Continuous, Not Categorical: Deep Embedded Clustering Analysis of Depression's Proteomic Architecture

Jacob Benjamine Ioffe
Department of Computer Science
Cornell-Tech
New York, NY 10044 USA
ji97@cornell.edu

Abstract

Major Depressive Disorder (MDD) is clinically heterogeneous, prompting efforts to uncover biologically defined subtypes that might inform diagnosis and treatment. Recent large-scale proteomic analyses, including those on UK Biobank (UKB) data, suggest that complex traits may be better characterized as continuous gradients rather than discrete categories. Here, we investigate whether MDD-related proteomic variation forms distinct clusters or lies along a continuum. Using data from over 53,000 UKB participants, we apply Improved Deep Embedded Clustering (IDEC) implemented in PyTorch to identify latent structures in 2,900 protein expression profiles per individual. Compared against baseline methods (PCA+k-means and UMAP+HDBSCAN), IDEC robustly uncovers broad demographic signals, particularly those related to sex, yet reveals no stable proteomic subtypes of depression across the full dataset, sex-stratified subsets, or an MDD-enriched sample. Adjusting the number of clusters, tuning model complexity, and focusing on more homogeneous subgroups do not isolate distinct MDD-related clusters. Instead, MDD prevalence and associated patterns remain diffusely distributed, suggesting that depression's proteomic architecture does not neatly partition into categorical subtypes. These findings underscore the importance of embracing dimensional, gradient-based frameworks when probing the molecular underpinnings of psychiatric disorders, potentially guiding more nuanced approaches to biomarker discovery and personalized interventions.

1 Introduction

Major Depressive Disorder (MDD) is clinically heterogeneous, prompting extensive efforts to identify biologically defined subtypes. Yet, systematic reviews report limited success in finding replicable, discrete clusters, suggesting that depression-related biology may manifest along continuous dimensions. Such gradient-based perspectives are increasingly prominent in psychiatry, where dimensional models often capture subtle biomarker-symptom relationships more effectively than categorical frameworks.

The UK Biobank (UKB) Plasma Proteomics dataset—over 53,000 participants with nearly 3,000 quantified proteins—offers a powerful setting to examine depression's proteomic architecture. Recent research using this dataset demonstrated that biological aging follows continuous gradients Argentieri et al. (2024), raising the question of whether depression-related signals similarly disperse across smooth landscapes.

Here, we employ three complementary clustering approaches: (1) Improved Deep Embedded Clustering (IDEC), our primary method implemented in PyTorch Guo et al. (2017), (2) Principal Component Analysis with k-means (PKM), and (3) UMAP with Hierarchical Density-Based Clustering (UH). We examine the full cohort, a subset with in-patient clinical diagnosis of MDD, and sex-stratified groups. Our results consistently show that proteomic variation associated with depression does not yield stable, distinct clusters, instead revealing

a gradient-based architecture that may better inform biomarker discovery and personalized interventions.

2 Methods

2.1 Data and Preprocessing

The UK Biobank (UKB) Plasma Proteomics dataset comprises 53,018 participants with 2,917 quantified proteins per individual. Our analysis pipeline examined three cohorts: (1) the full dataset (discovery cohort n=22,601, held-out replication n=22,601), (2) an MDD-enriched subset of clinically diagnosed individuals (n=1,616, ICD-10 codes with PHQ-9 and GAD7 assessments), and (3) sex-stratified subsets (male: n=10,416; female: n=12,185). We used $\log_2$-transformed NPX values with mean imputation for missing values.

2.2 Clustering Methodology

We implemented three complementary clustering approaches to analyze proteomic patterns. Our primary method, IDEC, uses a deep learning framework to simultaneously learn data representations and cluster assignments. We compared this against two baseline approaches: PKM (Principal Component Analysis with k-means) and UH (UMAP with HDBSCAN). This multi-method strategy enables robust validation of identified patterns across different algorithmic paradigms.

2.3 Model Configuration

Our IDEC implementation used a symmetric autoencoder architecture (Figure 1) with encoder dimensions d-500-500-2000-z and ReLU activations. Through empirical evaluation, we identified optimal parameters including latent dimension z=16, layer scaling factor ls=1.0, batch size 256, learning rate 0.0001, and reconstruction loss weight γ=0.1, using 10,000 pretraining and 25,000 training epochs. For baseline comparisons, PKM used 16 components, while UH employed 100 neighbors.

