An Exploration of High-Dimensional Data: Geometry, Spectra, and Structure

Itay Meiri

August 17, 2025

1 Introduction

This report summarizes my work as part of the High-Dimensional Probability (HDP) graduate course at Ariel University. My overarching goal was to select one or more high-dimensional datasets and conduct a thorough analysis to uncover their geometric, spectral, and statistical properties. The emphasis was not on predictive modeling for its own sake, but on understanding the structure of the data - for example, estimating intrinsic dimensionality, examining concentration-of-measure effects, analyzing spectral behavior via Random Matrix Theory (RMT), and exploring how these characteristics differ across datasets [2].

A unique aspect of this project is my deliberate and documented use of large language models (LLMs) as active tools in the research process. Rather than treat AI assistance as incidental, I integrated LLM queries into my workflow at multiple stages - from initial dataset selection, through the design of analysis pipelines, code refactoring, and improving the tone and delivery of this final report.

While this report focuses on the technical results, I am maintaining a separate log that documents the usage of LLM tools. To access, please visit the projects GitHub page [8].

2 The Datasets Under Investigation

I have chosen two distinct datasets for the analyses performed in this report, allowing for a direct comparison of geometric and spectral phenomena across different modalities and dimensional scales.

2.1 Covertype: A High-Dimensional Cartographic Challenge

The Covertype dataset is a large-scale benchmark widely used in machine learning [3, 4]. It contains 581,012 samples with 54 features and 7 classes corresponding to forest cover types (e.g., Spruce-Fir, Ponderosa Pine). The feature space is mixed: 10 continuous variables (elevation, aspect, slope, hydrological distances, and fire-point proximities), 4 binary features encoding wilderness areas (one-hot), and 40 binary features encoding soil types (one-hot). This structure necessitates preprocessing: the continuous variables are standardized, while the binary features are retained as-is. The dataset provides a testbed for studying geometry in mixed discrete–continuous high-dimensional data.

2.2 Tox21: A Non-Euclidean Chemical Space

The Tox21 dataset, part of the MoleculeNet collection [12], is a molecular property prediction benchmark. It comprises approximately 8,000 compounds with binary toxicity labels across 12 biological targets. Molecules are inherently graph-structured, and for this analysis, they are embedded into a 2048-dimensional binary vector space using Morgan fingerprints [10]. The Tox21 dataset thus provides a non-Euclidean, chemically meaningful case for geometric and spectral analysis, complementary to the Covertype dataset.

3 Analysis of the Covertype Dataset: A Story in Two Parts

I began my investigation with the Covertype dataset. This analysis is presented in two parts: the first part uncovers the challenges posed by high-dimensional geometry, and the second part presents a solution based on transitioning from the raw feature space to a learned embedding space.

3.1 Part 1: The Challenge of High-Dimensional Geometry

My first step was to analyze the distributions of pairwise distances to understand the inherent structure and separability of the classes. The central idea is to compare *intra-class* distances (between points in the same class) with *inter-class* distances (between points in different classes). In a

well-structured dataset, we would expect the mean intra-class distance to be smaller than the mean inter-class distance.

I estimated the distributions by sampling 300,000 random pairs from a 10,000-point subset to manage the computational load. The results are summarized in Table 1.

Table 1: Intra- vs. Inter-Class Distance Statistics for the Covertype Dataset.

Statistic	**Euclidean**	**Cosine**
Mean Intra-class Dist.	4.37	0.91
Mean Inter-class Dist.	4.73	0.96
Gap (Δ = Inter - Intra)	**0.36**	**0.054**
Overlap Coefficient	0.89	0.93
Silhouette Score (on subset)	-0.02	-0.11

