

AI-Driven Flowsheet Optimization via Multiphysics Surrogate Models

K. Pikounis¹, K. Loupos¹, A. Karydis¹

1. R&D Department, INLECOM INNOVATION, Athens, Greece.

Abstract

The recovery of Platinum Group Metals (PGMs) and other strategic critical metals from industrial waste catalysts is essential for a sustainable circular economy. While multiphysics simulations accurately model advanced electrochemical recovery technologies, like Gas-Diffusion Electro-crystallization (GDEx) and Electroleaching (ELX), they are computationally intensive and unsuitable for dynamic, system-wide optimization. To address this computational bottleneck, we present an AI-driven digital tool powered by surrogate models trained on high-fidelity COMSOL Multiphysics® datasets. For data-rich processes, Multi-Layer Perceptron (MLP) neural networks were trained, while processes with less data available were modeled using physics-informed analytical functions. In total, seven surrogate models were developed. The trained MLPs demonstrated exceptional predictive accuracy, achieving a coefficient of determination (R^2) exceeding 0.9999 and a Mean Absolute Percentage Error (MAPE) of approximately 0.56%. By transforming heavy multiphysics data into a lightweight framework, the tool successfully optimizes multi-model recycling flowsheets in a fraction of the time (ranging from seconds to a few hours) with minimal RAM utilization. This approach provides a highly efficient, trustworthy decision-support system for iterative scenario testing and real-time process optimization.

Keywords: Artificial Intelligence (AI), Circular Economy, Flowsheet Optimization, Machine Learning (ML), Multiphysics, Platinum Group Metals, Surrogate Modeling.

Introduction

The chemical industry heavily relies on strategic metal catalysts to drive efficient and scalable chemical processes. Recovering these valuable metals from industrial waste, spent, or off-specification catalysts is of paramount importance for establishing a sustainable circular economy. Despite the economic and environmental imperatives, recovering these critical and strategic metals efficiently and cleanly continues to pose significant technical challenges.

The FIREFLY project [1] addresses these challenges directly by developing sustainable, electro-driven recycling flowsheets. The project focuses on recovering metal catalysts from industrial waste utilizing a combination of integrated advanced electro-driven technologies.

To maximize recovery efficiencies and minimize energy consumption, the interconnected operational variables of these electrochemical processes must be rigorously evaluated and optimized. While multiphysics simulations accurately capture the complex physical, electrochemical, and fluid-dynamic behavior of these systems, they are computationally heavy and highly time-intensive. Consequently, executing full multiphysics simulations is unsuitable for dynamic, system-wide optimization, which often requires evaluating thousands of operational scenarios.

This work focuses on bridging that computational gap by training Machine Learning (ML) and

analytical surrogate models on massive datasets generated via COMSOL Multiphysics® simulations. By integrating these surrogate models into a comprehensive web-based application utilizing a Neo4j Knowledge Graph, this digital tool can instantaneously predict the recovery efficiencies and energy consumption of each individual process. Ultimately, this framework allows for the rapid optimization of multi-step recycling flowsheets against user-defined criteria and additional constraints, replacing computationally heavy simulations with a lightweight, high-fidelity decision-support tool.

Data Generation and Preprocessing

The development of robust machine learning surrogate models requires extensive and well-structured training datasets. This section outlines the methodology utilized to generate and extract the necessary high-fidelity data for the targeted electrochemical processes, bridging the gap between computationally heavy multiphysics simulations and the subsequent AI-driven optimization pipeline.

Sourcing Multiphysics Simulation Data and Dataset Extraction

While the experimental work on the FIREFLY electro-driven technologies yielded the crucial baseline data required to build high-fidelity simulation models, the physical data volume was insufficient for directly training Machine Learning (ML) architectures. Therefore, the optimal approach to generate massive, high-quality training datasets

was to leverage these detailed simulations. These high-fidelity datasets were generated using mainly COMSOL Multiphysics® by consortium partners within the FIREFLY project. To capture the physical, electrochemical, and fluid-dynamic behavior of these systems, specific physics interfaces were employed. The scale of data extraction varied significantly across the three core processes modeled using detailed simulations:

