Geometry of caspase representation in ESM2

Marco Andreis - s243116
Technical University of Denmark

Abstract

Caspases are a family of cysteine-aspartic proteases that play essential roles in apoptosis and inflammation. We extract the caspase representations from ESM2 and learn a low dimensional representation on them by fitting a VAE. In particular, the structured latent space is used to compute geodesics between proteins and show that geodesic lengths better reflect evolutionary distances compared to euclidean distances.

1 Overview

1.1 The caspase protein family

Caspases are cysteine-aspartic acid endoproteases, a family of enzymes whose function is the catalysis of protein cleavage through hydrolyzation of the peptide bond. Caspases are the main actors in the initiation and execution of the apoptosis pathway, a mechanism of programmed-cell-death necessary for the body to dispose of aberrant cells. For this reason mutation in caspase-coding genes have been linked to multiple diseases related to either uncontrolled proliferation or degeneration of specific cells, including cancer, cardiovascular and autoimmune diseases [1, 2].
Due to their importance, caspases and caspase-like proteins are present all across animal kingdom, with many species having developed multiple variants responsible for different stages of the apoptotic pathways leading to the current classification in 14 subfamilies [3].

1.2 Variational Autoencoders as manifold fitters

Variational Autoencoders (VAE) [4] are generative models, build on the assumption that the data x has many redundant features and thus can be represented by a lower dimensional latent representation z.
VAEs consist in an encoder $q_\theta(z \mid x) = \mathcal{N}(\mu_\theta(x), \sigma^2_\theta(x))$ approximating the true posterior distribution $p(z|x)$ mapping the input into a latent distribution; and a decoder learning $p_\theta(x|z)$, modelling the probability of reconstructing the x from a latent sample z.
The two components of the VAE are trained jointly by maximizing the Evidence Lower Bound (ELBO):

$$\text{ELBO} = \underbrace{E_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{Reconstruction term}} + \underbrace{D_{\text{KL}}(q_\phi(z|x)\|p(z))}_{\text{Regularization term}} \tag{1}$$

optimizing for good reconstructions while keeping the latent space shaped similarly to the prior. This balance gives structure to the latent space, allowing for decoding meaningful samples from latent region non specifically covered by the training data.

We are interested in VAEs, specifically in the decoder, as a surface generator fitting a stochastic manifold described by the predicted μ and σ on the data space. The decoder mapping from the latent space to the data manifold induces (if the decoder is smooth) a Riemannian metric on the latent space [5]:

$$\overline{\mathbf{M}}_{\mathbf{z}} = \left(\mathbf{J}_{\mathbf{z}}^{(\mu)}\right)^T \left(\mathbf{J}_{\mathbf{z}}^{(\mu)}\right) + \left(\mathbf{J}_{\mathbf{z}}^{(\sigma}\right)^T \left(\mathbf{J}_{\mathbf{z}}^{(\sigma}\right) \tag{2}$$

which we can use to compute geodesics by minimizing the curve energy:

$$E[\dot{f}(\gamma_t)] = \frac{1}{2}\int_0^1 \dot{\gamma}_{\mathbf{t}} \overline{\mathbf{M}}_\gamma \dot{\gamma}_{\mathbf{t}} \tag{3}$$

which is an essentially easier task than minimizing length, because it avoid problems related to different speed reparameterizations of optimal curves (refer to [6] for more details) as well as simpler and smoother derivatives by avoiding the square root in the length formulation.

2 Methods

2.1 Data collection

The amino acid sequences of all available caspases were downloaded from the CaspBase database (as of September

2025) [7], for a total of 2,879 sequences.
We filter the dataset, by removing sequences for which the subfamilies information was not available, and those longer than 600 aminoacids to avoid necessity of excessive padding in further processing. Additionally we focus on the subfamilies for which more than 100 sequences were present, caspase 1-2-3-6-7-8-9-10-14, which also represent the most studied subfamilies with functionally established roles. This left us with 2541 proteins distributed as show in *Figure 1*.

