Response to Referee 3

Dear Editor and Referee 3:

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 3 comments: This manuscript proposes PHYGNET, a GPU-accelerated surrogate model combining a physics-informed graph neural network and a residual MLP for nonlinear reactive transport simulations. The approach is interesting and potentially valuable. However, major concerns remain regarding benchmark fairness, temporal generalizability, long-term numerical stability, and manuscript organization. Detailed comments are provided below.

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

1. The manuscript compares GPU-based PHYGNET with CPU-based PHREEQC, but the benchmark lacks essential information, including the versions of COMSOL and PHREEQC, the coupling interface, CPU and GPU models, core counts, clock frequency, memory, and software environment. More importantly, PHYGNET appears to be compared with serial PHREEQC. This is not an equivalent benchmark because the USGS provides PhreeqcRM, which supports OpenMP and MPI parallelization for reactive transport simulations. A GPU-parallel versus CPU-serial comparison may substantially overestimate the reported speedup. The authors should provide complete software and hardware specifications and include comparisons with both serial PHREEQC and multicore PhreeqcRM. Wall-clock time, initialization, data-transfer cost, memory use, training cost, and inference cost should be reported separately.

Response: We sincerely thank the reviewer for this valuable suggestion. 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. We will also use

PHREEQCRM to conduct the same efficiency comparison in Section 4.3. However, although PHREEQCRM supports parallel computing, its computations are CPU-based and do not utilize the GPU. Since parallel computing using the GPU is a major advantage of machine learning, it is foreseeable that PHYGNET will still be more efficient than PHREEQCRM.

2. All major experiments use a fixed time step of (Δ t = 1 h). It remains unclear whether PHYGNET merely learns a one-hour state-transition mapping. Can the trained model be directly applied to other time steps, such as (0.5), (2), or (5 h), or would it need to be retrained?

Response: We highly appreciate this insightful question. To directly answer: if optimal accuracy is required for a different base reaction time step (e.g., 0.5 h or 2 h), the model indeed needs to be retrained.

However, in standard operator-splitting reactive transport modeling, physical simulation processes follow a temporal hierarchy: flow time step $\geq$ transport time step $\geq$ reaction time step. Because geochemical kinetics are numerically stiff, maintaining a fine, fixed reaction step (i.e., Δ t = 1 h) is a standard practice. Therefore, for larger global transport steps (i.e., 5 h), the current 1-hour PHYGNET can be applied using a sub-stepping strategy (i.e., sequentially invoking the surrogate 5 times) without any need for retraining.

Furthermore, if a slight sacrifice in absolute precision is acceptable, the currently trained model can actually be directly applied to other Δ t inputs without retraining. This flexibility stems from our decoupled architecture: the front-end P-GNN predicts instantaneous parameters, which the subsequent physical kinetic layer explicitly scales according to the newly assigned Δ t. The resulting minor deviations would originate solely from the final MLP residual blocks, as their weights were specifically optimized to compensate for exactly 1-hour numerical integration residuals.

To ensure clarity for future readers, we will add a discussion in the revised manuscript explaining this temporal hierarchy, the sub-stepping application, and the architectural trade-offs regarding variable Δ t configurations.

3. The manuscript adopts a hybrid strategy in which PHREEQC is invoked every 10 time steps for physical correction. However, such periodic correction may constrain mass conservation and other physical requirements without eliminating the systematic trajectory drift caused by PHYGNET prediction errors. This external intervention may therefore mask the true accumulation of errors when PHYGNET is used as an independent surrogate. The authors should provide a comparison without PHREEQC correction, i.e., continuous prediction using pure PHYGNET, to evaluate its intrinsic numerical stability, long-term error accumulation, and the actual contribution of the periodic correction to prediction accuracy.

Response: We sincerely thank the reviewer for this comment. To directly address this, we will conduct the requested continuous prediction experiment using pure PHYGNET without any external PHREEQC intervention. Furthermore, to comprehensively demonstrate the actual contribution of the periodic correction to prediction accuracy, we will expand this analysis to evaluate the spatial root-mean-square error (RMSE) time series across multiple variable recalibration frequencies (i.e., pure continuous prediction, as well as 5-step, 15-step, and 20-step intervals) against the fully physical baseline. These new evaluations may allow us to quantitatively track the trajectory drift of the multi-species transport and clearly define the model's intrinsic stability boundaries over extended simulation periods without external intervention.