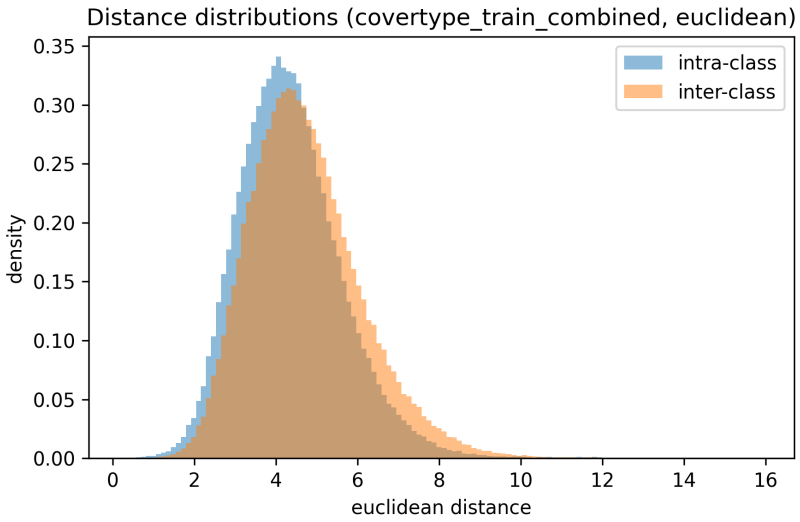

Figure 1: Density histograms of intra-class and inter-class Euclidean distances. The extensive overlap visually confirms the poor class separation.

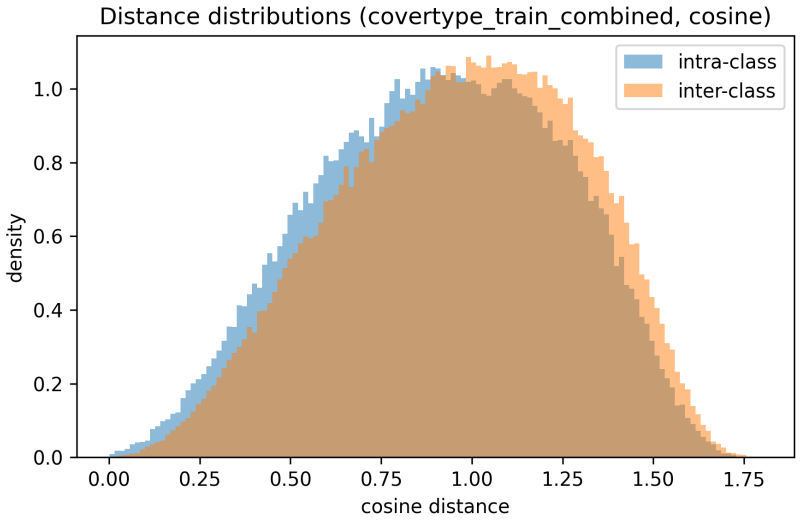

Figure 2: Density histograms of intra-class and inter-class Cosine distances. The overlap is even more pronounced than with the Euclidean metric.

The results clearly depicted a dataset with very poorly separated classes. The small separation gap, high overlap coefficient, and negative silhouette score [11] are classic symptoms of the **concentration of measure** phenomenon in high-dimensional spaces [2]. As dimensionality increases, the distances between any two points tend to become very similar, "washing out" the relative distance differences and making it challenging to distinguish clusters.

Given the poor separation observed, I next explore if a more sophisticated view of the data can reveal a more transparent structure. I investigate two extensions: using the **Mahalanobis Distance**[7] to account for feature correlations and applying Principal Component Analysis (PCA)[1] to the continuous features to see if the class structure resides in a lower-dimensional subspace.

3.1.1 Mahalanobis Distance

The Mahalanobis distance is a scale-invariant metric that accounts for the data's covariance matrix [7]. Transforming the space so that feature correlations are removed and variances are equalized (a process known as "whitening") can often reveal separations that are obscured in the standard Euclidean space.

As shown in Table 2, the results are apparent. The separation gap Δ under the Mahalanobis distance is **1.083**, nearly three times larger than the

Table 2: Mahalanobis Distance Analysis

Statistic	**Mahalanobis**
Mean Intra-class Dist.	8.26
Mean Inter-class Dist.	9.35
Gap (Δ = Inter - Intra)	**1.08**

gap observed with Euclidean distance (0.368). This is a significant finding: it suggests that the global covariance structure of the data holds crucial information for class separation. Once we account for how the features covary, the classes become considerably more distinct.

3.1.2 Dimensionality Reduction via PCA

Perhaps the separating signal lies in a lower-dimensional manifold. I test this by applying PCA only to the 10 continuous features, retaining enough principal components to explain 95% of their variance [1]. The 44 binary one-hot features are left unchanged.

