

A Generative-Fuzzy Approach for Transparent Rule Induction from Latent Embeddings

Matteo Grazioso^{*§}, Elisa Degara^{†§}, Marco S. Nobile^{*}, Andrea Tangherloni[‡]

^{*}Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

Email: matteo.grazioso@unive.it, marco.nobile@unive.it

[†]Department of Computing Sciences, Bocconi University, Milan, Italy

Email: elisa.degara@studbocconi.it, andrea.tangherloni@unibocconi.it

[‡]Bocconi Institute for Data Science and Analytics, Milan, Italy

Abstract—We propose a hybrid Generative-Fuzzy framework to reconcile predictive performance with interpretability in high-dimensional settings by coupling a Variational Autoencoder (VAE) with a Takagi-Sugeno Fuzzy Inference System (FIS) induced via pyFUME. The VAE compresses the original feature space into a smooth latent manifold suitable for fuzzy partitioning, while the FIS translates latent regions into compact human-readable IF-THEN rules. To bridge latent semantics to physical variables, we introduce a feature-attribution mechanism that combines the decoder's Jacobian sensitivity with latent traversal effects, thereby grounding rule antecedents in the original features. We validate the proposed approach on a single-cell transcriptomics benchmark (*FCGR3A*⁺ monocytes under *IFN- β* stimulation), showing that the induced rule base achieves performance comparable to a Random Forest trained on the full feature space, despite operating on a 10-dimensional latent embedding. In addition, the attribution analysis yields biologically coherent drivers (e.g., *ISG15*, *CXCL10*, *IFIT* family) and a near-perfect agreement between local (Jacobian) and global (traversal) effects for the principal activation axis in the latent space, supporting stable and faithful interpretability.

Index Terms—Artificial Intelligence, Deep Learning, Fuzzy Logic, Explainability, Interpretability

I. INTRODUCTION

The exponential growth of high-dimensional data across diverse scientific domains has created a critical tension between predictive performance and model interpretability. While Deep Learning (DL) architectures excel at capturing the complex, non-linear dependencies inherent in these large-scale datasets, they often operate as “black box” models, offering internal logic that is inaccessible to human experts. In high-stakes environments – such as medicine, finance, or genomics – a model that achieves high accuracy but cannot explain its underlying reasoning is of limited utility for decision support or discovery. Standard linear models provide transparency but often fail to capture the underlying manifold structure of complex data. In contrast, state-of-the-art ensemble methods offer robust classification but yield only static feature-importance rankings. These rankings identify which variables are relevant but fail to explain the linguistic logic that defines specific states. To bridge this gap, we propose a hybrid framework that

couple the generative and compression power of Variational Autoencoders (VAEs) with the linguistic transparency of Fuzzy Inference Systems (FIS). Our approach leverages the continuous latent manifold learned by a VAE to induce a set of first-order Takagi-Sugeno (TS) fuzzy rules via the pyFUME library. Unlike traditional classifiers, our framework achieves a massive reduction in dimensionality while maintaining performance comparable to established ensemble baselines. By anchoring fuzzy membership functions in a smooth, differentiable latent space, we transform abstract embeddings into a transparent “IF-THEN” logic. This architecture enables us to map abstract latent dimensions back to the original feature space via a novel feature-attribution mechanism based on Jacobian sensitivity and manifold traversal. We evaluate the efficacy of this general-purpose framework using a high-dimensional single-cell transcriptomics dataset as our primary use case. This domain serves as a benchmark due to its inherent sparsity, technical noise, and the continuous nature of biological state transitions. Our results show that our hybrid model achieves a classification accuracy exceeding 99%, matching the performance of a Random Forest (RF) baseline, while reducing the complex decision logic to a minimal set of human-readable rules. The proposed framework addresses the inherent trade-off between predictive power and interpretability in high-dimensional datasets.

Recent works in interpretable deep learning for single-cell omics have shifted from post-hoc general explainability methods to models that incorporate interpretability directly into representation learning. General explainability methods such as LIME, SHAP, and Integrated Gradients identify the most influential features using local surrogate models, additive importance scores, or sensitivity maps [1]–[3]; therefore, they provide per-sample or aggregated explanations, but do not formally represent the global decision logic of the model. In single-cell transcriptomics, generative models such as scVI [4] have been used to generate latent embeddings of cells, but they typically lack intrinsic interpretability. However, several extensions seek to incorporate some element of explainability. For example, LDVAE constrains the decoder to be linear, thus allowing interpretation of latent variables through gene loadings [5]; VEGA uses a latent variable model to structure the latent

[§]These authors contributed equally to this work.

space based on known gene modules so that pathway-level representations can be obtained [6]; topic modeling approaches have also provided methods for producing interpretable gene signature factors [7]; factor-analysis frameworks such as MOFA+ extend this to multi-condition and multi-modal settings [8]. Overall, interpretability has typically been captured via latent factors, gene loadings, and module activations; however, explicit decision rules have normally not been provided. The proposed coupling of VAE and FIS allows for the creation of a set of linguistic IF-THEN rules on the latent manifold, while linking them back to the original feature space through a dual attribution mechanism (via both decoder Jacobian sensitivities and latent traversal). The system produces an explicit rule base with cluster-level feature attributions, providing human-readable decision rules.

II. THE GENERATIVE-FUZZY HYBRID APPROACH

The proposed framework reconciles predictive power with interpretability through a three-stage hybrid architecture. This approach is designed to transform complex, non-linear data distributions into a structured set of human-readable fuzzy rules by coupling generative manifold learning with linguistic induction. The workflow can be divided into three steps:

- 1) **Manifold Extraction** (Sec. II-A). A VAE compresses the high-dimensional input space $\mathcal{X}$ into a continuous, smooth latent manifold $\mathcal{Z}$.
- 2) **Rule Induction** (Sec. II-B). The latent manifold is partitioned into overlapping fuzzy regions using `pyFUME`, generating a first-order TS fuzzy inference system.
- 3) **Feature Attribution** (Sec. II-C). A bidirectional flow of information is established by computing the Jacobian of the VAE decoder, mapping abstract latent antecedents back to physical feature sensitivities.

