NAML
NAVAL APPLICATIONS OF MACHINE LEARNING

Siamese Networks Effectively Learn Robust Features for Malware Attribution

Nathan Clark
Machine Learning Engineer
noblis

Introduction

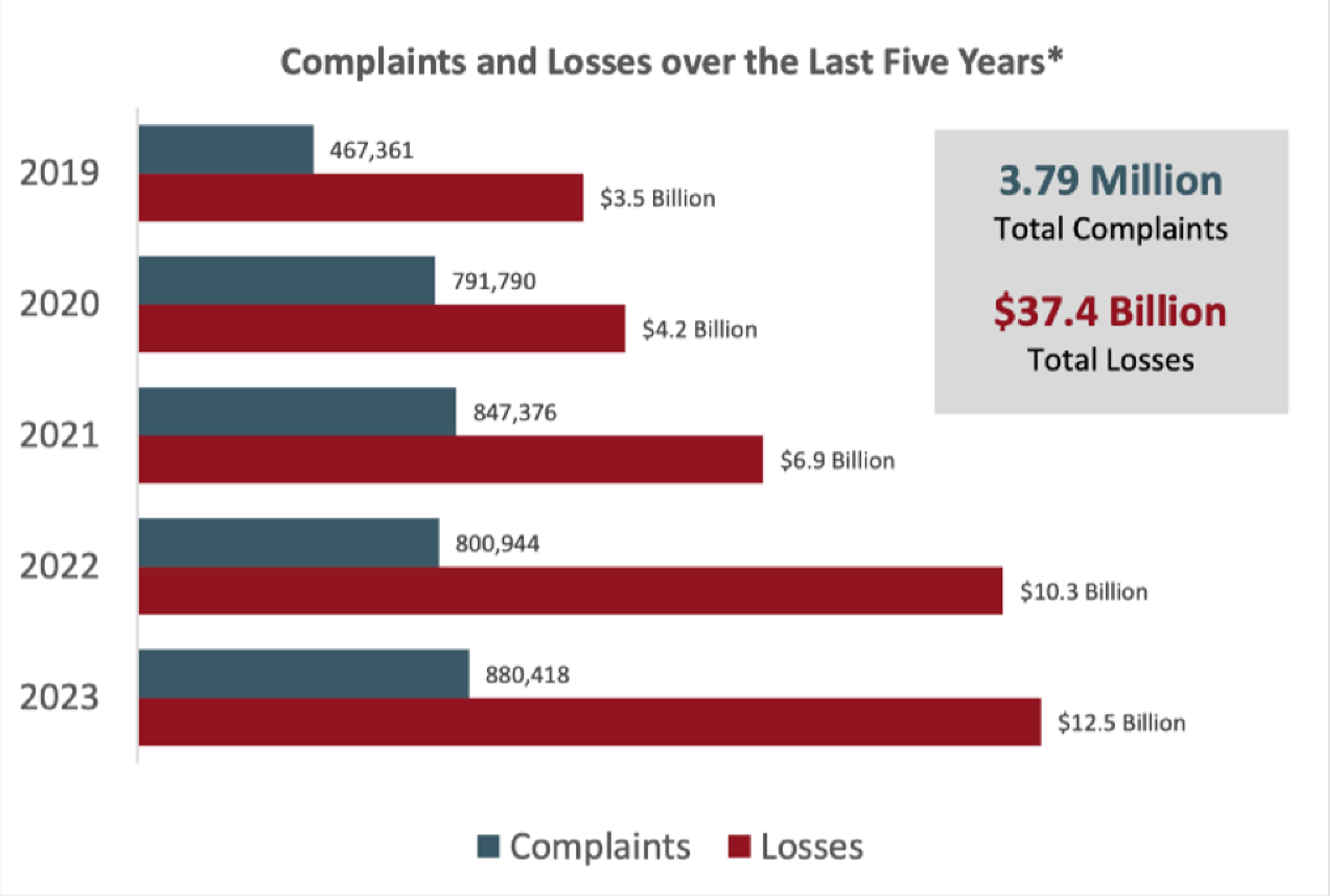
Cyberattacks on companies and infrastructure have grown significantly and will continue to evolve over time, posing a serious threat to national security.

Recent cyber expertise shortages present difficulty filling commercial cyber-focused positions.

Cyber forensic expertise is in short supply and agencies struggle to scale identification from cyberattack to threat actor.

LAST FIVE YEARS

Over the last five years, the IC3 has received an average of 758,000 complaints per year. These complaints address a wide array of Internet scams affecting individuals across the globe.[3]



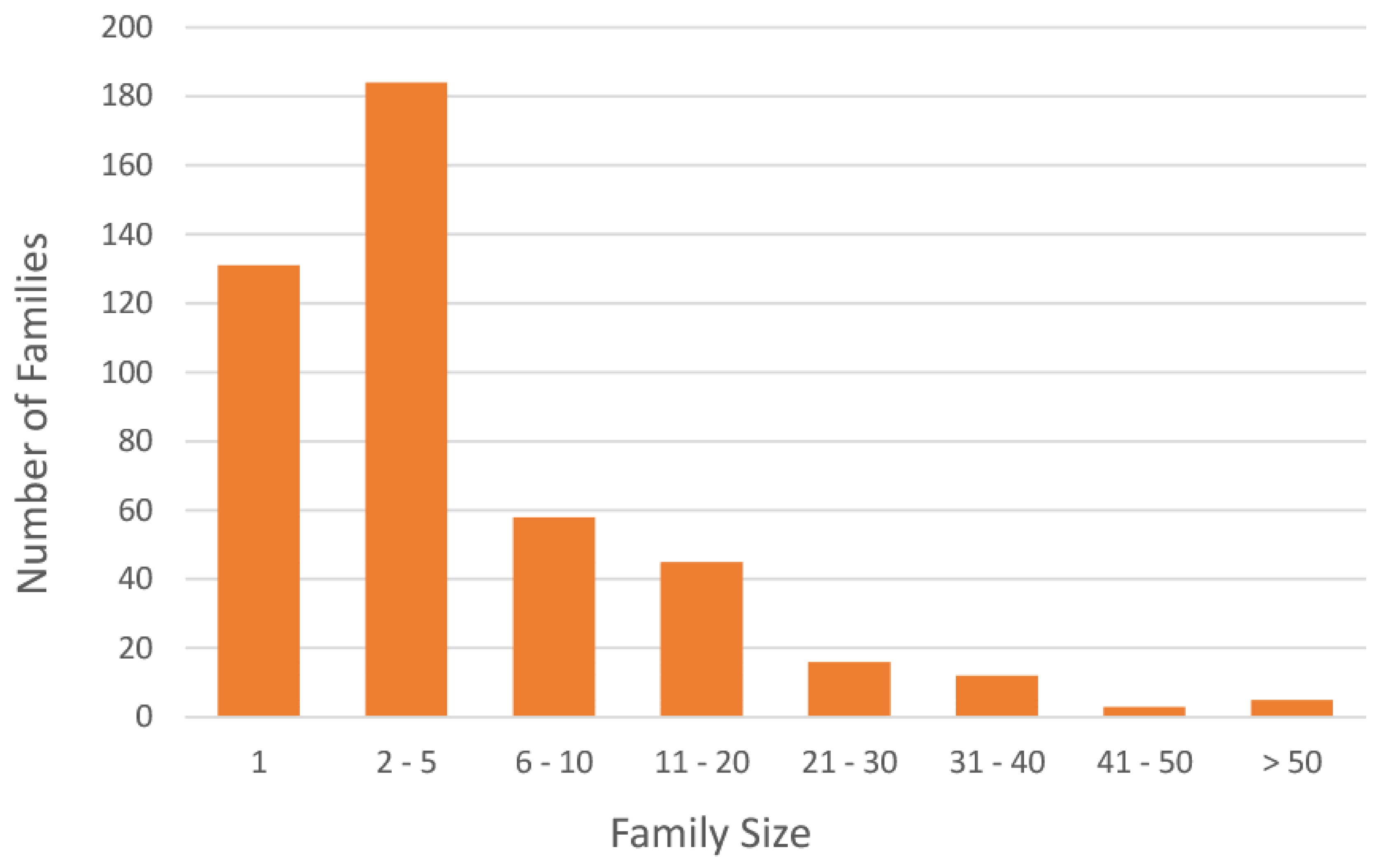
Introduction

We present a novel deep learning approach for automated classification of malware by attribution categories on the tasks of:

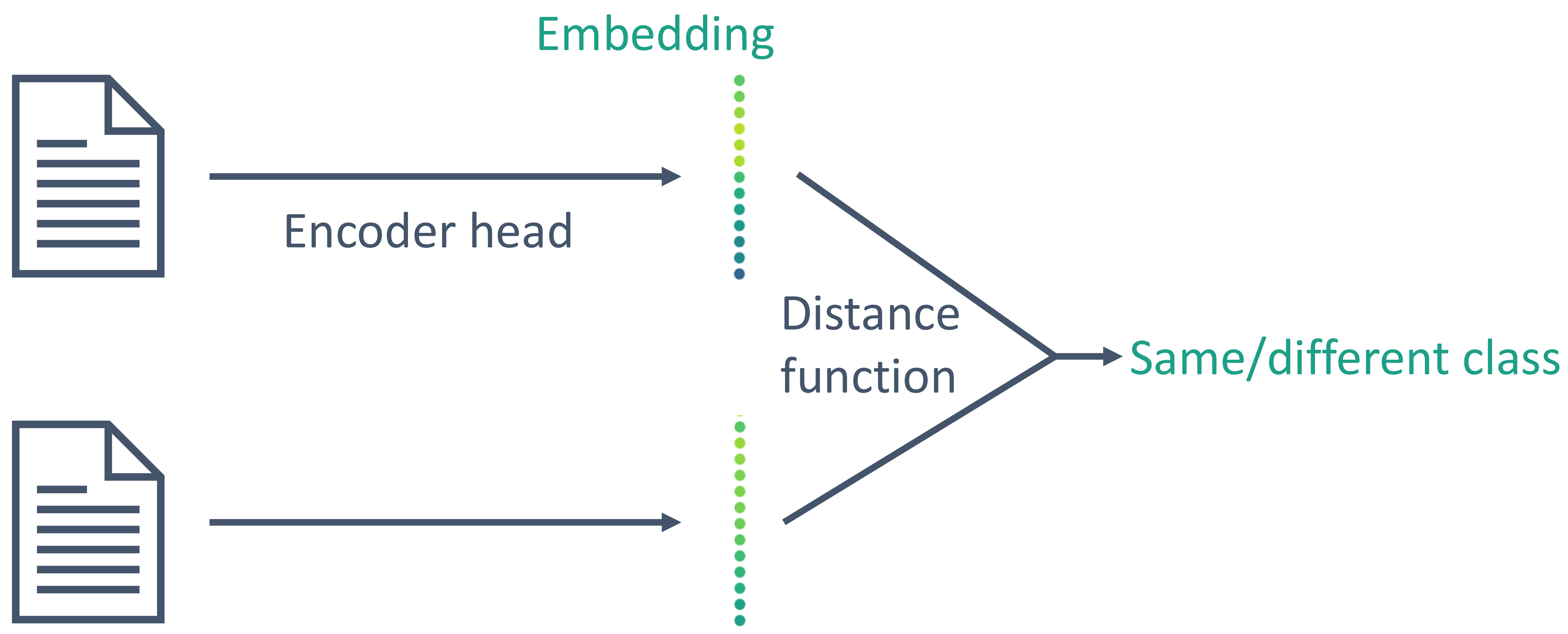
- pairwise classification of a pair of samples into same or different classes
- one-to-many classification of malware samples into attribution categories

Approach

We leverage the MOTIF malware dataset as a benchmark, with 3095 malware samples labelled into 454 categories.



We leverage a Siamese neural network architecture, modelling the problem as a similarity metric learning problem.



Motivation

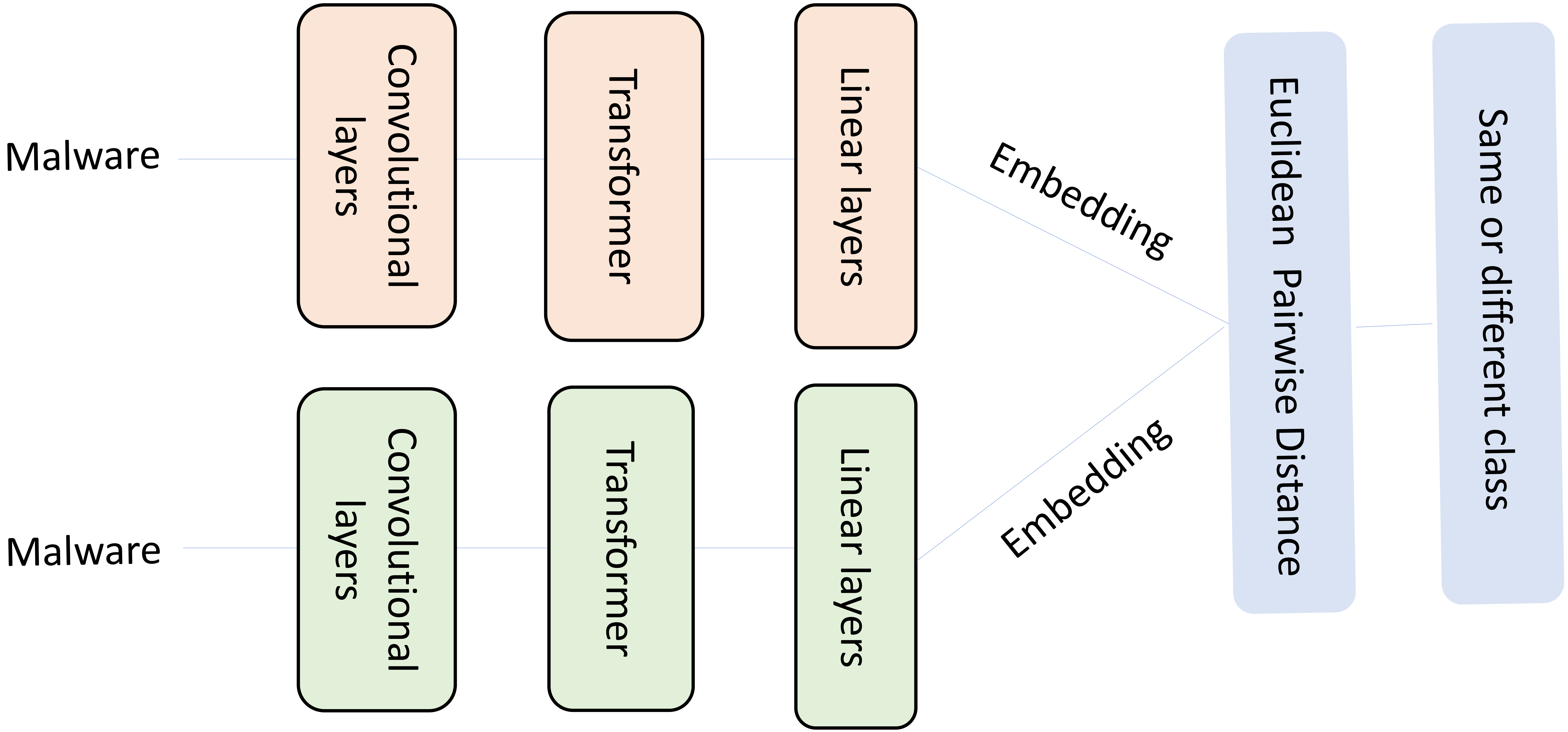
Dataset	Precision	Recall	F1 Measure	Accuracy
Drebin	0.954	0.884	0.918	Unknown
Malicia	0.949	0.680	0.792	Unknown
Malsign	0.904	0.907	0.905	Unknown
MalGenome*	0.879	0.933	0.926	Unknown
Malheur	0.904	0.983	0.942	Unknown
MOTIF-Default	**0.763**	**0.674**	**0.716**	**0.468**
MOTIF-Alias	0.773	0.700	0.735	0.506

Existing approaches (e.g. AVClass above) have poor accuracy on the MOTIF dataset.