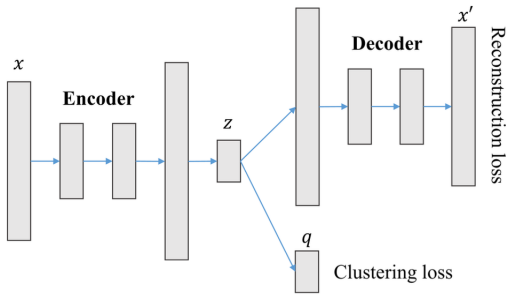


Figure 1: The network structure of IDEC. The encoder and decoder are composed of fully connected layers. Clustering loss scatters the embedded points z while reconstruction loss preserves local structure.

2.4 Evaluation Framework

We assessed clustering performance through technical metrics and biological validation. Technical evaluation used three complementary measures: Silhouette score (cluster cohesion), Davies-Bouldin index (cluster separation), and Calinski-Harabasz score (cluster density). For biological validation, we examined demographic distributions (age, sex, BMI) across clusters and analyzed MDD prevalence patterns using bootstrap confidence intervals and chi-square tests with Bonferroni correction.

Table 1: Comparison of Clustering Methods Across Different Cohorts

Cohort	Method	k	Silhouette↑	Davies-Bouldin↓	Calinski-Harabasz↑
Original (n=22,601)	PCA + k-means	2	0.225	1.562	19,195.0
	PCA + k-means	3	0.153	1.867	13,989.2
	PCA + k-means	4	0.142	1.815	11,743.4
	UMAP + HDBSCAN	–	0.196	0.633	858.3
	IDEC	2	**0.787**	**0.296**	**514,862.9**
	IDEC	3	0.637	0.456	248,971.3
	IDEC	4	0.542	0.595	79,353.2
Female-Only (n=12,185)	PCA + k-means	2	0.242	1.489	4,887.8
	PCA + k-means	3	0.156	1.841	3,474.4
	PCA + k-means	4	0.149	1.790	2,908.2
	UMAP + HDBSCAN	–	0.051	0.719	234.6
	IDEC	2	**0.633**	**0.496**	**40,269.1**
	IDEC	3	0.295	1.127	11,123.2
	IDEC	4	0.401	1.035	17,367.8
Male-Only (n=10,416)	PCA + k-means	2	0.242	1.495	4,074.8
	PCA + k-means	3	0.152	1.893	2,896.7
	PCA + k-means	4	0.148	1.798	2,418.7
	UMAP + HDBSCAN	–	0.190	0.825	410.3
	IDEC	2	**0.576**	**0.590**	**25,880.7**
	IDEC	3	0.282	1.409	3,674.2
	IDEC	4	0.357	1.253	11,118.7
MDD-Enriched (n=1,616)	PCA + k-means	2	0.226	1.560	594.6
	PCA + k-means	3	0.145	1.961	421.9
	PCA + k-means	4	0.145	1.800	364.8
	UMAP + HDBSCAN	–	**0.601**	**0.545**	**2,409.3**
	IDEC	2	0.161	2.069	353.1
	IDEC	3	0.106	2.186	248.8
	IDEC	4	0.095	2.342	205.1

All methods use a 16-dimensional representation for fair comparison. For PCA + k-means and IDEC, k is varied. UMAP + HDBSCAN does not require specifying k. Boldface indicates the best score in each cohort for each metric.

3 Results

As an initial validation, our IDEC implementation reproduced performance metrics from Guo et al. (2017) on MNIST (87.2% accuracy, 86.1% NMI).