- Gas-Diffusion Electro-crystallization (GDEx) [2]: This process was modeled as a gas-liquid system utilizing a laminar two-phase flow physics interface coupled with species transport [3]. For GDEx on PGMs, the simulations yielded massive datasets comprising over 1.9 million simulated data points across nearly 3,000 distinct experimental conditions. This extensive mapping provided the necessary granularity for ML applications.
- Electroleaching (ELX) [4]: This technology is designed for the recovery of metals from spent catalysts, involving an initial step where the solvleaching of the targeted metals takes place before they are recovered. The ELX simulations utilized an average of 250 experiments across 5 different experimental setups, featuring different combinations of Deep Eutectic Solvents (DES) and catalysts.
- Electrochemical Transformation in Molten Salts (ETMS) [5]: This process operates in batch mode, utilizing a unique electrolyte composed of a molten salt mixture to extract metals from their respective oxides at lower temperatures [6]. It is important to note that the ETMS process was modeled by a dedicated simulation tool other than COMSOL, generating a robust dataset of over 725 simulated experiments and approximately 133,000 data points.

Variable Mapping and Preprocessing

The data provided for each electrochemical process were systematically collected, homogenized, pre-processed, and then combined into a single dataset, which was used to construct the respective surrogate model. For GDEx and ETMS processes, where the datasets were data-rich, ML models were trained, while for the case of the five ELX processes, physics-informed analytical functions were fitted. During the initial data preparation phase, key input and target variables were explicitly defined. These variables include parameters directly related to the electrochemical processes (such as applied electric current or cell voltage at each time point), parameters capturing the concentration of specific elements in the output stream, and the overall energy consumption of the process. The inputs generally fall into two categories: Input Flow Parameters, which relate to the initial state of the catalyst or feed, and

Process Parameters, which dictate the operational conditions.

Table 1: Summary of input parameters, input flow parameters, and target variables for the modeled electrochemical processes.

Process	Process parameters	Input flow parameters	Target variables
GDEx	Current Density, Gas Flow Rate, Time	Initial Metal Concentrations: Pt, Pd, Rh	Metal Recovery Efficiencies, Energy Consumption
ETMS	Cell Voltage, Porosity, Time	Ti Oxide Feed	Metal Product Ti, Energy Consumption, Current, Current Efficiency, Cumulative Current Efficiency
ELX-1	Current Density, pH, Time	Input Metal Mass: Mo, V	Metal Recovery Efficiencies, Energy Consumption
ELX-2		Input Metal Mass: Mo, V, W	
ELX-3		Input Metal Mass: Co, Cu, Mo, Ni, V	
ELX-4		Input Metal Mass: Ag, Ir, Pd, Pt, Rh, Ru	
ELX-5		Input Metal Mass: Ag, Ir, Pd, Pt, Rh, Ru	

A critical step in the preprocessing pipeline involved managing the temporal resolution of the raw data. The entries in each simulation output file initially reflected the original sampling rate of the simulation. For example, data points for the ETMS process were generated every 25 seconds. However, this level of detail was too fine-grained for training ML models intended to predict the process over a total duration of up to 30 hours. Moreover, such high temporal resolution is not particularly meaningful in real-world applications, where the overall process spans hours. Therefore, the data were systematically resampled at defined intervals (e.g., at 10-minute intervals) to reduce the dataset size and align with practical use cases. This homogenization ensured that the resulting datasets were computationally viable for the subsequent machine learning training phases without sacrificing physical accuracy.

Surrogate Modeling

To replace the computationally expensive multiphysics simulations with lightweight, real-time predictive components, tailored surrogate modeling strategies were developed. The approach was defined based on data availability: Multi-Layer Perceptron (MLP) architectures were deployed for data-rich processes, while physics-informed analytical fitting was utilized for processes with less available data.

GDEx

For the data-rich GDEx process, a Multi-Layer Perceptron (MLP) neural network was designed to map the operational inputs to the recovery efficiencies of the target metals. To construct a robust model, the dataset was divided into training, validation, and test sets using a 70-15-15 ratio. This split was executed by entire experimental runs (complete time-series) rather than randomized individual data points. Forcing the model to predict the full temporal sequence based solely on the initial flow and process parameters prevents data leakage and ensures that the test and validation samples remain entirely uncontaminated. Even though this is a time-series prediction problem, Recurrent Neural Networks, like Long Short-Term Memory and Gated Recurrent Unit, yielded suboptimal results compared to simple MLPs with time included as an additional input parameter; thus this architecture was selected. The optimal regressor was determined to be an MLP with three hidden layers consisting of 50, 100, and 50 nodes, respectively. The output layer included three nodes to simultaneously predict the recovery efficiencies of Pt, Pd, and Rh. Formulating a unified network for simultaneous prediction allows the model to identify and exploit underlying correlations among the target variables, utilizing the data from one element to strengthen the predictive accuracy of the others.