A. Probabilistic Feature Embedding via VAE

To extract a low-dimensional representation that preserves the essential topological properties of the high-dimensional input space $\mathcal{X} \in \mathbb{R}^D$, we use a VAE [9], [10], which optimises a probabilistic mapping $p(\mathbf{z}|\mathbf{x})$. This formulation ensures that the resulting latent manifold $\mathcal{Z} \in \mathbb{R}^d$ is characterised by continuity and local smoothness. Such properties are a mathematical prerequisite for the subsequent induction of stable fuzzy Membership Functions (MFs) via `pyFUME`, as they guarantee that proximal points in the latent space represent semantically similar observations. The architecture follows a bottleneck topology composed of an inference network, named *encoder*, $q_\phi(\mathbf{z}|\mathbf{x})$ and a generative network, known as *decoder* $p_\theta(\mathbf{x}|\mathbf{z})$. The encoder creates a non-linear projection of the input vector $\mathbf{x}$ through a hierarchy of layers. In our case, we used dense layers with *Rectified Linear Units* (ReLU) activation function to model complex, non-linear feature interdependencies. *Batch Normalisation* is applied to stabilise the internal covariate shift during the learning of the manifold. In contrast, the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ employs *Gaussian Error Linear Units* (GELU) in its hidden layers. The smoothness of GELU ensures that the *in silico* perturbation is performed in an informative and stable

way (see Section II-C). To transform the discrete inputs into a continuous posterior distribution, the final hidden layer of the encoder parametrises the latent space by yielding two vectors $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}^2$, which represent the parameters of a Multivariate Normal distribution with a diagonal covariance matrix. Note that calculating the inverse and determinant of a full covariance matrix Σ is computationally expensive ($O(d^3)$), where d is the number of latent dimensions. To simplify this step, VAEs typically assume that the latent space dimensions are linearly independent. An important consideration in training VAEs is the stochasticity of the latent bottleneck. Direct sampling from the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is a non-differentiable operation, hindering backpropagation. To resolve this, we employ the *reparameterisation trick* [11], which reformulates the sampling process by expressing the random variable $\mathbf{z}$ as a deterministic transformation of an auxiliary noise variable ϵ . The latent embedding $\mathbf{z}$ is thus expressed as:

$$\mathbf{z} = \boldsymbol{\mu} + \exp\left(\frac{1}{2} \log \boldsymbol{\sigma}^2\right) \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (1)$$

In this formulation, the stochasticity is decoupled from the network parameters ϕ . The mapping from the input $\mathbf{x}$ to the latent representation $\mathbf{z}$ becomes a fully differentiable operation with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. This differentiability is crucial for our proposed framework, as it ensures that the gradient $\nabla_{\mathbf{z}} f_\theta(\mathbf{z})$ (i.e., the Jacobian matrix) is well-defined and stable. Consequently, we can backpropagate through the generative decoder to ground our abstract fuzzy rules in the original physical features (Section II-C). The learning process is mathematically formulated as the minimisation of a composite cost function $\mathcal{L}$, representing the negative Evidence Lower Bound (ELBO). This objective optimises network parameters $\{\phi, \theta\}$ to balance reconstruction quality with latent geometric regularisation:

$$\begin{aligned} \mathcal{L}(\phi, \theta; \mathbf{x}) &= \mathcal{L}_{MSE} + \beta \cdot \mathcal{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \\ &= \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \beta \cdot \mathcal{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \end{aligned} \quad (2)$$

The first term, $\mathcal{L}_{MSE}$, minimises the Mean Squared Error (MSE) between the input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$, ensuring the latent space $\mathcal{Z}$ retains maximal mutual information. The second term, the Kullback-Leibler (KL) divergence, regularises the latent distribution toward a standard normal prior $\mathcal{N}(0, \mathbf{I})$. Optimisation is performed using the Adam algorithm, which adaptively navigates the non-convex loss landscape to ensure convergence to a stable, continuous manifold. In this work, we adopt the β -VAE formulation [12], specifically setting $\beta = 0.001$, after a preliminary evaluation (data not shown). This low-magnitude weighting prioritises a high-fidelity mapping of the data manifold. By relaxing the bottleneck constraint (where $\beta \ll 1$), we ensure that the latent embeddings preserve the fine-grained topological nuances and non-linear variances of the original feature space. The specific architectural configuration and training hyperparameters are detailed in Table I. Ultimately, subsequent Linguistic Rule induction enables the fuzzy membership functions to capture subtle transitions within the data distribution precisely.

TABLE I: VAE hyperparameters

Hyperparameter	Configuration
Encoder Layers	512, 256, 10 (d)
Latent Dimensions (d)	10
Activation (Encoder)	ReLU (Hidden)
Activation (Decoder)	GELU (Hidden), Linear (Out)
Regularization	Batch Norm
β Coefficient	0.001
Optimizer	Adam
Batch Size	128
Epochs	60

B. Linguistic Rule Induction via pyFUME

FIS provide a paradigm that encodes complex relationships into a transparent, rule-based architecture. This makes FIS inherently interpretable, as each rule corresponds to a local region of the input space with an associated output function. The core of our interpretability approach utilises a first-order TS model. Within this framework, the continuous d -dimensional latent space $\mathcal{Z}$ is partitioned into K overlapping fuzzy regions, each governed by an antecedent-consequent rule structure. Formally, a rule R_k is defined as:

$$R_k : \text{IF } z_1 \text{ is } A_{k1} \wedge \dots \wedge z_d \text{ is } A_{kd} \text{ THEN } y_k = f_k(\mathbf{z}) \quad (3)$$

where:

- $\mathbf{z} = (z_1, \dots, z_d)$ represents the input vector derived from the VAE latent space.
- A_{ki} denotes the fuzzy set (linguistic variable) associated with the i -th latent dimension for the k -th rule.
- y_k is the consequent function. In this work, we utilise a first-order linear consequent $y_k = \mathbf{a}_k^\top \mathbf{z} + b_k$, although zero-order approximations are also supported.

