Response to Referee 1

Dear Editor and Referee 1:

We are very grateful to referee for reviewing the paper so carefully and we will incorporate of the suggestions made by the referee. Those changes will be highlighted with revision mode in the manuscript. Please see below, in blue, for a point-by-point response to the referees' comments and concerns.

Referee 1 comments: The authors build a decoupled hybrid architecture combining graph attention units, a differentiable Monod physics layer and a lightweight residual corrector, and integrate it into the mature CPqPy COMSOL-PHREEQC platform to resolve high computation costs and convergence defects of traditional geochemical solvers. Overall, this work presents a physically consistent surrogate workflow that reconciles computational speed with mechanistic accuracy for stiff biogeochemical reaction systems. However, several limitations merit further attention:

Response: Thank you very much for your approval of our work. The manuscript will be improved by us according to your comments, and the specific responses are as follows.

1. In the Methodology, P-GNN predicts dynamic $k_{max,j}$ values, which are passed to the differentiable physical layer for Multi-Monod kinetic calculations. In field-scale long-term bioremediation, microbial biomass varies by multiple orders of magnitude: abundant substrates drive rapid growth, while toxic metabolites cause severe die-off. Since training data only covers limited concentration and biomass ranges, P-GNN may output unreliable $k_{max,j}$ under out-of-distribution biomass conditions and induce numerical failures in the differentiable kinetic layer. Existing benchmarks only examine spatial scalability to 10^5 grid nodes, lacking quantitative analysis of the surrogate's long-term robustness against out-of-range biomass variations. It is recommended to add quantitative assessments of the framework's extrapolation performance under extreme biomass fluctuations over long simulation horizons within the Discussion section.

Response: We sincerely thank the reviewer for this valuable suggestion. We agree that evaluating the extrapolation capability of the surrogate model under out-of-

distribution (OOD) conditions is important for assessing its practical applicability in long-term field-scale bioremediation.

Because the current PHYGNET framework takes hydrochemical variables rather than microbial biomass as model inputs, we may not directly perturb biomass values. Instead, we will use Case A to perform an additional OOD robustness assessment by progressively enlarging the concentrations of all reactive species from 1.0× to 3.0× the original training range, thereby generating biochemical conditions beyond those represented in the training dataset. The surrogate predictions will be then systematically compared with the corresponding PHREEQC simulations. The new results will be incorporated into Section 4.3 (Discussion) together with a new Figure.

2. Within the Methodology section, to mitigate long-term drift in predicted pH and pe values, the authors adopt a hybrid solving scheme that invokes full PHREEQC geochemical recalibration every ten timesteps throughout transport simulations. Fractured aquifers have highly heterogeneous flow and steep reaction fronts, making fixed recalibration intervals inefficient: smooth transport zones waste computing resources, while steep-gradient areas suffer delayed error accumulation. The authors are advised to discuss the feasibility of an adaptive recalibration strategy triggered by spatial residual variability or abrupt chemical gradients as a potential future improvement within the Discussion section.

Response: We sincerely thank the reviewer for the suggestion regarding the hybrid solving scheme. We agree that a rigorous quantitative evaluation of the recalibration frequency is essential, and that exploring an adaptive strategy represents a vital step forward for applications in highly heterogeneous fractured aquifers. However, due to space limitations, we will include this section of the discussion in the supplement. The following text is the section we are about to add:

“ While the fixed 10-step interval provides a robust baseline for current simulations, it presents inherent limitations when applied to highly heterogeneous fractured aquifers characterized by steep reaction fronts. In such complex domains, fixed intervals can lead to redundant computations in smooth transport zones while delaying necessary error corrections in steep-gradient areas. Therefore, as a crucial

future improvement, this framework will benefit significantly from an adaptive recalibration strategy. Instead of relying on predefined steps, future iterations of the model could dynamically trigger the full-physics solver based on real-time spatial residual variability or abrupt chemical gradients across the graph network, thereby maximizing both accuracy and computational resource allocation."

