# Summary of Personal Computational Project

Hu Hongrui

Nanyang Technological University

*Matriculation No. G2508031J*

*Biomedical Data Science*

July 18, 2025

# Contents

# 【I】 Kaggle RNA 3D Folding Challenge: A Computational Approach to Structural Prediction

- **The Challenge: RNA's 3D Puzzle**
  - RNA plays a vital role in biological processes (e.g., gene expression, catalytic reactions).
  - Core challenge: Predicting RNA's 3D structure accurately and efficiently remains a key hurdle in computational biology.
- **Our Goal:** To precisely predict the three-dimensional coordinates of RNA sequences using advanced deep learning models.
- **My Role:**
  - Computational pipeline setup and data preprocessing.
  - Integration and optimization of multi-model prediction results.

# 【I】 Kaggle RNA 3D Folding Challenge: A Computational Approach to Structural Prediction

## Our Core Strategy

- Recognizing the limitations of a single model, we adopted a dual-model parallel prediction strategy, integrating two leading Diffusion Models: Boltz-1 and Protenix.
- Briefly on Diffusion Models: Imagine starting with a random ”point cloud” (disordered structure); the model refines it step-by-step through a multi-stage ”denoising” process into a clear, meaningful 3D structure.

# 【I】 Kaggle RNA 3D Folding Challenge: A Computational Approach to Structural Prediction

## Data Flow & My Contributions

- Input Data Standardization: I was responsible for converting raw RNA sequence text files into the specific structured YAML format required by the models. This ensured data could be accurately read and processed by deep learning models, serving as the foundation for model inference.
- Merging & Optimization: I developed post-processing scripts to precisely extract the 3D coordinates of the C1' atom for each nucleotide from the raw outputs of both models.
- An intelligent merging strategy was implemented: the effective prediction results of Boltz-1 were prioritized, while Protenix's predictions were utilized to fill in any potential prediction gaps of Boltz-1, significantly enhancing the robustness and completeness of the final outcome.

# YAML Configuration Generation Script

## Python script for generating YAML configs

```python
import pandas as pd
import os

# Ensure the output directory exists
os.makedirs('/kaggle/working/inputs_prediction', exist_ok=True)

# Loop through target IDs and sequences from your test data
# 'names' and 'sequences' would typically come from loading '
    test_sequences.csv'
for tmp_id, tmp_sequence in zip(names, sequences):
    with open(f'/kaggle/working/inputs_prediction/{tmp_id}.yaml
    ', 'w') as f:
        f.write("constraints: []\n")
        f.write("sequences:\n")
        f.write("- rna:\n")
        f.write("    id:\n")
        f.write("    - A1\n")
        f.write(f"    sequence: {tmp_sequence}")
```

# 【II】Research on Three-Factor Return Modeling and Hedging Strategies of Crude Oil Futures

## Data Engineering Innovation

Dynamic Splicing Technology for Futures Main Contracts (High-Frequency Data Processing)
Convenient Proxy Variable Design for Returns：

$$\widetilde{CY}_t = \frac{F_t^{\text{close}} - F_t^{\text{settle}}}{F_t^{\text{settle}}} - r_t$$

Dynamic Modeling of Volatility GARCH(1,1) ：

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

Calibration Process through Maximum Likelihood Estimation (MLE)
Technical Highlights: Capturing Volatility Clustering Effects and Asymmetric Market Effects

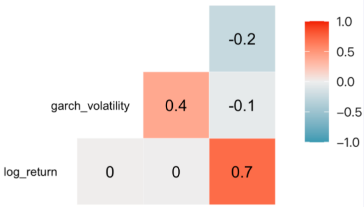
# 【II】Research on Three-Factor Return Modeling and Hedging Strategies of Crude Oil Futures

## Three-factor model calculation implementation

Multi-factor regression architecture — OLS model：

$$R_t = \alpha + \beta_1 CY_t + \beta_2 \sigma_t + \beta_3 r_t + \epsilon_t$$

Robustness analysis: There was no significant multicollinearity. The Bruchy-pagan test showed that the model residuals did not have significant heteroscedasticity. And residual autocorrelation diagnosis.



