$  iterations, (ii) the number of ranking changes, and (iii) the average Spearman’s correlations between class rankings at consecutive iterations. Each experiment is also repeated 3 times in order to ensure robustness of results.

In Table 1 we can notice that SiMEC rarely gives rise to ranking and prediction changes, while SiMExp, especially with  $\eta = 10$ , a hyperparameter choice which effectively intensify its effects, almost always produces a number of prediction changes significantly different from 0. Since also probabilities lower than the maximum may change during the SiMExp exploration process, ranking changes are usually more frequent than prediction changes. For both metrics, results in the *all* and *q2* settings give more evident results, no matter how strong the attribution of the selected patch/token

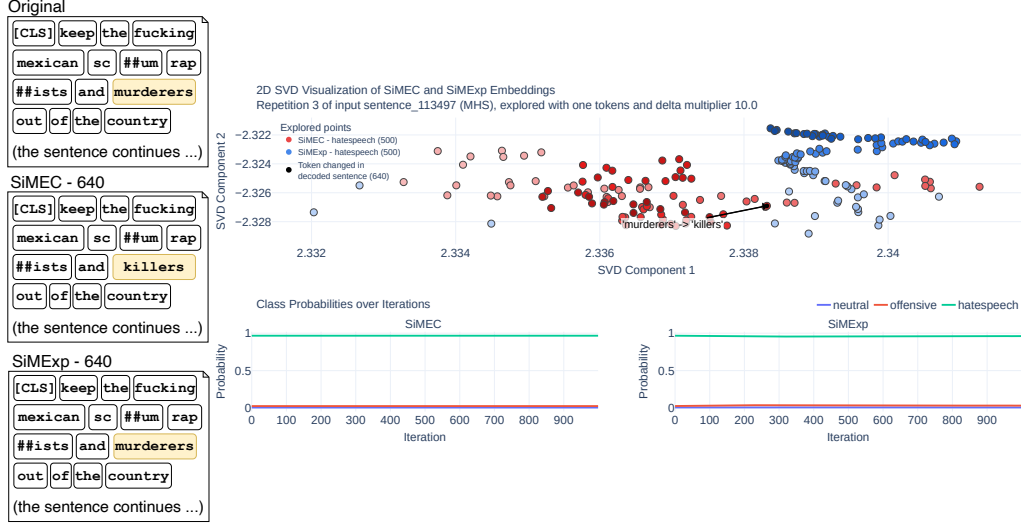


Figure 3: Example of exploration on an MHS sentence using SiMEC and SiMExp. *Left*: Original sentence, followed by interpretation outputs of  $x_{640}$  from SiMEC (middle) and SiMExp (bottom). *Top right*: SVD projection of the explored points  $x^{(1)}, \dots, x^{(K)}$  for SiMEC (red) and SiMExp (blue), where color intensity encodes iteration progress (darker colors correspond to later iterations), and point shapes indicate predicted class labels. *Bottom right*: Evolution of class probabilities over iterations, for SiMEC (middle) and SiMExp (right).

is. Results on textual data, however, give some insights even in the *one* configuration. Ranking correlations between consecutive iterations, finally, follow the path of ranking changes.

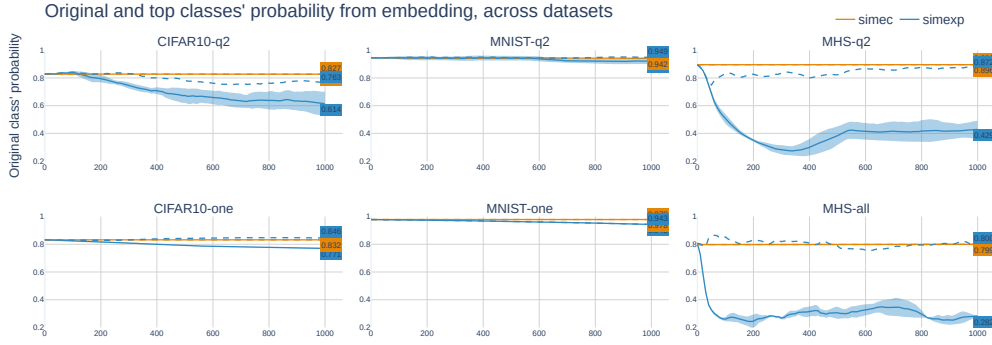


Figure 4: Probability values for the original class (full lines) and the top predicted class (dashed lines) based on embeddings, through the iterations. Results are divided per dataset and patch/token configuration, while  $\eta$  is set to 10 to highlight the effects that, with  $\eta = 1$ , appear more slowly.