3. The manuscript only reports results of the hybrid CPgPy workflow (PHYGNET + periodic PHREEQC recalibration), without long-term comparisons of three core benchmarks: original CPqPy, standalone PHYGNET, and the proposed hybrid scheme. No quantitative error curves of drift-sensitive variables including pH, pe and vinyl chloride over $\geqslant$500 h are provided, so the necessity and optimal value of the 10-step calibration cannot be verified. The authors are advised to supplement sensitivity tests with 5, 10 and 20-step recalibration frequencies, quantify the magnitude of long-term prediction drift under non-recalibrated conditions, plot comparative results of all three schemes in unified figures, and incorporate this analysis into Section 4.3 on computational efficiency.

Response: Thanks for your suggestion regarding quantitative evaluation of the recalibration frequency. To fully address this comment, we will conduct the requested sensitivity tests across multiple full-physics recalibration intervals (5, 10, 15, 20 steps and without calibration). However, due to the fact that PHREEQC diverges when computing long time series, the baseline model CPqPy was unable to complete the 500-hour calculation. Therefore, we plan to analyse the evolution of the RMSE over time for each calibration scheme over a 240-hour period. After that, we will analyse the impact of different calibration strategies on these results to further substantiate the scientific validity of our current choice of a 10-step calibration strategy.

4. The model is trained on 50000 single-step state transitions generated by PHREEQC (Δt = 1 h). However, when embedded in COMSOL for 240-hour RTM simulations, the surrogate performs autoregressive sequential predictions. The paper only reports spatial residuals at the final time step (Fig. 8), without quantifying how prediction errors evolve over time or whether they saturate or diverge.

Response: We sincerely thank the reviewer for raising this critical point. To thoroughly address this concern, we will extract and visualize the spatial residual error maps at multiple intermediate time steps (t = 10, 50, 100, 150, 200, and 240 h) for both primary concentration variables (e.g., HC, NO_3^-, O_2) and thermodynamic environment variables (pH, pe). We will also add these comprehensive evolution plots and their corresponding discussions to the revised supplement.

5. The impressive 3524-fold speedup is a central claim, yet the manuscript lacks detailed specifications of the hardware (CPU/GPU models), software versions (PHREEQC, PyTorch, COMSOL), batch size for PHYGNET, and whether PHREEQC was run in serial or parallel mode. These details are essential for reproducibility.

Response: We agree with the reviewer that detailed computational specifications are essential for ensuring the reproducibility of the reported 3524-fold speedup. To fully address this comment, we will add a dedicated paragraph in the revised manuscript (Section 3.2.2) detailing the exact hardware configurations, software versions, and execution modes used for the performance comparison. Here are the detailed information:

Hardware: Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU.

Software Versions: COMSOL Multiphysics version 6.1, PHREEQC version 3.8.6, and PyTorch version 2.4.1.

ML Surrogate Settings: The PHYGNET model was trained and executed with a batch size of 256 on the GPU.

PHREEQC Execution Mode: The traditional geochemical model was executed in serial mode. To optimize its standard execution efficiency, all sample conditions were compiled into a single, unified batch-input script and solved sequentially by the PHREEQC engine on the CPU.

The following text is the section we are about to add in Section 3.2.2:

"To ensure the reproducibility of the computational performance, all simulations were executed on a standardized workstation equipped with an Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU. The software environment was

configured with COMSOL Multiphysics (v6.1), PyTorch (v2.4.1), and PHREEQC (v3.8.6).

During the model development phase, the PHYGNET was trained on the GPU using a batch size of 256. However, during the actual reactive transport simulations, the trained surrogate fully leveraged GPU acceleration to perform massively parallel predictions across all spatial grid nodes simultaneously. In contrast, the baseline physical simulations utilizing the PHREEQC engine were conducted in serial mode on the CPU. To represent a highly optimized traditional workflow, all independent reaction scenarios across the grid were compiled into a unified batch-input script, which was then solved sequentially without parallel multi-processing overhead."

Thank you for your warm and timely work earnestly. This will be very helpful for improving the manuscript and our future research. We hope that our responses and the revisions will be satisfactory, while we will be happy to work with you to resolve any remaining issues.