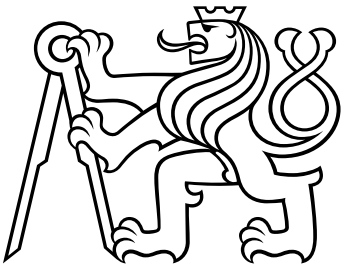
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Physics



Master's Thesis

Self-supervised feature extraction from break junction traces for enhanced clustering

Oliver Klimt

klimtoli@fel.cvut.cz
https://icluto.oklimt.com

Supervisor: Ing. Ladislav Sieger CSc.
Supervisor specialist: Jaroslav Vacek, Ph.D., RNDr. Jindřich Nejedlý, Ph.D.

Study programme: Cybernetics and Robotics

May 2026

Compiled on 19. 4.2026 21:49

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Klimt Oliver** Personal ID number: **499151**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Measurement**

Study program: **Cybernetics and Robotics**

II. Master’s thesis details

Master’s thesis title in English:

Self-supervised feature extraction from break junction traces for enhanced clustering

Master’s thesis title in Czech:

Extrakce příznaků z křivek experimentu break-junction metodou self-supervised learningu pro následné klastrování

Name and workplace of master’s thesis supervisor:

Ing. Ladislav Sieger, CSc. Department of Physics FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **23.01.2026** Deadline for master's thesis submission: **22.05.2026**

Assignment valid until: **by the end of summer semester 2026/2027**

Head of department’s signature

Vice-dean´s signature on behalf of the Dean

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work.
The student must produce his thesis without the assistance of others, with the exception of provided consultations.
Within the master’s thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Bc. Klimt Oliver

Student’s signature

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Klimt Oliver** Personal ID number: **499151**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Measurement**

Study program: **Cybernetics and Robotics**

II. Master’s thesis details

Master’s thesis title in English:

Self-supervised feature extraction from break junction traces for enhanced clustering

Master’s thesis title in Czech:

Extrakce příznaků z křivek experimentu break-junction metodou self-supervised learningu pro následné klastrování

Guidelines:

1) Research machine learning architectures for auto-encoders, transformers and learned descriptors.
2) Research applicable loss functions.
3) Propose an encoder that maps traces into a latent/descriptor space.
4) Based on the latent space cluster traces with K-Means with large K (e.g. K=30) and compare the clusters with previous clustering runs on histogram features for the same K.

Bibliography / sources:

1] BENAVIDES-CESAR, Llinet; MANSO-CALLEJO, Miguel-Ángel; CIRA, Calimanut-Ionut. Methodology Based on BERT (Bidirectional Encoder Representations from Transformers) to Improve Solar Irradiance Prediction of Deep Learning Models Trained with Time Series of Spatiotemporal Meteorological Information. Forecasting, 2025, 7.1: 5.
2] SHIEH, Jin; KEOGH, Eamonn. i SAX: disk-aware mining and indexing of massive time series datasets. Data Mining and Knowledge Discovery, 2009, 19.1: 24-57.
3] PAN, Zhichao, et al. Data-Driven Insights in Single-Molecule Break Junction Studies: A Comprehensive Review of the Data Analysis Methods. Langmuir, 2025, 41.36: 24152-24174
4] MUTSCHLER, Christopher, et al. (ed.). Unlocking Artificial Intelligence: From Theory to Applications. Springer, 2024.
5] LIN, Luchun, et al. Spectral clustering to analyze the hidden events in single-molecule break junctions. The Journal of Physical Chemistry C, 2021, 125.6: 3623-3630.

Abstract

Ahoj **TODO: Add abstract text**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At.

Abstrakt (CZ)

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At.

Acknowledgement

I would like to express my sincere gratitude to Dr. Starý for the opportunity to become a member of his research group at IOCB Prague. I am deeply indebted to Dr. Nejedlý for providing the molecular data and for his invaluable mentorship regarding data interpretation.

My thanks also go to Dr. Sieger for his steadfast guidance and mentorship, and to Dr. Vacek for his expert consultation on the clustering workflow and insightful feature suggestions. Finally, we acknowledge the High Performance Computing group (HPCg) at IOCB Prague for providing the computational resources and cluster access that enabled the training of many large neural networks.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, 22. 05. 2026

.....

14 TODOs remaining

10 notes

Contents

Section 1

Introduction

1.1 Motivation

The Mechanically Controllable Break Junction (MCBJ) experiment is a foundational technique for measuring the electrical conductance of single molecules. By providing insight into molecular charge transport, this method is essential for the advancement of molecular electronics—a field aiming to utilize single molecules as active electronic components. While organic electronics are already ubiquitous in technologies such as Organic Light-Emitting Diodes (OLEDs), the transition to unimolecular devices requires understanding quantum phenomena at the nanometer scale.

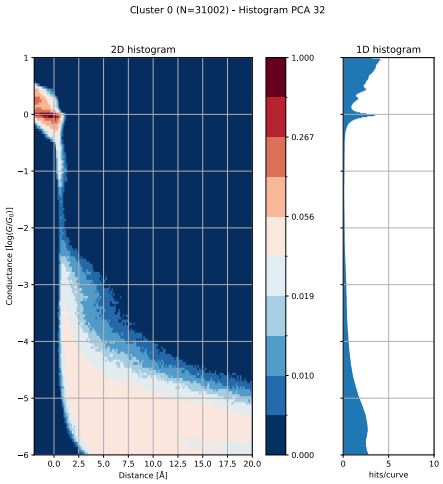
The theoretical framework for molecular electronics was first proposed by Aviram and Ratner **TODO: Tohle jsem citoval v BP: Aviram, A., & Ratner, M. A. (1974). Molecular rectifiers. Chemical Physics Letters, 29(2), 277-283.**, who conceptualized a molecular rectifier. Today, this field relies on the synthesis of novel, complex architectures, such as those developed by the Ivo Starý research group at the Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences in Prague. To evaluate these candidates for organic electronics, the break junction setup measures electrical conductance as a function of electrode displacement. This process generates massive datasets characterized by stochastic quantum events, necessitating advanced computational pipelines for accurate analysis.