4. The sawtooth-like oscillations in (pe) during (t=10-60 h) in Fig. 8f suggest local overfitting and high-frequency numerical artifacts in this strongly nonlinear transition regime. However, the manuscript provides neither model performance across different datasets nor sufficient details of the training procedure. The authors should report the proportions and splitting strategy used for the training, validation, and independent test sets, together with quantitative metrics such as (R^2) and RMSE for each dataset. Please also clarify whether overfitting-control strategies, such as smoothness regularization, early stopping, or cross-validation, were employed, and explain and correct the nonphysical oscillations observed in Fig. 8f.

Response: Thanks for raising these important questions regarding model training and prediction stability. We address the concerns regarding the physical interpretation

of Fig. 8f and the machine learning training procedures respectively below.

(1). Clarification on the "Sawtooth-like Oscillations" in Fig. 8f

We must respectfully clarify that the fluctuations observed in pe during the t = 10-60h interval are not numerical artifacts or indicative of local overfitting, but rather reflect highly dynamic, genuine physical and geochemical processes. Specifically, Fig. 8f illustrates the time-series fitting of pH and pe at a single specific probe point with spatial coordinates (0.15, 0.5). This exact location is situated directly at the complex interface between the discrete fracture network (DFN) and the surrounding porous rock matrix. The t = 10-60h temporal window corresponds precisely to the period when the main reactive contaminant plume migrates through this specific interfacial region. Because of the extreme contrasts in advective flow velocities and mass transfer rates between the highly fractured zones and the less permeable matrix, sharp and rapid localized shifts in concentration gradients occur. These continuous mixing-controlled reactive transport dynamics at the boundary naturally cause the redox state (pe) to exhibit jagged, sawtooth-like variations. Therefore, PHYGNET is accurately capturing the physical complexity of the multi-component reactions at this specific fracture-matrix interface.

(2). Details on Model Training, Data Splitting, and Overfitting Control

We apologize for not providing sufficient training details in the previous version. To comprehensively evaluate the model's performance and generalization capability, the overall dataset was randomly partitioned using an 80:20 splitting strategy (80% of the data allocated for the training set and 20% reserved strictly for the independent test set). Regarding the concern of local overfitting, we respectfully direct the reviewer's attention to the training and validation loss trajectories previously presented in Fig. 2. The curves clearly demonstrate that both training and testing losses decrease synchronously and stabilize at a low level without diverging or showing any "U-shaped" rebound characteristic of overfitting. This confirms that the model generalizes well to unseen data. To ensure full transparency as requested, we will incorporate all necessary quantitative metrics (R^2 and RMSE for both datasets) and a detailed description of the dataset allocation into the revised manuscript.

5. In Fig. 8e, the prediction errors for major species, including (NO_3^-) and (O_2), increase noticeably after approximately 200 steps (200 h). However, the simulation is terminated at 240 steps, before the system reaches steady state. This early termination may conceal further error accumulation and long-term prediction instability. The authors should extend the simulation well beyond 240 steps until steady state or reaction completion is reached, explain the increase in errors after 200 steps, and evaluate the numerical stability of PHYGNET over longer prediction horizons.

Response: We thank the reviewer for this thoughtful comment regarding the simulation horizon and the long-term numerical behavior of CpgPy. We appreciate the opportunity to clarify the rationale behind selecting the 240-hour simulation window and how we address the long-term stability evaluation.

(1). Rationale for the 240-Hour Simulation Horizon

As visualized in Fig. 6, the spatial morphology of the reactive contaminant plume remains virtually unchanged between t = 120h and t = 240h. During this temporal window, the fast reactive dynamic transition phase has largely concluded, and the system enters a representative quasi-steady state where spatial species distributions are predominantly governed by slow matrix diffusion, release, and kinetic attenuation. Therefore, the 240-hour window effectively captures both the dynamic transient evolution and the subsequent quasi-steady regime.

Moreover, a primary constraint in extending the simulation far beyond 240 steps stems from the numerical limitations of the conventional coupled baseline model (CPqPy). Traditional geochemical reaction solvers are inherently prone to severe numerical non-convergence and divergence issues when integrated over extended temporal horizons under low concentration conditions. While CPgPy itself possesses the intrinsic capacity to evaluate further forward steps stable without breaking, running simulations far beyond 240 steps would lack a reliable physical ground truth for quantitative accuracy verification. Without valid CPqPy benchmark outputs, any accuracy metrics (such as RMSE or R^2) computed for extended periods would be unsubstantiated.