# 【II】Research on Three-Factor Return Modeling and Hedging Strategies of Crude Oil Futures

## Dynamic Hedging Algorithm and Backtesting

- Dynamic Hedging Algorithm: alculation of Hedging Ratio:

$$h_t^* = \frac{\mathrm{Cov}(R_t^s, \widehat{R_t^f})}{\mathrm{Var}(\widehat{R_t^f})}$$

  Real-time Prediction Framework: Daily update of three factors $\rightarrow$ Prediction of return rate $\rightarrow$ Dynamic portfolio adjustment
- Backtesting System Design:
  Performance Indicators: Hedging Effectiveness (HE) = 46.92%, Volatility Compression from 1.94% to 1.42%, Sharpe Ratio Optimization
- Technology Stack: R Language

# 【III】Analysis of the Impact of Meteorological Factors on Urban Air Quality

## Research Framework & Technical Approach

- Scientific Question
  - Quantify nonlinear impact of meteorological factors on pollutants
  - Reveal synergistic mechanisms of temperature-precipitation-wind
- Technical Pathway: Data Integration → Feature Engineering → Random Forest Modeling → Policy Optimization
- Technology Stack: R Language (randomForest(ntree=100), ggplot2, car)

# 【III】 Analysis of the Impact of Meteorological Factors on Urban Air Quality

## Model Architecture & Key Findings

- Model Architecture
  - Six independent RF regressions

$$\text{Air Quality} = RF(X_1, X_2, \cdots, X_k) + \epsilon$$

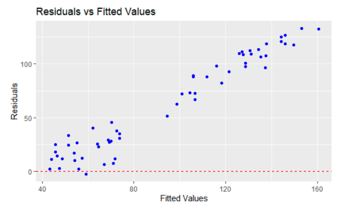
Among them, Air Quality represents the six pollutants: PM2.5, PM10, SO2, NO2, CO, and O3.

- High temp $\rightarrow$ photochemical O3 generation
- Low temp $\rightarrow$ PM accumulation
- Limited rain cleansing effect
- Strong temperature correlation

# 【III】Analysis of the Impact of Meteorological Factors on Urban Air Quality

## Statistical Validation

- Residuals reject normality ($p = 0.000138$) $\rightarrow$ Uncaptured key factors




- VIF confirms temperature collinearity

## Policy Recommendations

- High-temp O  early warning system
- Low-temp mobile source control
- Meteorology-pollution coordinated response

# 【IV】Research on Risk Assessment Models for Small Enterprises

## Research Framework & Technical Approach

- Scientific Objective
  - Build sustainable catastrophe insurance: Balance insurer profitability and policyholder affordability
  - Minimize coverage gaps: Reduce uninsured risk exposure
- Technical Pathway: Multi-source Data Integration $\rightarrow$ Risk Prediction $\rightarrow$ Insurance Pricing $\rightarrow$ Payment Capacity Assessment $\rightarrow$ Profit and Loss Decision

# 【IV】Research on Risk Assessment Models for Small Enterprises

## Core ML Tech

- ARIMA Time Series Forecasting: Dynamic Modeling of Disaster Frequency (Optimization of p, d, q Parameters)
- Random Forest Regression: Prediction of Disaster Loss Magnitude (Feature Importance Analysis)
- Fuzzy Comprehensive Evaluation (FCE): Multi-dimensional Quantification of Engineering Risks

## Technology Stack:

- R language, Python, SPSS
- forecast(ARIMA), randomForest, AHP

## Data Analysis and Visualization

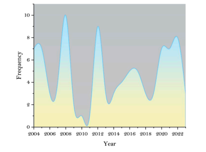


**Figure4:Florida Hurricane Frequency**

(a)

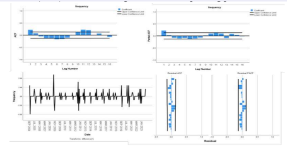


**Figure5:ACF&PACF&First order difference**

(b)

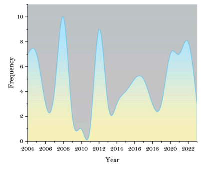


**Figure6:Shanghai typhoon frequency**

(c)

After multiple rounds of fitting, the optimal ARIMA model was obtained as ARIMA (0,0,1). Based on the variable Frequency, the parameters of the model indicate its significance.