The GDEx surrogate model demonstrated exceptional predictive capabilities. As shown in Figure 1, the convergence behavior throughout the training phase indicated stable learning, with both training and validation losses decreasing smoothly without divergence, confirming the absence of overfitting. To prevent the network from memorizing stochastic noise in the later stages of training, optimal model weights were extracted from the onset of convergence. As the validation MAPE stabilized into a plateau around epoch 650, the model state at epoch 660 (the local minimum in the plateau) was selected. Although training continued to 1,200 epochs to observe long-term behavior, subsequent improvements were negligible, confirming that the model had successfully trained by epoch 660. The statistical performance across the testing subsets further validated the model's excellent predictive

power. As detailed in Table 2, the model consistently achieved an R^2 exceeding 0.9999 overall, proving its readiness to replace the rigorous COMSOL solvers in system-wide optimizations.

To verify physical intuition alongside mathematical metrics, the model's temporal predictions were plotted against the ground-truth COMSOL data. Figure 2 illustrates the "worst-case" experiment from the test set for the combined Pt and Pd subset. Even in this "worst-case" scenario, the prediction remains excellent.

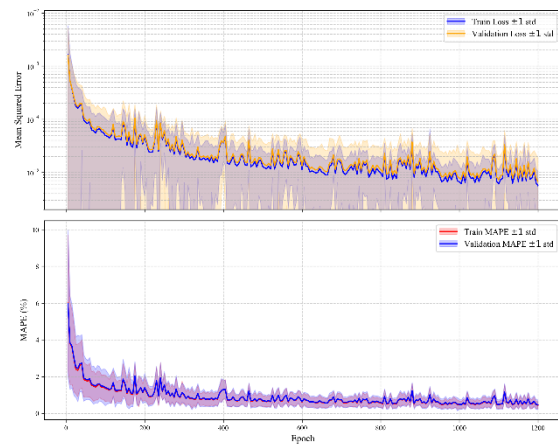


Figure 1. Convergence and performance tracking for the optimal GDEx MLP architecture, showing the Mean Squared Error loss on a logarithmic scale (top) and the Mean Absolute Percentage Error (bottom) over the training epochs.

Table 2. Statistical performance of the optimal MLP for predicting simultaneous Pt, Pd, and Rh recovery efficiencies. Values represent the mean $\pm$ standard deviation across the experiments within each dataset.

Metric	Training Set	Validation Set	Test Set
R^2	0.99995 ± 0.00005	0.99994 ± 0.00007	0.99994 ± 0.00009
MAPE	0.55 ± 0.25	0.56 ± 0.30	0.56 ± 0.30

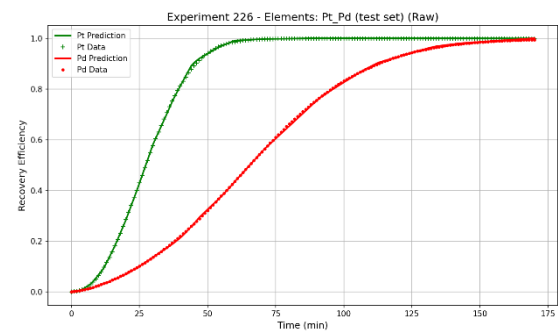


Figure 2. Comparison of the simulated data and the MLP surrogate model predictions for the simultaneous temporal evolution of Pt and Pd recovery efficiencies (Experiment 226 from the test set).

ELX

Since the ELX process exhibited less complex underlying dynamics, fewer experimental scenarios were simulated. Given this limited dataset size (averaging 250 experiments per setup, where each full time-series yields a single kinetic data point), applying ML risked severe overfitting. Instead, a two-step, physics-informed analytical fitting approach was utilized.

First, the temporal dynamics of the metal recovery efficiency, $E(t)$, were modeled. The time-series data exhibited a classic saturation kinetics behavior, which was fitted to an exponential rise function:

$$E(t) = 1 - e^{-a(pH,I)t}$$

where a represents the characteristic rate parameter governing the leaching kinetics.

In the second step, the operational dependency of the process was modeled by expressing the extracted rate parameter a as a direct function of the primary input variables: the acidity (pH) and the applied current (I). The phenomenological relationship was empirically defined as:

$$a(pH, I) = \frac{C}{10^{pH}} (1 + b \cdot 10^{pH} \cdot I)$$

where C and b are metal-specific characteristic constants. Non-linear least squares optimization (via the Levenberg-Marquardt algorithm) was employed to fit these functions to the experimental space and derive the precise values of C and b for each element. This analytical methodology was systematically applied across the five diverse ELX experimental setups. The robust nature of this approach is illustrated by the Pd recovery model for the Pd-on-carbon spent catalyst using DES3.

