Response to Referee 2

Dear Editor and Referee 2:

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 2 comments: This paper proposes a Physics-Constrained Graph Neural Network surrogate model (PHYGNET) to address the issues of low computational efficiency in microbial kinetic reactions, high computational costs of traditional geochemical solvers, and susceptibility to numerical convergence failures in existing Reactive Transport Models (RTMs). The method maps the microbial reaction network into a graph structure, introduces a Monod kinetics physical layer to ensure mass conservation, and incorporates a residual correction network to improve prediction accuracy. The model is further coupled with COMSOL to enable rapid simulation of complex reactive transport processes. Experimental results demonstrate that the proposed model significantly improves computational efficiency while maintaining prediction accuracy, and exhibits good interpretability. I have raised numerous comments below, including unscientific wording, unclear key steps that are not adequately explained, repetitive expressions, and contradictory statements throughout the manuscript.

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. Lines 38 – 40 and Lines 49 – 51: The content expressed in these two sections is highly repetitive.

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 retain only the discussion specific to the unique characteristics of sequential microbial reaction networks.

2. Line 43: The reference of "The cost" is unclear and ambiguous.

Response: Adopted. Thank you for this suggestion. To improve clarity, we will replace it with a more explicit description referring to the computational cost of solving the geochemical reaction module. Here is the specific revision:

"The computational cost of solving the geochemical reaction module escalates dramatically in large-scale or highly resolved simulations and is frequently accompanied by numerical instabilities, including convergence failures under extreme chemical gradients."

3. Lines 49 – 73: The introduction section lacks clarity regarding the development trajectory of AI-based methods and the research motivation. In the current text, terms such as AI, machine learning (ML), neural networks (NN), and graph neural networks (GNN) are used interchangeably, but a unified conceptual definition and clear hierarchical relationship among these terms are missing. The manuscript first points out that traditional neural networks are difficult to apply to the simulation of complex reaction networks, and then directly proposes GNN as a solution. However, GNN itself is also a type of neural network, which makes the logical transition problematic.

Response: Adopted. Thank you for this suggestion. To improve clarity, we will revise the manuscript by adopting machine learning (ML) as the unified term throughout the Introduction. We will describe graph neural networks (GNNs) as a specific ML architecture for graph-structured data and avoid using the term "neural networks (NNs)" separately, thereby improving the logical flow and clarifying the motivation for selecting GNNs.

4. Line 85: The phrase "not only a theoretically appealing" and the subsequent "powerful choice for surrogate modelling of biogeochemical reaction systems" appear to convey largely the same meaning. It seems that the intended emphasis here should be on the advantage of embedding physical mechanisms into the network. The wording should be revised to clarify this point.

Response: Adopted. Thank you for this suggestion. We will focus on the advantage of embedding physical mechanisms into the network. This will be the specific revision:

“By explicitly embedding these physical mechanisms into the network architecture, GNNs provide a uniquely suited framework for the surrogate modelling of biogeochemical reaction systems.”

5. Line 99: The phrase "kinetic logic of given biogeochemical system" is missing an article. It should be revised to, for example, "the kinetic logic of a given biogeochemical system" or "the kinetic logic of the given biogeochemical system."

Response: We agree with the reviewer’s assessment. We will take the latter one in the revised article based on your comment.

6. Line 106: The authors only used two reaction networks to validate the model, so it is inappropriate to describe this as a "dual-validation approach." Additionally, the expression "ready extensibility to" seems problematic and should be rephrased for clarity and accuracy.

Response: We agree with the reviewer’s assessment. We will remove the phrase “dual-validation approach” to avoid overstatement, and rephrase the sentence to more accurately state that evaluating the model on these two distinct networks suggests its potential applicability to other subsurface processes, rather than claiming “ready extensibility”. This will be the specific revision:

“Evaluating the framework across the two networks suggests its potential applicability to other critical subsurface processes, such as microbially induced heavy metal reduction (e.g., U(VI) or Cr(VI) bioremediation) and nutrient cycling.”

7. Line 120: The subscripts *m* and *f* should be typeset in mathematical italics rather than in Roman or default text style.

Response: Adopted. Thank you for this suggestion. We will correct this error based on your comment.

8. Line 130: The phrase "which are" needs to be checked for grammatical agreement. Also, regarding "reaction terms (S)": is the term singular or plural? If singular, it should be "term"; if plural, it should correspond to multiple (S) entries. The current usage of "terms" with a single (S) is inconsistent.