1.2 Research Goals

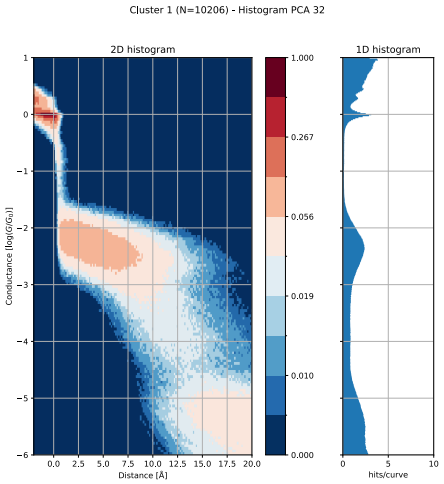
In our previous work we combined supervise and unsupervised machine learning for analysing big amounts of data[1].

We have developed a computational framework and pipeline that utilizes supervised machine learning to identify ‘snap-back’ events within molecular traces. The pipeline performs Principal Component Analysis (PCA) on both the raw traces and their respective histograms, followed by unsupervised clustering using K-means or DBSCAN algorithms. Our results demonstrate that this approach effectively distinguishes between traces with and without molecules, and can further differentiate between varying levels of molecular bonding to a metal contact. However, we observed that linear PCA can lead to the loss of rare, individual events. These stochastic features may collapse into the latent space, rendering them impossible to recover or accurately assign to a specific cluster.

To address the limitations of linear dimensionality reduction, we transitioned from Principal Component Analysis (PCA) to a learned feature representation. By employing a deep learning approach—specifically designed to capture non-linear dependencies—we aim to preserve the high-dimensional variance associated with rare, stochastic events. We anticipate that learning the feature set directly from the raw data will prevent the collapse of these unique signatures into an unrecoverable latent space, thereby enabling the detection and classification of individual molecular events that were previously obscured.



(a) Cluster A - no molecule



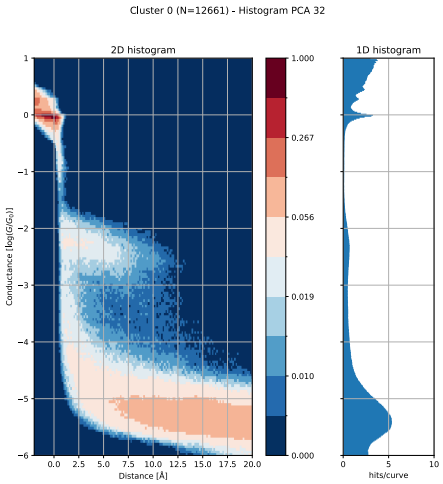
(b) Cluster B - molecule present

Figure 1: Example of clusters obtained in our previous work using linear PCA and K-means clustering for *hist32* feature that was computed for the range of 10^{-4} to 10^{-1} G/G_0.

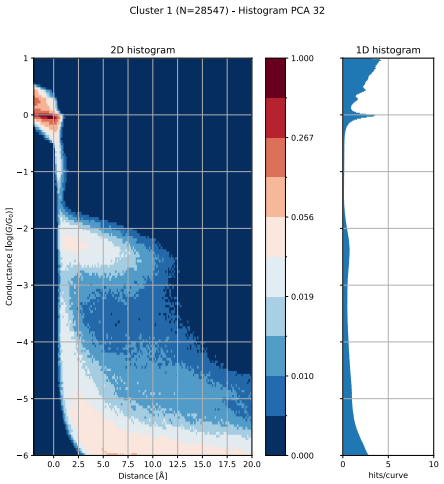
The efficacy of the previous methodology was highly dependent on the specified conductivity range. While the approach demonstrated high performance when the parameters were finely tuned (e.g., 10^{-4} to 10^{-1} $\frac{G}{G_0}$, as shown in Figure 1), the default configuration yielded suboptimal results (Figure 2), as it was mainly focused in the limit area of the conductance values. This discrepancy stems from the method's lack of molecular awareness; it relied exclusively on expert-driven hyperparameter optimization to differentiate trace clusters rather than intrinsic chemical properties.

1.3 Thesis Structure

TODO: update



(a) Cluster A



(b) Cluster B

Figure 2: These two clusters were obtained using the same method as in Figure 1, but with the default range of 10^{-6} to 10^{-1} G/G_0.

- **Chapter Section 2** provides the theoretical background and state-of-the-art.
- **Chapter Section 3** discusses Self-Supervised Learning and architectural choices.

Section 2

Theoretical Background

2.1 Single-Molecule Electronics

2.2 Mechanically Controllable Break Junction

2.3 Data Analysis Challenges

2.3.1 The "Curse of Dimensionality" and the "Uniform Effect"

In our previous work we used PCA for dimensionality reduction. This approach assumes a linear relationship between the data points, which is not always the case in single-molecule conductance traces. Traces are, visualised in logarithimc scale, which is a non-linear transformation, but in logarithmic scale an exponenational decay is observed as a linear trend. A dimensionality reduction is necessary to in order to obtain a reduced representation of the data, on which clustering algorithms can be applied in finite time.

