Exploring the Platonic Representation Hypothesis Beyond In-Distribution Data

Aryasomayajula Ram Bharadwaj
Independent Researcher
ram.bharadwaj.arya@gmail.com

20th Oct 2024

Abstract

The Platonic Representation Hypothesis (PRH) [1] suggests that models trained with different objectives and on various modalities can converge to a shared statistical understanding of reality. This paper explores whether PRH holds when models are exposed to out-of-distribution (OOD) data. We analyze the alignment of model representations using ImageNet-O, a dataset designed with OOD images, and compare these results with those from the original experiment where the evaluation data is within the pretraining data distribution and also with purely random data. Our findings indicate that PRH holds in OOD settings but not for random data, highlighting the importance of underlying structure in data for shared model representations.

1 Introduction

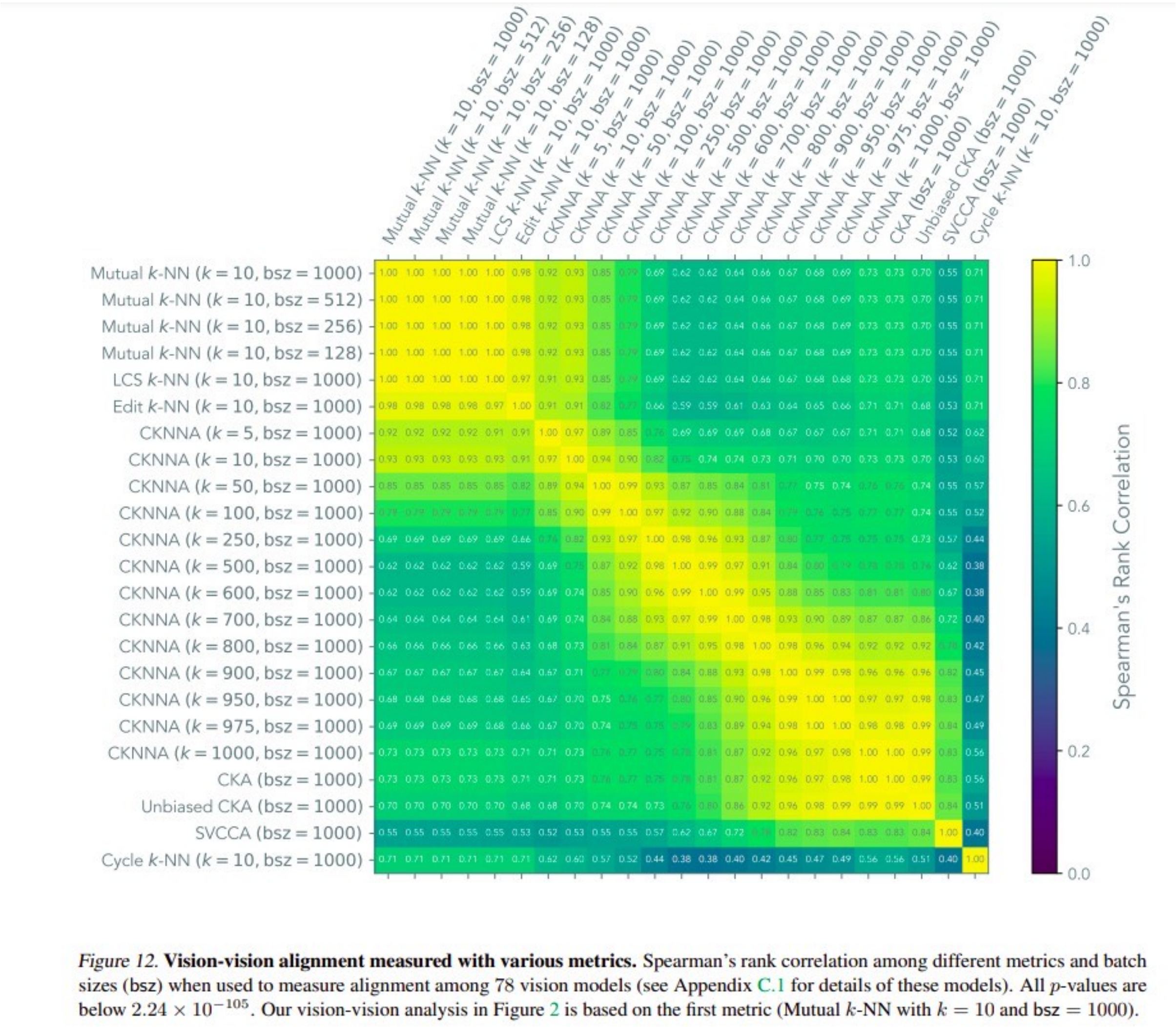
The PRH proposes that models converge toward a shared statistical representation of reality, regardless of training specifics such as dataset or training objectives. Previous evidence largely focused on available datasets which are in-distribution with respect to the training data of the models (e.g., Places365's validation dataset). This hypothesis is extended here to include OOD scenarios using ImageNet-O and random noise data. The ImageNet-O dataset, as introduced in [2], contains images that lie outside the distribution of ImageNet, making it a valuable resource for testing generalization.