Table 3: PCA on Continuous Features and Subsequent Analysis

Analysis Step	**Result**
Original Continuous Dimensions	10
PCA Components for 95% Variance	8
New Total Dimensionality	52 (from 8 PCs + 44 binary features)
Euclidean Gap (Δ) in Reduced Space	0.363824

The analysis shows that the 10 continuous features can be compressed to 8 dimensions while losing only 5% of their variance, as seen in Table 3. However, when we recalculate the Euclidean distances in this new 52-dimensional space, the separation gap Δ is 0.3638, virtually identical to the original gap of 0.3676. This indicates that the (unsupervised) reduction via PCA did not improve class separability. The directions of highest variance, which PCA finds, are not necessarily the directions of highest class discrimination. The information that separates the classes appears to be distributed across the original feature space rather than being concentrated in a few principal components.

3.2 Part 2: Finding Structure in Learned Representations

The initial analysis established that Covertype classes are poorly separated in the original feature space. This makes tasks like Out-of-Distribution (OOD) detection challenging. I hypothesize that a neural network, trained on a classification task, could project the data into a lower-dimensional *embedding space* where the classes form more compact and separable manifolds.

3.2.1 Experimental Setup for OOD Detection

Let us reframe the problem as an OOD detection task. The first five cover types (classes 1-5) were designated as in-distribution (ID) data. I defined two types of OOD data:

- **Hard OOD:** Samples from the held-out classes 6 and 7.
- **Easy OOD:** Synthetic Gaussian noise, as a sanity check.

I trained a simple Multi-Layer Perceptron (MLP) on the ID data, achieving a validation accuracy of 86.69%. I then extracted the 32-dimensional embeddings from its penultimate layer and fit a Kernel Density Estimator (KDE) to model the density of the ID data [9].

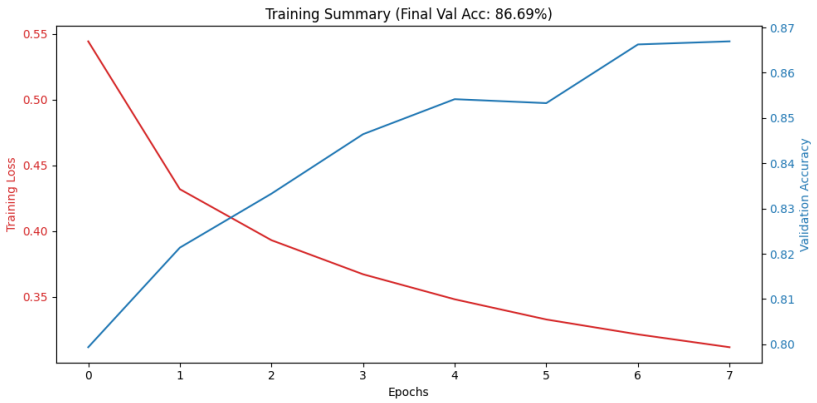

Figure 3: Training history of the MLP classifier on in-distribution data. The model reaches a final validation accuracy of 86.69%, indicating it has learned a useful representation.

3.2.2 Results: From Raw Data to Learned Embeddings

First, I established a baseline by applying the KDE directly to the original 54-dimensional feature space. As predicted, this approach struggled, achieving

an AUROC of only 0.7310 for the hard OOD case.

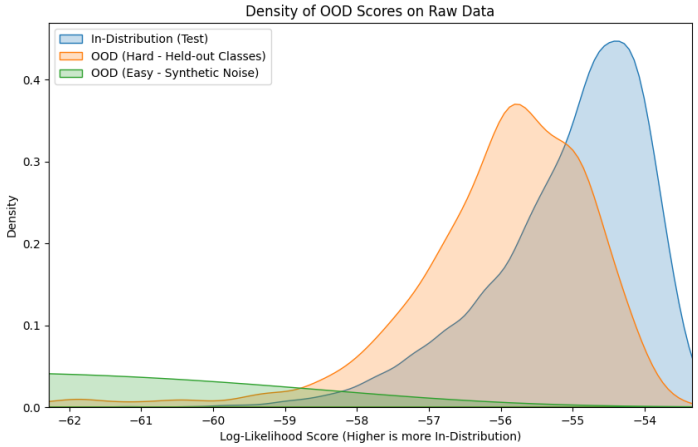

Figure 4: Density of OOD scores on raw data. The significant overlap between the ID (blue) and hard OOD (orange) distributions confirms poor separability.