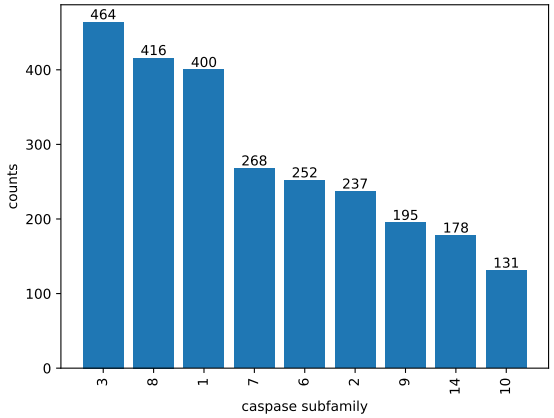

Figure 1: *Sample distribution across caspase subfamilies.*

2.2 Comparing protein encodings

ESM2 is large-protein model inspired from BERT in both architecture and training. In its introductory paper, the model representations are shown to provide a base for highly performant atomic level structure prediction, therefore we investigate the model as a possibility to create highly informative continuos proteins embeddings [8].
We compare the quality of onehot encodings (standard practice for representing protein sequences) with 480d and 1280d encodings from the ESM2 model. The former are obtained by representing each position as a 21d onehot encoded vector (canonical aminoacid and a padding token), importantly the sequences were not aligned before encoding, this is because alignment (MAFFT with auto settings [9]) resulted in sequences longer than 2000 aminoacids which would have been difficult to train upon due to the size of the flattened encoding and amount of padding. The ESM2 encodings were obtained from the (unofficial) JAX implementation of the model [10], to reduce the dimensionality of the encodings, we average them along the sequence length.

We train a VAE on of each encoding, while keeping a comparable model capacity. Once verified that the latent space looks promising, we fit knn classifier and use its accuracy on as proxy for the quality of the encoding.

Trying to maintain the capacity comparable across models and still get some meaningful learning, proved difficult for the onehot encoding, due to the large weight matrix of the first and last layer. To solve this two option where available, either scale up significantly the ESM2 models or give an edge to the onehot model making it larget than the others. The latter was chosen to reduce computation resources and to avoid unnecessarily scaling networks which already show good learning.

2.3 Standard VAE

The VAEs utilized for comparing encodings have 4 layer fully connected layers with GELU activation, both as encoder and decoder. Each model is trained for 500 epochs with a learning rate of 10^{-4}, warming up the regularization loss linearly during the first 100 epochs.
The prior on the latent space is gaussian. For onehot encoding, the last layer outputs the logits of a categorical distribution which are used to compute a cross-entropy reconstruction loss. While in the case of the ESM2 encodings, each feature is modelled with a gaussian distribution parametrized by a mean and variance predicted by the model, making the reconstruction loss the negative log-likelihood of the input.

2.4 Variance adjusted VAE

The validity of the uncertainty quantification given by the predicted variance of a standard VAE has been shown to be unreliable and not respect the desired behaviour of low variance in regions dense with data and high and regions where data is scarce. This has severe implication on the quality of the metric induced by the decoder on the latent space. We apply the solution proposed by Detlefsen et al. [11], consists of adjusting the predicted variance as follows:

$$\sigma^2(x_0) = (1 - \nu(\delta(x_0)))\hat{\sigma}^2_\theta + \eta\nu(\delta(x_0)) \qquad (4)$$

where $\delta(x_0) \in [0, +\infty)$ is a measure of how distant the latent point is from other data, approximated as the distance from the closest centroid of a k-means ($k = 100$) optimized at the beginning of every epoch. $\nu : [0, +\infty) \to [0, 1]$ is a scaled-and-translated sigmoid $\nu(x) = \text{sigmoid}(\frac{x+a}{\gamma})$ making the adjusted variance a convex combination of the actual predicted value and an arbitrary high value $\eta = 10$ to which the variance will tend towards when far from data. The scale γ is a learned parameter which determines how smoothly the transition happens, while a is fixed to -6.9077γ assuring that $\nu(0) \approx 0$.
To get the model to train effectively, the variance is kept fixed for the initial 250 epoch to first learn accurately to reconstruct the data, and only further in training the model can focus on refining the variance predictions.

2.5 Geodesic optimization

We are interested in calculating geodesics between latent points utilizing the metric induced by the decoder. To avoid computing geodesics between every two points, following Detlefsen et al. [12], we focus on the representative subset composed by the 100 points closest to the optimized k-means clusters optimized during training. The curves are constructed as piecewise linear approximations of 100 points, the coordinates of which are optimized for 500 iterations and a learning rate of 0.005 based on gradient compute w.r.t. the discrete curve energy:

$$E_{\text{disc}}(\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}) = \sum_{k=0}^{N-1} (\mathbf{z}_{k+1} - \mathbf{z}_k)^T \, \overline{\mathbf{M}}_{\mathbf{z}_k} \, (\mathbf{z}_{k+1} - \mathbf{z}_k) \tag{5}$$

with $(\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}$ being the intermediate latent points approximating the curve.

