are represented by degree-one nodes, which are connected to other local minima through a single saddle point. The saddle points connecting different minima are represented by degree-three nodes (each connecting two local minima and one other saddle point). Loss functions often display many shallow local minima with low barriers (i.e., the value difference between the minima and the connecting saddle point is small) corresponding to "short-lived" connected components that merge quickly with other connected components.

In our work, we use the merge tree to extract the underlying structure of a loss landscape. We then use this extracted information to construct our topological landscape profile representations. Since the merge tree can be computed for an arbitrary dimensional loss landscape, we can use it to construct our representation for higher-dimensional loss landscapes, which would otherwise be difficult to visualize.

2.2. Topological Landscape Profiles

To enable the visualization of higher-dimensional loss landscapes, we introduce a new topological landscape profile representation that captures the minima and saddle points encoded by merge trees. This work builds upon Oesterling et al. (2013), who first introduced the idea of representing high-dimensional data clusters (and their nesting) as hills in a landscape, where the height, width, and shape of each hill encodes the coherence, size, and stability of each cluster. To construct the landscape profile, they first use a merge tree to encode the distribution (or density) of the data points. They then use this merge tree to construct the landscape profile, by representing maxima in the merge tree as hills in the landscape, where the size and shape of each hill are determined by characteristics like persistence and the number of points along the corresponding branch. In the context of loss functions, we are more interested in minima than maxima, so here we introduce a new version of this topological landscape profile, using the metaphor of valleys (or basins) rather than hills.

3. Methods

To construct our new topological landscape profile representations, we build on traditional loss landscape sampling approaches and leverage tools from TDA to capture the underlying shape (or topology) of the sampled loss landscapes. First, we select n vectors ($n \leq m$) to define an n-dimensional subspace (Figure 1.1), where m is the number of parameters in the model. We then sample a set of points from this subspace, where each point corresponds to a distinct set of parameters. We evaluate the loss for each set of parameters and represent the set of points (and their associated loss values) as an unstructured grid (Figure 1.2). We then compute a merge tree to capture the topology of the n-dimensional loss landscape (Figure 1.3), and we construct our final topological landscape profile based on this merge tree (Figure 1.4). In this section, we go into more detail about each of these steps.

3.1. Loss Landscape Construction and Representation

In this work, we limit our analysis to Hessian-based loss landscapes. We calculate the top n Hessian eigenvectors using PyHessian (Yao et al., 2020) (Figure 1.1) and then sample along the subspace spanned by these directions (Figure 1.2). The idea is that by using the eigenvectors associated with the top n largest eigenvalues, we can visualize the most

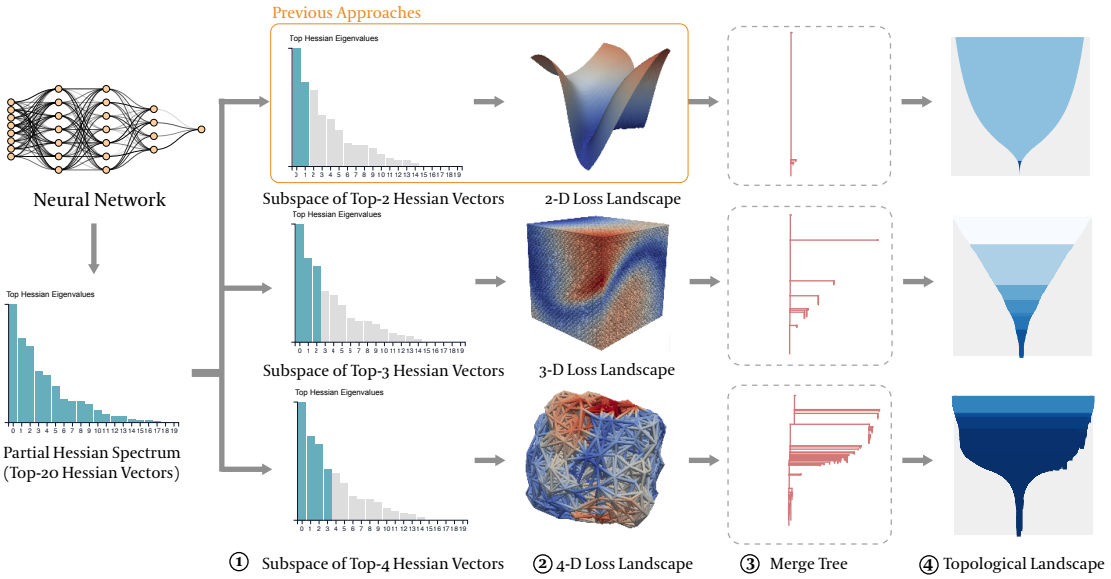

Figure 1: Our topological landscape profiles enable the visualization of higher-dimensional loss landscapes by capturing their underlying shape (or topology). Here we show loss landscapes based on the top n Hessian eigenvectors. See Section 3 for details.