Response: We will correct the grammatical error and changed “reaction terms” to the singular form “reaction term” to maintain consistency with the variable (S), and update the corresponding verb to “is handled”.

9. Line 144: The abbreviation "e.g.," is used to introduce an example, but only one example is provided. If multiple examples are not given, "e.g.," may not be appropriate; consider using "i.e.," or rephrasing accordingly.

Response: We will correct the grammatical error and changed “e.g.,” to “i.e.,” to maintain consistency.

10. Lines 145 – 146: The use of bold and non-bold "C" to represent different parameters is highly confusing and prone to misinterpretation. In addition, the use of conjunctions (e.g., "and") is problematic, and there are excessive commas throughout the sentence. The sentence following Equation (5) that defines the variables is overly long and structurally unclear, making it difficult for readers to follow.

Response: We thank the reviewer for pointing out the confusing notation and structural issues. To eliminate the ambiguity between the chemical product “C” and the concentration vector “C” (and to clarify the bold vs. non-bold usage), we will replace the dummy product symbols in Equation 5 with “Y” and “Z”. Additionally, we will split the overly long sentence into several shorter, independent sentences to improve grammatical accuracy and overall readability.

11. Line 155: The phrase "network-level regulatory feedbacks" is mentioned here, but it is unclear where this concept was previously introduced or defined in the manuscript. The authors should clarify its origin or provide an appropriate reference.

Response: To avoid confusion, we will replace the undefined phrase “network-level regulatory feedbacks” with “inhibitory feedbacks among species” to ensure consistency with our earlier terminology, and explicitly link it to the inhibition term.

12. Lines 144 – 165: The definitions of the formulas in this section are highly disorganized. Different parameters are defined using the same letter symbols, and several parameters are defined repeatedly throughout the text. This creates significant confusion and should be thoroughly revised for consistency and clarity.

Response: We sincerely apologize for the confusion caused by the disorganized parameter definitions. We will thoroughly revise the text between Lines 144 – 165 to ensure that each variable is uniquely defined only upon its first appearance, and we will standardized the notation (e.g., using **C** for the concentration vector and C_i for individual species concentrations) to eliminate ambiguity.

13. Lines 179 – 181: The meanings expressed in these two sentences are redundant and overlap with each other. The authors should consolidate or remove the repetitive content.

Response: We will remove the latter sentence "Having clarified the mechanistic dependencies of the biochemical system, this stage involves translating these established relationships into a parametric graph architecture" to improve the flow of the text.

14. Line 188: The term "latent embeddings" is used here without any prior definition or explanation. The authors should provide a clear definition of what latent embeddings refer to in the context of this work.

Response: We thank the reviewer for pointing this out. We will add a brief explanation to clarify that "latent embeddings" refer to the low-dimensional vector representations that capture the topological and chemical context of the nodes within the graph.

15. Line 193: The output of the MPL (Message Passing Layer) is described as defining "the outputs (concentrations, pH, and pe)," yet earlier the prediction target was stated to be "($k_{max,j}$)." This raises a critical inconsistency: which one is actually the output layer of the neural network? The authors need to clarify the exact output of the model.

Response: We thank the reviewer for highlighting this apparent inconsistency. To clarify, PHYGNET is a modular framework composed of different sub-networks rather than a single neural network with one output. The Graph Neural Network (P-GNN) component outputs the kinetic parameters ($k_{max,j}$). These parameters are then fed into the hard-coded Kinetic Layer, which computes the preliminary states. Finally, the Multilayer Perceptron (MLP) component takes these preliminary results and outputs

the final, residual-corrected concentrations, pH, and pe. We will revise the text to make the sequential flow of outputs from each specific component explicitly clear. This will be the specific revision:

"Through message passing, the P-GNN component learns latent embeddings (low-dimensional vector representations that encapsulate the topological and chemical context of each node) and uses them to predict biological kinetic parameters (i.e., $k_{max,j}$) for each pathway. These predicted parameters are then passed into the explicit kinetic physics layer, which deterministically calculates preliminary species concentrations, pH, and pe to ensure fundamental mass conservation and stoichiometric consistency. Finally, a lightweight multilayer perceptron (MLP) takes these preliminary physics-based results and outputs the final refined concentrations, pH, and pe by correcting residual errors. This decoupled architecture preserves interpretability while improving accuracy all target variables."