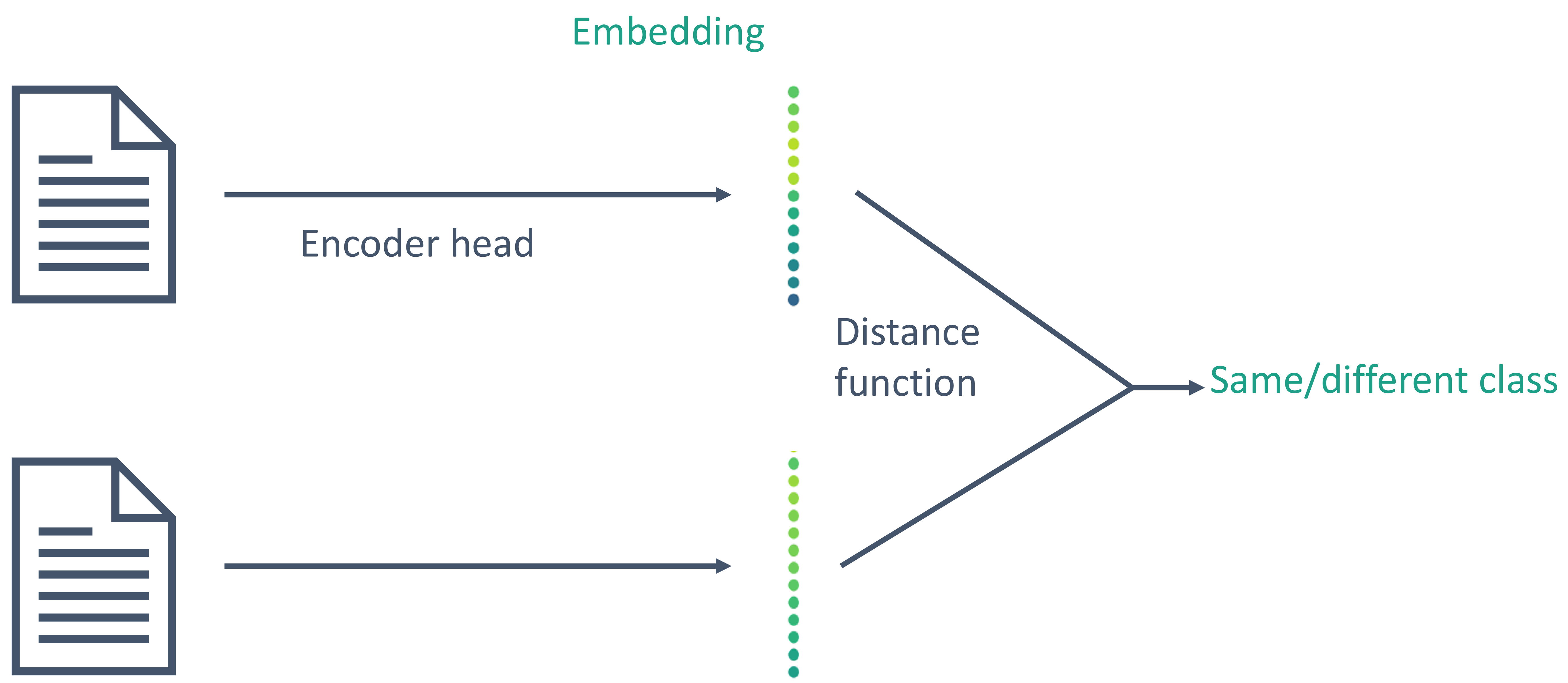
Further, these approaches lack the ability to scale to classification categories and datasets beyond the limited available training data.

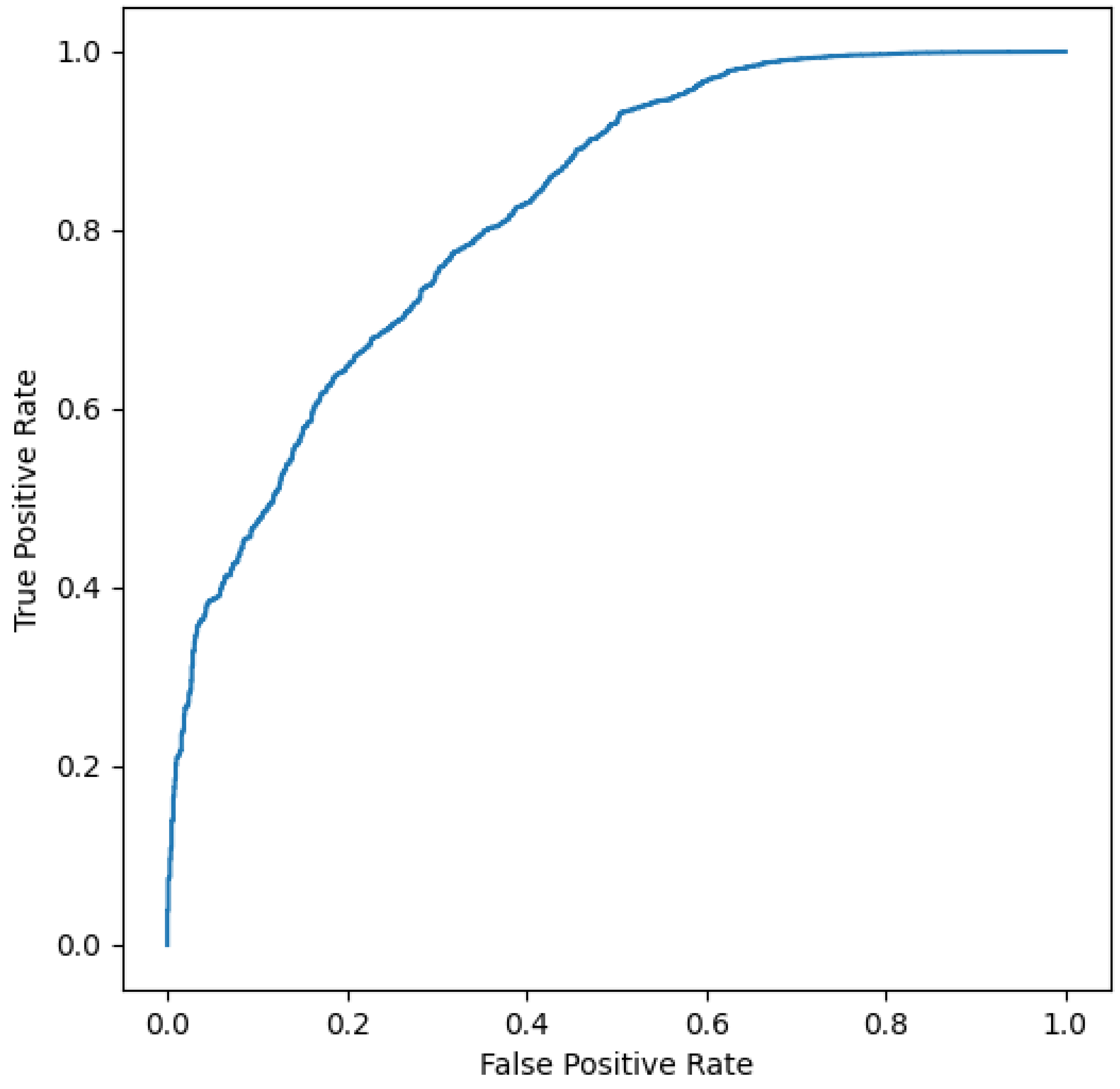
We propose:
- leveraging the learned similarity metric of the Siamese network combined with clustering to solve the open set problem
- applying malware perturbations as data augmentation to increase effective data diversity and reduce reliance on unimportant features