To evaluate the predictive capability of the derived physics-informed equation, a Leave-One-Out Cross-Validation (LOOCV) strategy was implemented. For a dataset of N experiments, the characteristic constants (C and b) were fitted using $N - 1$ samples. The resulting equation was then used to predict the kinetic rate parameter a for the single omitted condition, iterating this process until every data point had been predicted as an unseen test case.

As shown in Figure 3, even under strict cross-validation, the predicted a values aligned exceptionally well with the actual rate parameters extracted from the raw data, achieving a highly accurate linear fit ($R^2 = 0.9992$).

Ultimately, substituting these analytically derived $a(pH, I)$ parameters back into the exponential rise function yielded a comprehensive and robust surrogate model. As visually confirmed by Figure 4, this final analytical model accurately predicts the full temporal evolution of the Pd recovery efficiency across a diverse matrix of operational conditions. While ML architectures excel in data-rich environments, these results demonstrate that this

two-step, physics-informed analytical framework is the optimal modeling strategy for data-sparse processes. By explicitly anchoring the mathematical equations to the underlying saturation kinetics, the approach prevents overfitting and guarantees robust generalization. This surrogate model is computationally light and at the same time it retains the accuracy of the underlying data, so it can be deployed in practical applications to bypass the need for running new, computationally intensive multiphysics simulations whenever high-speed, high-fidelity predictions are required.

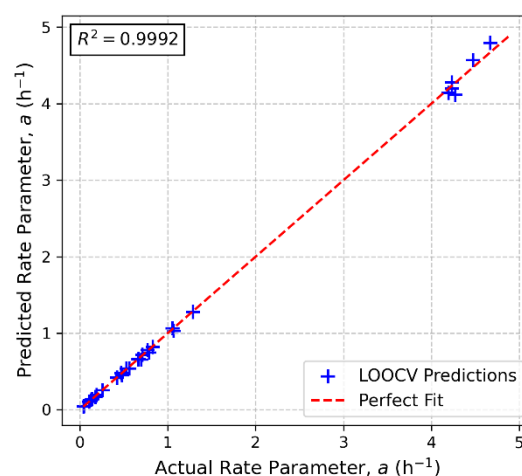


Figure 3. Leave-One-Out Cross-Validation (LOOCV) comparing the predicted and actual kinetic rate parameters (a) for Pd recovery using DES3.

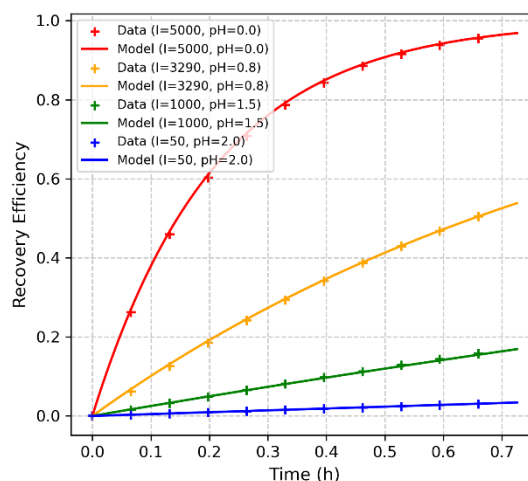


Figure 4. Analytical surrogate model predictions versus simulation data for the temporal evolution of Pd recovery efficiency across representative experiments (applied currents and pH levels).

Flowsheet Optimization

The flowsheet optimization framework was built upon the excellent predictive power of the surrogate

models. This framework is designed to identify the optimal operational parameters based on three primary objective functions:

- minimum energy consumption (evaluated for the flowsheet as a whole, rather than for individual processes),
- maximum efficiency (focusing on the total recovered mass of a specific target element or the combined mass of multiple elements) and
- maximum selectivity (defining the purity of the yield as the mass fraction of the target metal or metals relative to the total recovered mass in a multi-element feedstock).

To provide greater utility and operational flexibility, the framework allows for the application of constraints on a criterion not selected as the optimization target. For instance, a user can ask to minimize the energy consumption while imposing a constraint on the recovery efficiency of specific elements to remain above 90%. Notably, when energy minimization is selected as the primary objective, defining an additional constraint is strictly required; otherwise, the solver will trivially converge to an operation time of $t = 0$, yielding zero energy consumption but failing to recover any material.