significant local loss fluctuations for a given model. Given the n orthogonal directions, we generalize the approach taken by Li et al. (2018) by expanding the subspace beyond two dimensions. Formally, we perturb trained model parameters along the n directions and evaluate the loss $\mathcal{L}$ as follows:

$$f(\alpha_1 ... \alpha_n) = \mathcal{L}(\theta + \Sigma_{i=1}^{n} \alpha_i \delta_i), \tag{1}$$

where $\alpha_1 ... \alpha_n$ are the coordinates in the n-dimensional subspace, δ_i is the i-th direction in that subspace, and θ is the original model. As such, each coordinate corresponds to a point associated with a computed loss value, and the collection of loss values forms an n-dimensional loss landscape. In this work, we use an equally spaced grid by taking each α_i to be the set of equally spaced integers between 0 and r, where r is the resolution of each dimension in the grid. Here we use $r = 41$, such that the center of the grid corresponds to the original model, i.e., $\Sigma_{i=1}^{n} \alpha_i \delta_i = 0$.

Given an n-dimensional loss landscape, we can represent the sampled points as an unstructured grid, where each vertex in the grid is associated with n coordinates and a scalar loss value. Before we can characterize how the loss changes throughout the landscape (i.e., as parameters are perturbed from one vertex to the next), we need to define the spatial proximity (or connectivity) of vertices in the grid based on the similarity of their coordinates. Here we use a scalable, approximate nearest neighbor algorithm to construct a neighborhood graph representation of the loss landscape (Dong et al., 2011). The *neighborhood graph*, proposed by Jaromczyk and Toussaint (1992), of a dataset D is a graph $G = (D, E)$ where two points u and v are connected by an edge $(u, v) \in E$ if they are *similar*. Here we focus on the k-nearest neighbor graph, where each point is connected to the k most similar points.

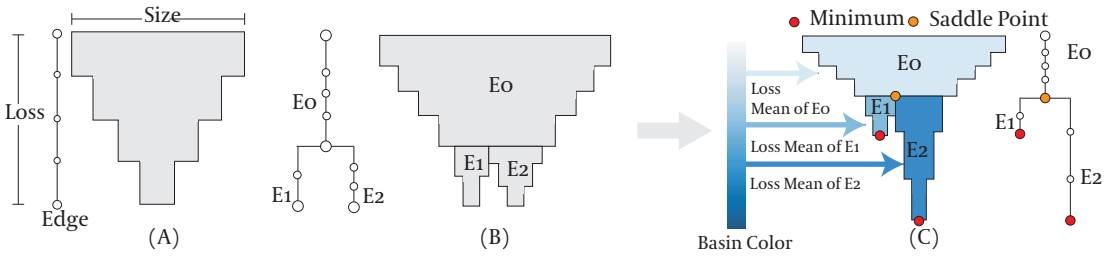

Figure 2: Representing the merge tree as a topological landscape profile. In (A) we show a single basin corresponding to a merge tree with a single branch, and in (B) we show multiple basins corresponding to multiple branches. In (C) we color the basins based on their average loss.

We also use a *symmetric* version of this graph, where points are only considered neighbors if each point is a neighbor of the other. In this case, an edge (u, v) is pruned from the graph if u is not one of the k nearest points to v, or vice versa. We note that this approach involves selecting an appropriate value for the k parameter. Here we use $k = 4 \times n$, such that the connectivity is similar to the spatial proximity of pixels in an image (i.e., each pixel having $k = 8$ neighbors, corresponding to the left, right, top, bottom, and all four corners).

3.2. Topological Structures and Landscape Profiles

After defining the subspace and computing the loss landscape, we perform topological data analysis to extract and summarize the most important features. In this work, we use a merge tree to extract key information from the loss landscape, which we then use to define our topological landscape profile. We compute the merge tree for each loss landscape using the Topology ToolKit (TTK), developed by Bin Masood et al. (2021).

Given a merge tree, we then construct the topological landscape profile using the method proposed by Oesterling et al. (2013). In this representation, each branch (in the merge tree) ending in a local minimum is represented by a basin (in the landscape profile), and each sub-branch ending in a saddle point is represented as a sub-basin, below which other basins are placed. In either case, each basin (or sub-basin) is represented by a set of rectangles encoding the cumulative size of the branch (or sub-branch), from bottom to top, such that the top of the basin is as wide as the number of points found along the corresponding branch in the merge tree.

We introduce this topological landscape profile representation of loss functions to effectively capture more information from higher-dimensional loss landscapes, in such a way that can still be visualized. While this topological representation and the merge tree used to create it both capture important features of the high-dimensional space, it also discards some important information by design. Here, we reincorporate some of this discarded information back into our representation, for example, by using the loss values to color the different basins. As shown in Figure 2.C, we compute the average loss across the points in each basin, and we use darker blues to represent lower average loss values. Thus, deeper basins are represented by a darker blue color, evoking the idea of deeper ocean depths. In