2 Methodology

To test PRH in OOD settings, we used various metrics for measuring model alignment, such as mutual k-NN and CKNN-A. The data was divided into three categories: (1) in-distribution data (ImageNet), (2) OOD data (ImageNet-O), and (3) random noise.

3 Results

3.1 Vision-Vision Alignment on Places365's Validation Dataset



	Mutual k-NN ($k = 10$, bsz = 1000)	Mutual k-NN ($k = 10$, bsz = 512)	Mutual k-NN ($k = 10$, bsz = 256)	Mutual k-NN ($k = 10$, bsz = 128)	LCS k-NN ($k = 10$, bsz = 1000)	Edit k-NN ($k = 10$, bsz = 1000)	CKNNA ($k = 5$, bsz = 1000)	CKNNA ($k = 10$, bsz = 1000)	CKNNA ($k = 50$, bsz = 1000)	CKNNA ($k = 100$, bsz = 1000)	CKNNA ($k = 250$, bsz = 1000)	CKNNA ($k = 500$, bsz = 1000)	CKNNA ($k = 600$, bsz = 1000)	CKNNA ($k = 700$, bsz = 1000)	CKNNA ($k = 800$, bsz = 1000)	CKNNA ($k = 900$, bsz = 1000)	CKNNA ($k = 950$, bsz = 1000)	CKNNA ($k = 975$, bsz = 1000)	CKNNA ($k = 1000$, bsz = 1000)	CKA (bsz = 1000)	Unbiased CKA (bsz = 1000)	SVCCA (bsz = 1000)	Cycle k-NN ($k = 10$, bsz = 1000)
Mutual k-NN ($k = 10$, bsz = 1000)	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.93	0.85	0.79	0.69	0.62	0.62	0.64	0.66	0.67	0.68	0.69	0.73	0.73	0.70	0.55	0.71
Mutual k-NN ($k = 10$, bsz = 512)	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.93	0.85	0.79	0.69	0.62	0.62	0.64	0.66	0.67	0.68	0.69	0.73	0.73	0.70	0.55	0.71
Mutual k-NN ($k = 10$, bsz = 256)	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.93	0.85	0.79	0.69	0.62	0.62	0.64	0.66	0.67	0.68	0.69	0.73	0.73	0.70	0.55	0.71
Mutual k-NN ($k = 10$, bsz = 128)	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.93	0.85	0.79	0.69	0.62	0.62	0.64	0.66	0.67	0.68	0.69	0.73	0.73	0.70	0.55	0.71
LCS k-NN ($k = 10$, bsz = 1000)	1.00	1.00	1.00	1.00	1.00	0.97	0.91	0.93	0.85	0.79	0.69	0.62	0.62	0.64	0.66	0.67	0.68	0.68	0.73	0.73	0.70	0.55	0.71
Edit k-NN ($k = 10$, bsz = 1000)	0.98	0.98	0.98	0.98	0.97	1.00	0.91	0.91	0.82	0.77	0.66	0.59	0.59	0.61	0.63	0.64	0.65	0.66	0.71	0.71	0.68	0.53	0.71
CKNNA ($k = 5$, bsz = 1000)	0.92	0.92	0.92	0.92	0.91	0.91	1.00	0.97	0.89	0.85	0.76	0.69	0.69	0.69	0.68	0.67	0.67	0.67	0.71	0.71	0.68	0.52	0.62
CKNNA ($k = 10$, bsz = 1000)	0.93	0.93	0.93	0.93	0.93	0.91	0.97	1.00	0.94	0.90	0.82	0.75	0.74	0.74	0.73	0.71	0.70	0.70	0.73	0.73	0.70	0.53	0.60