The degree to which an input $\mathbf{z}$ satisfies the conditions of rule R_k is quantified by the firing strength β_k , calculated as $\beta_k(\mathbf{z}) = \min(\mu_{A_{k1}}(z_1), \dots, \mu_{A_{kd}}(z_d))$, where $\mu_{A_{ki}}(z_i)$ represents the membership function for the fuzzy set of feature i , with $i = 1, \dots, d$, on cluster k , with $k = 1, \dots, K$. The final output of the system, y^* , is determined through the weighted average of all active rules $y^* = \frac{\sum_{k=1}^K \beta_k(\mathbf{z}) y_k(\mathbf{z})}{\sum_{k=1}^K \beta_k(\mathbf{z})}$.

To automate the generation of these rules, we utilise pyFUME (Python Fuzzy Model Estimation), a library designed for the data-driven induction of TS models and generation of FIS [13], enabling a conversion of numerical datasets into rule-based representations. The workflow is divided into the following five phases:

- 1) **Data Characterization.** Ingestion and normalisation of the latent embeddings $\mathcal{Z}$, to ensure numerical stability during clustering.
- 2) **Clustering – Latent Manifold Partitioning.** Identification of topological centres of the data by partitioning the latent manifold. This is typically achieved using a variant of Fuzzy C-Means (FCM) [14] powered by Fuzzy Self-Tuning Particle Swarm Optimisation [15].
- 3) **Antecedent Estimation.** Estimation of MFs that define the fuzzy sets A_{ki} for each input variable. By fitting convex

envelopes to the membership values derived during the previous stage, pyFUME induces linguistic boundaries that reflect the actual distribution of the latent space.

- 4) **Consequent Estimation.** Estimation of parameters of the linear functions a_k and b_k via least-squares optimisation, ensuring the rules accurately approximate the VAE’s output.
- 5) **Simplful Integration.** Implementation of the resulting fuzzy model using the *Simplful* library, which enables the design and assessment of the fuzzy rules in natural language.

Although pyFUME includes additional advanced functionalities – such as imputation, feature selection [16], automatic rule simplification and model size reduction [17] – these capabilities were not exploited in the present work and are thus omitted from further discussion.

C. Perturbation and Gradient-Based Feature Attribution

While the FIS induced via pyFUME provides a transparent decision logic within the latent space $\mathcal{Z}$, the latent dimensions z_i represent abstract, compressed representations that lack direct semantic correspondence to the original features in $\mathcal{X}$. To bridge this gap and achieve full model interpretability, we propose a gradient-based feature-attribution method that maps the linguistic boundaries of fuzzy rules back to the original feature space. The continuity and differentiability of the latent space $\mathcal{Z}$, guaranteed by the reparameterisation trick discussed in Section II-A, enable *in silico* perturbations. We treat the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ as a differentiable generative function $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^D$. To quantify the influence of each original feature x_l , with $l = 1, \dots, D$, on the local geometry of the manifold, we compute the Jacobian matrix $\mathbf{J}_{f_\theta}(\mathbf{z})$ at specific latent coordinates:

$$\mathbf{J}_{f_\theta}(\mathbf{z}) = \nabla_{\mathbf{z}} f_\theta(\mathbf{z}) = \begin{bmatrix} \frac{\partial \hat{x}_1}{\partial z_1} & \dots & \frac{\partial \hat{x}_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{x}_D}{\partial z_1} & \dots & \frac{\partial \hat{x}_D}{\partial z_d} \end{bmatrix}. \quad (4)$$

Each entry $J_{li} = \partial \hat{x}_l / \partial z_i$ is a partial derivative that captures the *local sensitivity* of the l -th reconstructed feature with respect to the i -th latent dimension. This property is particularly relevant to our hybrid generative–fuzzy framework, as it enables the evaluation of feature relevance at precise locations in latent space, namely at the centroids $\mathbf{c}_k$, with $k = 1, \dots, K$, associated with the fuzzy regions. The linguistic transparency of each fuzzy rule R_k is anchored in the latent space $\mathcal{Z}$. To map a fuzzy rule R_k back to the original input space $\mathcal{X}$, we evaluate the Jacobian of the decoder at the latent centroid $\mathbf{c}_k$, identified during manifold partitioning (Section II-B). To do so, we define the importance score $S_{l,k}^{(i)}$ for the l -th physical variable associated with rule k as the magnitude of the partial derivative of the decoder output with respect to the i -th latent dimension:

$$S_{l,k}^{(i)} = \left| \frac{\partial [f_\theta(\mathbf{z})]_l}{\partial z_i} \right|_{\mathbf{z}=\mathbf{c}_k}. \quad (5)$$

Ranking input features according to $S_{l,k}^{(i)}$ identifies the *driver features* governing the local manifold region associated with the fuzzy rule R_k . This attribution mechanism transforms abstract latent antecedents (e.g., “IF z_i is High”) into semantically grounded descriptions in the original input space (e.g., “IF Feature l is upregulated”). As a result, the global fuzzy

model remains explicitly anchored to the physical variables of the dataset, ensuring interpretability without sacrificing expressiveness. The feature attribution process is partitioned as an *in silico* perturbation study through the following steps:

- 1) **Centroid Evaluation.** For each induced rule R_k , the corresponding fuzzy cluster centre c_k is utilised as the representative state of the system for that local region.
- 2) **Differentiable Synthesis.** A deterministic forward pass is executed through the decoder f_θ , followed by an automated differentiation pass (backpropagation) to compute the sensitivities of all D features with respect to the d latent coordinates.
- 3) **Local Descriptor Extraction.** Features are ranked by their sensitivity scores $S_{l,k}^{(i)}$, and a Top- N selection is applied to isolate the most significant descriptors for each rule.