Next, I repeated the analysis using the 32-dimensional embeddings from the trained MLP. The performance improved markedly. In the full 32-D embedding space, the AUROC for hard OOD detection jumped to **0.8393**.

An even more striking result emerged when I applied PCA to the embeddings, reducing them to just six dimensions. OOD detection performance in this PCA-reduced space improved dramatically, with the hard OOD AUROC reaching an impressive **0.9427** and the False Positive Rate (at 95% TPR) falling to just 16.79%.

Table 4: OOD Detection Performance Comparison.

Space	**OOD Type**	**AUROC**	**FPR at 95% TPR**
Raw Data (54-D)	Hard (Classes 6,7)	0.7310	85.76%
Full Embedding (32-D)	Hard (Classes 6,7)	**0.8393**	**60.86%**
PCA-Reduced (6-D)	Hard (Classes 6,7)	**0.9427**	**16.79%**

In the full 32-D embedding space and especially after PCA reduction, the separation improves substantially (see Table 4).

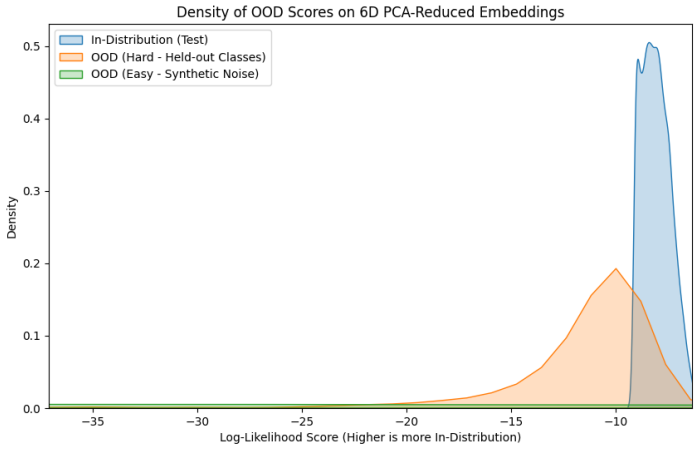

Figure 5: Density of OOD scores on the 6-D PCA-reduced embeddings. The separation is now much more distinct, leading to the best OOD detection performance.

This experiment confirmed my hypothesis: a discriminatively trained network organizes data into a more structured embedding space. Furthermore, the success of PCA suggests that projecting onto the principal components of this space acts as a powerful denoising step, creating a more compact representation of the in-distribution data and enabling superior OOD detection.

4 Analysis of the Tox21 Dataset: Uncovering Latent Structure

I now focus on the Tox21 dataset, where molecules are represented as 2048-dimensional binary fingerprints. My core hypothesis was that despite the vast ambient space, the chemical rules governing molecular structure would constrain the data to a much lower-dimensional manifold.

4.1 Probing the Manifold: Intrinsic Dimension Estimation

My first question was: What is the "true" dimensionality of the chemical space spanned by the Tox21 molecules? I employed a suite of modern estimators to find out.

Table 5: Intrinsic Dimension Estimation Results for Tox21. All estimators were applied using the Jaccard distance on binary molecular fingerprints; Euclidean distance is not meaningful for this representation.

Estimator	**Estimated Dimension**
k-NN	5.00
TwoNN	9.52
MOM (Method of Moments)	26.91

The results confirmed my hypothesis. All three estimators pointed to an intrinsic dimension (ranging from 5 to 27) that is orders of magnitude smaller than the ambient dimension of 2048 [5]. This suggests the entire dataset lies on a thin, highly constrained manifold.

4.2 Isolating Signal from Noise with Random Matrix Theory

Next, I investigated the features' structure using RMT. By comparing the eigenvalue spectrum of the feature correlation matrix to the theoretical Marčenko-Pastur distribution [6], I could distinguish meaningful correlations (signal) from statistical noise.