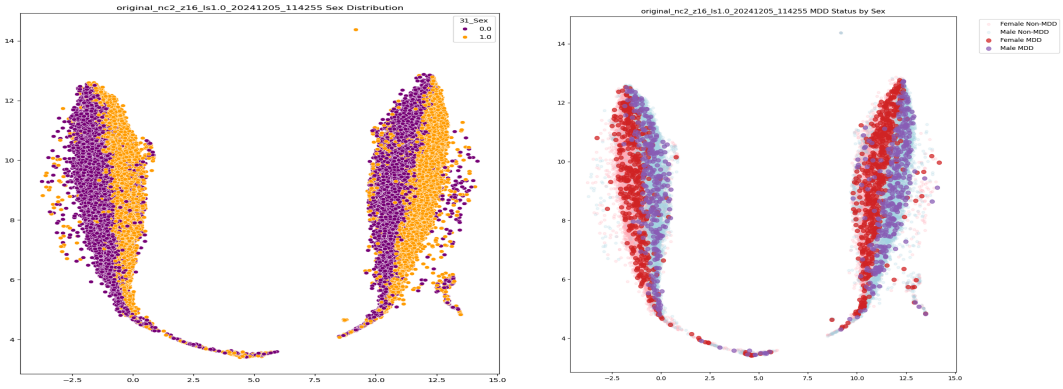
3.1 Full Dataset Analysis

Analysis of the discovery cohort (n=22,484) revealed substantial performance differences across methods (Table 1). IDEC with k=2 achieved superior scores across all quality metrics, significantly outperforming both baseline approaches. Notably, clustering quality decreased monotonically with increasing k, suggesting the data naturally separates into two primary groups.

The optimal two-cluster solution revealed biological sex as the dominant organizing factor (χ^2=6.056, p=1.386e-02). UMAP visualization of the 16-dimensional IDEC embeddings (Figure 2) shows two distinct regions with internal sex-based gradients. Despite strong technical clustering performance, MDD-related patterns showed minimal between-cluster variation (6.5-7.7%). Instead, we observed continuous demographic gradients, particularly in BMI-MDD relationships that transcended cluster boundaries.

3.2 Sex-Stratified Analysis

To investigate whether sex differences masked subtler depression-specific signals, we analyzed female-only (n=12,185) and male-only (n=10,416) cohorts independently. Across both populations, IDEC with k=2 maintained superior clustering performance compared



(a) UMAP colored by sex

(b) UMAP colored by MDD Status and Sex

Figure 2: UMAP projections of IDEC embeddings (k=2) for the full UKB proteomics dataset (n=22,601). (a) Shows cluster organization colored by sex. (b) Shows the same embedding space colored by MDD status.

Table 2: Cluster Numbers Ablations for Sex-Stratified Analysis.

	Male Cohort			Female Cohort		
Metric	k=2	k=3	k=4	k=2	k=3	k=4
Cluster Characteristics						
Size Range	5170-5176	1313-4568	253-3556	5949-6189	3765-4244	2371-3483
Balance Ratio	0.999	0.287	0.071	0.961	0.887	0.681
Depression Patterns						
MDD Rate Range (%)	5.6-6.0	4.3-6.3	5.1-6.5	7.3-9.1	7.1-9.5	6.9-9.6
Max Difference (%)	0.4	2.0	1.4	1.8	2.4	2.7
CI Overlap (%)	92.4	83.7	76.5	88.2	79.4	72.8

Balance Ratio = smallest/largest cluster size. CI Overlap = percentage of cluster pairs with overlapping MDD rate confidence intervals. Max Difference = largest absolute difference in MDD rates between any pair of clusters. All configurations use z=16, ls=1.0, batch=256, pretrain=10k, train=25k, lr=0.0001, γ=0.1

to baseline methods (Table 1). However, examination of MDD prevalence within these refined clusters revealed limited evidence for biologically distinct subtypes. As shown in Table 2, increasing cluster numbers did not produce meaningful differences in MDD rates, with prevalence varying by less than 3% even at higher k values. Moreover, increased granularity often led to imbalanced cluster sizes, particularly in the male cohort, reducing interpretability.

3.3 MDD-Enriched Analysis

Analysis of the clinically homogeneous MDD-enriched subset (n=1,616) revealed different methodological challenges. Unlike larger cohorts, this smaller sample favored UH, which achieved superior clustering scores compared to both PKM and IDEC (Table 1). This methodological shift likely reflects UH's robustness to limited sample sizes, where deep architectures typically struggle. However, even with this more suitable approach, clustering patterns primarily reflected demographic gradients rather than depression-specific subtypes, supporting the gradient-based nature of depression's proteomic architecture.