16. Line 194: The terms "decoupled residual block" and "tasks" are mentioned but have not been defined or introduced anywhere previously. The authors should provide clear explanations for these terms.

Response: The terms will be clarified in the revised manuscript. We will explicitly define the "decoupled residual block" as the architectural separation of the residual MLP from the physical kinetic engine, and replace the ambiguous word "tasks" with "target variables" to refer to the simultaneous prediction of concentrations, pH, and pe.

17. Line 196: The manuscript states that P-GNN is "for reaction rate estimation," but the main text says it predicts "the kinetic parameters (kmax,j)." These two descriptions are not equivalent — reaction rate estimation and kinetic parameter prediction are different objectives. The authors should reconcile this inconsistency. In addition, there is no full name of "AI RTM".

Response: The inconsistency will be resolved by updating the figure caption to state that the P-GNN is used for "kinetic parameter prediction", strictly aligning it with the main text. Additionally, AI RTM" will be replaced by "ML RTM" to keep consistency with the Introduction.

18. Line 188: The term "full-physics recalibrations" is ambiguous—what exactly does it refer to? Does it mean that the MLP refines the outputs (concentrations, pH, and pe) by correcting residual errors, or does it refer to the enforcement of fundamental mass conservation and stoichiometric consistency? Neither of these interpretations seems to qualify as "full-physics" in the strict sense. The authors should clarify what they mean by this term or consider using more precise wording.

Response: The term “full-physics recalibrations” does not refer to the internal components of PHYGNET (such as the MLP or the kinetic layer). Rather, it refers to the hybrid coupling strategy detailed in Section 2.4, where the traditional geochemical solver (PHREEQC) is called every 10 simulation time steps to perform a complete thermodynamic and kinetic simulation. To prevent any confusion, we will replac this ambiguous phrase with “periodic recalibrations using the PHREEQC solver” in the figure caption.

19. Line 197: The phrase "To balance physical consistency with predictive accuracy" implies a trade-off between the two objectives. However, the manuscript does not previously mention or justify why such a trade-off exists in this context. The authors should explain the rationale behind this balance and clarify the potential conflict between physical consistency and predictive accuracy.

Response: We thank the reviewer for this insightful comment. We agree that the word “balance” was misleadingly used. Our actual intention was to explain how the framework integrates the two: the hard-coded physics layer enforces strict mass conservation, while the ML architecture captures the thermodynamic subtleties to maintain high predictive accuracy. We will revise the text to explicitly clarify this synergy rather than portraying it as a trade-off:

“To seamlessly integrate physical consistency with predictive accuracy, PHYGNET adopts a multi-task learning strategy that simultaneously captures chemical dynamics and system states.”

20. Line 200: PHYGNET is described as employing a multi-task learning approach that simultaneously captures chemical kinetic processes and system states. However, the subsequent description of "joint prediction" as a "unified loss function"

is conceptually problematic: joint prediction is an output/result, whereas a unified loss function pertains to the training/optimization process of the network. These are distinct concepts and should not be conflated.

Response: We will revise the sentence to clearly separate the structural capability of joint prediction from the optimization process driven by the unified loss function. This will be the specific revision:

"Specifically, the model performs joint prediction of species concentrations, pH, pe, and the kinetic parameters, and this multi-task learning process is optimized during training through a unified loss function."

21. Lines 201 – 204: The explanation provided here does not clearly articulate how the multi-task learning is actually implemented, nor does it specify how the balance between physical consistency and predictive accuracy is achieved. The authors should provide a more detailed and rigorous description of the multi-task learning framework and the specific mechanisms used to balance these two objectives.

Response: We agree that the conceptual overview in this paragraph lacked the necessary technical depth. Thus, we will add a forward reference to Section 3.2.2 (Equation 18), where the rigorous mathematical formulation of the unified loss function is fully detailed.

22. Line 218: The term "every node" is not clearly explained. Since the graph contains both species and reactions, which entities correspond to species nodes and which correspond to reaction nodes? Furthermore, the specific organization of the heterogeneous graph has not been clearly defined; the manuscript has not previously stated that this is a heterogeneous graph neural network. The authors should further clarify the node types, edge definitions, and connection relationships within the graph structure.