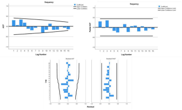


**Figure7:ACF&PACF**

The formula for obtaining the model is:

$$y_{(t)} = 2.471 + 0.7 * \varepsilon_{(t-1)}$$

(d)

# 【IV】 Research on Risk Assessment Models for Small Enterprises

## Core Models & Empirical Findings

- Actuarial Pricing Model:

$$C = X + Y = np \times \left(\frac{1}{1+i}\right)^{t} \times (1 + \alpha + \beta + \gamma)$$

- ML Prediction Engine:ARIMA(1,0,0)

$$\widehat{y}_t = \mu + \emptyset_1 y_{t-1} + \ldots + \emptyset_p y_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q}$$

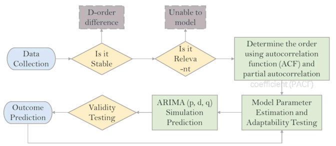



**Figure3:The operational logic of ARIMA**

## Loss degree model

- Multiple linear regression model

$$y = 5.684 \times 10^{-16} + 0.3827x_1 + 0.5682x_2 + 0.3828x_3 + 0.6852x_4$$

- K-means Model

**Table2: Clustering centers**

| Cluster types | DAMAGE PROPERTY | INJURIES INDIRECT | DEATHS INDIRECT | INJURIES DIRECT | Count | DEATHS DIRECT |
|---|---|---|---|---|---|---|
| 1 | 142118455.690 | 1.457 | 0.199 | 78.371 | 94.80 | 5.057 |
| 2 | 2687065009.366 | 2.333 | 0.999 | 1090.666 | 225 | 96 |
| 3 | 5111544738.749 | 12.333 | 5 | 1984.666 | 392 | 194.333 |
| 4 | 1302903562.679 | 18.125 | 1.75 | 521.25 | 256.875 | 41.75 |

- Random Forest Loss Prediction:

**Table3:Model evaluation effect**

| | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| Training set | 0.047 | 0.216 | 0.113 | 5.783 | 0.939 |
| Test set | 0.431 | 0.657 | 0.281 | 13.311 | -1.205 |

# 【IV】 Research on Risk Assessment Models for Small Enterprises

## From Insurance to Heritage Preservation

- Fuzzy Comprehensive Evaluation Method(FCE) and Analytic Hierarchy Process(AHP)

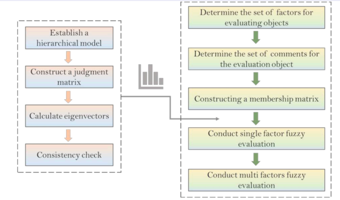



**Figure8:FAHP**

# 【IV】Research on Risk Assessment Models for Small Enterprises

- Construct a judgment matrix to implement AHP analysis

$$A = \begin{bmatrix} \frac{W_1}{W_1} & \cdots & \frac{W_1}{W_n} \\ \vdots & \ddots & \vdots \\ \frac{W_n}{W_1} & \cdots & \frac{W_n}{W_n} \end{bmatrix}$$

- calculate the eigenvectors

$$\begin{cases} M_i = \prod_{j=1}^{n} a_{ij} \quad (i = 1, 2, \cdots, n) \\ \overline{W_i} = \sqrt[n]{W_i} \quad (i = 1, 2, \cdots, n) \\ W_i = \overline{W_i} / \sum_{i=1}^{n} \overline{W_i} \quad (i = 1, 2, \cdots, n) \\ W = (W_1, W_2, \cdots, W_n)^T \end{cases}$$

# 【IV】 Research on Risk Assessment Models for Small Enterprises

**Table 7:index weight**

| Primary indicators | Weight | Secondary indicators | Weight |
|---|---|---|---|
| Natural factors | 0.405 | Extreme weather frequency | 0.276 |
| | | Extreme weather damage level | 0.431 |
| | | Geographic location | 0.292 |
| Social factors | 0.384 | Per Capita Disposable Income | 0.380 |
| | | Urban supporting facilities | 0.251 |
| | | Market supply and demand relationship | 0.257 |
| | | Macroeconomic policies | 0.112 |
| Engineering factors | 0.211 | Construction cost | 0.594 |
| | | Project quality | 0.406 |

Thanks!