(2). Explanation of Error Trends and Quantitative Stability Evaluation

The modest rise in prediction errors for major species (NO_3^- and O_2) observed around t = 200h is primarily attributed to the low concentration magnitudes characteristic of the matrix-dominated attenuation stage, where small residual errors naturally yield slightly higher relative metrics.

To rigorously demonstrate CPgPy's long-term stability within the valid baseline horizon, we will extract the continuous spatial root-mean-square error (RMSE) distribution and time-series evolution for all major species from t = 0 to t = 240h, and these newly added figures will be discussed in the Supplement to figure out how the error accumulation evolves during simulation.

6. The primary role of the final paragraph of the Introduction (Lines 85-96) should be to clearly delineate the study scope, experimental design, and evaluation objectives, namely, what the study does and aims to investigate, rather than prematurely presenting the results. I recommend that the authors restructure this concluding paragraph by clearly explaining the proposed model, the validation scenarios, and the evaluation criteria. Specifically, the paragraph should state: (1) what model is proposed, including PHYGNET and its core physics-constrained mechanisms; (2) the reactive transport scenarios in which it is validated, such as multispecies biogeochemical kinetic systems; and (3) the specific performance dimensions evaluated, such as long-term temporal stability, physical consistency, and computational speedup. This revision would allow the paragraph to serve as an effective roadmap for the remainder of the manuscript.

Response: We appreciate this valuable suggestion. We will restructure the final paragraph of the Introduction to serve as a roadmap of the study rather than a summary of results. Specifically, we will clarify (1) the architecture and physics-constrained mechanisms of PHYGNET, (2) the reactive transport scenarios used for validation, including multispecies biogeochemical kinetic systems and COMSOL-coupled simulations, and (3) the evaluation criteria focusing on predictive accuracy, temporal stability, physical consistency, and computational efficiency. All result-oriented statements have been removed from this paragraph and moved to the Results section.

This would be the final paragraph of the Introduction:

"Based on these considerations, we propose PHYGNET (Physically GNN Network), a physics-integrated GNN framework designed to accelerate microbially mediated kinetic reactive transport simulations while preserving mechanistic consistency. PHYGNET represents biogeochemical reaction systems as directed graphs, where chemical species and reaction pathways are encoded as graph components, and integrates a Monod-based kinetic physics layer to explicitly enforce stoichiometric constraints and mass conservation. A residual correction module is further incorporated to capture unresolved system-specific variations. To evaluate its applicability to complex reactive transport problems, PHYGNET is validated against representative multispecies biogeochemical kinetic systems, including thermodynamically ordered competitive degradation and sequential reductive dechlorination networks, and is further coupled with a COMSOL-based reactive transport framework for spatially distributed simulations. The performance of PHYGNET is systematically assessed in terms of predictive accuracy, long-term temporal stability during autoregressive simulations, physical consistency of the simulated geochemical states, and computational efficiency relative to conventional geochemical solvers. This study aims to establish a generalizable physics–ML framework for accelerating large-scale reactive transport modelling while retaining the fundamental chemical constraints required for reliable environmental predictions."

7. Lines 38-48 and 49-59 contain substantial overlap and redundancy. These passages should be merged and condensed to improve the clarity and conciseness of the manuscript.

Response: Adopted. Thank you for this suggestion. To improve the conciseness and logical flow of the Introduction, we will remove the repetitive statements in the second paragraph and retained only the discussion specific to the unique characteristics of sequential microbial reaction networks.

8. The authors should move the methodological description of SHAP from Section 4.3 to Section 2 (Methodology), for example, around Line 204. Section 4.3 should retain only the SHAP results and their physical interpretation. This revision would improve the manuscript's structural consistency and logical flow.

Response: Adopted. Thank you for this suggestion. Accordingly, we will move the description of the SHAP methodology from Section 4.4 to Section 2 (Methodology) and place it after the description of the PHYGNET framework (around Line 204). Section 4.4 will be revised accordingly to focus only on the SHAP results and their physical interpretation.

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.