Response: We agree that the graph structure required clearer definition. We will explicitly introduce it as a heterogeneous bipartite graph. To make the organization entirely transparent, we will add a new figure (using the competitive degradation of HC as a concrete example) to visually define the graph structure. The revised text will explicitly detail the two node types (chemical species and reaction processes) and the

three specific edge types (reactant-to-reaction, reaction-to-product, and inhibitory edges) connecting them. This will be the specific revision:

"To process this network, the P-GNN models the biogeochemical system as a graph comprising two distinct node types: chemical species nodes and reaction process nodes. As illustrated using the competitive degradation of HC (Fig. 2), directed edges connect these nodes to define their biogeochemical relationships. These connections are explicitly categorized into three edge types: reactant-to-reaction edges (solid black lines) indicating mass consumption, reaction-to-product edges (solid green lines) indicating mass generation, and inhibitory edges (dashed purple lines) representing regulatory signals (e.g., O_2 inhibiting sulfate reduction and methanogenesis).

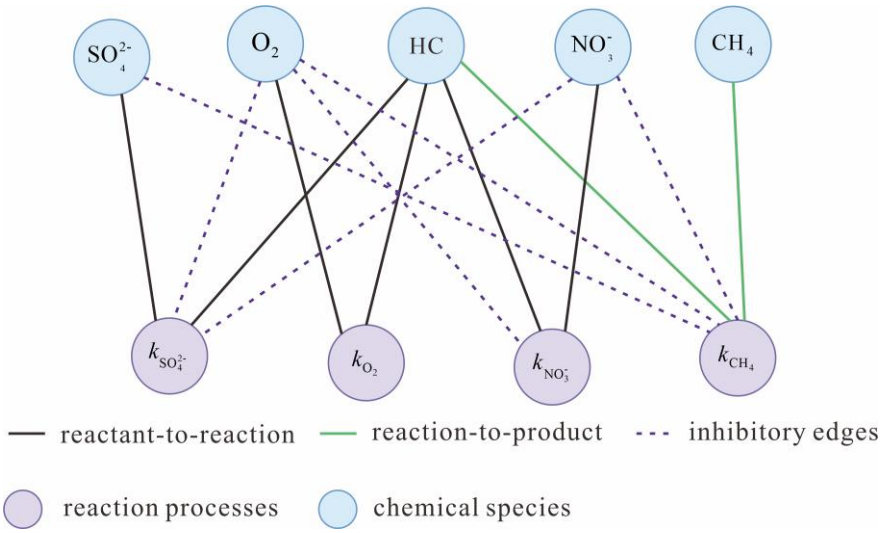

Figure 2 Schematic of PHYGNET's reaction graph encoding in the P-GNN module.

The P-GNN applies 3 stacked GAT layers to this specifically organized heterogeneous graph. At each layer, every individual node, whether representing a chemical species or a reaction process, receives messages from its incoming neighbours, weighted by the attention mechanism described above."

23. Line 211: The text states "A final readout or MLP head," which raises the question: which one is actually the output layer? Why is it presented as an "or" (i.e., a choice between the two)? The authors should specify the exact architecture of the output layer.

Response: It will be revised in the comment 22.

24. Line 222: The term "global environment outputs" has not appeared or been defined previously. Also, what are the "relevant nodes" mentioned here—are they species nodes or reaction nodes? How exactly is the aggregation performed over these nodes?

Response: It will be revised in the comment 22.

25. Line 223: What is meant by "learned embedding"? This term has not been clearly defined in the manuscript.

Response: It will be revised in the comment 22.

26. Line 227: The text mentions "Encoding P-GNN into PHYGNET." However, PHYGNET already incorporates P-GNN as a core component—so what does it mean to "encode" P-GNN into PHYGNET? This phrasing is confusing and should be clarified.

Response: We will replace the section heading with “Integrating P-GNN with the Differentiable Kinetics Layer” to clarify that we are specifically integrating the P-GNN with the differentiable physics layer.

27. Line 233: The statement that data-driven models "will violate mass conservation" is overly absolute. Not all data-driven models necessarily violate mass conservation; the authors should qualify this claim or provide supporting evidence.

Response: We will remove the statement “which is a common pitfall in naive deep learning surrogates that inevitably leads to mass balance violations” to keep overly absolute description.

28. Line 234: What are "localized kinetic parameters"? And what are "dynamically predicted parameters"? These terms are introduced without clear definitions.

Response: We will delete the word “localized” and “dynamically” to keep clear statement.

29. Line 237: Equations (6) – (8) do not appear to include any explicit mass conservation formula. The authors should check whether the mass conservation constraint is actually represented in these equations.