The surrogate models are considered valid and trustworthy when each variable assumes values within the range used in its training. To ensure reliable results, the optimization algorithms search exclusively within the valid phase space of the variables of the surrogate models (e.g., the range of the pH variable in the dataset used to train the ELX-2 surrogate model was $[0, 2]$; consequently, the minimization algorithm strictly confines the pH within this same range). However, the framework permits users to restrict these boundaries to simulate specific operational scenarios. Continuing the example above, a user can constrain the search space of the pH variable in the $[1, 2]$ range to evaluate maximum efficiency under less acidic conditions.

In total, four distinct flowsheet configurations were constructed utilizing a subset of the seven validated surrogate models, comprising both single-process and coupled dual-process pipelines. For standalone processes with a three-variable phase space, a brute-force grid search algorithm was implemented. Given the low dimensionality, this approach exhaustively evaluates all discrete points across a finely resolved grid, guaranteeing the discovery of the absolute global optimum.

Conversely, for coupled flowsheets, the combined phase space expands to six variables. In a six-dimensional space, creating a sufficiently fine grid results in exponentially high execution times and

unreasonable RAM utilization. To overcome this computational bottleneck, a multi-start gradient descent approach was employed. To mitigate the intrinsic randomness of gradient descent and the risk of convergence upon local minima, the minimization is executed eight independent times, each initiating from a randomly selected point within the phase space. Furthermore, the algorithm incorporates a trajectory-tracking mechanism: if a newly generated random starting point falls within close proximity to a previously explored gradient path, it is discarded, and a new origin is drawn to ensure a diverse exploration of the phase space. Upon completion, the algorithm compares the converged values of all successful runs and outputs the absolute minimum.

The architecture and computational efficiency of this framework are best demonstrated through a coupled ELX-GDEx flowsheet case study. In this configuration, the sequential pipeline ingests the user inputs (initial metal masses, optimization criteria, and constraints) and processes them through the physics-informed analytical functions of the ELX stage and the MLP neural network of the GDEx stage to deliver the optimized outputs. The gradient descent approach maintains exceptionally modest memory requirements across all runs, with peak RAM utilization remaining strictly under 700 MB.

To assess the stability of the multi-start algorithm against intrinsic randomness, 318 distinct optimization scenarios were executed 10 times each. The resulting execution time distribution shown in Figure 5 reveals a distinct three-peaked pattern that directly correlates with the stringency of the user-defined constraints. Scenarios with absent or very loose constraints converge rapidly, forming the dominant peak at the lower end of the time scale. When constraints are moderately tight, the algorithm requires deeper exploration of the phase space, clustering around an average runtime of approximately 15 minutes. Finally, optimizations under highly restrictive constraints demand the most extensive gradient tracking and rejection sampling, forming the third peak with an absolute maximum runtime of up to 2 hours. This demonstrates that the solver's computational demand scales logically with problem complexity while maintaining absolute stability.

Ultimately, this framework provides a highly efficient and scalable tool for process engineers to rapidly explore multi-variable parameter spaces, enabling dynamic, system-level decision making and process optimization.

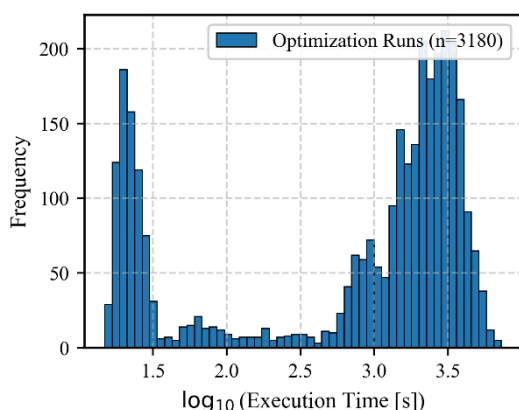


Figure 5. Execution time distribution for 3,180 optimization iterations (318 distinct scenarios executed 10 times each) using the multi-start gradient descent algorithm.

Digital App Architecture and UI

The predictive tool is a web-based application [7] created to provide a user-friendly interface for exploring, configuring, and executing optimization workflows. It considers domain-specific knowledge in electrochemistry, including catalysts, elements, technologies, and related parameters. This layered architecture ensures modularity, scalability, and extensibility. It is built around four core components: the User Interface (UI), the Backend Orchestrator (BE), the Knowledge Graph (KG), and the Keycloak authentication service.