This synthesis of generative manifold learning and fuzzy inference enables a dual-objective architecture that maintains global predictive fidelity while achieving granular, local linguistic interpretability. The resulting framework establishes a pipeline that transforms high-dimensional, non-linear data distributions into structured, human-readable explanations, bridging the gap between DL performance and transparent decision support.

III. RESULTS AND VALIDATION

We evaluate the proposed generative-fuzzy framework using a high-dimensional single-cell transcriptomics dataset. This domain serves as an ideal benchmark due to the high sparsity, technical noise, and complex non-linear dependencies inherent in biological systems. We demonstrate that our approach preserves the global topology of cellular states while providing a transparent, rule-based decomposition of the underlying gene regulatory logic.

A. Experimental Setup: Single-Cell Transcriptomics

To evaluate the efficacy of the proposed framework, we utilised a publicly available Single-Cell RNA-sequencing (scRNA-seq) dataset of Peripheral Blood Mononuclear Cells (PBMC), provided by *10x Genomics* [18]. The dataset comprises 14 039 cells across two experimental conditions: *control* and *stimulated* (interferon- β stimulation).

To maintain a high signal-to-noise ratio and isolate cell-type-specific regulatory programs, we focused our analysis on the $FCGR3A^+$ Monocyte sub-population for a total of 1 388 cells. This subset serves as a rigorous benchmark for linguistic rule induction, as the transcriptomics shift induced by stimulation represents a continuous transition within a specific immune lineage, rather than a discrete cluster separation.

Data integrity was maintained through a stringent quality control pipeline. Features detected in fewer than 200 cells were discarded to mitigate the influence of technical artefacts and the inherent sparsity (dropout) characteristic of scRNA-seq data. Following filtration, the expression matrix was subjected to library-size normalisation (target sum of 10^4) and a log transformation ($\log(1 + x)$) to stabilise the variance of the raw count distribution. This preprocessing is relevant to the numerical stability of the VAE, as it maps highly skewed

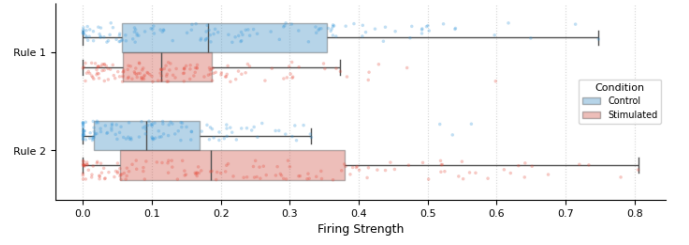


Fig. 1: Firing strength of the two fuzzy rules determined by pyFUME, evaluated on the test set.

transcriptomics counts to a range suitable for gradient-based optimisation (post-normalisation range: $[0.0, 8.73]$).

Finally, the data was partitioned into an 80% training set for manifold discovery and rule induction, and a 20% held-out test set to evaluate the predictive performance and generalisability of the induced fuzzy rules on unseen biological states.

B. VAE Manifold Learning and Feature Compression

The first stage of our framework compresses the transcriptomic feature space onto a low-dimensional generative manifold. Given the high-dimensional nature of the input ($D = 12\,762$ genes), we implemented a deep bottleneck architecture, as detailed in Table I. This configuration ensures a robust latent representation that prioritises reconstruction fidelity over aggressive prior regularisation.

The model was trained for 60 epochs with a learning rate of 5×10^{-4} and a low β -coefficient of 0.001. This specific β weighting was chosen to minimise the KL divergence pressure, thereby ensuring that the 10-dimensional latent space $\mathcal{Z}$ captures the fine-grained variances associated with the monocyte activation program rather than collapsing into a standard normal distribution.

The training converged without evidence of posterior collapse, yielding a smooth, continuous manifold. This continuity is important for the subsequent stage, as it provides a stable coordinate system for pyFUME to induce overlapping fuzzy clusters.

C. Linguistic Rule Induction and Latent Variable Interpretation

Following the manifold extraction, we utilised pyFUME to induce a first-order TS FIS. Given the binary nature of the experimental condition (*control* vs. *stimulated*), the system was configured to identify 2 principal fuzzy clusters within the 10-dimensional latent space $\mathcal{Z}$. This choice is grounded in the binary nature of the two experimental conditions. Each cluster serves as the prototypical centre for a linguistic rule that characterises a specific functional state of the $FCGR3A^+$ monocytes. The induced rules are defined by the membership functions across all latent dimensions, capturing the *fuzziness* of cellular states as they transition from unstimulated to activated configurations. The resulting rule base is expressed as follows:

- Rule 1: IF (z_1 is C_1) $\wedge \dots \wedge$ (z_{10} is C_1) THEN $y_1 = f_1(\mathbf{z})$
- Rule 2: IF (z_1 is C_2) $\wedge \dots \wedge$ (z_{10} is C_2) THEN $y_2 = f_2(\mathbf{z})$

The consequent MFs (f_1, f_2) provide a local linear approximation of the manifold’s response to the $IFN-\beta$ stimulus. For instance, in Rule 1 (the control-dominant region), the output function is characterised by the hyperplane $y_1 = 0.541 - 0.084z_1 + \dots + 0.004z_{10}$, whereas Rule 2 (the stimulation-dominant region) yields $y_2 = 0.237 - 0.047z_1 + \dots + 0.083z_{10}$. The variation in coefficients between these two functions, particularly in dimensions such as z_3 (increasing from 0.28 in Rule 1 to 0.51 in Rule 2) and z_4 , highlights the specific latent axes that most heavily influence the transition to a stimulated state.