The problem with PCA is that it is very limited in capturing rare events in traces, these events are often lost in the latent space, which makes it impossible to cluster them. So Principal component analysis allowed us to overcome the curse of dimensionality, but it introduced a new problem, the loss of rare events. This loss of rare events is often called the uniform effect.

NOTE: The stochastic Challenge of Single-Molecule Electronics

In our proposed pipeline

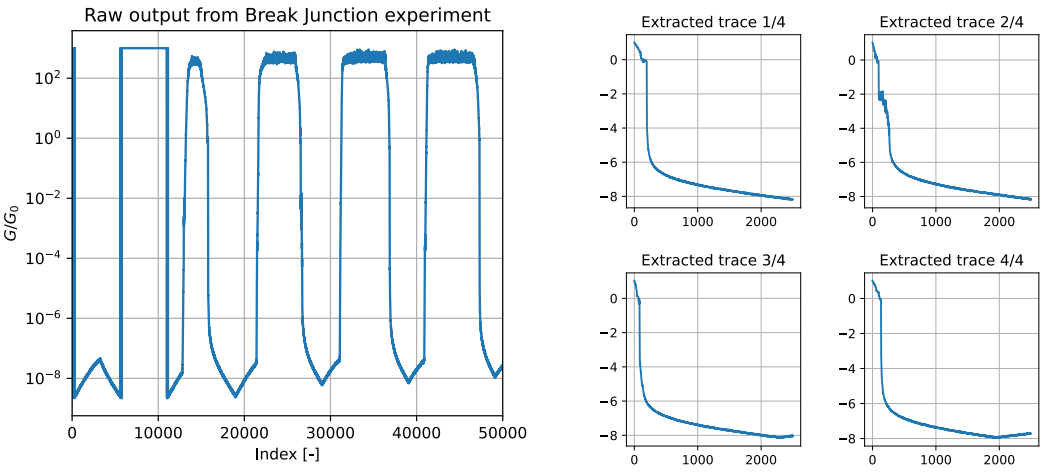
NOTE: traces -> filter -> feature extraction -> PCA -> clustering

TODO: make an image from the previous note. We could compress the data using other non-linear method, but using UMAP or t-SNE is not challenging enough for a diploma thesis. **TODO: rewrite in more academic way :)**.

NOTE: Mention UMAP, t-SNE, ..., what they do

2.4 Improvements over previous work

During our research we have improved iCluto on several levels. The first major improvement is the implementation of extracting traces from raw .txt files.



(a) Raw input (b) Extracted traces

Figure 3: Preprocessing workflow: raw conductance data as recorded by the STM-BJ setup, and four traces extracted from the raw data for further processing.

Section 3

Self-Supervised Learning

To overcome the limitations of supervised snap-back detection, we have developed a framework that learns feature representations without the need for human-labeled data. While our initial models were restricted to a small subset of manually annotated traces, this new approach leverages the full scale of our dataset (hundreds of thousands of traces). Moving beyond manual labeling mitigates the risks of subjective bias and inter-annotator inconsistency. More importantly, it allows the model to discover objective features within the molecular conductance data that transcend human intuition, potentially revealing rare events that traditional manual inspection might miss.

In our case we use unlabeled raw traces for training. A simplest form of self-supervised learning is to use the data itself as labels as shown in figure Figure 4. This approach is suitable to overcome the curse of dimensionality, but most of the trace is influenced by bulk and limit, these segments represent 80% of the data. Which means that such autoencoder will focus on learning these segments, and will not be able to learn the molecular conductance traces. This feature extraction serves as a simple demonstration of the concept of self-supervised learning, but is not suitable to overcome the uniform effect. Keep in mind that the dimensionality of the input and output is the same, for all traces regardless of snapback and limit. This limits our ability to learn the molecular conductance traces.

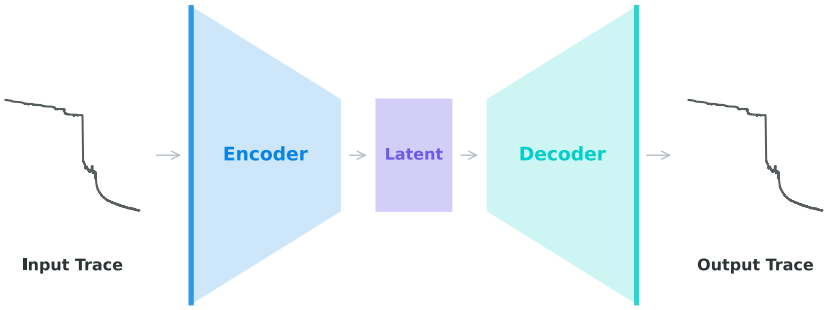


Figure 4: Conceptual diagram of an autoencoder architecture.

TODO: garbege in garbage out

A possible improvement is mask a part of the input trace and force the autoencoder to reconstruct the masked part. These “masked autoencoders” are often abbreviated as MAEs. **TODO: cite this!**

TODO: Explain the concept of self-supervised learning in more detail Autoencoders and MAEs learn a representation of the whole trace, regardless of the presence of

molecules. In this thesis we want to go a step deeper, where we want to identify segments of traces and perform clustering or similarity search based on those segments.

3.1 Deep Learning Architectures

A novel approach to self-supervised learning is called self distillation with no labels (DINO)**TODO: cite this! meta DINO v1-v3**.