Looking at the probabilities of the original and top classes through the iterations, depicted in Figure 4, it is possible to appreciate how SiMEC keeps the prediction probability constant, while SiMExp makes the probability of the original class and that of the most probable class clearly detach. Together with the results in Table 1 this empirically shows that, besides the configurations displayed in Section 4 of the paper, the expected behavior holds true also for other configurations.

We have already noticed how interpretation introduces noise into the process, with the effect that predictions over interpreted images/texts seem to lag behind the ones over pure embeddings produced by SiMExp. Table 2 presents extensive results that support this insight. Indeed, both ranking and prediction changes are less frequent than the ones measured on predictions over embeddings. Indeed, even on SiMExp output, when it undergoes the interpretation phase, the count of prediction changes



Dataset	Algorithm	$\eta$	Patches/tokens	Pred. changes	Ranking changes	Ranking correlations
CIFAR10	SiMEC	1	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	all	0.000 (0.000)	0.200 (0.400)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.167 (0.373)	1.000 (0.000)
	SiMExp	1	all	0.000 (0.000)	3.733 (3.750)	1.000 (0.001)
			one	0.100 (0.300)	0.100 (0.300)	1.000 (0.000)
			q2	0.233 (0.803)	<b>3.967 (3.167)</b>	1.000 (0.000)
MHS	SiMEC	1	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
	SiMExp	1	all	<b>0.897 (0.672)</b>	<b>1.436 (0.928)</b>	0.993 (0.005)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	<b>0.538 (0.499)</b>	<b>0.744 (0.669)</b>	0.996 (0.003)
MNIST	SiMEC	1	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.100 (0.300)	1.000 (0.000)
		10	all	0.000 (0.000)	0.067 (0.249)	1.000 (0.000)
			one	0.000 (0.000)	0.033 (0.180)	1.000 (0.000)
			q2	0.000 (0.000)	0.100 (0.300)	1.000 (0.000)
	SiMExp	1	all	0.000 (0.000)	<b>2.067 (1.672)</b>	1.000 (0.000)
			one	0.000 (0.000)	0.100 (0.300)	1.000 (0.000)
			q2	0.000 (0.000)	<b>1.767 (1.726)</b>	1.000 (0.000)
WinoBias	SiMEC	1	target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			target-word	0.000 (0.000)	0.125 (0.331)	1.000 (0.000)
	SiMExp	10	target-word	0.000 (0.000)	0.125 (0.331)	1.000 (0.000)
			target-word	1.083 (1.766)	<b>3.250 (2.411)</b>	0.997 (0.002)
			target-word	1.083 (1.766)	<b>3.250 (2.411)</b>	0.997 (0.002)
		10	target-word	0.000 (0.000)	0.125 (0.331)	1.000 (0.000)
			target-word	1.083 (1.766)	<b>3.250 (2.411)</b>	0.997 (0.002)
			target-word	1.083 (1.766)	<b>3.250 (2.411)</b>	0.997 (0.002)

Table 1: Prediction changes from embeddings: means and standard deviations for counts, ranking changes and ranking Spearman’s correlations for each experimental configuration. Statistically significant results at  $\alpha = 0.05$  are highlighted in bold.

is never statistically significant, while ranking changes are significant only in a couple of experiments. Such behavior is consistent with our previous considerations on the effects of interpretation.