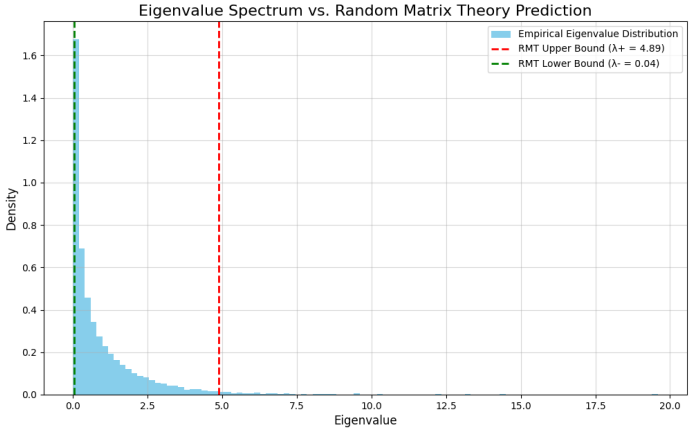

Figure 6: The empirical eigenvalue density of the feature correlation matrix. A distinct "tail" of 48 eigenvalues lies far outside the theoretical bound for noise (red line), representing significant, non-random correlational structures.

The analysis identified **48** eigenvalues lying far outside the noise bound.

These 48 factors represent the dominant, independent patterns of co-varying chemical features—the core "chemical concepts" in the dataset.

4.3 Practical Validation: Improved Predictive Modeling

The analyses converged on a powerful conclusion: Tox21 data has a hidden low-dimensional structure. However, does this insight have practical value? To answer this, I trained a neural network classifier on a binary prediction task (NR-AR activity) using three different feature sets.

Table 6: Model Performance Comparison on the Tox21 Task.

Model	**Feature Dimension**	**Test AUC**	**Test Accuracy**
Baseline (Full Data)	2048	0.6744	0.9642
RMT-Reduced	**48**	**0.7517**	**0.9704**
Autoencoder-Reduced	32	0.7292	0.9656

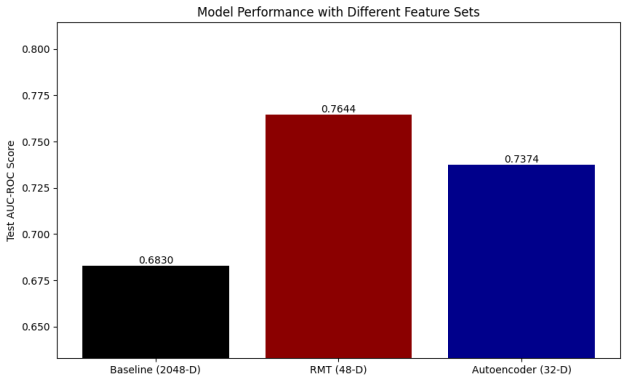

Figure 7: A comparison of model performance (AUC score) on the test set. Both dimensionality reduction techniques significantly outperform the baseline model. The RMT-based feature set yielded the best performance.

The outcome was clear: reducing the dimensionality based on my RMT analysis did not just maintain performance - it *significantly improved* it. By projecting the data onto the 48 dimensions RMT identified as signal-carrying, I created a feature set that was not only 40x smaller but also richer in relevant information. This is a robust, practical validation of the

RMT analysis; removing the noise allowed the model to learn the underlying patterns more effectively.

Decoding the Principal Axes: What Chemical Features Define the Latent Space?

The RMT analysis successfully identified 48 dimensions of non-random structure. While this confirms the existence of a low-dimensional latent space, it leaves a critical question unanswered: what do these dimensions *mean* in chemical terms? Each 48 "signal" eigenvector represents the dataset's principal axis of variation. We can translate abstract mathematical vectors into intuitive chemical concepts by inspecting the features and the specific fingerprint bits that define these eigenvectors.

To achieve this, I performed a chemical motif enrichment analysis on the top five signal eigenvectors. For each eigenvector, I identified the fingerprint bits with the largest positive and negative "loadings" (i.e., the bits that contribute most significantly to that dimension). I then statistically determined which chemical motifs (e.g., "aromatic ring," "nitro group") were over-represented in the molecules corresponding to those bits. The results, summarized below, reveal that these principal axes neatly partition the chemical space along clear structural and functional lines.