The Knowledge Graph, built on Neo4j, serves as the primary database for the application. It stores structured definitions of catalysts, elements, technologies, input parameters, ML models, and output data, along with their relationships. A graph database was selected for this application because it is specifically optimized for managing highly connected and relationship-centric data. This makes it well-suited for translating intermediate outputs between sequentially linked ML models in a multi-step flowsheet.

The Backend Orchestrator, implemented using FastAPI, acts as the central execution and communication hub. It manages the flow of data between the UI, the Knowledge Graph, and the machine learning layer. Given that complex flowsheet optimizations may take time to complete, the Backend system handles these triggers asynchronously. It creates a new execution node in the Knowledge Graph to track the state and metadata of the execution, immediately returning a Universally Unique Identifier (UUID) to the UI. Once the ML models finish processing, the

orchestrator enriches the corresponding node with the output results.

The User Interface was developed using React.js and TypeScript to catch errors at an early stage and make the tool more reliable. Coupled with Chakra UI, it provides a simple, contemporary, and visually pleasing experience. This interface enables users to initiate complex tasks without waiting for them to complete in real time. Through a dedicated History page, the UI periodically fetches all execution nodes marked as "Finished" to display the optimized potential flowsheets, efficiencies, and total energy consumption, ensuring transparency and traceability through the database.

Conclusions

The transition toward a sustainable circular economy requires not only advanced metal recovery technologies but also the digital infrastructure to optimize them rapidly and reliably. This work bridged the computational gap between high-fidelity, time-intensive multiphysics simulations and the need for dynamic, system-wide flowsheet optimization. By employing a dual-track surrogate modeling strategy, the framework demonstrated exceptional adaptability to varying data volumes. For data-rich processes (like GDEx and ETMS), ML architectures (MLPs) provided extraordinary predictive accuracy, achieving a validation MAPE of approximately 0.56% and an R^2 exceeding 0.99 without overfitting. For data-sparse environments (like the ELX processes), the implementation of a two-step, physics-informed analytical fitting strategy successfully bypassed the risks of machine learning overfitting, strictly anchoring predictions to underlying saturation kinetics.

Integrating these validated surrogate models into a multi-start gradient descent optimization algorithm effectively solved the dimensionality bottlenecks associated with coupled, six-variable flowsheets. The tool successfully bounded peak RAM utilization strictly under 700 MB while returning optimized operational parameters within minutes to a maximum of two hours (under highly restrictive constraints).

Wrapped in a modern, scalable digital architecture utilizing a Neo4j Knowledge Graph, a FastAPI orchestrator, and a React-based user interface, this framework transcends theoretical modeling. Ultimately, it delivers a computationally lightweight, high-fidelity decision-support system that empowers process engineers to iteratively test, evaluate, and optimize complex electro-driven

recycling flowsheets with unprecedented speed and reliability.

References

- [1] "FIREFLY Project," [Online]. Available: <https://www.firefly-project.eu/>. [Accessed 3 8 2026].
- [2] O. Martinez-Mora, G. Pozo, L. F. Leon-Fernandez, J. Fransaer and X. Dominguez-Benetton, "Synthesis of platinum group metal nanoparticles assisted by CO₂ reduction and H₂ cogeneration at gas-diffusion electrodes," *RSC Sustainability*, vol. 1, no. 3, p. 454–458, 2023.
- [3] M. I. Leon Sotelo and S. Thayumanasundaram, "Simulation study of the gas-liquid system in the GDEx process for PGMs recovery," in *COMSOL Conference 2025*, Amsterdam, The Netherlands, 2025.
- [4] A. I. Siriwardana, A. U. Iturbe, C. D. R. Gaztelurrutia, J. H. Betanzos, G. A. P. Zamora, E. A. García and Y. B. I. a. E. M. Ugarte, "An electrochemical process for recovering platinum group metals (PGMs) from solid residues". European Patent Office Patent EP4490346A1.
- [5] G. Z. Chen, D. J. Fray and T. W. Farthing, "Direct electrochemical reduction of titanium dioxide to titanium in molten calcium chloride," *Nature*, vol. 407, p. 361–364, 2000.
- [6] C. Osarinmwian, I. Mellor and E.P.L. Roberts, "Titanium production in rotationally symmetric electrochemical reactors," *Electrochimica Acta*, vol. 164, pp. 48-54, 2015.
- [7] "FIREFLY AI-based predictive application," INLECOM INNOVATION, [Online]. Available: <https://firefly.inlecom.eu/>. [Accessed 6 8 2026].

Acknowledgements

Funded by the European Union under Grant Agreement No 101091715. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.