Model: Siamese Transformer



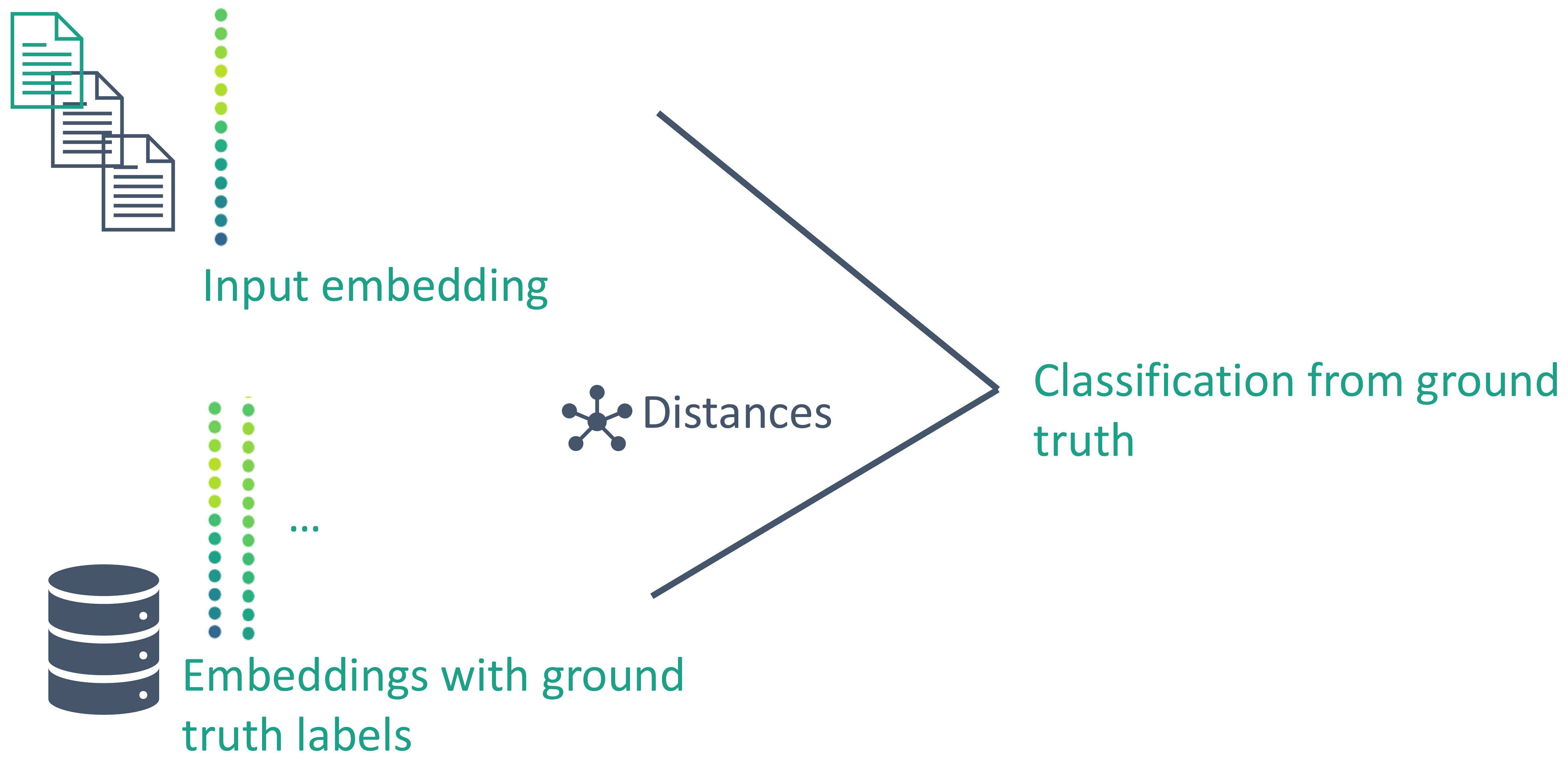
Results

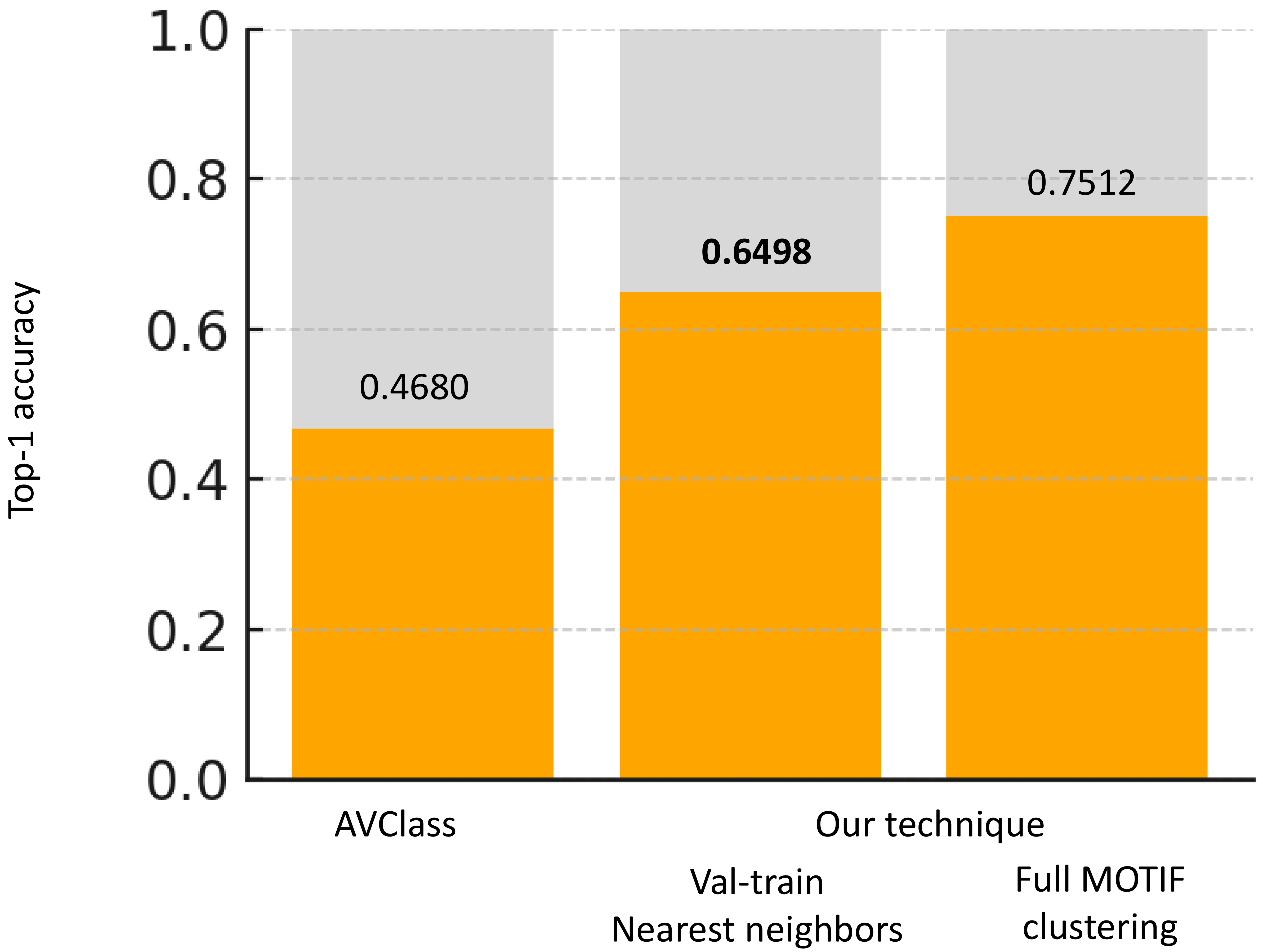




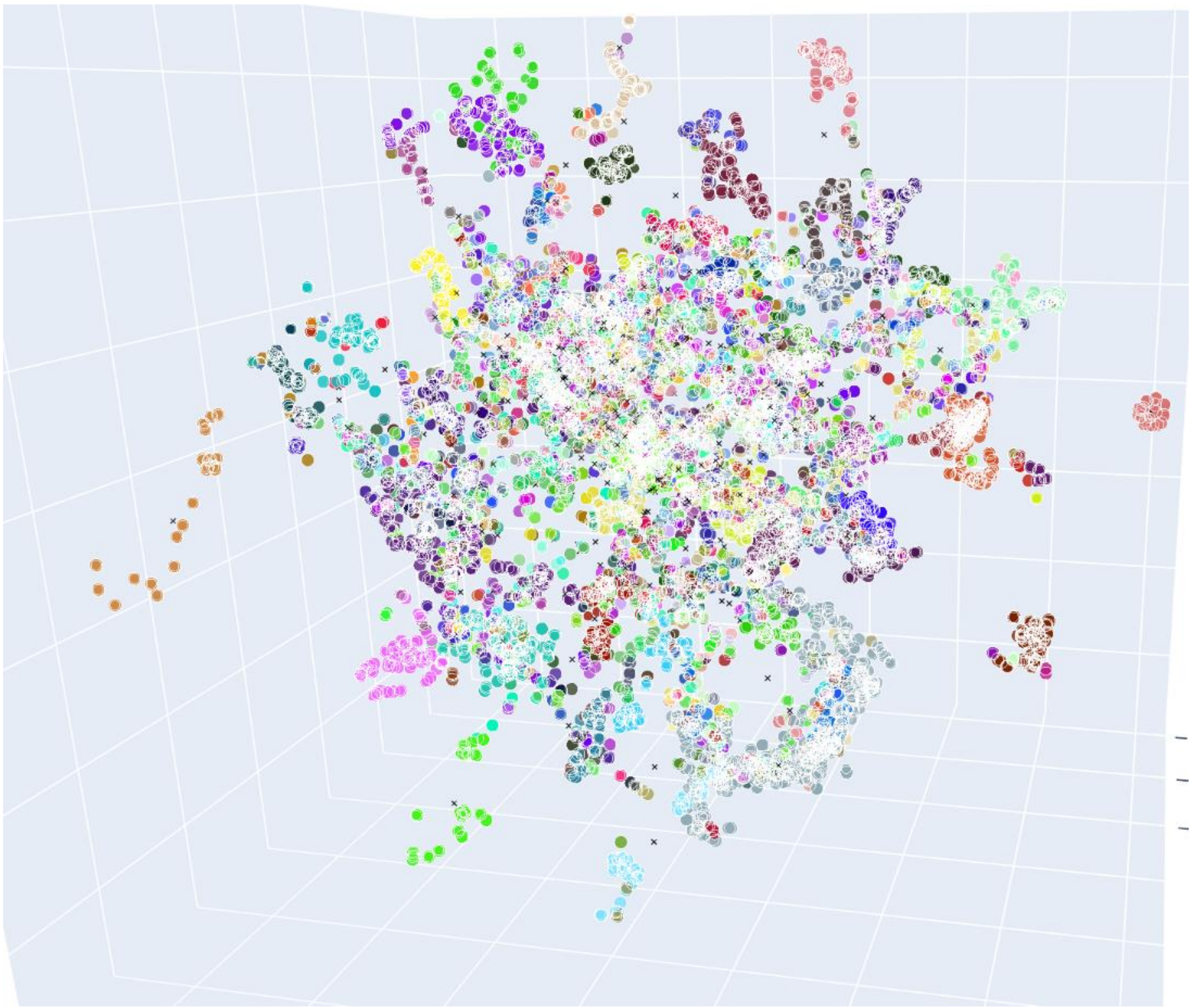
Validation ROC curve for Siamese transformer model on pairwise comparison task (auc = 0.8209)

Results


Input embedding
Distances
Classification from ground truth
...
Embeddings with ground truth labels


Top-1 accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.4680