Table 7: Chemical Interpretation of the Dominant Eigenvectors. Each eigenvector acts as an axis separating molecules based on distinct chemical features.

Eigenvector	**Positive Loading Features Enriched For...**	**Negative Loading Features Enriched For...**
1	Nitro groups, Halogens, Aromatic Rings	Hydroxyl groups, Carbonyls, Amides
2	Heteroaromatic Rings, Amides	Nitro groups, Hydroxyl groups
3	Hydroxyl groups, Heteroaromatic Rings	Aromatic Rings, Halogens, Nitro groups
4	(General Carbonyl and Amide Structures)	Aromatic and Heteroaromatic Rings
5	(General Amide and Carbonyl Structures)	Aromatic Rings, Halogens, Heteroaromatic Rings

The analysis reveals a clear pattern. **Eigenvector 1**, the most dominant axis of variation, acts as a polarity and functional group separator. It effectively distinguishes between molecules containing electron-withdrawing

groups like nitro and halogens (positive loadings) and those with hydrogen-bond donating/accepting motifs like hydroxyls and carbonyls (negative loadings). This is a chemically fundamental distinction that governs properties like solubility and reactivity.

Subsequent eigenvectors capture more nuanced structural differences. **Eigenvector 2** appears to separate molecules based on the type of cyclic system they contain, contrasting specific heteroaromatic amide structures with molecules defined by nitro and hydroxyl groups. **Eigenvectors 4 and 5** both strongly separate molecules based on the presence of aromaticity, with negative loadings heavily associated with various aromatic and heteroaromatic systems. The entire result could be seen in my GitHub page. [8]

This analysis provides robust, tangible evidence for the underlying structure of this chemical dataset. The 48 dimensions discovered by RMT are not random, abstract constructs; they are chemically meaningful axes that correspond to the core rules of molecular design and functional group compatibility. By isolating these dimensions, we have created a feature space that is more compact and aligned with the fundamental principles of chemistry itself. This alignment is why the RMT-reduced model achieved superior predictive performance: it learned from features representing true chemical concepts, not high-dimensional noise.

5 Conclusion

This project told a tale of two datasets, each offering a distinct but complementary perspective on the challenges and opportunities within high-dimensional data. Our analysis of the Covertype dataset served as a classic case study in the "curse of dimensionality," where the geometry of the raw, 54-dimensional feature space rendered classes nearly indistinguishable. However, we demonstrated a powerful resolution: by training a neural network for a supervised task, we transformed the data into a 32-dimensional embedding space where the data's latent structure became evident. Projecting these embeddings onto just six principal components dramatically enhanced class separability, enabling highly effective OOD detection. This highlights a key principle: supervised learning is a potent tool for dimensionality reduction, capable of untangling complex manifolds into more structured and practical representations. Conversely, our investigation of the Tox21 dataset revealed a different narrative. The data, despite its vast 2048-dimensional binary representation, was found to lie on a remarkably low-dimensional manifold. Intrinsic dimension estimators suggested a true dimensionality between 5 and 27, and RMT identified only 48 dimensions of meaningful signal amidst

a sea of noise. This insight was not merely theoretical; it was profoundly practical. A predictive model trained on these 48 signal-carrying dimensions significantly outperformed a baseline model trained on the full dataset, confirming that principled dimensionality reduction can serve as a powerful form of noise removal, leading to more efficient and accurate models. Across both studies, a unified theme emerges: the raw ambient dimension of a dataset is often a misleading indicator of its true complexity. The most crucial insights lie in understanding the data's underlying geometric and spectral structure. Whether that structure is revealed by untangling a complex manifold with a learned embedding or by isolating a low-dimensional subspace of the signal, the goal is to move beyond the superficial high-dimension.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[3] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[4] Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

[5] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[6] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[7] Goeffrey J McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

[8] Itay Meiri. HDP-Final. GitHub Repository, 2024. URL: https://github.com/ItayMeiri/HDP-Final.

[9] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[10] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[11] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[12] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.