CKNNA ($k = 50$, bsz = 1000)	0.85	0.85	0.85	0.85	0.85	0.82	0.89	0.94	1.00	0.99	0.93	0.87	0.85	0.84	0.81	0.77	0.75	0.74	0.76	0.76	0.74	0.55	0.57
CKNNA ($k = 100$, bsz = 1000)	0.79	0.79	0.79	0.79	0.79	0.77	0.85	0.90	0.99	1.00	0.97	0.92	0.90	0.88	0.84	0.79	0.76	0.75	0.77	0.77	0.74	0.55	0.52
CKNNA ($k = 250$, bsz = 1000)	0.69	0.69	0.69	0.69	0.69	0.66	0.76	0.82	0.93	0.97	1.00	0.98	0.96	0.93	0.87	0.80	0.77	0.75	0.75	0.75	0.73	0.57	0.44
CKNNA ($k = 500$, bsz = 1000)	0.62	0.62	0.62	0.62	0.62	0.59	0.69	0.75	0.87	0.92	0.98	1.00	0.99	0.97	0.91	0.84	0.80	0.79	0.78	0.78	0.76	0.62	0.38
CKNNA ($k = 600$, bsz = 1000)	0.62	0.62	0.62	0.62	0.62	0.59	0.69	0.74	0.85	0.90	0.96	0.99	1.00	0.99	0.95	0.88	0.85	0.83	0.81	0.81	0.80	0.67	0.38
CKNNA ($k = 700$, bsz = 1000)	0.64	0.64	0.64	0.64	0.64	0.61	0.69	0.74	0.84	0.88	0.93	0.97	0.99	1.00	0.98	0.93	0.90	0.89	0.87	0.87	0.86	0.72	0.40
CKNNA ($k = 800$, bsz = 1000)	0.66	0.66	0.66	0.66	0.66	0.63	0.68	0.73	0.81	0.84	0.87	0.91	0.95	0.98	1.00	0.98	0.96	0.94	0.92	0.92	0.92	0.78	0.42
CKNNA ($k = 900$, bsz = 1000)	0.67	0.67	0.67	0.67	0.67	0.64	0.67	0.71	0.77	0.79	0.80	0.84	0.88	0.93	0.98	1.00	0.99	0.98	0.96	0.96	0.96	0.82	0.45
CKNNA ($k = 950$, bsz = 1000)	0.68	0.68	0.68	0.68	0.68	0.65	0.67	0.70	0.75	0.76	0.77	0.80	0.85	0.90	0.96	0.99	1.00	1.00	0.97	0.97	0.98	0.83	0.47
CKNNA ($k = 975$, bsz = 1000)	0.69	0.69	0.69	0.69	0.68	0.66	0.67	0.70	0.74	0.75	0.75	0.79	0.83	0.89	0.94	0.98	1.00	1.00	0.98	0.98	0.99	0.84	0.49
CKNNA ($k = 1000$, bsz = 1000)	0.73	0.73	0.73	0.73	0.73	0.71	0.71	0.73	0.76	0.77	0.75	0.78	0.81	0.87	0.92	0.96	0.97	0.98	1.00	1.00	0.99	0.83	0.56
CKA (bsz = 1000)	0.73	0.73	0.73	0.73	0.73	0.71	0.71	0.73	0.76	0.77	0.75	0.78	0.81	0.87	0.92	0.96	0.97	0.98	1.00	1.00	0.99	0.83	0.56
Unbiased CKA (bsz = 1000)	0.70	0.70	0.70	0.70	0.70	0.68	0.68	0.70	0.74	0.74	0.73	0.76	0.80	0.86	0.92	0.96	0.98	0.99	0.99	0.99	1.00	0.84	0.51
SVCCA (bsz = 1000)	0.55	0.55	0.55	0.55	0.55	0.53	0.52	0.53	0.55	0.55	0.57	0.62	0.67	0.72	0.78	0.82	0.83	0.84	0.83	0.83	0.84	1.00	0.40
Cycle k-NN ($k = 10$, bsz = 1000)	0.71	0.71	0.71	0.71	0.71	0.71	0.62	0.60	0.57	0.52	0.44	0.38	0.38	0.40	0.42	0.45	0.47	0.49	0.56	0.56	0.51	0.40	1.00