Furthermore, we measured the statistics about first occurring prediction change in each SiMExp experiment, for predictions over embeddings as well as interpretations. Table 3 shows from such perspective that the predictions over interpretations tend to change later than those over embeddings. Indeed, only a couple of cases (CIFAR10 dataset with  $\eta = 1$  and all patches, and MHS dataset with  $\eta = 10$  and all tokens) present an higher number of prediction changes over interpretations, and even in the most numerous of these two cases, changes on interpretations tend to occur later on average.

Dataset	Algorithm	$\eta$	Patches/tokens	Pred. changes	Ranking changes	Ranking correlations
CIFAR10	SiMEC	1	all	0.000 (0.000)	0.733 (2.294)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	all	0.000 (0.000)	1.800 (3.718)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.400 (1.306)	1.000 (0.000)
	SiMExp	1	all	0.233 (0.920)	1.933 (2.407)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.167 (0.582)	1.233 (1.521)	1.000 (0.000)
MHS	SiMEC	1	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	all	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
	SiMExp	1	all	0.103 (0.496)	0.462 (1.677)	0.998 (0.008)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.026 (0.158)	0.026 (0.158)	1.000 (0.001)
MNIST	SiMEC	1	all	0.000 (0.000)	0.600 (2.304)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.967 (3.167)	1.000 (0.000)
		10	all	0.000 (0.000)	3.367 (6.534)	1.000 (0.001)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	2.067 (3.405)	1.000 (0.000)
	SiMExp	1	all	0.000 (0.000)	1.867 (2.232)	1.000 (0.000)
			one	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
			q2	0.000 (0.000)	0.800 (1.108)	1.000 (0.000)
WinoBias	SiMEC	1	target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
	SiMExp	1	target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
		10	target-word	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)

Table 2: Prediction changes from interpretations: means and standard deviations for each experimental configuration.

Following the previous considerations, we hypothesize the existence of a *catch-up effect*, in which predictions over interpretations align with predictions over embeddings with a delay  $\Delta k$ . Table 4 presents the statistics covering such phenomenon. In most cases, the catch-up effect doesn't apply, since there was no prediction change. In cases in which it does apply, however, it is still very rare, achieving 70% of occurrences only in the set of experiments over MHS dataset with all tokens changing and  $\eta = 10$ . In other datasets, it achieves at most 30% of occurrences in experiments over CIFAR10. When it occurs, we observe a delay in the order of  $10^2$  iterations, so it may be conjectured that running longer experiments, depending on the data at hand, may make the catch-up effect manifest itself on the long run.

Dataset	Algorithm	$\eta$	Patches/tokens	Embedding/ Interpretation	Count	Mean	Std.Dev.		
CIFAR10	SiMExp	1	all	Int	7	444.3	274.7		
			one	Emb	3	650.0	70.0		
			q2	Emb Int	7 5	395.7 470.0	322.1 457.0		
		10	all	Emb Int	29 18	593.8 376.7	325.6 231.1		
			one	Emb	6	130.0	91.4		
			q2	Emb Int	45 11	474.2 547.3	283.1 353.7		
		MHS	SiMExp	1	all	Emb Int	35 4	372.2 917.5	258.2 95.0
					q2	Emb Int	21 1	400.0 970.0	135.5 -
10	all				Emb Int	143 175	433.7 451.1	318.3 266.0	
	one			Emb	29	538.2	275.4		
	q2			Emb Int	70 54	353.0 578.5	279.2 240.7		
MNIST	SiMExp			10	all	Emb	3	813.0	289.0
					q2	Emb	4	552.5	309.9
WinoBias	SiMExp			10	target-word	Emb	52	575.4	223.8

Table 3: Iteration at which the first prediction change occurs: count of inputs for which a change occurs, mean and standard deviation of the first prediction change.

Finally, for what concerns the comparison between exploration results and interpretation results, we can measure probability differences over all output classes, not restricting ourselves to the most probable or the original class. Measuring such difference amounts to computing Wasserstein distance (here with order  $p = 1$  for the sake of interpretability), which results in the statistics gathered in Table 5. Again, in SiMEC the class probabilities never differ significantly between the embedding and its interpretation. In the case of SiMExp, instead, interpretation has a non-negligible effect on class probabilities, with peaks at 0.8 distance in the same experiments on MHS dataset for which we observed the catch-up effect and a high number of prediction changes. However, since results are not totally consistent across datasets, one may be cautious in drawing general conclusions about the distortion that interpretation may induce.