Response: Mass conservation is rigorously and intrinsically enforced during the integration of Equation (6) via the stoichiometric matrix $v_{i,j}$, which strictly couples the

consumption of reactants to the generation of products. We will revise the text to precisely state that the kinetic layer enforces mass balance through this stoichiometric coupling, rather than through a standalone algebraic conservation equation.

This will be the specific revision:

"These predicted parameters are injected into the differentiable kinetics layer, which explicitly executes the multi-Monod rate laws and inhibition functions Equations (7) to (8), and rigorously enforces mass conservation through the stoichiometric matrix ($v_{i,j}$) defined within the governing ODE system (Equation (6))."

30. Line 241: The text states that "this code" can encode "elemental mass," but the preceding paragraph refers to "stoichiometric mass." These two terms seem to refer to different concepts. The authors should reconcile the terminology.

Response: We will reconcile the terminology by uniformly referring to 'mass conservation' (which is enforced via stoichiometric relationships) to ensure consistency and precision throughout this section. This will be the specific revision:

"This specific integration strategy guarantees that regardless of the specific biogeochemical reaction network being modelled, whether involving the parallel thermodynamic competition of electron acceptors (as seen in BTEX oxidation) or the sequential, inhibitory chain reactions of chlorinated solvents (such as PCE dechlorination), mass conservation is maintained."

31. Line 245: What are "updated environmental features"? This term has not been mentioned previously in the manuscript.

Response: "The updated environmental features" means the pH and pe, here we will add "(**E**)" which has been explained in the former text to keep it clearly described.

32. Line 246: The claim that the MLP can maintain mass conservation because the residual is computed for concentrations, pH, and pe raises the question: where do the ground-truth values for these residuals come from? Residuals are calculated from predicted values versus measured/true values. If such ground-truth data are available, then in theory, one could directly build a data-driven model using these data. The authors should explain why this is not the case and justify the use of residual correction.

Response: We sincerely thank the reviewer for pointing out this critical issue. We agree with your underlying concern: purely data-driven components like MLPs lack inherent physical constraints and will inevitably introduce some degree of deviation to mass conservation in rigid kinetic systems. However, the impact of the MLP on mass conservation is confined within a highly controllable range, preventing severe macroscopic deviations. As seen in Figure 4(i), which validates the total mass balance of the C_2 backbone species (Σ(PCE, TCE, DCE, VC, Ethene)), the predicted total molarity data points (blue dots) do exhibit slight scatter and fluctuations around the absolute 1:1 identity line. We acknowledge that these minor deviations are a direct result of the MLP's residual correction applied on top of the physical layer. However, these fluctuations are entirely bounded within a highly stringent ± 0.1 log unit error band (indicated by the dashed lines). The statistical metrics (R^2 = 0.994, RMSE = 1.98 × 10^{-5} mol/L^{-1}) demonstrate that the overall mass balance remains acceptable. In summary, although the MLP has some impact on the conservation of mass, this impact remains within an acceptable range due to the presence of the preceding kinetic layer.

Here will be the revised texts:

Section 2.3.3

"Furthermore, real-world hydrogeochemical systems often exhibit subtle, unmodeled thermodynamic fluctuations. While the ground-truth target values for training are generated via comprehensive batch simulations using PHREEQC (detailed in Section 3.2.1), we deliberately avoid using a purely data-driven model to predict the final states directly, as such black-box models notoriously fail to maintain mass conservation in stiff kinetic systems. Instead, the physics-constrained concentration estimates are concatenated with the updated environmental features(E) and passed through a shallow Multi-Layer Perceptron (MLP) corrector. By restricting the MLP to learning only the residual differences between the physical approximations and the PHREEQC labels, this architecture ensures that the final predicted state is fundamentally anchored by strict mass balance laws, while still possessing the flexibility to capture complex environmental feedbacks."

Section 3.3

"Crucially, while resolving trace-level intermediate products poses a significant challenge, the PHYGNET framework completely eradicates a fatal flaw of traditional pure data-driven surrogates: the violation of mass conservation. As demonstrated in Fig. 5i, we evaluated the global mass balance of the Case B network by tracking the total molarity of all two-carbon (C_2) species (Σ(PCE, TCE, DCE, VC, Ethene)). Because the sequential dechlorination process merely strips chlorine atoms while leaving the C_2 carbon backbone intact, the sum of these species must theoretically remain constant throughout the temporal integration. As Fig. 5i illustrates, the predicted total molarities align with the 1:1 ground truth line, constrained within an exceptionally tight error margin of ± 0.1 log units. Despite the slight perturbation to mass conservation induced by the MLP layer, the magnitude of this impact is acceptable. Therefore, the proposed architecture represents an optimal compromise between computational precision and physical fidelity."

