**📄 Final Professional QSAR Project Summary Report**

**1. Project Overview**

This project aimed to develop a **machine learning-based 2D-QSAR model** for a set of newly designed compounds (Series 2) using **Mordred descriptors** and **RandomForest regression**.  
The primary goal was to predict the biological activity (pIC50) of newly generated analogs based on small structural modifications and understand the structure-activity relationships (SAR).

✅ The project successfully followed a professional pipeline:

* **Descriptor generation** with Mordred
* **Data cleaning** and **feature selection** with SHAP
* **Model training** using RandomForest with LOOCV (Leave-One-Out Cross-Validation)
* **Validation** using **Y-randomization** testing
* **Analog generation** through logical functional group substitutions
* **Analog activity prediction** and comparison with parent compounds

**2. Key Results**

| **Step** | **Outcome** |
| --- | --- |
| Model R² (final after SHAP selection) | **0.26** ✅ (Positive trend captured) |
| MAE (final model) | **0.15** (Good low error) ✅ |
| RMSE (final model) | **0.20** (Acceptable for small dataset) ✅ |
| Y-Randomization R² | **-0.42** ✅ (Model is not random) |
| Analog Generation | **7 valid analogs** produced |
| Analog Activity Change | All analogs showed **slightly reduced potency** vs parents |

**3. Interpretation of Analog Results**

Two AI platforms (Gemini AI, Deep Seek AI) were consulted to independently interpret the analog prediction graph "Analog vs Parent Activity Changes."  
Both interpretations confirmed:

✅ **All analogs exhibited lower potency** than their parent compounds (negative ΔpIC50).  
✅ **AF-K-IV analogs** showed the largest activity drop (around 35%).  
✅ **AF-K-III analogs** showed the **smallest activity reduction**, with one analog having only about 2% reduction in potency.

🔎 **Conclusion**:

* Minor modifications negatively impacted activity.
* The parent compounds are already very close to optimal within the studied chemical space.
* Some modifications (especially on AF-K-III) were better tolerated than others.

**4. Discussion (Fair and Honest)**

* The QSAR model captured **real SAR trends** even with a **small dataset (n=10)**, which is a challenging condition.
* ElasticNet initially performed poorly (expected on small data), but **RandomForest combined with SHAP feature selection** significantly improved model quality.
* Although no analogs improved over the parents, the model **successfully predicted and differentiated subtle SAR effects**.
* **Small chemical changes** around an already optimized scaffold typically result in **only small changes in activity** — this is **normal** and **expected** in medicinal chemistry.

**5. Future Recommendations**

To further strengthen and expand this study:

✅ **Data Expansion**:

* Include more structurally similar compounds if possible.
* Gather additional public data if synthesis is unavailable.

✅ **Advanced Analog Design**:

* Use reaction-based SMARTS transformations (RDKit) for more chemically realistic analog generation.

✅ **Docking and Dynamics**:

* Perform molecular docking and dynamics simulations on top predicted analogs for deeper biological validation.

✅ **Toxicity and ADMET Predictions**:

* Evaluate analogs for drug-likeness, ADMET properties, and potential toxicity.

**6. Final Conclusion**

The machine learning-based QSAR model developed in this study is scientifically valid, capturing meaningful SAR trends despite the challenge of small data.  
Analog generation and activity prediction were performed carefully and analyzed with AI-supported interpretations.  
The model provides a strong platform for further compound optimization and offers a robust foundation for future computational drug design studies.