Putting these results in the form of chart in Figure 5, we aim at exploring the possible relationships between noise induced from interpretation, difference of interpreted output from one iteration to the next, and exploration pace, i.e.  $\delta^{(k)}$ . As expected, in SiMEC patch/token differences are almost constant, and so is the exploration pace. This is a direct effect of the constant class probabilities obtained by SiMEC, and the low distance observed between prediction on embeddings and prediction on interpretations (distributions depicted in the boxplots) suggests that the direction taken by SiMEC is one of exploration of a neighborhood of the initial input. As for SiMExp, these relationships are more intricate, and no obvious pattern involving the three measures displayed in the figure seems to emerge.

## 7 Explored hypervolume

In this section, we provide a more detailed analysis of the explored hypervolumes and compare the extent of the space covered by the two algorithms, SiMEC and SiMExp. Table 6 presents the Welch t-test statistics on  $\rho_V$  and the corresponding p-values for each dataset and delta multiplier  $\eta$ .

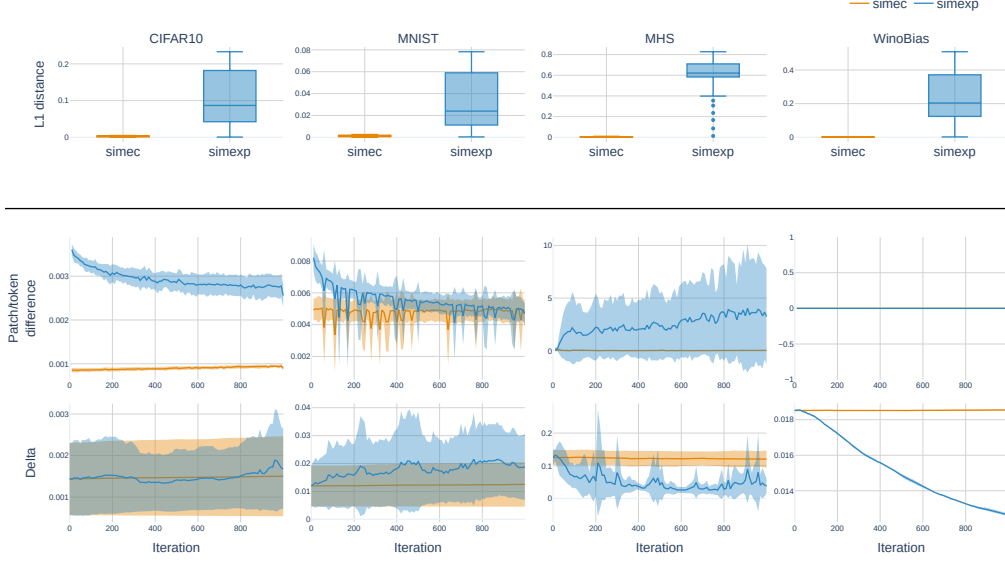


Figure 5: On row 1: boxplots of the prediction probability differences (in terms of Wasserstein distance with order  $p = 1$ ) between predictions performed on embeddings, which is the pure output of SiMEC and SiMExp, and their corresponding interpretations. As stated in the paper, in SiMEC only slight changes are induced, whereas the effect on the output of SiMExp is non-negligible. On row 2: difference in terms of patches/tokens from the original input image/text. On row 3: delta at each iteration. Results here are averaged across all configurations available for each dataset.

Although values of the  $\rho_V$  ratios are generally in the order of tens, it is important to emphasize that the magnitude of the Welch t-test statistics is not directly comparable across datasets. This is due to differences in the shape and size of the maximum feasible exploration volume, which are determined by the bounds of the embedding space specific to each Transformer model. Nonetheless, since all Welch t-test statistics are greater than 1 in a statistically significant way, we can confidently conclude that SiMExp consistently explores larger regions of the embedding space compared to SiMEC. This observation aligns with SiMExp’s intended objective of seeking out alternative equivalence classes.