0.6498
0.7512
AVClass
Our technique
Val-train Nearest neighbors
Full MOTIF clustering

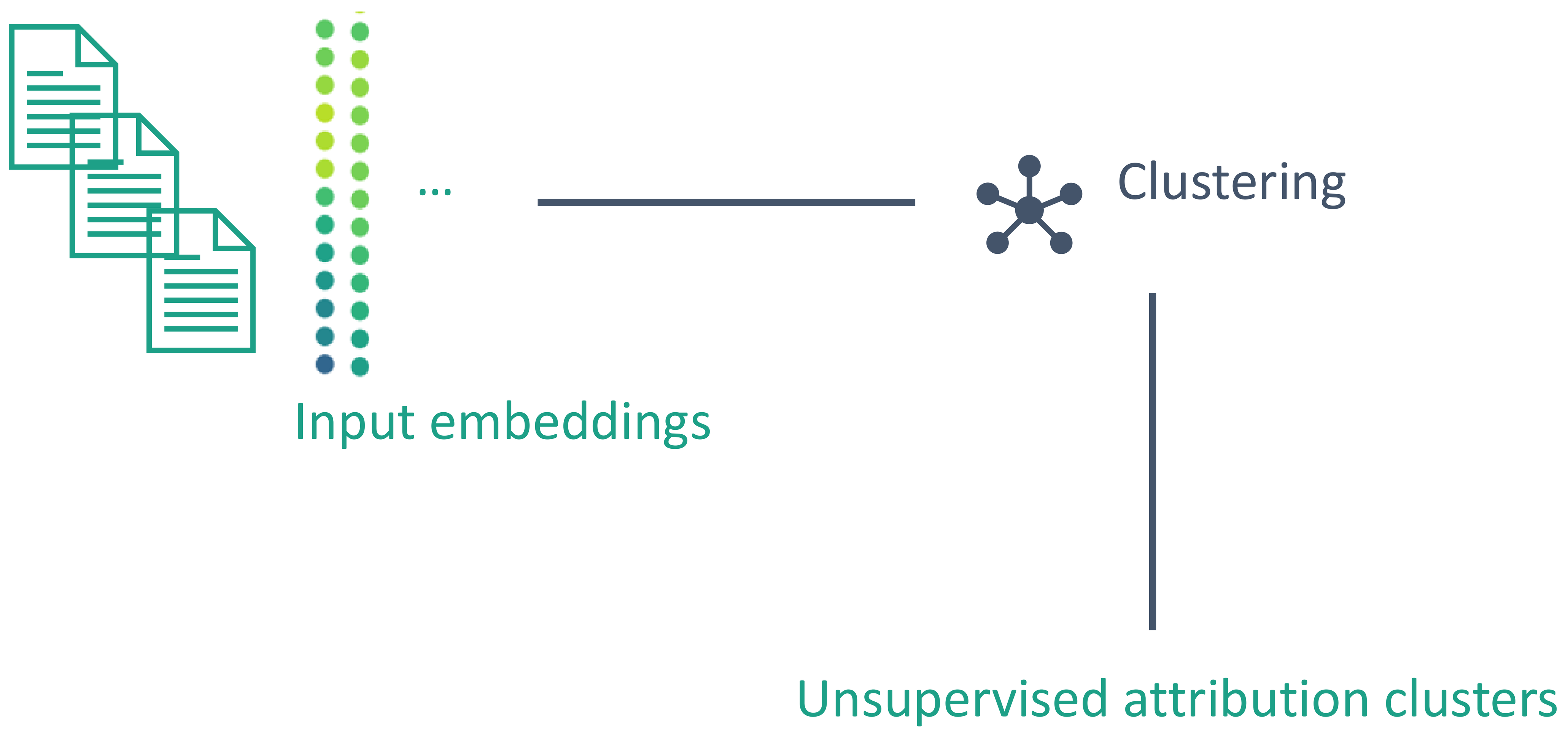
Results



3D t-SNE plot produced from MOTIF malware embeddings

We further find the embedding space produced by our model is meaningful; particularly, nearby clusters by our model in the embedding space represent closely related threat actors.