Figure 12. **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All p-values are below 2.24×10^{-105}. Our vision-vision analysis in Figure 2 is based on the first metric (Mutual k-NN with $k = 10$ and bsz $= 1000$).

Figure 1: Vision-vision alignment measured with various metrics. Spearman's rank correlation among different metrics and batch sizes when used to measure alignment among 78 vision models.

3.2 Vision-Vision Alignment on ImageNet-O

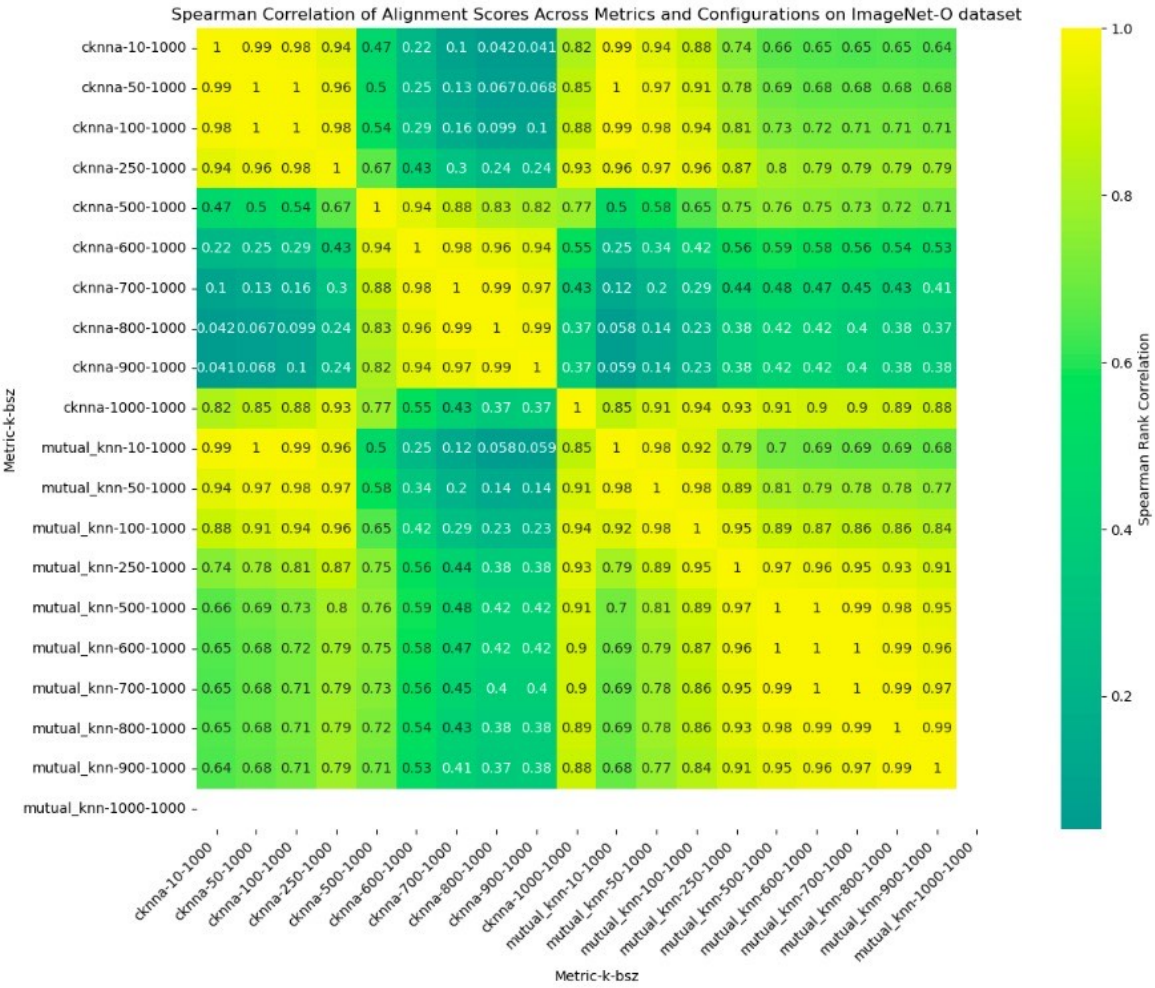
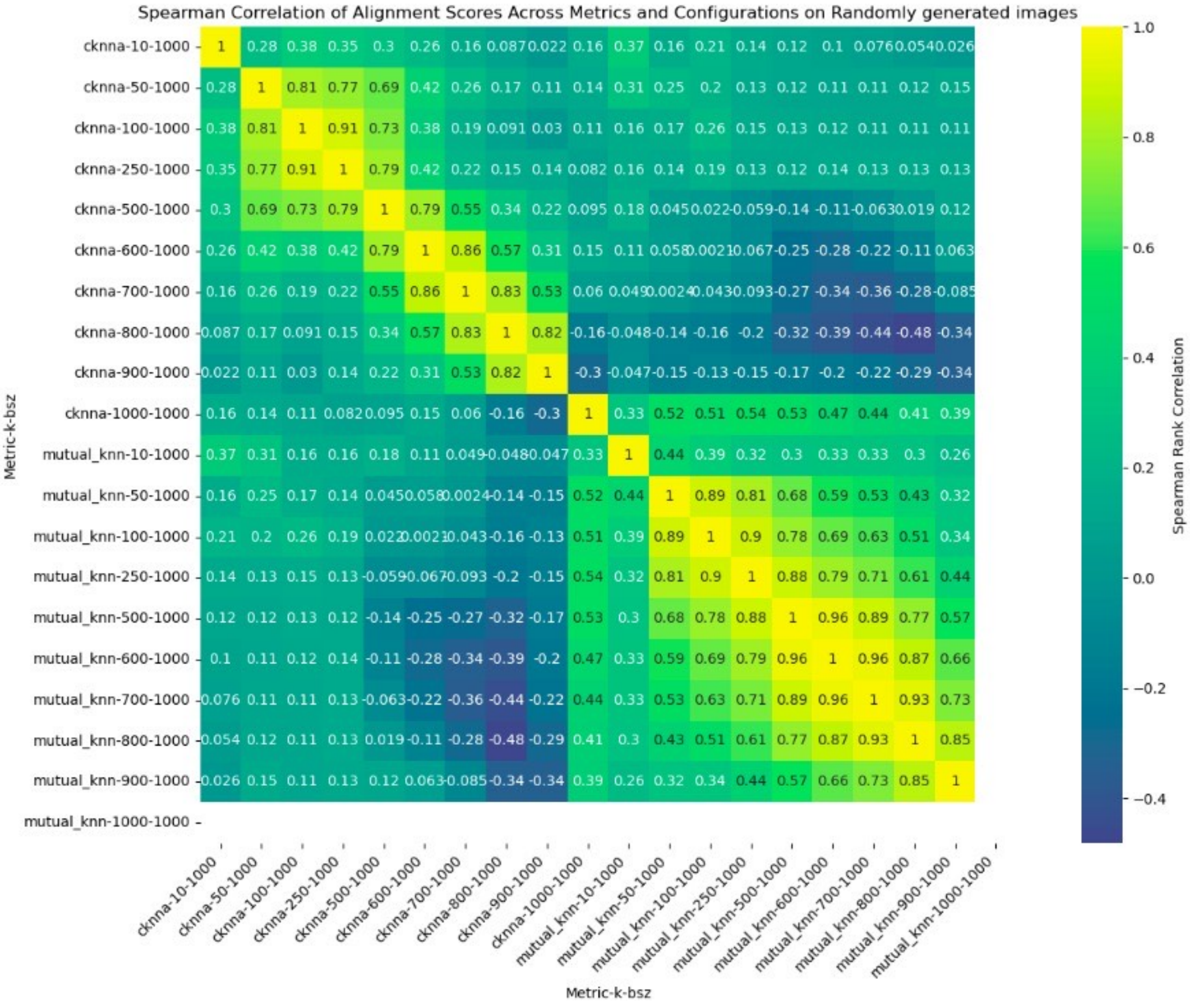