Figure 1 shows the firing strength of both rules calculated over the samples of the training set. By analysing the distributions, we observe that the firing strength of Rule 1 is higher for cells in the control set, whereas the strength of Rule 2 is higher for cells in the stimulated set. Figure 2 illustrates the induced MFs for the most discriminative latent variables. The overlapping regions of these sets effectively model the biological “grey zones”, *i.e.*, cells transitioning between a quiescent state and an interferon-activated state, providing a more nuanced representation than discrete Boolean logic.

D. Comparative Performance: Generative-Fuzzy vs. RF

To evaluate the predictive integrity of the induced FIS, we conducted a comparative analysis against an RF classifier. The RF provides a robust baseline for high-dimensional biological data, as it naturally captures non-linear interactions and feature dependencies. While the RF model was trained on the full high-dimensional feature space D , our hybrid approach operates on the heavily compressed 10-dimensional latent manifold $\mathcal{Z}$, representing a 1276-fold reduction in dimensionality. The classification results on the 20% held-out test set are summarised in Table II. The RF baseline achieved near-perfect accuracy of 99.135% and a corresponding Area Under the ROC Curve of 0.9991. In comparison, the proposed Generative-Fuzzy framework attained an accuracy of 99.1%.

TABLE II: Classification Performance (FCGR3A+ Monocytes)

Model	Accuracy	MSE
Random Forest (Baseline)	99.135%	0.0113
Generative-Fuzzy (Proposed)	99.100%	0.0145

The performance differential ($\Delta_{Acc} = -0.035\%$) is statistically negligible, indicating that our hybrid model achieves performance comparable with the state-of-the-art ensemble method. This is a significant result as it shows that the VAE successfully preserved the discriminative variance of the interferon response within the latent bottleneck. Furthermore, the low MSE of the FIS suggests that the first-order TS rules provide a highly precise approximation of the manifold’s local topology. The scientific value of this comparison lies in the trade-off between complexity and interpretability. While the RF requires an ensemble of decision trees to achieve this accuracy – rendering the global decision logic opaque – our Generative-Fuzzy framework achieves the same result using a rule base

consisting of only two linguistic rules. This proves that high-dimensional cellular states can be accurately modelled and classified using a transparent, low-dimensional logic without sacrificing the predictive power of DL architectures.

E. Biological Grounding via Jacobian-Traversal Consensus

While the Takagi-Sugeno rules provide a transparent logic within the latent space, the dimensions z_i lack inherent biological meaning. To resolve this, we performed an *in silico* perturbation analysis by computing the Jacobian matrix $\mathbf{J}_{f_\theta}(\mathbf{z})$ of the VAE decoder. By evaluating the partial derivatives $\partial \hat{x}_1 / \partial z_i$ at the fuzzy cluster centroids $\mathbf{c}_k$, we quantified the sensitivity of each physical gene with respect to the latent antecedents. The core contribution of our hybrid framework is the ability to assign semantically rigorous biological meaning to abstract latent dimensions. To achieve this, we performed a dual-validation analysis on the 10D manifold: (i) a first-order Jacobian sensitivity analysis to identify local gene influence via gradients, and (ii) a latent space traversal to observe the global transcriptomic shift. In the latter, we perturb the latent coordinates across their range and measure the resulting displacement in the reconstructed feature space $\mathcal{X}$, effectively quantifying the global impact of each dimension.

To identify the main regulatory drivers, we quantified the condition-dependent shift along each latent axis $\mathcal{Z}_i$ by calculating the absolute difference in mean latent coordinates between the control and stimulated populations ($|\Delta \bar{z}_i|$). This displacement serves as a proxy for the dimension’s discriminative relevance. A high $|\Delta \bar{z}|$ demonstrates that the VAE bottleneck has effectively isolated the stimulus-induced variance into a dedicated coordinate, providing a stable antecedent space for the subsequent induction of fuzzy rules. As illustrated in Table III and confirmed by Figure 3, dimension z_5 emerges as the primary axis of immune activation, exhibiting a maximal separation of $|\Delta \bar{z}_5| = 1.908$. This statistical displacement signifies that z_5 captures the global transcriptomics state-transition induced by $IFN-\beta$ stimulation (see Figure 2).

TABLE III: Ranking of Latent Dimensions by Biological Separation ($|\Delta \bar{z}|$)

Dimension	Control $\bar{z}$	Stimulated $\bar{z}$	$ \Delta \bar{z} $	Top Driver Gene
z_5	-1.3017	0.6066	1.9083	<i>ISG15</i>
z_7	0.4447	-1.2974	1.7421	<i>ISG15</i>
z_1	-0.9806	-0.3392	0.6414	<i>CXCL10</i>
z_4	-0.4245	-0.0852	0.3393	<i>VMO1</i>
z_6	-0.0425	0.2493	0.2918	<i>CCL2</i>
z_3	0.1443	0.4094	0.2651	<i>CD63</i>
z_9	-1.2816	-1.4461	0.1645	<i>FTH1</i>
z_{10}	-0.1649	-0.0460	0.1189	<i>TNFSF10</i>

The ranking provides a quantitative justification for the linguistic rules induced in the subsequent phase. While all 10 dimensions were utilised by `pyFUME`, the magnitude of the shift in z_5 is nearly six times greater than that of z_4 . This indicates that the VAE successfully concentrated the stimulus-response variance into a specific, high-signal latent coordinate. The biological convergence further validates this ranking: the