2.6 Evolutionary distances

As comparison for the computed geodesics lengths, the evolutionary distances between for the considered points were calculated by aligning the sequences using MAFFT, constructing the phylogenetic tree using FastTree2 [13] with standard settings, and computing the patristic distances using the Phylo module provided by biopython [14].

3 Results

3.1 Comparing protein representations

Three VAEs were trained with 80-20 train-test split respectively on onehot, 480d ESM2 and 1280d ESM2 embeddings, and a KNN classifier was fitted on the latent space. As shown in *Figure 2*, the accuracy for all models is relatively stable with different k values, highlighting that proteins are highly similar within a subfamily and vastly different between them.
Despite the edge in model capacity, that the lack of alignment and the small latent space compared to the input dimensionality impacted significantly the onehot model, resulting in both ESM2 variants outperformed. Furthermore, the 1280d encodings provide an additional improvement on the 480d, for this reason we proceeded with the following experiments with just the larger ESM2 encodings.

3.2 Training variance adjusted VAEs

In *Figure 3* is shown how the predictive variance behave between the standard VAE and the one with adjusted variance. As can be seen, the standard VAE predictions do not follow the data distribution at all, with low variance far from data and high variance close to the center. This is reflected in the erratic behaviour of the metric.

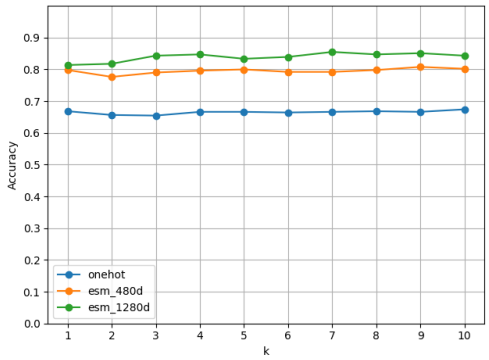

Figure 2: *KNN accuracy comparison across different encodings*

The opposite happens by adjusting the variance calculation, now the prediction are closely aligned to the distribution of the lalents. Despite this, the highly clustered structure of the latent space is slightly problematic, as it results in high-valued metric regions between cluster which complicates geodesics optimization, as the curves will be forced to traverse steep regions when connecting points from different clusters.
Experiments aimed at balancing the regularization loss and tuning the variance adjustment parameters (initial value of γ, a) to the get a more favourable metric behaviour were unsuccessful, this is indicative of the diversity between caspase subfamilies.

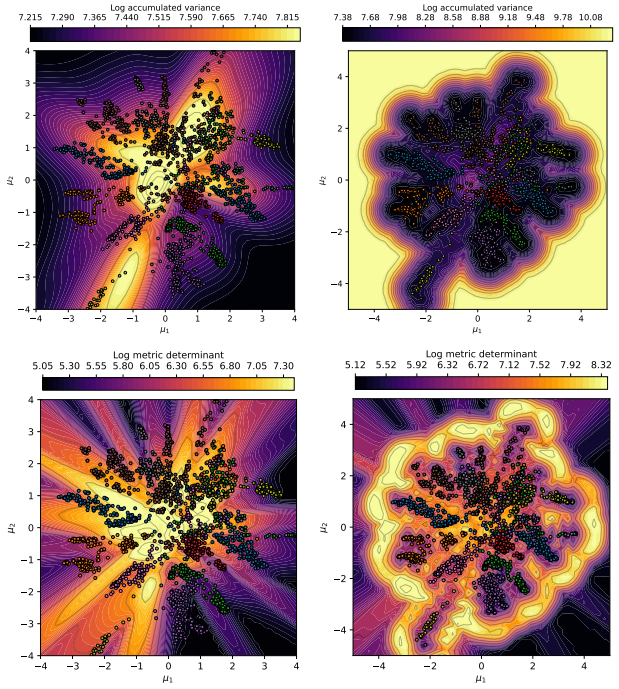

Figure 3: *Total predicted variance and metric volume in the latent space of standard VAE (left) and variance adjusted VAE (right). Points are training latents coloured by subfamily.*