Results



Running KNN on the produced embeddings compared to the ground truth labels, we obtain:

Normalized mutual information score: 0.5736

Rand index: 0.0779

Perturbations

```
0x40c78f    push ebp
0x40c790    mov ebp, esp
0x40c792    sub esp, 0xc
0x40c795    push esi
0x40c796    mov esi, dword ptr [0x41ac40]
0x40c79c    mov ecx, dword ptr [0x41abdd]
0x40c7a2    sub esi, ecx
0x40c7a4    xor esi, dword ptr [esi+ecx*1]
0x40c7a7    sub ecx, edx
0x40c7a9    mov esi, 0xf89c85b9
0x40c7ae    mov dword ptr [ebp-0x8], esi
0x40c7b1    sub dword ptr [0x42b010], edi
0x40c7b7    mov dword ptr [ebp-0x4], 0xf89c85b8
0x40c7be    and dword ptr [0x42901c], 0x0
```

…81550b9951dad52aadcbb315
2ed7e0cb196f240f18bb328…

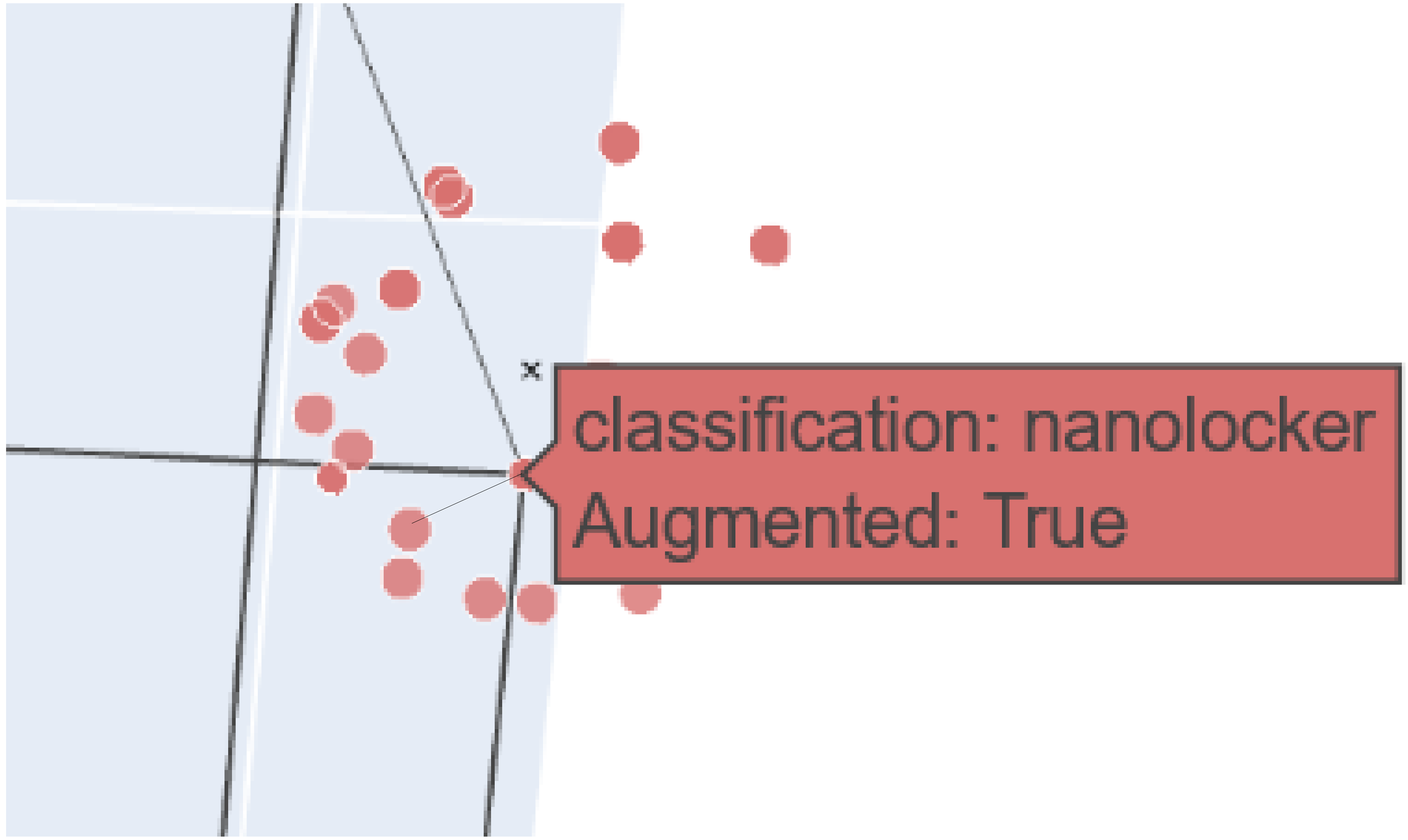
Perturbation

```
0x40c78f    push ebp
0x40c790    mov ebp, esp
0x40c792    sub esp, 0xc
0x40c795    push esi
            nop
0x40c796    mov esi, dword ptr [0x41ac40]
0x40c79c    mov ecx, dword ptr [0x41abdd]
0x40c7a2    sub esi, ecx
0x40c7a4    xor esi, dword ptr [esi+ecx*1]
0x40c7a7    sub ecx, edx
0x40c7a9    mov esi, 0xf89c85b9
0x40c7ae    mov dword ptr [ebp-0x8], esi
0x40c7b1    sub dword ptr [0x42b010], edi
0x40c7b7    mov dword ptr [ebp-0x4], 0xf89c85b8
0x40c7be    and dword ptr [0x42901c], 0x0
```

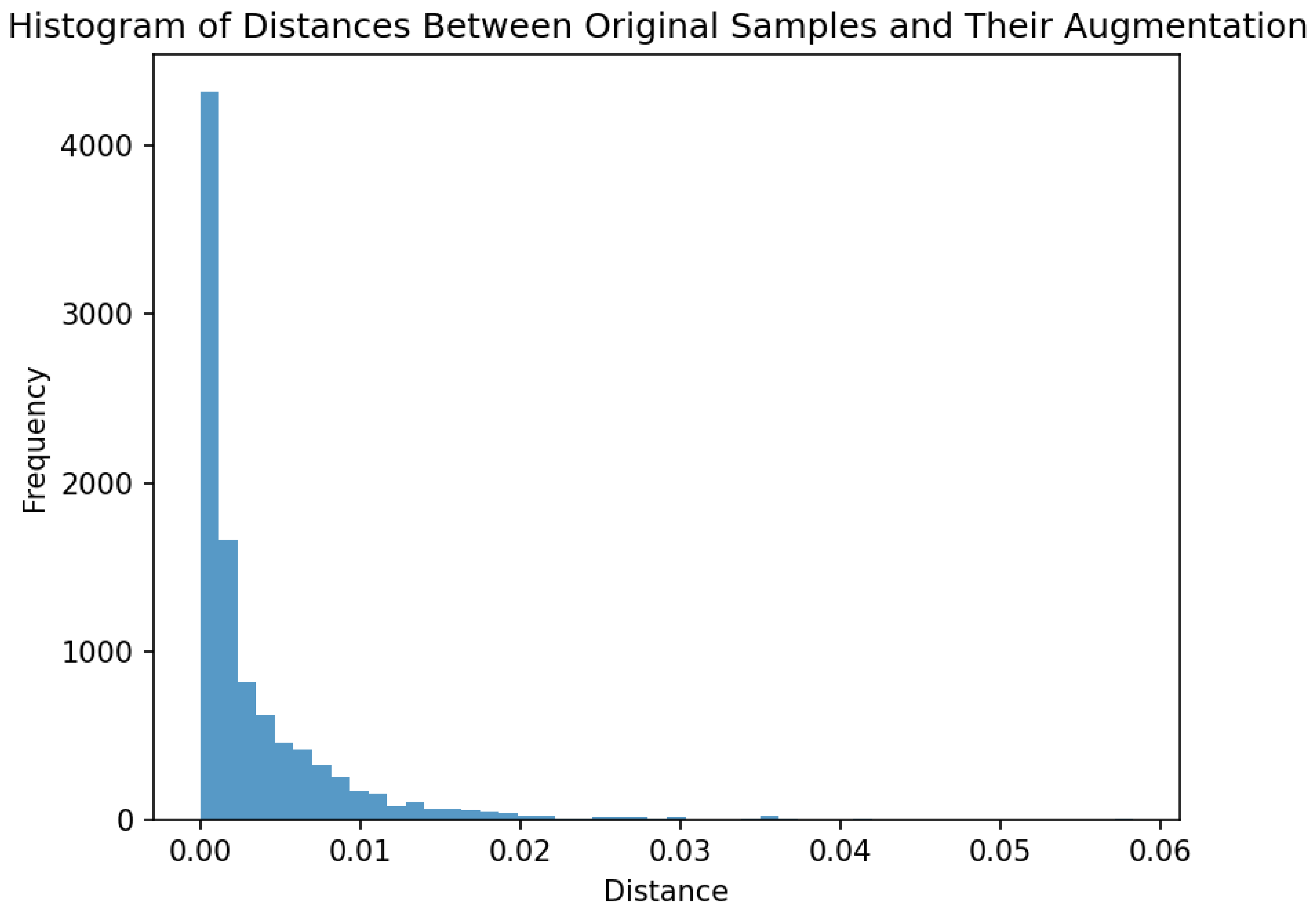
…815590 0b9951dad52aadcbb3
152ed7e0cb196f240f18bb3…