In this architecture we have two networks, a student and a teacher network. They share the same architecture, but their weights are different. Student is updated using standard backpropagation, while the teacher network is updated using an exponential moving average of the student network weights. A student is given a local crop of the trace, on the other hand the teacher is given a global crop of the same trace. Yet the student network is forced to match the output of the teacher network. Which is what where the *distillation* comes from.

This approach has the advantage that it can be used to learn features from the data without the need for human-labeled data. Both of our networks has to agree, by embeding the information into a shared latent space. We keep track of their prediction using an MSE loss.

Such network has only a training stage, hence, we do not have a labels we can validate against.

3.1.1 Data Augmentation

TODO: Describe data augmentation techniques

Section 4

Implementation

4.1 Data Preparation

4.1.1 Curating the Dataset

Training dataset consists of more than 400,000 traces, which include molecules synthesized by Dr. Nejedlý, namely JIN206, JIN466, JIN467, JIN536, and JIN537. The named molecules were dissolved in MesH solvent. **TODO: What solvent?**

As our validation dataset we have chosen 41000 traces of R296 molecules.

NOTE: R296 is a molecule synthesized by Jindra or R as Rybacek?

Thanks to their simple structure, they are a good candidate for a validation dataset and we know how much of a displacement they have (13 Å).

4.2 Model Architecture

The architecture and distillation process are illustrated in Figure 5.

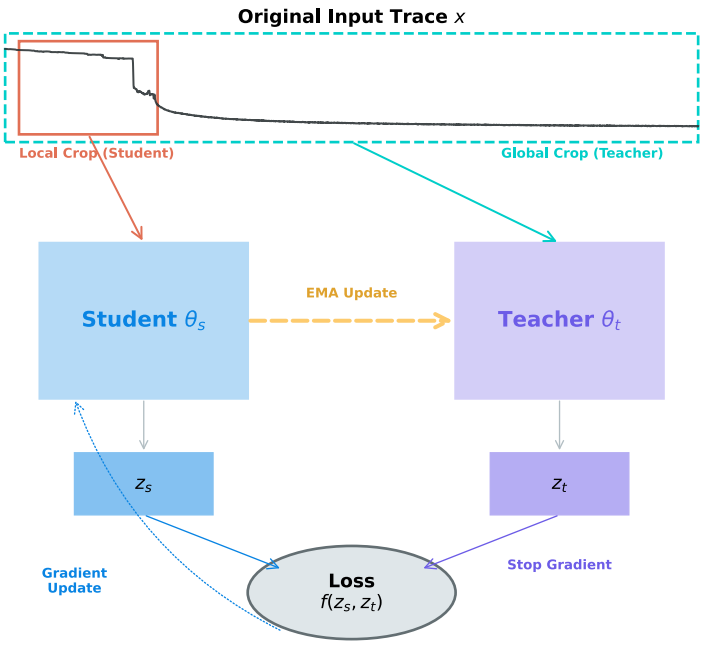


Figure 5: System architecture of the DINO framework for self-supervised trace representation learning. The Student network learns from local crops while being regularized by a Teacher network updated via an EMA path.

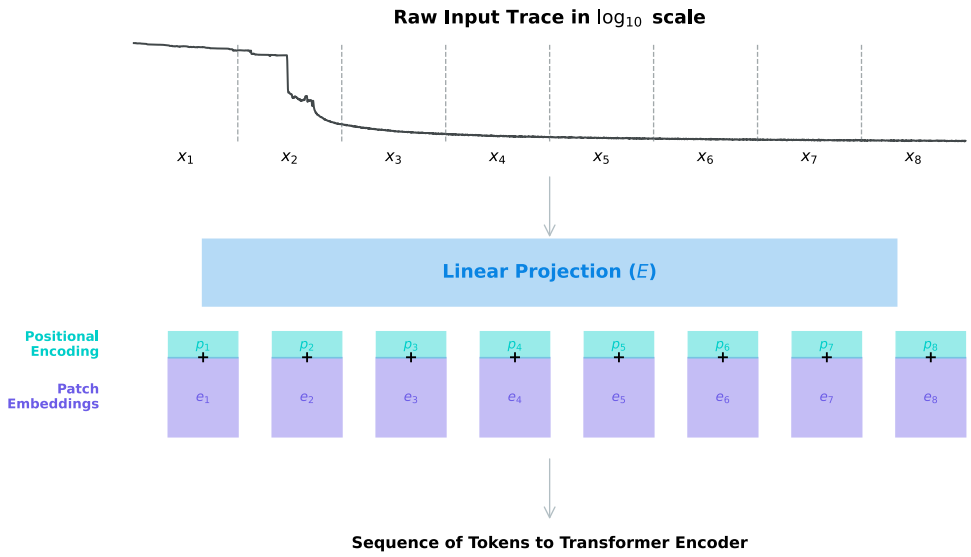


Figure 6: Illustration of trace segmentation into patches for feature embedding.

4.3 Implementing DINO

4.3.1 Hyperparameter Analysis

Apart from classical hyperparameters like learning rate, batch size, and number of epochs, DINO's crucial hyperparameters are patch size and embed dim. We evaluated the impact of these hyperparameters on the model's performance.