3.3 Geodesic for phylogenetic analysis

A sample of 256 geodesics from the total computed (4950) are shown in *Figure 4*. As expected the geodesics deviate significantly from the initialized straight lines underlying the non-euclidean structure of the latent space.
Most geodesics behave as expected and avoid traversing areas of high metric determinant. Some of the optimized curves, especially those connecting distant points, still move through steep regions, this is because the optimization had to compromise between the cost of traversing a costly region and the cost of circumnavigating it and found the latter to lead to lower energy.
To see whether the geodesics carry meaningful information about protein relationship, their lengths are compared with patristic evolutionary distances obtained from the phylogenetic tree built using FastTree2. Of particular interest is verifying how the geodesics compare to simple euclidean distances.
The results in *Figure 5* show that geodesics distance indeed correlate to evolutionary distances better than simple euclidean distances ($\sim 19.2\%$ improvement), which highlights the importance of respecting the geometry described by the Riemannian metric induced by the decoder.
The same experimental setup was applied to variance adjusted VAEs trained with slightly different hyperparameters, as mentioned in the previous paragraph. Despite them showing a latent space structure which respected less the latent representation distribution, the outcome of the comparison consistently showed the correlation of geodesic distances improving on the euclidean ones in the range of $15 - 25\%$.

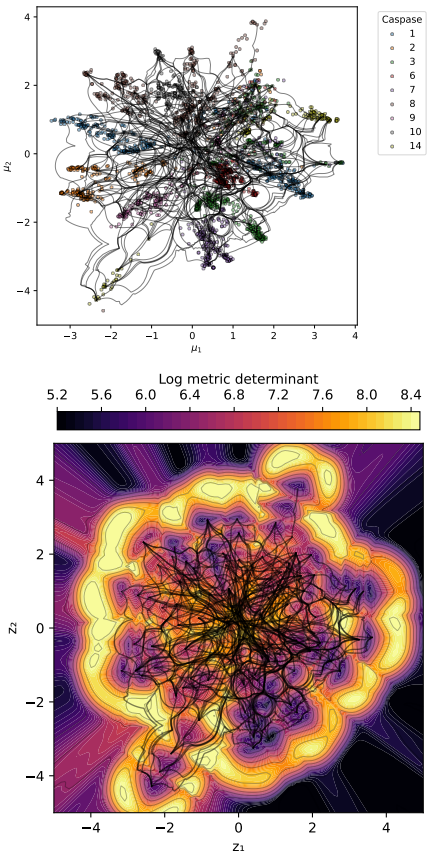

Figure 4: *Geodesics overlapped on latents (top) and on metric volume (bottom).*

3.4 Protein interpolation

Another task of major interest to perform in the latent space is the interpolation between different proteins, to produce a sequences coding for chimeras of the two endpoints, which can be measured in terms of sequence identity, folded structure through and chemical properties (ex. hydrophobic profile).To be able to reconstruct full protein sequences, the VAE needs to be trained on the whole encodings without averaging along the sequence dimension, this allows decoding the intermediate latent points of the geodesics first into the full ESM2 encodings, and then into actual protein sequences using the ESM2 head.
Unfortunately our efforts in tuning a VAE to be able to learn a satisfactory latent space from the whole ESM2 encodings failed. In particular the optimization wasn't able to find a good compromise between reconstruction and regularization, most often resulting in a latent space containing a smear of latents, indicating the model is not capable of clustering clearly the proteins belonging to different families and focuses on regions of the encodings which are shared. An example of such latent space is given in *Figure 6*.