4 Discussion and Conclusion

Our comprehensive analysis reveals that depression's proteomic landscape manifests as a continuous gradient rather than discrete clusters. This finding persisted across multiple analytical approaches, cohort compositions, and clustering granularities. Even IDEC, which

excels at uncovering complex latent structures, consistently identified broad demographic patterns rather than depression-specific subtypes. Quantitatively, MDD rates showed minimal variation between clusters (6.5-7.7%), with overlapping confidence intervals suggesting no distinct depression subtypes. Instead, we observed continuous demographic gradients, particularly in the relationship between BMI and MDD risk. High-BMI groups consistently showed elevated MDD rates compared to low-BMI groups across clusters (7.9-9.4% vs 5.0-6.0%, p=4.8e-12 - 7.3e-10), indicating that this biological association transcends cluster boundaries.

These results align with emerging perspectives in psychiatry that favor dimensional frameworks over categorical classifications. The persistent gradient-based organization we observed suggests that future biomarker discovery efforts might benefit from embracing this inherent continuity rather than seeking discrete subtypes. This shift in perspective could inform more nuanced approaches to personalized interventions, acknowledging the smooth transitions in biological states that appear to characterize depression at the proteomic level.

Our extensive experimentation with IDEC revealed significant challenges in parameter optimization that warrant discussion. The algorithm's sensitivity to hyperparameters necessitated comprehensive tuning across latent dimensions (2-32), number of clusters (2-10), batch sizes (16-256), learning rates (10^{-5}-10^{-3}), and loss weighting coefficients (0.01-0.5). The two-phase training process, involving autoencoder pretraining followed by end-to-end optimization, introduced additional complexity. Notably, increasing cluster numbers did not produce meaningful differences in MDD rates, with less than 3% variation in prevalence even at higher k values. Confidence intervals for MDD rates across clusters maintained substantial overlap (>72%), further supporting a continuous rather than discrete distribution. We found that while IDEC could effectively separate clusters in the latent space (achieving Silhouette scores of 0.787 for k=2), the resulting embeddings often appeared artificially compact without dimensionality reduction techniques like UMAP, suggesting potential over-compression of the continuous biological variation.

Several key technical limitations of IDEC emerged during our implementation. First, we observed that significant loss reduction required extraordinarily long training periods, with meaningful improvements continuing well beyond 20,000 epochs - substantially longer than reported in previous applications. This extended convergence time imposed considerable computational costs, particularly for large-scale proteomic datasets. Our attempts to optimize the architecture through scaling layers proved ineffective, suggesting limitations in the model's ability to adapt to varying data scales. The two-phase training process, while theoretically sound, introduced additional complexity in practice as the transition between pretraining and clustering optimization often led to instability. We found that while IDEC could achieve strong clustering metrics (Silhouette scores of 0.787 for k=2), these quantitative improvements did not necessarily translate to biologically meaningful separations.

Recent advances in proteomic analysis have increasingly demonstrated that many biological processes follow continuous trajectories rather than discrete states. This emerging pattern suggests a broader principle: complex biological processes, particularly in psychiatric disorders, may be better understood through gradient-based models rather than categorical frameworks. The consistency of this observation across different analytical approaches and biological domains strengthens the case for dimensional perspectives in biomarker development and precision medicine. Our work extends this paradigm specifically to psychiatric phenotypes, where the historical focus on categorical diagnoses may have inadvertently obscured important biological patterns.

The convergence of our findings with broader trends in biological machine learning suggests that future methodological development should prioritize approaches that can effectively model and interpret continuous variation. This could include extending deep clustering frameworks to explicitly accommodate gradient structures, developing hybrid methods that combine discrete and continuous representations, or exploring novel visualization techniques that better capture biological continuity. Such advances would be particularly valuable for psychiatric research.

References

M Austin Argentieri, Sihao Xiao, Derrick Bennett, Laura Winchester, Alejo J Nevado-Holgado, Upamanyu Ghose, Ashwag Albukhari, Pang Yao, Mohsen Mazidi, Jun Lv, et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nature medicine*, 30(9):2450–2460, 2024.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, pp. 1753–1759, 2017.