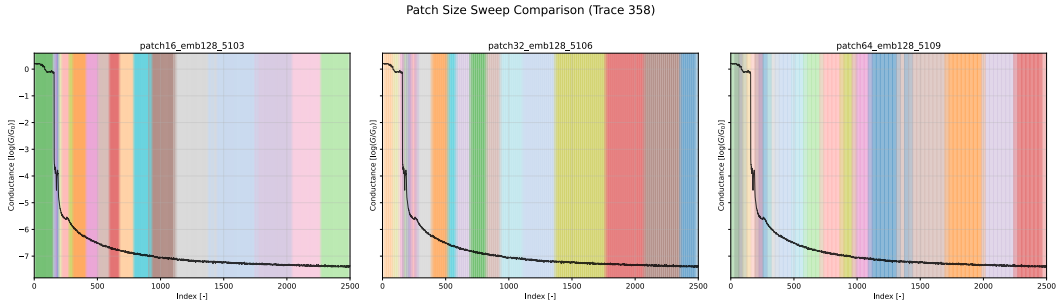
We have trained 15 different combinations of patch sizes and embedding dimensions. For few epochs to see the trend.

The smaller the patch size the finer details of the trace are captured, this allows us to segment a very small events like snapbacks (Compare the first and last column of Figure 7, mainly in the snapback region of 7a).

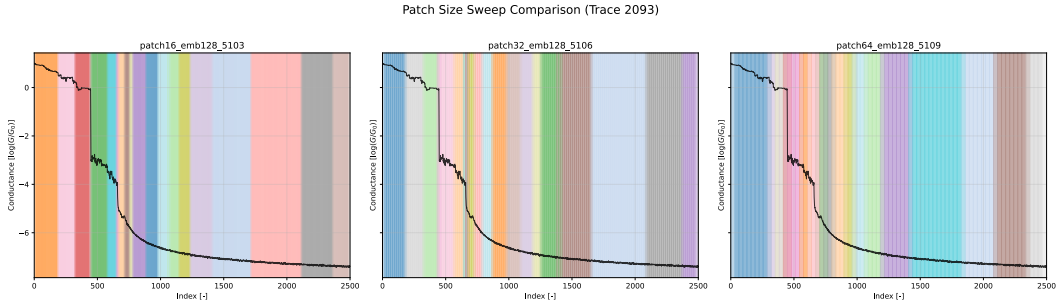
The larger the embedding dimension the more and finer information can be stored in the latent space, this allows to capture more complex patterns in the data. Our observations do not directly confirm this hypothesis, as we can see in Figure 8, the 256 dimension is not always better than 128 dimension, yet it is on par with it. Upon a closer inspection of Figure 8c we can see that the bulk segment is classified by 4 different labels in the 128 dimension, while it is classified by 3 labels in the 256 dimension, which is more consistent with the rest of the traces.

4.3.2 DINO training

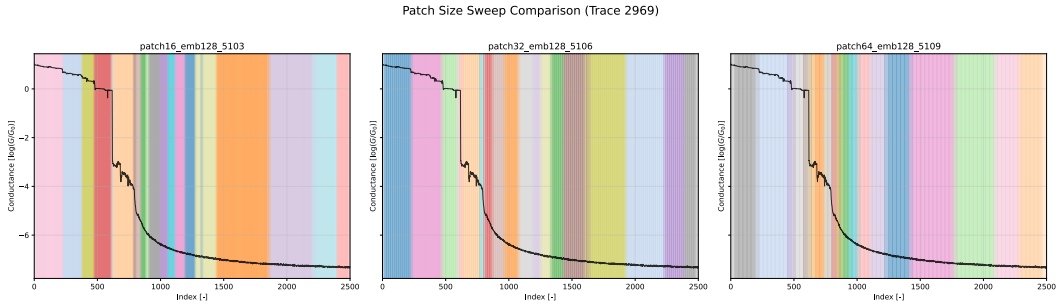
All DINO training runs were performed on the High-Performance Computing Cluster (HPCG), specifically utilizing the 'd' section computing nodes. These nodes are equipped with NVIDIA L40S GPUs featuring 48 GB of GDDR6 memory (CUDA architecture 89). The underlying system consists of dual AMD EPYC 9654 96-core processors running at 2.4 GHz and 363 GB of RAM, providing the necessary computational throughput for large-scale self-supervised learning on trace datasets.



(a) Trace 358



(b) Trace 2093



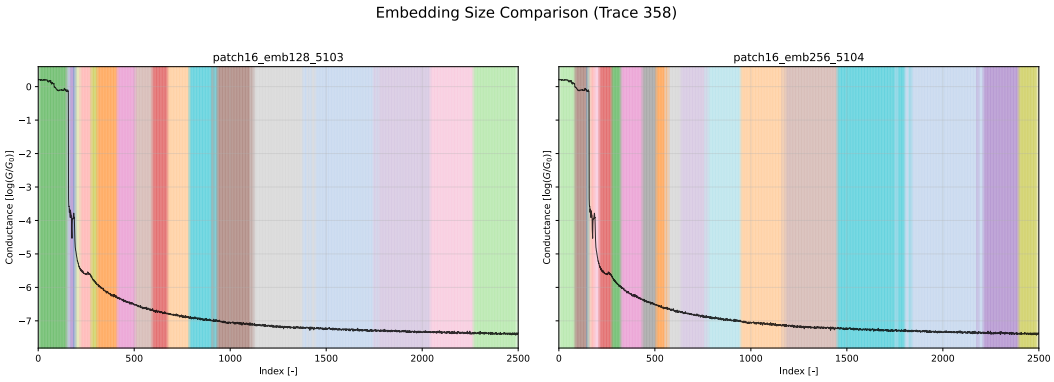
(c) Trace 2969

Figure 7: Evaluation of DINO performance during pretraining for different patch sizes on 3 randomly selected traces.

