# HydroCarbon: A Physics-Informed Machine Learning Model

# for Environmental Footprint Prediction in Fashion Products


## ABSTRACT

This report presents HydroCarbon, a novel physics-informed machine learning model designed to predict carbon and water footprints for fashion products. The model addresses a critical gap in sustainability research: the absence of large-scale, publicly available life cycle assessment (LCA) datasets for fashion products. By combining synthetic data generation using Large Language Models (LLMs) with physics-based footprint calculations and robust XGBoost regression, HydroCarbon achieves state-of-the-art performance with $R^2 > 0.999$ on complete data and maintains $R^2 > 0.93$ even when 40%% of input features are missing.


## Core Mathematical Formulas

**Carbon Material Footprint:**

$$C_{material} = \Sigma(weight \times material\_\% \times carbon\_factor)$$

**Carbon Transport Footprint:**

$$C_{transport} = (weight/1000) \times distance \times (weighted\_EF/1000)$$

**Total Carbon Footprint:**

$$C_{total} = C_{material} + C_{transport}$$

**Water Footprint:**

$$W_{total} = \Sigma(weight \times material\_\% \times water\_factor)$$

## Model Performance

| Target | R² (Complete) | MAE | R² (40% Missing) | MAE |
|---|---|---|---|---|
| Carbon Material | 0.9999 | 0.041 kgCO■e | 0.936 | 0.29 kgCO■e |
| Carbon Transport | 0.9998 | 0.001 kgCO■e | 0.968 | 0.001 kgCO■e |
| Carbon Total | 0.9999 | 0.044 kgCO■e | 0.936 | 0.29 kgCO■e |
| Water Total | 0.9998 | 115.3 L | 0.902 | 772 L |

## Model Architecture

**Input Layer (129 features):**
• Contextual Features (93): Gender, Category, Parent Category
• Physics Features (36): Weight, Distance, Material percentages

**Feature Engineering:**
• Formula features injected from physics calculations
• One-hot encoding for categorical variables
• Log transformation for numerical stability

**Model Core:**
• XGBoost Multi-Output Regressor
• 1000 estimators, max depth 8
• GPU acceleration (CUDA)
• Custom physics-constrained objective

**Outputs (4 predictions):**
• carbon_material (kgCO■e)
• carbon_transport (kgCO■e)
• carbon_total (kgCO■e)
• water_total (liters)

## Key Innovations

1. **Synthetic Data Generation:** 900,000+ products generated using Google Gemini 2.5 Flash

2. **Physics-Informed ML:** Hybrid architecture combining formulas with XGBoost learning

3. **Robustness Training:** Feature dropout augmentation handles 40% missing data

4. **Performance:** $R^2 > 0.999$ on complete data, $R^2 > 0.93$ with missing data

## Usage Example

```
from hydrocarbon import FootprintPredictor predictor =
FootprintPredictor("trained_model/robustness") results =
predictor.predict( gender="Male", category="Jeans",
weight_kg=0.934, materials={"cotton_conventional": 0.92,
"elastane": 0.08}, total_distance_km=12847 ) # Output: Carbon:
2.26 kgCO■e, Water: 7,888 liters
```

## Data Sources

• **Material Factors:** TU Delft Idemat 2026 database (34 materials)

• **Transport Factors:** CE Delft STREAM 2020 with multinomial logit modal split

• **Water Factors:** Water Footprint Network studies

• **Synthetic Products:** Google Gemini 2.5 Flash generation

Generated: 2026-01-22 13:25:13

Version: 2.0 | Status: Proof of Concept

Repository: https://github.com/Avelero/Avelero_HydroCarbon