33. Line 254: The text says "see Fig. B," but no such figure is referenced or appears to be included. Please check and correct the reference.

Response: Sorry for that mistake, it should be "see Fig. 1b", it will be correct in the revised paper.

34. Line 262: The manuscript states that PHREEQC is used for correction. Does this correction apply to the reaction rates at the current time step? The subsequent text mentions correction based on the "latest COMSOL output," but PHREEQC typically simulates reactions over a certain time period. Is this period equivalent to the ten time steps mentioned? This is not clearly explained. Moreover, if PHREEQC is used for correction, this implies that PHREEQC is assumed to be correct. If so, why not simply use PHREEQC directly instead of the surrogate model?

Response: We appreciate the reviewer's thorough evaluation and apologize for the lack of clarity. To clarify, PHREEQC does not retroactively simulate the entire 10-step period; rather, it is invoked to simulate only the current single time step (Δt) to re-establish a highly accurate thermodynamic and kinetic baseline, thereby preventing long-term error accumulation. Regarding why we do not use PHREEQC exclusively: as demonstrated in Section 4.3, running traditional solvers at every time step for large-

scale grids incurs prohibitive computational costs. The hybrid strategy represents a deliberate balance between predictive accuracy and computational efficiency. PHYGNET handles the bulk of the temporal integration to provide massive speedups, while periodic PHREEQC calls ensure the simulation remains physically tethered over long durations without sacrificing this efficiency.

35. Line 265: Where do the "dynamic reaction rates" come from? Can COMSOL actually output dynamic rates, or does it only output a fixed numerical value? This should be

Response: We thank the reviewer for pointing this out. This was a typographical error in the manuscript. The phrase was intended to refer to the predefined “reaction rates for each species” that are scripted within the PHREEQC input files, rather than dynamic rates outputted by COMSOL. We will correct this typo to avoid any further confusion. This will be the specific revision:

“The input files for PHREEQC are automatically generated based on the latest COMSOL outputs and include both solution conditions and the defined reaction rates for each species, while the outputs are parsed back into the pipeline.”

36. Line 328: The text says "For Case A (BTEX), The sampled spac"—this sentence appears to be cut off or incomplete. Please revise.

Response: We thank the reviewer for catching this typographical error. We will correct the capitalization of “the” and ensured the sentence is complete, with all chemical species properly formatted. This will be the specific revision:

“For Case A (BTEX), the sampled space included independent log-uniform distributions for HC, O2, NO3-, SO42-, and CH4, ensuring the model experiences scenarios ranging from fully aerobic to strictly methanogenic conditions.”

37. Line 333: The phrase "a localized batch simulation" is used in reference to PHREEQC. It is unclear what is meant by "localized batch simulation" in this context. The authors should explain how PHREEQC performs such simulations.

Response: We thank the reviewer for pointing this out. The word ‘localized’ was indeed unnecessary and confusing in this context. We will remove it to clarify that it is simply a standard batch simulation. This will be the specific revision:

"For each sampled initial state, a batch simulation was executed using the standard geochemical code PHREEQC."

38. Line 336: The description of data partitioning is insufficient. The authors only mention sampling at fixed time intervals but do not specify how the data are split into training, validation, and test sets.

Response: We thank the reviewer for pointing out this omission. We will update the manuscript to explicitly detail the data partitioning strategy. Specifically, the generated dataset was randomly divided into a training set comprising 80% of the data and a testing set comprising the remaining 20%. This will be added in the revised manuscript:

"Following data generation, the entire dataset was randomly partitioned to train and evaluate the model. Specifically, 80% of the data was allocated to the training set, while the remaining 20% was reserved as the testing set to validate model performance."

39. Line 342: The term "Contaminant concentrations" is used here. However, in the context of this study, are these truly "contaminants" or simply chemical species of interest? The authors should verify whether this terminology is appropriate.

Response: We appreciate the reviewer's careful attention to scientific terminology. While the primary hydrocarbons (e.g., BTEX) and chlorinated solvents (e.g., PCE/TCE) are contaminants, the other entities in our reaction networks, such as electron acceptors and metabolic products, are not contaminants. We will replace the phrase "contaminant concentrations" with "chemical species concentrations" throughout the manuscript to ensure technical precision.

