# Analysis of CVAE Implementation

## Coulomb Explosion Atomic Reconstruction

A.Ghanaatian Supervisor: Prof. Caragea

November 11, 2024

# Dataset Overview

- **Input Features** (Position Coordinates):
  - Carbon (cx, cy, cz)
  - Oxygen (ox, oy, oz)
  - Sulfur (sx, sy, sz)
- **Target Features** (Momenta):
  - Carbon (pcx, pcy, pcz)
  - Oxygen (pox, poy, poz)
  - Sulfur (psx, psy, psz)
- **Goal:** Solving inverse problem

# Data Preprocessing

**Data Split**

- Training: 70% (Primary training)
- Validation: 15% (Checking overfitting)
- Test: 15% (Final evaluation)

**Normalization**

- Position: StandardScaler
- It has inverse function
- Momenta: None!

# CVAE Architecture Overview

- **Encoder Network**:
  - Input dimension: 9
  - Hidden layers: [256, 512, 1024]
  - Output: $\mu$ and $\log \sigma$ vectors (1024 each)
- **Latent Space**:
  - Dimension: 1024
  - Reparameterization trick
  - Gaussian prior (N(0,1))
- **Decoder Network**:
  - Input: Concatenated [z, condition]
  - Hidden layers: [512, 256]
  - Output dimension: 9

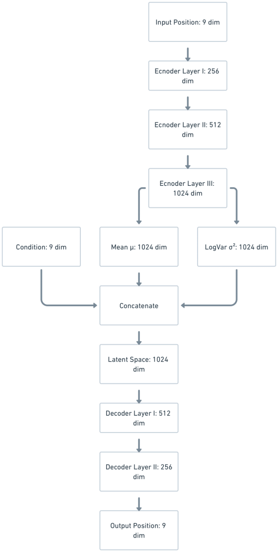
# Detailed Layer Configuration

- **Layer Sizes**:

$$\begin{aligned} \text{Input} &\rightarrow 9 \\ \text{Encoder}_1 &\rightarrow 256 \\ \text{Encoder}_2 &\rightarrow 512 \\ \text{Encoder}_3 &\rightarrow 1024 \\ \text{Latent} &\rightarrow 1024 \\ \text{Decoder}_1 &\rightarrow 512 \\ \text{Decoder}_2 &\rightarrow 256 \\ \text{Output} &\rightarrow 9 \end{aligned}$$

- **Activation**: Sigmoid
- **Parameter Sharing**: None between encoder and decoder

# Model Architecture



# Hyperparameters

**Basic Parameters**

- Epochs: 200
- Batch Size: 256
- Learning Rate: $1 \cdot 10^{-4}$
- Encoders Layer Size: [256, 512, 1024]
- Decoders Layer Size: [512, 256]
- Number of Hidden Layers: 3

**Advanced Parameters**

- Latent Dimension: 1024
- Early Stopping Patience: 100
- Minimum Delta: $1 \cdot 10^{-5}$
- L1 Lambda: *False*
- L2 Lambda: $1 \cdot 10^{-2}$
- Gradient Clip: 5.0

# Loss Function Components

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{L1}}$$

$$\mathcal{L}_{\text{recon}} = \text{MSE}(x_{\text{recon}}, x)$$

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2N} \sum (1 + \log \sigma^2 - \mu^2 - \sigma^2)$$

$$\mathcal{L}_{\text{L1}} = \sum |\theta|$$

where:

- $x_{\text{recon}}$ is the reconstructed output
- $\mu, \sigma$ are latent space parameters
- $\theta$ represents model parameters
- $\lambda = 1.41 \cdot 10^{-4}$ (L1 regularization strength)

# L2 Regularization Deep Dive

**Mathematical Formulation:** $L_{\text{reg}}(\theta) = L(\theta) + \lambda \sum_{i=1}^{n} \theta_i^2$

- **Properties**:
  - Creates circular/spherical constraint region
  - Assumes Gaussian prior on weights
  - Differentiable everywhere
- **Mathematical Characteristics**:
  - Gradient proportional to parameter value
  - Strictly convex optimization
  - Handles multicollinearity effectively
- **Effects**:
  - Shrinks weights proportionally
  - Distributes importance across correlated features
  - Produces unique solutions

# Additional Regularization Techniques

- **Early Stopping**:
  - Monitors validation loss
  - Patience: 100 epochs
  - Minimum improvement: $1 \cdot 10^{-5}$
- **Gradient Clipping**:
  - Maximum norm: 5.0
  - Prevents exploding gradients
  - Stabilizes training
- **Data Normalization**:
  - StandardScaler for Positions
  - No Normalization for Momanta
  - Prevents scaling issues
  - Consider inverting to original scale for Evaluation

# Training Algorithm

**Algorithm** CVAE Training Process

**for** epoch in range(EPOCHS) **do**
  **for** batch in train_loader **do**
    $x$, condition $\leftarrow$ batch
    $\mu, \log \sigma \leftarrow$ encoder($x$)
    $z \leftarrow \mu + \epsilon * \exp(0.5 * \log \sigma)$
    $x_{\text{recon}} \leftarrow$ decoder($z$, condition)
    Compute loss $\mathcal{L}_{\text{total}}$
    Backward pass and update weights
    Clip gradients at 5.0
  **end for**
  Validate and check early stopping
**end for**

# Training Monitoring

- **Loss Tracking**:
  - Training loss per epoch
  - Validation loss per epoch
  - Individual loss components
- **Model Checkpointing**:
  - Save best model state
  - Based on validation loss
  - Restore for testing
- **Early Stopping Logic**:
  - Monitor validation loss
  - Reset counter on improvement
  - Stop when patience exceeded

# Evaluation Metrics

- **Mean Relative Error (MRE)**:

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_{\text{pred},i} - y_{\text{true},i}|}{|y_{\text{true},i}| + \epsilon}$$

- **Latent Space Statistics**:
  - Best Test MRE: 48%

# Other Metrics

| Dataset | MRE | MSE |
|---|---|---|
| Training | 0.553 | 5.128 |
| Validation | 0.489 | 5.152 |
| Test | 0.486 | 5.263 |

Table: Model Performance Metrics

MRE: Mean Relative Error, MSE: Mean Squared Error

# Learning Curve Plot