Figure 2: Spearman correlation of alignment scores across metrics and configurations on the ImageNet-O dataset.

3.3 Vision-Vision Alignment on Random Data



Metric-k-bsz	cknna-10-1000	cknna-50-1000	cknna-100-1000	cknna-250-1000	cknna-500-1000	cknna-600-1000	cknna-700-1000	cknna-800-1000	cknna-900-1000	cknna-1000-1000	mutual_knn-10-1000	mutual_knn-50-1000	mutual_knn-100-1000	mutual_knn-250-1000	mutual_knn-500-1000	mutual_knn-600-1000	mutual_knn-700-1000	mutual_knn-800-1000	mutual_knn-900-1000	mutual_knn-1000-1000
cknna-10-1000	1	0.28	0.38	0.35	0.3	0.26	0.16	0.087	0.022	0.16	0.37	0.16	0.21	0.14	0.12	0.1	0.076	0.054	0.026	
cknna-50-1000	0.28	1	0.81	0.77	0.69	0.42	0.26	0.17	0.11	0.14	0.31	0.25	0.2	0.13	0.12	0.11	0.11	0.12	0.15	
cknna-100-1000	0.38	0.81	1	0.91	0.73	0.38	0.19	0.091	0.03	0.11	0.16	0.17	0.26	0.15	0.13	0.12	0.11	0.11	0.11	
cknna-250-1000	0.35	0.77	0.91	1	0.79	0.42	0.22	0.15	0.14	0.082	0.16	0.14	0.19	0.13	0.12	0.14	0.13	0.13	0.13	
cknna-500-1000	0.3	0.69	0.73	0.79	1	0.79	0.55	0.34	0.22	0.095	0.18	0.045	0.022	-0.059	-0.14	-0.11	-0.063	0.019	0.12	
cknna-600-1000	0.26	0.42	0.38	0.42	0.79	1	0.86	0.57	0.31	0.15	0.11	0.058	0.0021	0.067	-0.25	-0.28	-0.22	-0.11	0.063	
cknna-700-1000	0.16	0.26	0.19	0.22	0.55	0.86	1	0.83	0.53	0.06	0.049	0.0024	0.043	0.093	-0.27	-0.34	-0.36	-0.28	-0.085	
cknna-800-1000	0.087	0.17	0.091	0.15	0.34	0.57	0.83	1	0.82	-0.16	-0.048	-0.14	-0.16	-0.2	-0.32	-0.39	-0.44	-0.48	-0.34	
cknna-900-1000	0.022	0.11	0.03	0.14	0.22	0.31	0.53	0.82	1	-0.3	-0.047	-0.15	-0.13	-0.15	-0.17	-0.2	-0.22	-0.29	-0.34	
cknna-1000-1000	0.16	0.14	0.11	0.082	0.095	0.15	0.06	-0.16	-0.3	1	0.33	0.52	0.51	0.54	0.53	0.47	0.44	0.41	0.39	
mutual_knn-10-1000	0.37	0.31	0.16	0.16	0.18	0.11	0.049	-0.048	0.047	0.33	1	0.44	0.39	0.32	0.3	0.33	0.33	0.3	0.26	
mutual_knn-50-1000	0.16	0.25	0.17	0.14	0.045	0.058	0.0024	-0.14	-0.15	0.52	0.44	1	0.89	0.81	0.68	0.59	0.53	0.43	0.32	
mutual_knn-100-1000	0.21	0.2	0.26	0.19	0.022	0.0021	0.043	-0.16	-0.13	0.51	0.39	0.89	1	0.9	0.78	0.69	0.63	0.51	0.34	
mutual_knn-250-1000	0.14	0.13	0.15	0.13	-0.059	0.067	0.093	-0.2	-0.15	0.54	0.32	0.81	0.9	1	0.88	0.79	0.71	0.61	0.44	
mutual_knn-500-1000	0.12	0.12	0.13	0.12	-0.14	-0.25	-0.27	-0.32	-0.17	0.53	0.3	0.68	0.78	0.88	1	0.96	0.89	0.77	0.57	
mutual_knn-600-1000	0.1	0.11	0.12	0.14	-0.11	-0.28	-0.34	-0.39	-0.2	0.47	0.33	0.59	0.69	0.79	0.96	1	0.96	0.87	0.66	