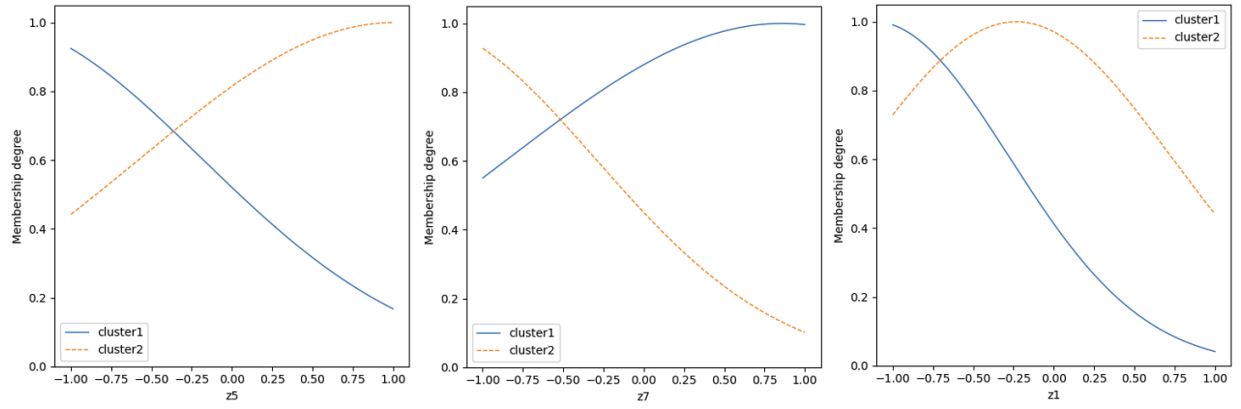


Fig. 2: Induced MFs for the most significant latent dimensions. The two fuzzy sets (Cluster 1 and Cluster 2) represent the linguistic partitions corresponding to the *control* and *stimulated* populations, respectively.

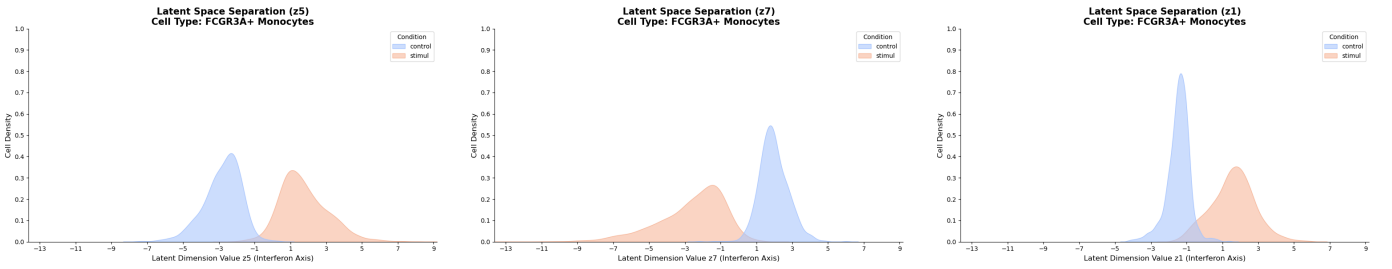


Fig. 3: Latent space separation between the two classes across selected latent dimensions.



Fig. 4: Jacobian-based gene sensitivity across the most discriminative latent dimensions. Heatmap reporting the Jacobian sensitivity $|\partial \hat{x}_l / \partial z_i|$ of the decoder for the top-ranked genes (columns) associated with the selected latent dimensions (rows; e.g., primary/secondary activation axes). Sensitivity analyses are performed at representative latent states (e.g., cluster centroids) and identify the genes most strongly influenced by each latent coordinate, revealing coherent interferon-response signatures.

dimensions with the largest $|\Delta \bar{z}|$ (i.e., z_5, z_7) are precisely those that exhibit the highest sensitivity to classical interferon markers like *ISG15* and *CXCL10*. This alignment ensures that our framework prioritises the most biologically relevant transitions when generating human-readable explanations.

The primary regulatory drivers of the z_5 axis are detailed in Table IV. To ensure the robustness of our feature attribution, we use a Consensus Rank, calculated as the arithmetic mean of the individual ranks derived from Jacobian sensitivity (local gradient) and latent traversal (global manifold shift). Fig. 4

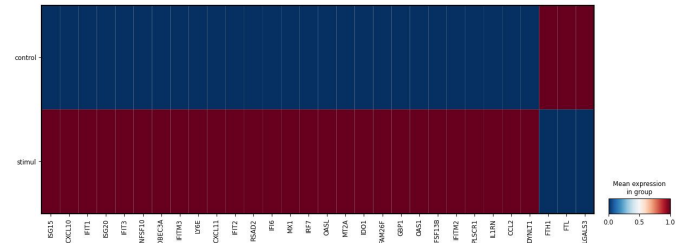


Fig. 5: Biological validation of latent drivers via differential expression patterns. Mean normalised log-expression of the genes highlighted by the Jacobian analysis shown separately for control and stimulated *FCGR3A*⁺ monocytes. The induced drivers exhibit strong condition-dependent modulation, supporting that the latent axes and their fuzzy antecedents are grounded in biologically meaningful transcriptional programs.

summarises the Jacobian sensitivity profiles for the most discriminative latent axes, showing that a small subset of genes dominates the local decoder response $|\partial \hat{x}_l / \partial z_i|$ and that these drivers form coherent immune-response modules. Fig. 5 provides biological grounding by showing that the same Jacobian-selected genes are strongly upregulated under stimulation, confirming that the discovered latent drivers correspond to genuine transcriptional programs rather than generative artefacts. Finally, the near-perfect agreement between Jacobian