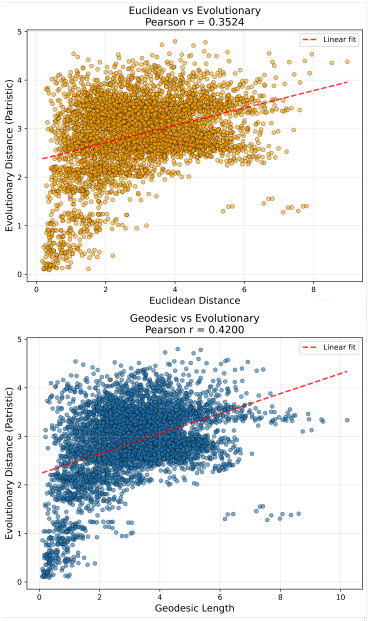

Figure 5: *Correlation between geodesics, euclidean and patristic distances*

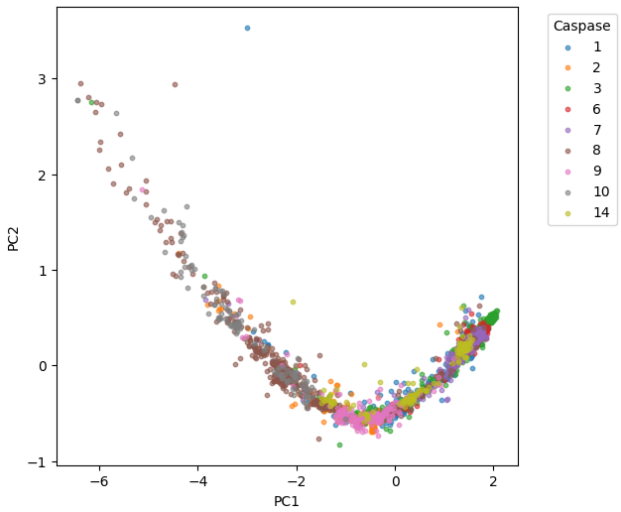

Figure 6: *Latent space of VAE trained on full ESM2 embeddings.*

4 Methodological improvements

While we succeeded in showing the biological relevance of geodesics, the correlation obtained are in the 0.35-0.45 range. We suspect that these could be improved with more time at our disposal as there are numerous places in our pipeline that could benefit from additional tuning. For example, while the accumulated variance of the utilized VAEs decoder follows the latent distribution quite well the resulting metric are far from perfect and could be improved by modifying the way the variance correction is done. A interesting approach proposed by Arvanitidis et al. [5] consitsts in mdoeling precision of the decoder using RBF neural networks. Another area of improvement is the parametrization of the geodesics. Polynomial parametrization as well as the use of splines are already proven to be viable methods which avoid limitation due to the segmented nature of the piecewise linear approximation [12].

The problems encountered while training on the full ESM2 embeddings are also documented in the literature. They are due to the network focusing on feature regions shared by many examples in the training set effectively learning to reconstruct the average representation of a caspase subfamily, which could justify why even in our failed attempts there is some reasonable class separation. This issue has been solved by switching to conditional VAEs and using as input a concatenation of the ESM2 embeddings with onehot positional encodings, allowing sampling position-specific encodings [15].

5 Conclusions

In this project, we highlight show that protein encodings from the ESM2 model averaged along the sequence length allow for a relatively low-dimensional representation (ex. compared to flattened onehot encodings) while being highly informative, which can reduce computational resources for tasks that do not require full sequence reconstruction.
We show the adoption of variance adjusted VAEs provide a significantly better behaved latent space, additionally our results indicate that respecting the geometry learned by the decoder, by computing geodesics using the induced metric, is necessary to obtain more reliable information about evolutionary relationships between proteins, when compared to simple euclidean distances.

Code availability

The code and data utilized to perform the experiments are available in this github repository.

References

1. Nadendla, E. K., Tweedell, R. E., Kasof, G. & Kanneganti, T.-D. Caspases: structural and molecular mechanisms and functions in cell death, innate immunity, and disease. *Cell Discovery* (2025).
2. Dho, S. H., Cho, M., Woo, W., Jeong, S. & Kim, L. K. Caspases as master regulators of programmed cell death: apoptosis, pyroptosis and beyond. *Experimental & Molecular Medicine* (2025).
3. Lamkanfi, M., Declercq, W., Kalai, M., Saelens, X. & Vandenabeele, P. Alice in caspase land. A phylogenetic analysis of caspases from worm to man. *Cell Death & Differentiation* (2002).
4. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* 2022.
5. Arvanitidis, G., Hansen, L. K. & Hauberg, S. *Latent Space Oddity: on the Curvature of Deep Generative Models* 2021.
6. Hauberg, S. *Differential geometry for generative modeling* https://www2.compute.dtu.dk/~sohau/weekendwithbernie/Differential_geometry_for_generative_modeling.pdf.
7. Grinshpon, R. D., Williford, A., Titus-McQuillan, J. & Clay Clark, A. The CaspBase: a curated database for evolutionary biochemical studies of caspase functional divergence and ancestral sequence inference. *Protein Science* (2018).
8. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* (2022).
9. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* (2013).
10. Irhum, S. *esmjax: ESM2 in JAX* 2022. https://github.com/irhum/esmjax.
11. Detlefsen, N. S., Jørgensen, M. & Hauberg, S. *Reliable training and estimation of variance networks* 2019.
12. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nature Communications* (2022).
13. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* (2009).
14. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009).
15. Zainchkovskyy, Y. *Probabilistic Generative Models for Automatic Guided Drug Discovery* PhD Thesis (Technical University of Denmark, 2024).