mutual_knn-700-1000	0.076	0.11	0.11	0.13	-0.063	-0.22	-0.36	-0.44	-0.22	0.44	0.33	0.53	0.63	0.71	0.89	0.96	1	0.93	0.73	
mutual_knn-800-1000	0.054	0.12	0.11	0.13	0.019	-0.11	-0.28	-0.48	-0.29	0.41	0.3	0.43	0.51	0.61	0.77	0.87	0.93	1	0.85	
mutual_knn-900-1000	0.026	0.15	0.11	0.13	0.12	0.063	-0.085	-0.34	-0.34	0.39	0.26	0.32	0.34	0.44	0.57	0.66	0.73	0.85	1	
mutual_knn-1000-1000																				

Figure 3: Spearman correlation of alignment scores across metrics and configurations on randomly generated images.

4 Discussion

The results suggest that the Platonic Representation Hypothesis (PRH) holds true even in out-of-distribution (OOD) settings, as demonstrated with the ImageNet-O dataset. Despite being exposed to outlier data, the models maintain high alignment in their representations, indicating that they might develop a shared underlying structure in their perception of reality, even when the training data distribution shifts. However, this alignment breaks down when models are exposed to purely random noise, revealing the limits of the PRH. This distinction is intriguing: on randomly generated data, models do not converge on a shared statistical understanding of reality. Yet, on OOD data like ImageNet-O, the models exhibit strong correlation in their representations, even though their predictions are consistently incorrect but made with high confidence. This suggests that, while the models fail on outlier data, their failure is systematic and predictable.

5 Conclusion

We have shown that PRH extends beyond in-distribution data, with models maintaining alignment even in OOD settings. However, the presence of underlying structure in data is crucial for such alignment. Future work should focus on developing better methods to test the generalizability of PRH, finding scenarios where the hypothesis might be falsified, and exploring cross-modality applications of PRH.

The code used for the experiments and analysis in this paper is available on GitHub at https://github.com/rokosbasilisk experiments.

References

[1] Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis. *International Conference on Machine Learning.*

[2] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural Adversarial Examples. *CVPR.* arXiv:2405.07987.