TABLE IV: Top 10 gene descriptors for latent dimension z_5

Gene	Jacobian Sensitivity	Traversal Δ	Consensus Rank
<i>ISG15</i>	0.9698	4.9223	1.0
<i>CXCL10</i>	0.9275	4.7221	2.0
<i>IFIT1</i>	0.7022	3.5589	3.0
<i>ISG20</i>	0.6886	3.4811	4.0
<i>IFIT3</i>	0.6525	3.3267	5.0
<i>TNFSF10</i>	0.6503	3.2592	6.0
<i>APOBEC3A</i>	0.6225	3.1572	7.0
<i>IFITM3</i>	0.5769	2.8833	8.5
<i>LY6E</i>	0.5702	2.8976	8.5
<i>IFIT2</i>	0.5566	2.8227	10.5

and traversal rankings for the principal axis (Pearson correlation $r = 0.9996$ for z_5) supports that local gradients are faithful descriptors of global manifold shifts under the chosen smooth decoder parametrisation. This near-perfect correlation is due to the mathematical smoothness of the GELU activation function in the decoder, which ensures that local gradients accurately reflect global manifold trends. As evidenced by the Consensus Rank, this analysis identifies a signature of classical Interferon-Stimulated Genes (ISGs), including *ISG15*, *CXCL10*, and the *IFIT* family. Furthermore, our analysis reveals a hierarchy of latent encoded functions. While z_5 and z_7 ($r = 0.997$) govern the primary interferon response, other dimensions capture auxiliary biological signals: z_4 and z_6 are associated with chemotaxis and cellular metabolism (e.g., *CCL2* and *VMO1*), while z_3 identifies structural and lysosomal markers (e.g., *CD63* and *FTL*). This multi-dimensional mapping demonstrates that the hybrid Generative-Fuzzy framework effectively deconstructs the complex, high-dimensional response of *FCGR3A*⁺ monocytes into independent, human-readable regulatory modules. By providing a direct link between the linguistic antecedent “IF z_5 is High” and the physical upregulation of *ISG15*, we achieve a level of biological interpretability that surpasses standard differential expression methods.

F. Cross-Model Validation and Feature Consistency

To evaluate the reliability of our gradient-based feature attribution, we performed a cross-model consistency analysis. We compared the top- N genes identified by the Jacobian analysis on the top three latent components (i.e., z_5, z_7, z_1) with the feature importance scores derived from the RF baseline. While the RF identifies global importance through supervised Gini impurity reduction, our framework identifies drivers via the unsupervised gradient of the generative manifold, grounded by fuzzy linguistic rules. As shown in Table V, we observed a remarkable degree of intersection between these two independent methodologies. The primary activation axis,

TABLE V: Feature overlap analysis: VAE-Jacobian (z_k) vs. RF importance

Latent Dim. z_k	Top- k	Overlap (n)	Overlap %	Common Biological Drivers
z_5	50	43	86.0%	<i>ISG15, CXCL10, BST2, IFIT1</i>
z_7	50	43	86.0%	<i>APOBEC3A, CXCL11, DDX58</i>
z_1	50	36	72.0%	<i>EPSTI1, FAM26F, LY6E</i>

z_5 , shows an 86% overlap with the RF top-features, sharing 43 out of 50 genes. This intersection includes the core ISG signature family (e.g., *ISG15*, *CXCL10*, *BST2*, *IFIT1*) identified in Table IV. Even the secondary interferon axis, z_1 , maintains a high consistency of 72%.

The high level of consensus observed between these two distinct methodologies is significant for two reasons. First, the 86% overlap provides a rigorous validation of the latent manifold, confirming that the VAE successfully concentrated the stimulus-response signal into the z_5 axis without introducing generative artefacts. Second, our framework offers greater explanatory value than RF. While the latter provides only a static ranking of importance, our approach contextualises these genes within the induced fuzzy rules. By explaining how the upregulation of specific common genes, such as *ISG15*, drives the linguistic classification of a cell as “Stimulated” through the membership functions of z_5 , the hybrid Generative-Fuzzy framework establishes itself as a robust alternative to traditional classifiers. Ultimately, this architecture harmonises the predictive precision of state-of-the-art ensemble models with the linguistic transparency required for deep insight.

G. Biological Validation: Differential Expression and Pathway Enrichment

To provide a definitive biological validation of our framework, we cross-referenced the genes driving the latent manifold with a standard Differential Expression (DE) gene analysis. We used the Wilcoxon rank-sum test to identify genes significantly modulated between the control and stimulated conditions. This analysis identified 559 significant genes meeting the criteria of an adjusted p-value ($p_{adj} < 0.05$) and a magnitude of change ($|\log FC| > 0.25$). Notably, the population exhibited a strong bias toward upregulation ($N = 317$), consistent with a robust transcriptional response to *IFN- β* stimulation. As shown in Table VI, our framework shows a remarkable quantitative alignment with these DE results. For the primary activation axis z_5 , the top 50 genes identified by the Jacobian analysis exhibit a 90% overlap with the top 50 DE genes. Even more striking is that for both z_5 and z_7 , the top 100 VAE-identified genes show a 100% overlap with the significantly differentially expressed genes. A direct comparison of gene

TABLE VI: Quantitative alignment: VAE-Jacobian vs. Differential Expression

Latent Dim.	Δz (Shift)	Top 50 Overlap	Top 100 Overlap
z_5	+1.908	45 (90%)	100 (100%)
z_7	-1.742	45 (90%)	100 (100%)
z_1	+0.641	48 (96%)	95 (95%)

rankings (see Table VII) reveals that the Generative-Fuzzy framework correctly prioritises the most important immune responders. For instance, *IFIT1* and *ISG15* occupy nearly identical ranks in both the VAE importance and DE significance lists. This statistical convergence confirms that the fuzzy rules reported in Section III-C are based on the most robust transcriptomics signals of the monocyte activation system.