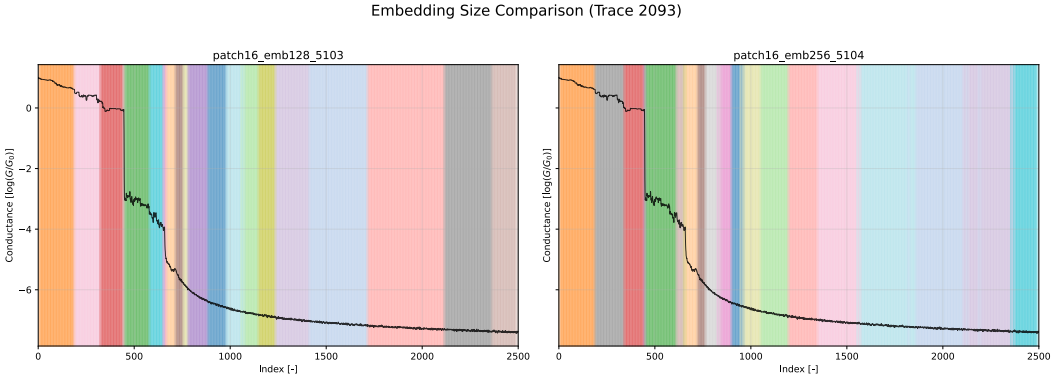
After determining the optimal hyperparameters, we considered training the model on patches of size 8, 16, 32, and 64 and embedding dimension of 128 and 256.

4.3.2.1 Validation

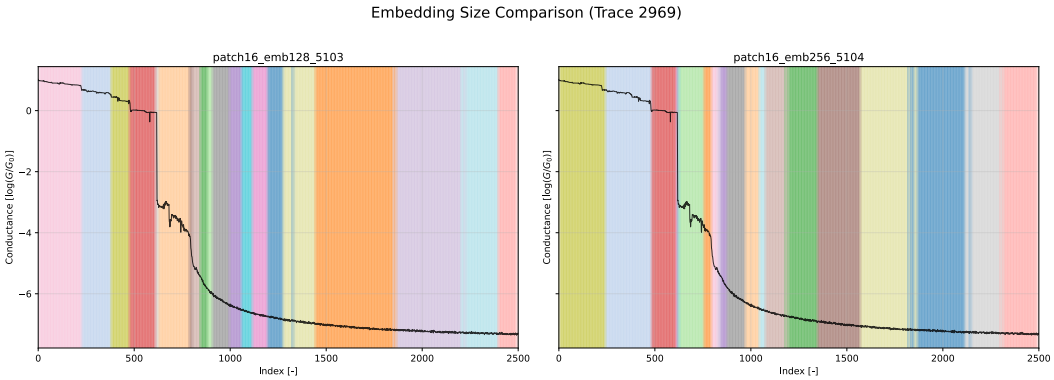
NOTE: Add also batch center_max and center_mean?



(a) Trace 358



(b) Trace 2093



(c) Trace 2969

Figure 8: Evaluation of DINO performance during pretraining for different embedding dimensions on 3 randomly selected traces.

Configuration (Patch, Dim)	Backbone Parameters	Projection Head	Total Parameters
Patch 64, Dim 128	811,776	722,176	1.53 Million
Patch 64, Dim 256	3,196,416	787,712	3.98 Million
Patch 32, Dim 128	817,664	722,176	1.54 Million
Patch 32, Dim 256	3,208,192	787,712	4.00 Million
Patch 16, Dim 128	835,584	722,176	1.56 Million
Patch 16, Dim 256	3,244,032	787,712	4.03 Million
Patch 8, Dim 128	874,624	722,176	1.60 Million
Patch 8, Dim 256	3,322,112	787,712	4.11 Million

Table 1: Parameter counts for different DINO model configurations.

Run Configuration	Final Loss	Min Loss	Best Epoch	Max Epochs	Duration (Days)
Patch 16, Dim 128	3.8004	2.5942	19	152	4.850
Patch 16, Dim 256	3.9032	2.2992	16	82	4.847
Patch 8, Dim 128	4.1394	2.6277	16	72	4.849
Patch 8, Dim 256	3.4008	2.7912	6	39	4.848

Table 2: Summary of DINO training runs executed on the HPCG infrastructure.

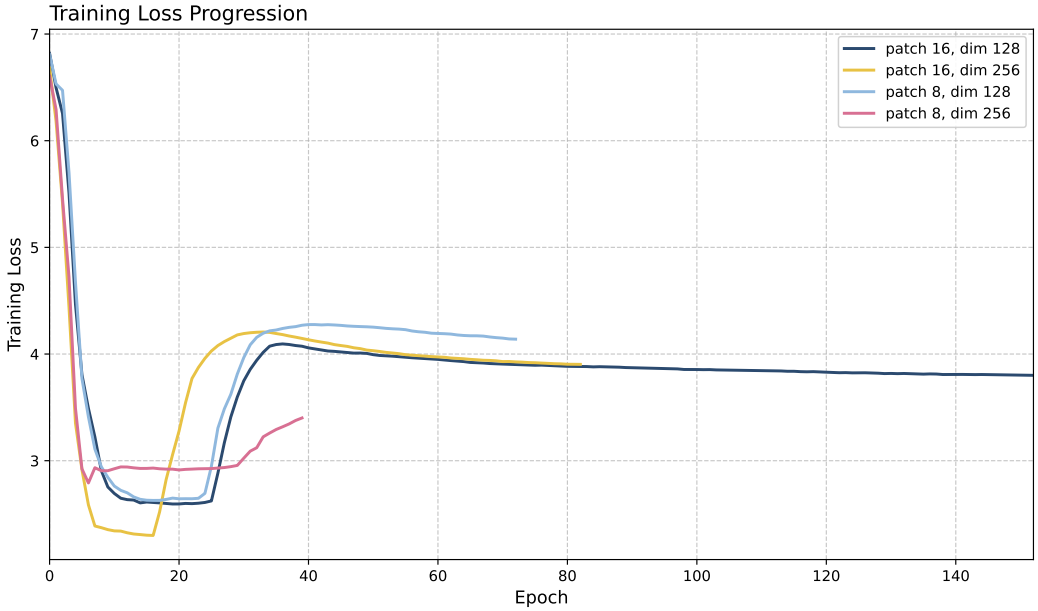


Figure 9: Progression of DINO training loss over epochs.