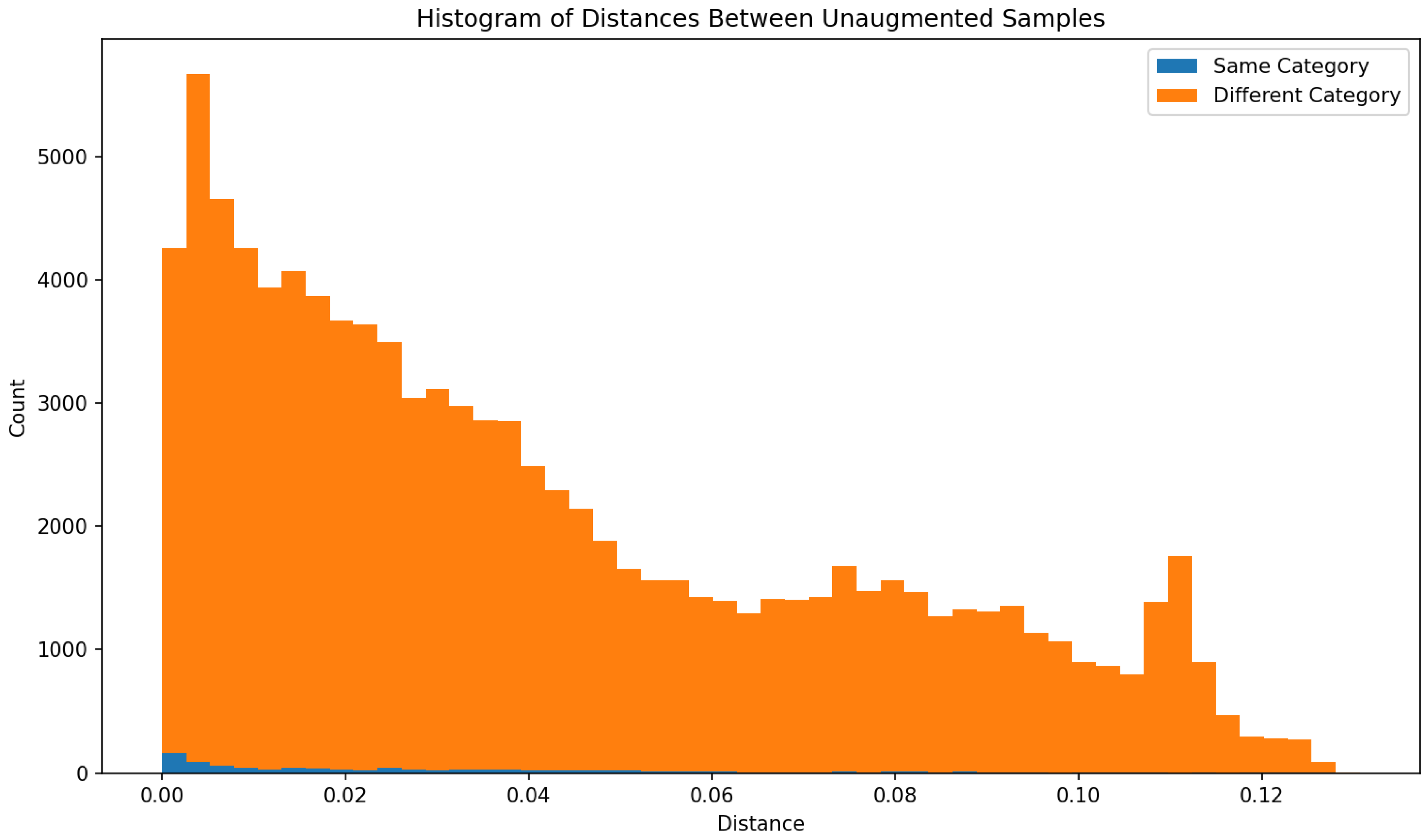
Perturbations

Our embedding space shows our model is ***robust to perturbations***

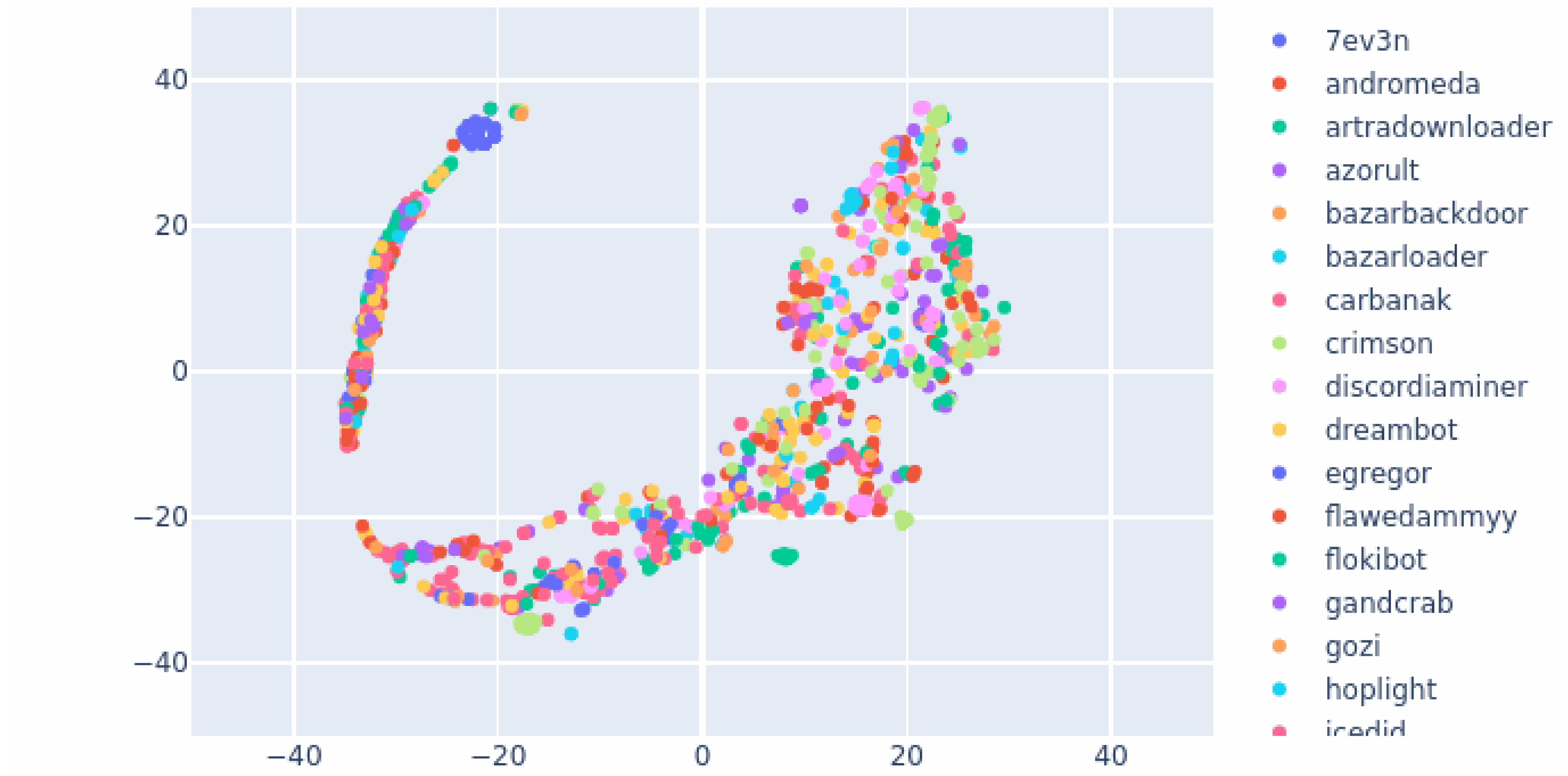


Perturbations





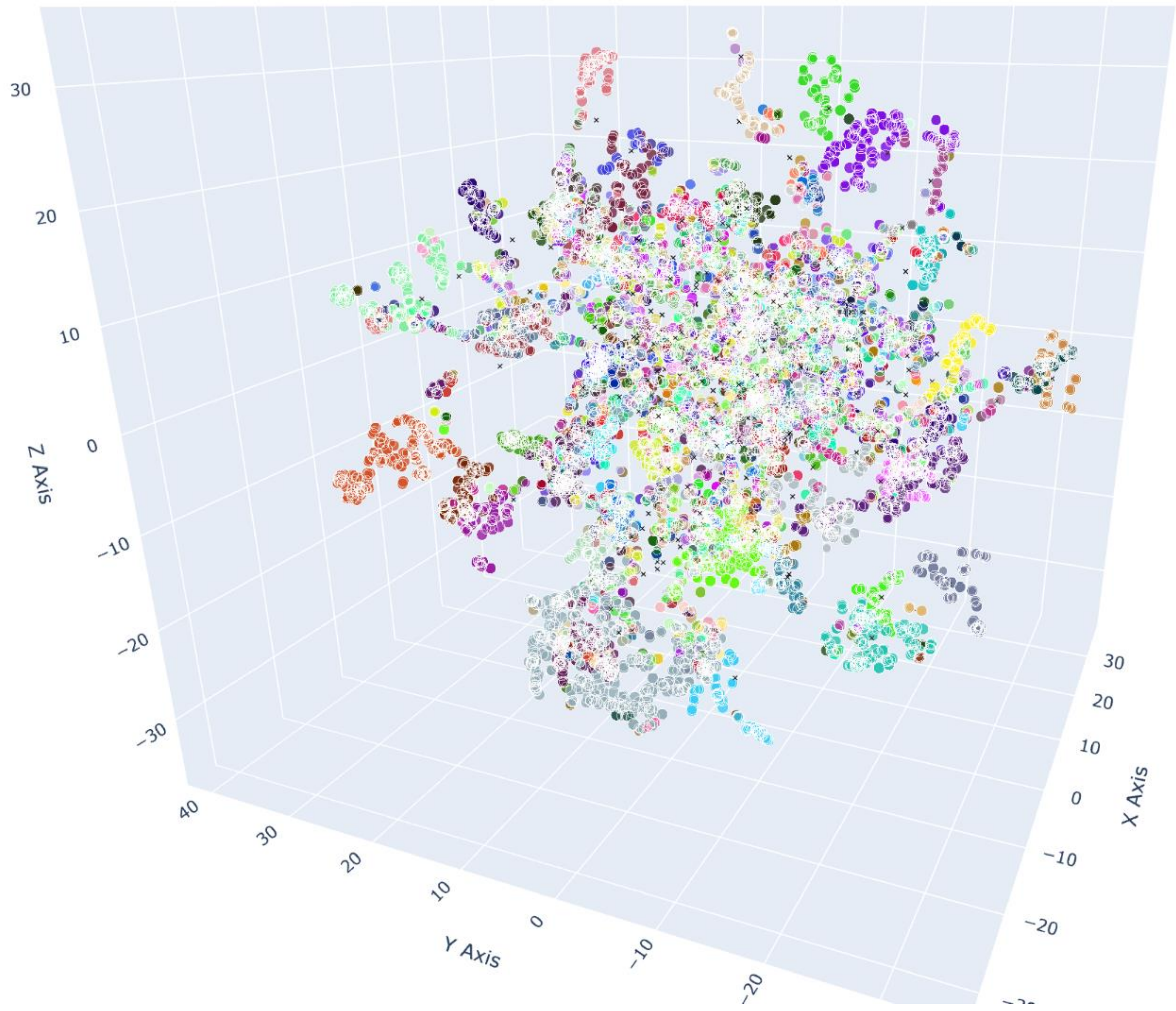
Explainability



Animation of 2D t-SNE over training iterations

```
[basic block] 0x402126:0x40216b -> 0x40268d
0x402126    mov eax, dword ptr [0x42e313]
0x40212b    mov eax, dword ptr [eax+0x194]
0x402131    mov ecx, dword ptr [eax]
0x402133    mov eax, dword ptr [ecx+0x3c]
0x402136    push dword ptr [0x42e3eb]
0x40213c    mov eax, dword ptr [eax+ecx*1+0x28]
0x402140    push dword ptr [0x42e197]
0x402146    add eax, ecx
0x402148    mov ecx, dword ptr [0x42e313]
0x40214e    mov ecx, dword ptr [ecx+0x194]
0x402154    push dword ptr [ecx]
0x402156    call eax
0x402158    mov ecx, dword ptr [0x42e313]
0x40215e    mov ecx, dword ptr [ecx+0x1cc]
0x402164    mov dword ptr [ecx], eax
0x402166    jmp 0x40268d
```

Feature attribution example on a sample of malware



3D t-SNE plot of embeddings learned by the model

References

R. Joyce, D. Amlani, Charles Nicholas and Edward Raff, "MOTIF: A Large Malware Reference Dataset with Ground Truth Family Labels," (2021). https://doi.org/10.48550/arXiv.2111.15031.

Malhotra, V., Potika, K. & Stamp, M. A comparison of graph neural networks for malware classification. *Journal of Computer Virology and Hacking Techniques*, vol 20, pgs 53–69 (2024). https://doi.org/10.1007/s11416-023-00493-y.

G. Montavon, A. Binder, S. Lapuschkin, W. Samek, KR. Müller. "Layer-Wise Relevance Propagation: An Overview," In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700, (2019). https://doi.org/10.1007/978-3-030-28954-6_10.

S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 539-546 (2005). doi: 10.1109/CVPR.2005.202.

Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, (2014). doi: 10.1109/CVPR.2014.220.