## 8 Computation times

In Table 7 we provide full details about computation time in all settings. No significant difference between SiMEC and SiMExp is observed in this scope.

Dataset	Algorithm	$\eta$	Patches/ Tokens	Catch-up effect freq.	$\Delta k$	Pearson's Correlation	Correlation p-value
CIFAR10	SiMEC	1	all	0.0	-	0.097 (0.618)	0.108
			one	0.0	-	-	-
			q2	0.0	-	0.048 (0.701)	0.059
		10	all	0.0	-	0.002 (0.660)	0.104
			one	0.0	-	-	-
			q2	0.0	-	-0.079 (0.736)	0.023
	SiMExp	1	all	0.0	-	0.655 (0.497)	0.057
			one	0.0	-	0.221 (0.223)	0.125
			q2	0.1	73 (12)	0.517 (0.538)	0.025
		10	all	0.2	28 (36)	0.537 (0.442)	0.042
			one	0.0	-	-0.450 (0.090)	0.000
			q2	0.3	142 (198)	0.455 (0.462)	0.069
MHS	SiMEC	1	all	0.0	-	-0.119 (0.331)	0.277
			one	0.0	-	-	-
			q2	0.0	-	-0.025 (0.500)	0.255
		10	all	0.0	-	0.100 (0.397)	0.184
			one	0.0	-	-0.051 (0.000)	0.611
			q2	0.0	-	0.194 (0.254)	0.151
	SiMExp	1	all	0.1	735 (5)	-0.083 (0.328)	0.181
			one	0.0	-	-	-
			q2	0.0	500 (0)	-0.006 (0.379)	0.122
		10	all	0.7	302 (251)	0.047 (0.327)	0.156
			one	0.1	256 (61)	0.126 (0.408)	0.030
			q2	0.4	490 (195)	-0.036 (0.329)	0.132
MNIST	SiMEC	1	all	0.0	-	0.097 (0.583)	0.026
			one	0.0	-	-	-
			q2	0.0	-	-0.393 (0.531)	0.027
		10	all	0.0	-	0.113 (0.562)	0.068
			one	0.0	-	-	-
			q2	0.0	-	-0.034 (0.624)	0.027
	SiMExp	1	all	0.0	-	0.552 (0.557)	0.044
			one	0.0	-	-	-
			q2	0.0	-	0.834 (0.310)	0.000
		10	all	0.0	-	0.497 (0.497)	0.062
			one	0.0	-	0.870 (0.018)	0.000
			q2	0.0	70 (0)	0.391 (0.443)	0.123
WinoBias	SiMEC	1	target-word	0.0	-	-	-
		10	target-word	0.0	-	-	-
	SiMExp	1	target-word	0.0	-	-	-
		10	target-word	0.1	90 (102)	0.242 (0.449)	0.032

Table 4: Presence/absence of and statistics about the *catch-up effect*, i.e. the case in which prediction over the embedding modified by SiMEC/SiMExp algorithm changes from class  $C_A$  to  $C_B$  at an iteration  $\hat{k}$  and prediction over the interpretation of the embedding undergoes the same change  $C_A \rightarrow C_B$  a number of iterations  $\Delta k$  later. Here we present the number of times such effect is recorded in our experiments, the average delay  $\Delta k$ , and Pearson’s correlation, with its p-value for testing if it is equal to 0, between probabilities predicted over embeddings and those predicted over interpretations. Standard deviations, where applicable, are given in brackets. Rows with no data indicate there was no change, neither on embeddings nor on interpretations, in that setting.