40. Line 351: The authors have not adequately explained why a dual-objective optimization approach is necessary, rather than simply using a single loss function to complete the training. The rationale for this design choice should be clarified.

Response: We appreciate the reviewer's insightful question. While we briefly introduced the rationale for this design in Lines 349 – 350, we agree that the underlying motivation requires a more explicit explanation. If we relied on a single loss function targeting only the final output, the neural network could adopt a "lazy" optimization path: the P-GNN might predict completely unphysical, arbitrarily large kinetic rates,

leaving the MLP corrector to mathematically force the final concentrations to match the labels. This would completely bypass the physical constraints of the kinetic layer. By employing a dual-task objective, Task 1 strictly forces the P-GNN to learn physically realistic kinetic parameters that independently satisfy mass conservation laws, while Task 2 restricts the MLP to making only minor residual corrections for thermodynamic subtleties. We will expand the explanation in the manuscript to clarify this vital design choice:

"To strictly enforce the physics-informed inductive bias while allowing for minor thermodynamic flexibility, PHYGNET is trained end-to-end using a dual-task objective function. A single loss function targeting only the final output would allow the network to bypass physical laws; the P-GNN could potentially output absurdly large or unphysical maximum reaction rates, relying entirely on the MLP corrector to mathematically fit the data."

41. Line 355: It is incorrect to state that MSE "ignores" low-concentration species, as the MSE is by definition computed over all samples. What the authors likely intend to express is that the error contribution from low-concentration species is imbalanced (i.e., overshadowed by high-concentration species). Furthermore, this imbalance is not due to the continuity of groundwater dynamics, but rather because different species concentrations span multiple orders of magnitude, and the successive reactions lead to uneven concentration distributions of intermediate products. The explanation should be revised accordingly.

Response: We sincerely thank the reviewer for this precise mathematical correction. The reviewer is correct that MSE computes over all species, and the issue is not that it "ignores" them, but that the loss contributions from low-concentration species are numerically overshadowed by those of higher concentrations. Furthermore, we agree that this imbalance stems directly from the multi-log-order disparity in concentration distributions caused by successive reactions, rather than just the "sequential nature" of the kinetics. We will revise the explanation to accurately describe this numerical imbalance and its true physical and mathematical causes:

"Because species concentrations in these successive reaction networks inherently span multiple orders of magnitude, a standard Mean Squared Error (MSE) loss suffers from severe numerical imbalance. The loss contributions are heavily dominated by high-concentration parent compounds, which disproportionately overshadows the errors associated with trace intermediate products."

42. Line 362: What are the specific values or settings of the weights used in the loss function? The authors should provide detailed information on how these weights are determined.

Response: We thank the reviewer for pointing out this omission. The weights were determined based on the sequential length of the reaction chain to progressively penalize error accumulation in trace downstream daughter products. We will explicitly add the specific weight vectors used for both Case A and Case B in the revised manuscript:

"Where ω_i is a manually assigned species-specific weight vector. Specifically, for the BTEX competitive system (Case A), the weight vector for [HC, O_2, NO_3^-, SO_4^{2-}, CH_4] was set to [1.0, 1.0, 1.0, 1.0, 2.0]. For the sequential PCE dechlorination system (Case B), to strongly penalize "VC stall" mispredictions, the weight vector for [PCE, TCE, DCE, VC, Ethene] was scaled progressively as [1.0, 1.0, 2.0, 3.0, 3.0]."

43. Line 368: The figure caption appears to mention the split between training and validation sets for the first time, but no details are provided regarding how this split was performed. The authors should describe the data partitioning strategy clearly.

Response: It will be revised in the comment 38.

44. Line 372: The description of the relationship between the surrogate model error and the reactive transport model error needs clarification. Improving the prediction accuracy of PHYGNET is intended to reduce the error propagation from the surrogate model when coupled with the RTM, rather than reducing the modeling error of the RTM itself. The authors should revise this statement to avoid misunderstanding.

Response: We thank the reviewer for this precise and valuable clarification. We will revise this statement to accurately reflect this relationship and avoid any potential misunderstanding:

"To apply PHYGNET to RTMs, the first step is to ensure that the surrogate itself achieves highly accurate simulation results. This high baseline accuracy is essential to minimize the propagation of surrogate-induced errors during the iterative spatial and temporal integration when coupled with the transport model."

45. Line 400: The phrase should read "the concentration of CH_4 is" (grammatical correction).

Response: We thank the reviewer for this precise clarification. We will revise this sentence.