X. Xu, S. Feng, Y. Ye, G. Shen, Z. Su, S. Cheng, G. Tao, Q. Shi, Z. Zhang and X. Zhang, "Improving Binary Code Similarity Transformer Models by Semantics-Driven Instruction Deemphasis," *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis* (2023). https://doi.org/10.1145/3597926.3598121.

References

K. Lucas, M. Sharif, L. Bauer, M. Reiter and S. Shintre, “Malware Makeover: Breaking ML-based Static Analysis by Modifying Executable Bytes,” (2021). https://doi.org/10.1145/3433210.3453086.
X. Li, Y. Qu, and H. Yin. “PalmTree: Learning an Assembly Language Model for Instruction Embedding,” *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (2021). https://doi.org/10.1145/3460120.3484587.
D. Kingma, J. Ba. “Adam: A Method for Stochastic Optimization,” *3rd International Conference for Learning Representations* (2014). https://doi.org/10.48550/arXiv.1412.6980. D. Chicco. “Siamese Neural Networks: An Overview.” *Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology,* vol 2190 (2020). https://doi.org/10.1007/978-1-0716-0826-5_3. S. Lee, D. Cook, N. Silva, U. Laa, E. Wang, N. Spyrison and H. Sherry Zhang, “A Review of the State-of-the-Art on Tours for Dynamic Visualization of High-dimensional Data,” (2021). https://doi.org/10.48550/arXiv.2104.08016. “Cybersecurity Alerts & Advisories,” (2024). https://www.cisa.gov/news-events/cybersecurity-advisories. SC. Hsiao, DY. Kao, ZY. Liu and R. Tso, “Malware Image Classification Using One-Shot Learning with Siamese Networks,” *Procedia Computer Science*, vol 159, (2019). https://doi.org/10.1016/j.procs.2019.09.358.
A. Choudhary, “Understanding Siamese Networks: A Comprehensive Introduction.” *Analytics Vidhya*, 5 Nov. 2023, www.analyticsvidhya.com/blog/2023/08/introduction-and-implementation-of-siamese-networks/.
K. Reese, “Source Code.” *Securing Our Underlying Resources in Cyber Environments*, 26 Oct. 2023, www.iarpa.gov/images/OA-Slicksheets/SOURCE_CODE_SlickSheet_10262023.pdf.