Finally, to translate these gene lists into functional biology,

TABLE VII: Consensus gene ranking: VAE-Jacobian vs. DE

Gene	VAE Rank	DE Rank	Gene	VAE Rank	DE Rank
<i>ISG15</i>	1	1	<i>IDO1</i>	19	30
<i>CXCL10</i>	2	12	<i>GBP1</i>	20	33
<i>IFIT1</i>	3	3	<i>EPSTI1</i>	26	27
<i>IFIT3</i>	5	4	<i>IFI35</i>	29	24
<i>APOBEC3A</i>	7	13	<i>DYNLT1</i>	32	18
<i>CXCL11</i>	11	19	<i>GBP4</i>	40	47
<i>IFI6</i>	13	5	<i>BST2</i>	43	28
<i>FAM26F</i>	18	39	<i>IFI44L</i>	45	40

we performed Gene Set Enrichment Analysis (GSEA) [19] on the z_5 drivers. The functional relevance of the z_5 latent axis is confirmed by its strong enrichment metrics. An adjusted p-value of 4.8×10^{-24} statistically guarantees that the genes driving this dimension are not random, but part of a dedicated Type-I Interferon (α/β) signalling program. Within our fuzzy framework, the MFs on z_5 partition these transcriptomic states: the fuzzy set associated with Rule 2 represents the activated antiviral regime. Thus, a cell's membership value in this set acts as a direct mathematical proxy for its biological activation level, providing a transparent, rule-based interpretation of the immune response.

IV. CONCLUSION

This work introduced a general-purpose Generative-Fuzzy pipeline that transforms complex, high-dimensional data into a compact, interpretable rule system by learning a smooth VAE latent manifold and inducing a Takagi-Sugeno FIS with pyFUME. On FCGR3A⁺ monocytes (control vs *IFN- β* stimulated), the fuzzy model operating on a 10D embedding matched the predictive performance of an RF trained on the original feature space, demonstrating that discriminative biological variance is preserved under strong compression. Crucially, we provided biological grounding of the latent antecedents via a dual attribution strategy: Jacobian sensitivity (local gradients) and latent traversal (global effects). The two views were highly consistent on the main activation axes, enabling robust identification of interferon-stimulated drivers and supporting a transparent interpretation of the learned rules as proxies for activation state. Future work will extend the framework to (i) multi-class and multi-condition settings with more than two fuzzy clusters/rules, (ii) count-aware likelihoods (e.g., NB, ZINB) and uncertainty-aware attributions, and (iii) rule-base simplification in pyFUME to further reduce rule complexity while preserving fidelity. In addition, we will validate transferability across donors and batches and explore whether latent antecedents can be aligned with known pathway activities to support discovery across broader biomedical domains.

ACKNOWLEDGMENT

This work was also supported by the high-performance computing infrastructure developed under the project “CONVECS”, funded by the PR Veneto FESR 2021-2027 program, Priority 1 – Specific Objective 1.1 – Action 1.1.2.

REFERENCES

- [1] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘‘Why should I trust you?’’: Explaining the predictions of any classifier,’’ in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.
- [2] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,’’ in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [3] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,’’ in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.
- [4] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, “Deep generative modeling for single-cell transcriptomics,’’ *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, 2018.
- [5] V. Svensson, A. Gayoso, N. Yosef, and L. Pachter, “Interpretable factor models of single-cell RNA-seq via variational autoencoders,’’ *Bioinformatics*, vol. 36, no. 11, pp. 3418–3421, 2020.
- [6] L. Seninge, I. Anastopoulos, H. Ding, and J. Stuart, “VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics,’’ *Nature Communications*, vol. 12, no. 1, p. 5684, 2021.
- [7] Y. Zhao, H. Cai, Z. Zhang, J. Tang, and Y. Li, “Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data,’’ *Nature Communications*, vol. 12, no. 1, p. 5261, 2021.
- [8] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle, “MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data,’’ *Genome Biology*, vol. 21, no. 1, p. 111, 2020.
- [9] L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto, “Variational autoencoder,’’ in *Variational methods for machine learning with applications to deep networks*. Springer, 2021, pp. 111–149.
- [10] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,’’ *arXiv preprint arXiv:1312.6114*, 2013.
- [11] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,’’ in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.
- [12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning basic visual concepts with a constrained variational framework,’’ in *International conference on learning representations*, 2017.
- [13] L. Bacciu, P. Cazzaniga, C. Gallese, C. Fuchs, U. Kaymak, D. M. Papetti, and M. S. Nobile, “Our fruitful relationship with sugeno inference from FUMOSO to pyFUME,’’ in *Proceedings of the 20th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2024)*, Lisbon, Portugal, 2024.
- [14] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,’’ *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.
- [15] M. Grazioso, C. Gallese, L. Vanneschi, and M. S. Nobile, “A survey of modern hybrid particle swarm optimization algorithms,’’ in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*. Springer, 2025, pp. 107–128.
- [16] A. Tangherloni, P. Cazzaniga, N. Stranieri, F. M. Buffa, and M. S. Nobile, “A fast feature selection for interpretable modeling based on Fuzzy Inference Systems,’’ in *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (IEEE CIBCB)*. IEEE, 2024, pp. 1–8.
- [17] C. Fuchs, U. Kaymak, and M. S. Nobile, “Building interpretable and parsimonious fuzzy models using a multi-objective approach,’’ in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–8.
- [18] 10x Genomics, “3k PBMCs from a Healthy Donor [dataset],” 10x Genomics, 2017.
- [19] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander et al., “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,’’ *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.