Dataset	Algorithm	$\eta$	Patches/Tokens	Embedding-Interpretation Prediction Distance
CIFAR10	SiMEC	1	all	0.000 (0.000)
			one	0.000 (0.000)
			q2	0.000 (0.000)
		10	all	0.003 (0.002)
			one	0.000 (0.000)
			q2	0.008 (0.005)
	SiMExp	1	all	0.047 (0.026)
			one	0.005 (0.003)
			q2	0.016 (0.007)
		10	all	0.166 (0.129)
			one	0.092 (0.044)
			q2	0.203 (0.120)
MHS	SiMEC	1	all	0.003 (0.001)
			one	0.000 (0.000)
			q2	0.000 (0.000)
		10	all	0.012 (0.008)
			one	0.000 (0.000)
			q2	0.001 (0.001)
	SiMExp	1	all	0.525 (0.280)
			one	0.037 (0.021)
			q2	0.288 (0.158)
		10	all	0.795 (0.190)
			one	0.258 (0.120)
			q2	0.820 (0.239)
MNIST	SiMEC	1	all	0.000 (0.000)
			one	0.000 (0.000)
			q2	0.000 (0.000)
		10	all	0.002 (0.001)
			one	0.000 (0.000)
			q2	0.004 (0.002)
	SiMExp	1	all	0.009 (0.004)
			one	0.002 (0.001)
			q2	0.006 (0.004)
		10	all	0.062 (0.050)
			one	0.026 (0.017)
			q2	0.041 (0.024)
WinoBias	SiMEC	1	target-word	0.000 (0.000)
		10	target-word	0.001 (0.001)
	SiMExp	1	target-word	0.027 (0.014)
		10	target-word	0.244 (0.142)

Table 5: Mean and standard deviation of Wasserstein distance (order  $p = 1$ ) between prediction probabilities on embeddings and on interpretations, across datasets, algorithms and hyperparameters.

Dataset	$\eta$	$\rho_V$	Welch t-tests pvalue on $\rho_V$
CIFAR10	1	158.202	$< 10^{-3}$
	10	82.504	$< 10^{-3}$
MHS	1	8.760	$< 10^{-3}$
	10	2.480	$7 * 10^{-3}$
MNIST	1	99.287	$< 10^{-3}$
	10	28.178	$< 10^{-3}$
Winobias	1	37.574	$< 10^{-3}$
	10	15.084	$< 10^{-3}$

Table 6: Welch t-tests on explored hyper-volumes ratio  $\rho_V = (\Pi\Delta_{SiMExp}^{(K)}/\Pi\Delta_{SiMEC}^{(K)})^n$  and p-values, for each experiment configuration. SiMExp explores much more space than SiMEC, and every test is statistically significant.

Dataset	Batch size	All patches/tokens	Half patches/tokens	One patch/token
CIFAR10	16	11.27 (0.38) s	5.77 (0.15) s	0.29 (0.02) s
MNIST	16	1.41 (0.18) s	0.75 (0.02) s	0.14 (0.01) s
WinoBias	16	-	-	0.30 (0.02) s
MHS	8	6.48 (2.30) s	2.56 (0.68) s	0.31 (0.07) s

Table 7: Mean and standard deviation of computation times (in seconds) for SiMEC and SiMExp across datasets and patch selection strategies. “Half patches/tokens” refers to input patches/tokens whose attribution exceeds the median attribution value for that input, corresponding to our  $q2$  setting, while “One patch” refers to either the patch/token with the highest attribution value or, in the WinoBias case, the token given considering the bias detection task. The batch size is lower for MHS dataset due to the higher length of sentences as well as size of the model.

## References

- [1] N. Bourbaki. *Variétés différentielles et analytiques: Fascicule de résultats*. Springer-Verlag Berlin Heidelberg, 1967.
- [2] R. L. Fernandes. *Lectures on Differential Geometry*. WORLD SCIENTIFIC, 2024. doi: 10.1142/12733. URL <https://www.worldscientific.com/doi/abs/10.1142/12733>.
- [3] T. Pirttimäki. A survey of Kolmogorov quotients, 2019. arXiv:1905.01157 [math.GN].
- [4] L. W. Tu. *An Introduction to Manifolds*. Springer-Verlag New York, 2 edition, 2011. doi: 10.1007/978-1-4419-7400-6.