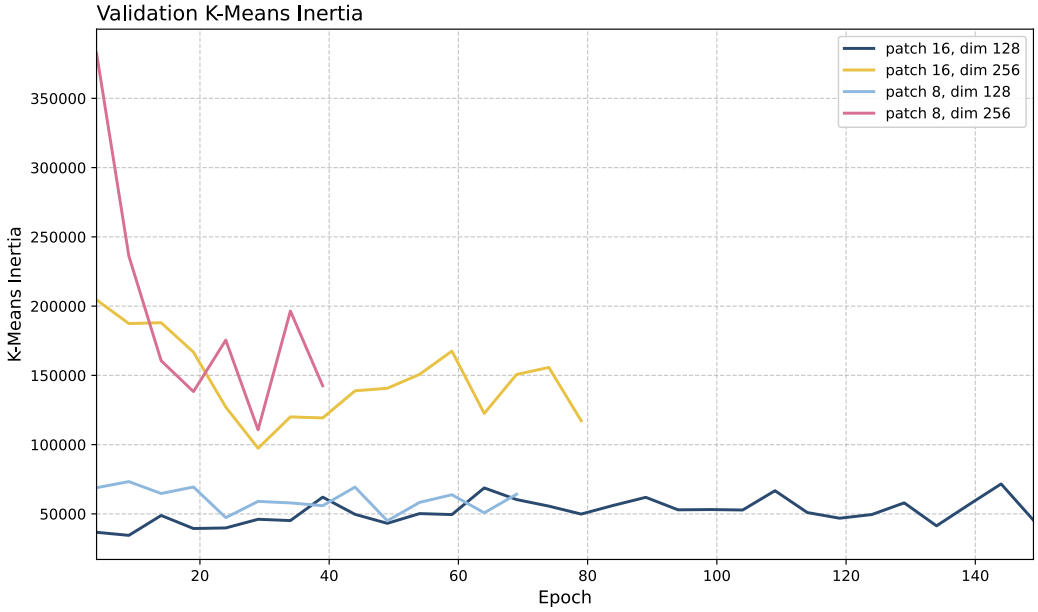


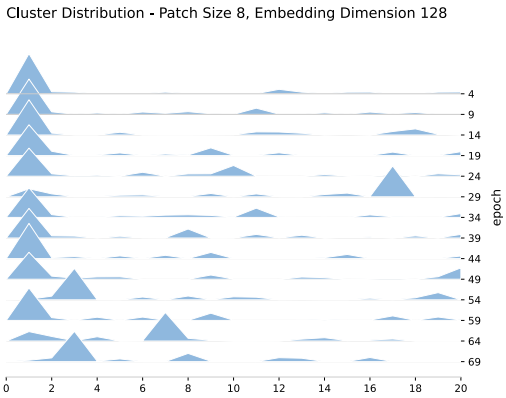
Figure 10: Validation K-Means inertia indicating cluster compactness over training.

4.4 iCluto toolkit

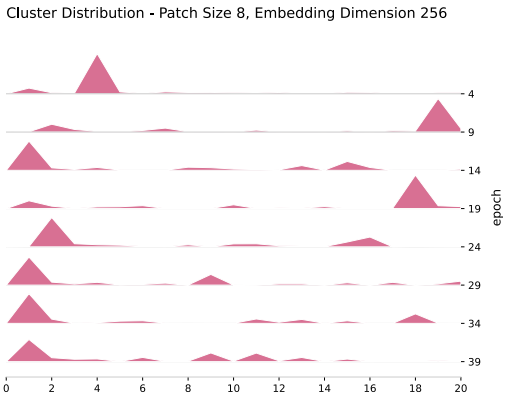
NOTE: How to install?

NOTE: How to use?

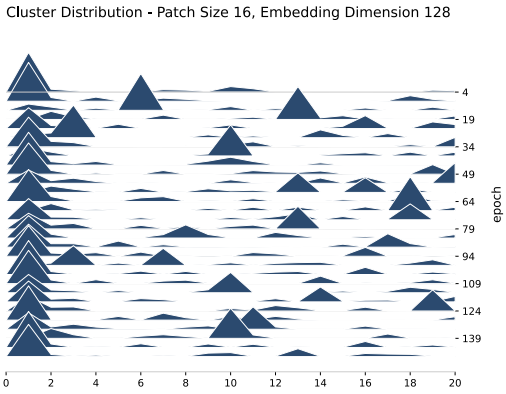
NOTE: Tutorials?