46. Line 406: The statement that PHREEQC data show "values near or below analytical detection thresholds" may not be appropriate or accurate in this context. The authors should reconsider this wording.

Response: We sincerely thank the reviewer for pointing out this inaccurate phrasing. Because PHREEQC is a numerical solver, its outputs are governed by computational and floating-point precision limits, not physical "analytical detection thresholds" (which apply to laboratory instruments). We will revise the text to accurately state that these discrepancies occur at extreme trace concentration levels that approach the numerical precision limits of the solver:

"Despite these visible log-scale deviations at the lower bounds, the discrepancies are largely confined to extreme trace concentration levels approaching the numerical precision limits of the solver, where computational round-off and discretization effects naturally dominate (Hammond et al., 2014)"

47. Line 414: The summation symbol Σ should encompass the entire expression within the parentheses. Please check and correct the mathematical notation.

Response: We thank the reviewer for pointing out this typographical error in our mathematical notation. We will correct the expression to ensure the summation symbol properly encompasses the entire group of C_2 species.

48. Throughout the manuscript: The spelling and capitalization of "ethene" should be checked and unified throughout the text.

Response: We appreciate the reviewer's careful attention to detail. We have thoroughly checked the manuscript and unified the spelling and capitalization of

“ethene”. It will be consistently written in uppercase (“Ethene”) throughout the running text, except where standard formatting requires capitalization (such as at the beginning of a sentence or within specific figure labels)

49. Line 433: The word "Specially" appears to be used incorrectly. Did the authors mean "Specifically"? Please correct if applicable.

Response: We appreciate the reviewer’s careful attention to detail. However, this is “Specifically”, but “Specially”, because we have deliberately chosen a complex fracture medium as our example.

50. Line 439: The format "Figure: 5." is incorrect. It should be written as "Figure 5" or "Fig. 5" following standard formatting conventions.

Response: Thank you for your careful review. We will correct all instances of this throughout the manuscript.

51. Line 446: The phrase "are treated to be" is grammatically awkward. It should be revised to "are treated as" or "are considered to be."

Response: Thank you for your careful review. We will correct it based on your comment.

52. Line 448: The phrase "capture sharp hydraulic gradients" should be checked for clarity and accuracy in context.

Response: We thank the reviewer for pointing out this imprecise phrasing. The local mesh refinement around the fractures is primarily designed to accurately resolve the sharp transitions in flow velocity and the resulting localized concentration gradients at the fracture-matrix interfaces. We will revise this phrase to accurately reflect the physical processes being resolved and to ensure consistency with the description provided in the caption of Figure 6b:

“The mesh is generated with local refinement around the fractures to accurately capture the sharp velocity and concentration gradients at the fracture-matrix interfaces (see Fig. 5b).”

53. Line 485: The phrase "surrogate for the traditional solve" is missing an article. It should be revised to, for example, "surrogate for the traditional solver" or "surrogate for the traditional solution."

Response: We thank the reviewer for their careful reading and helpful suggestion. We have double-checked this sentence and ensured that it reads exactly as recommended to maintain grammatical precision. The phrase will be updated to “a surrogate for the traditional solver”.

54. Line 534: What is meant by "knowledge distillation"? This term has not been introduced or defined previously in the manuscript. The authors should clarify its meaning and relevance in this context.

Response: We agree with the reviewer that the term “knowledge distillation” is confusing and inappropriate in this context. Upon further reflection, we realized that this entire sentence is redundant and does not accurately capture our methodology. To avoid any misunderstanding and improve the clarity of the text, we will simply delete this sentence from the revised manuscript.

55. Lines 568 – 578: These two paragraphs appear to convey the same or very similar meaning, resulting in redundancy. The authors should consolidate or remove the repetitive content.

Response: We sincerely thank the reviewer for catching this oversight. To eliminate this redundancy and improve the flow of the manuscript, we will delete the first repetitive paragraph and retained the second one, which provides a more natural transition.

56. Line 589: The term "krate" appears to be used for the first time here. Has this variable been defined earlier? If not, it should be introduced and defined properly.

Response: Thank you for your careful review. It should be “k_{max}”. We will correct it based on your comment.

57. Line 621: The phrase "could alleviates" is grammatically incorrect—"could" must be followed by the base form of the verb. It should be corrected to "could alleviate."

Response: Thank you for your careful review. We will correct it based on your comment.

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.