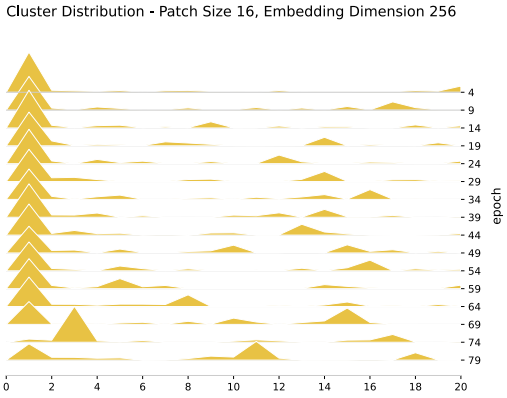
(a) Patch 8, Dim 128



(b) Patch 8, Dim 256



(c) Patch 16, Dim 128



(d) Patch 16, Dim 256

Figure 11: Comparison of cluster distribution evolution (waterfall plots) for different patch sizes (8 and 16) and embedding dimensions (128 and 256).

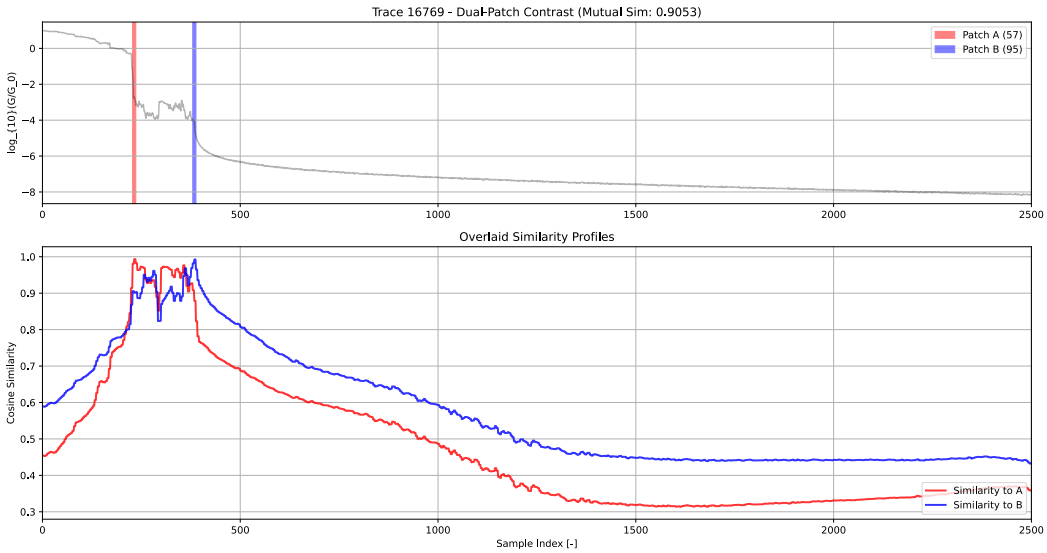


Figure 12: Demonstration of internal contrast for Trace 16769, illustrating the mutual cosine similarity (0.9684) between different patches of the same trace.

Section 5

Results

Show bovw clustering comment on an elbow method for finding the number of visual words.

Why it failed?

To evaluate the feature consistency of our model, we calculate the mutual cosine similarity between patch embeddings. For any two patches A and B, the mutual similarity $S_{\{A,B\}}$ is defined as the cosine similarity of their respective embedding vectors $\boldsymbol{e}_A, \boldsymbol{e}_B \in \mathbb{R}^d$:

$$\text{cosine}(\boldsymbol{e}_A, \boldsymbol{e}_B) = \frac{\boldsymbol{e}_A \cdot \boldsymbol{e}_B}{\|\boldsymbol{e}_A\|_2 \, \|\boldsymbol{e}_B\|_2 + \varepsilon}$$

TODO: Comment and compare iCluto's DINO with META's DINOv3 distilled model. TODO: What is a distelled model? TODO: Compare Parameters of DINOv3 and iCluto's DINO

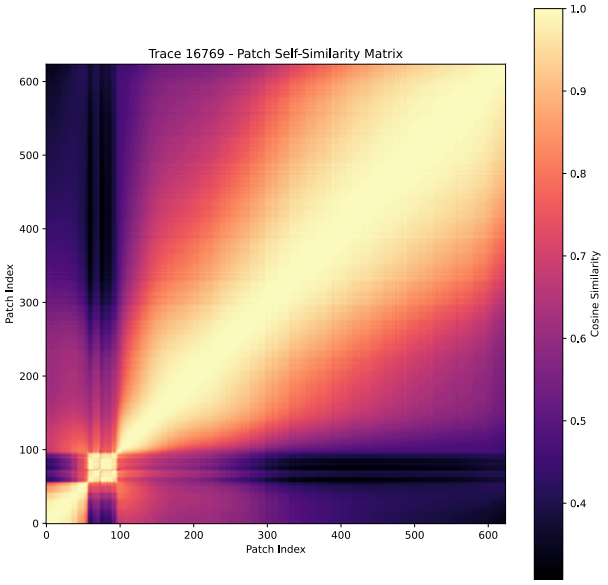


Figure 13: Self-similarity matrix for Trace 16769. The matrix highlights the cosine similarity between all patch embeddings within a single trace.

Section 6

Conclusion

6.1 Future Work

Developement of iCluto eventhough this section is called future work, we will discuss the development of iCluto here. icluto is `spec-driven` developed, this project started Created on October 10, 2022, it went through several iterations and improvements, reflecting the needs of Ivo Starý's group. some older parts of its code base were optimized for different purposes, for example validation and filetering of raw traces.

NOTE: iCluto DINO as foundational model.

NOTE: JEPA

Bibliography

[1] O. Klimt, “Break junction data clustering using supervised and unsupervised machine learning,” Bachelor's Thesis, Prague, Czech Republic, 2024. [Online]. Available: http://hdl.handle.net/10467/115225

Appendix A

An example appendix

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.