Skepthical review: *Mapping Interfacial Water States on Functionalized Graphene: A Machine Learning-Augmented Approach to Uncover Design Principles for Tunable Water Transport*

Summary

This manuscript analyzes a dataset of **91** pre-computed molecular dynamics (MD) simulations of water confined between functionalized graphene sheets, spanning five surface chemistries (UNFUNC, CH_3, CO, OH, COOH), several functional-group coverages, and NaCl loadings (Sec. II.1). From z-resolved water density profiles and existing trajectory analyses, the authors extract diffusion coefficients and a set of structural descriptors (primarily the first-layer density peak height and peak position). The paper then follows a pipeline of (i) exploratory and systematic parametric analysis (heatmaps and "extreme case" comparisons; Sec. II.2–II.3, Sec. III.1–III.3), (ii) K-means clustering to define discrete "interfacial water states" (Sec. II.4, Sec. III.4), and (iii) an interpretable gradient-boosting regression (XGBoost) with SHAP to rank the influence of design parameters (functionalization, coverage, salt) on diffusion (Sec. II.4, Sec. III.5).

The topic is timely and the overall workflow (EDA → clustering → interpretable regression) is a sensible approach to distill a multidimensional simulation sweep into design-oriented narratives. However, several issues currently limit scientific reliability and reproducibility: (1) contradictory diffusion statistics and unit/exponent inconsistencies across sections and clusters; (2) insufficient MD/provenance and diffusion-estimation details; (3) regression evaluation performed in-sample without cross-validation, risking overfitting while being used for design-rule claims; (4) clustering presented as a **10**-state "atlas" despite effectively using only two structural features due to parsing failures and limited robustness/justification of k; and (5) unclear definitions/units for "coverage" and "salt concentration," which reduces transferability and makes some figures hard to interpret. Addressing these would substantially strengthen confidence in the quantitative claims (e.g., "five-fold tunability"), the interpretability conclusions (SHAP rankings), and the generality claims about an "atlas" of interfacial water behavior (Abstract, Sec. I, Sec. IV).

Strengths

- Timely and relevant design question at the intersection of interfacial water physics, nanofluidics, and membrane/materials design, with a coherent narrative arc from physical observables to design principles (Sec. I, Sec. IV).
- Useful dataset breadth for a controlled sweep (**91** systems across functionalization type, coverage, and NaCl loading), enabling systematic comparisons beyond single-case studies (Sec. II.1–II.3).

- Clear visual summaries (heatmaps, density-profile comparisons, extreme-case contrasts) that help connect parameter choices to both structure (density layering) and dynamics (diffusion) (Sec. III.2–III.3).
- Interpretability-forward ML framing (SHAP with gradient boosting) aligned with the stated goal of ranking controllable design parameters rather than purely maximizing predictive accuracy (Sec. II.4, Sec. III.5).
- The core qualitative trends (diffusion suppression by increasing salt; strong slowing for COOH; secondary effects of coverage) are physically plausible and potentially useful as hypotheses for future simulation/experimental design (Sec. III.2–III.5).
- Limitations are at least partially acknowledged (e.g., failed RDF and bulk_density parsing), which provides a foundation for improving rigor in a revision (Sec. III.6).

Major issues

1. **Diffusion coefficient values, ranges, and even exponents are internally inconsistent across the manuscript, undermining the central quantitative claim of "∼five-fold tunability" and creating ambiguity about which diffusion definition is used (Sec. II.2/Table I vs Sec. III.1 vs Sec. IV; also cluster descriptions in Sec. III.4). For example, Table I reports 0.61–3.84 $\times 10^{-5}$ cm^2/s (mean 2.15 $\times 10^{-5}$), while Sec. III.1 and Sec. IV report 0.40–1.98 $\times 10^{-5}$ cm^2/s (mean 1.17 $\times 10^{-5}$); additionally, Cluster 9 is reported as 0.55 $\times 10^{-6}$ cm^2/s (Sec. III.4), conflicting with the global minima and the rest of the manuscript's 10^{-5} scale.**

   Recommendation: Recompute diffusion summary statistics (min/median/max/mean/SD, ideally also percentiles) directly from the exact master DataFrame used downstream (heatmaps, clustering, regression). Establish a single authoritative diffusion variable definition (units, scaling, and whether it is in-plane D_{xy} vs 3D D) and use it consistently in Table I (Sec. II.2), Sec. III.1, cluster summaries (Sec. III.4), and the Conclusions (Sec. IV). If multiple diffusion measures exist, label them explicitly (e.g., $D_{\parallel}$, D_{3D}) and keep their statistics separate. Verify and correct the Cluster 9 exponent and any plot/table scaling factors (e.g., whether axes display "$\times 10^{-5}$").

2. **Reproducibility and MD provenance are insufficiently described, and "public availability" is undermined by local absolute file paths (e.g., "/Users/osman_mbp/...") and missing simulation/analysis parameters (Sec. II.1). Key details needed to assess diffusion reliability in confinement are absent: force fields and water/ion models, geometry and channel height, boundary conditions, thermostat/barostat and ensemble, timestep, equilibration/production lengths, sampling cadence, functionalization protocol (random vs patterned; one-/two-sided), and how diffusion is computed (directionality, MSD fit window, drift removal, uncertainty estimation).**

Recommendation: Replace all machine-specific paths in Sec. II.1 with repository-neutral descriptions and provide an accessible archive (GitHub/Zenodo/DOI) containing (at minimum) analysis scripts and metadata mapping each system to its parameters and trajectory/source. Add a concise but complete MD methods paragraph in Sec. II.1 covering: system geometry/dimensions, graphene separation, functionalization placement protocol, water and ion models/parameters, force field details for graphene and functional groups, thermostat/barostat settings, timestep, equilibration/production durations, and any constraints. Add a clear diffusion-estimation protocol: whether D is computed parallel to the walls (recommended for confinement) or in 3D, the MSD time interval used for the linear fit, any block-averaging/CI or replicate strategy, and how uncertainties are handled (or explicitly state point estimates only). Temper any "publicly available" language if full data cannot be released.

3. **The supervised ML model (XGBoost) is evaluated in-sample (trained and assessed on the same 91 points) while being described as "excellent performance," and SHAP-based importance is used to support design principles without quantifying generalization or stability (Sec. II.4, Sec. III.4–III.5). With a small dataset and one-hot categorical variables, this creates a substantial overfitting/circularity risk: strong-looking predicted-vs-actual plots may reflect memorization rather than robust trends.**

   Recommendation: Introduce a validation protocol in Sec. II.4: at minimum repeated k-fold cross-validation (or LOOCV) with reported MAE/RMSE/R^2 on held-out folds. Recompute/aggregate SHAP results across folds (training-only per fold) and report stability of the feature ranking (e.g., rank frequencies or mean±SD SHAP importance across resamples). Consider adding a simple baseline model (e.g., linear/GLM with main effects and selected interactions) to show that the qualitative hierarchy (salt, COOH, coverage) is not an artifact of boosted trees. If predictive generalization is not a goal, reframe the regression explicitly as descriptive and soften claims tied to "performance."

4. **Clustering is presented as a 10-state "interfacial water atlas," but the implemented feature set effectively collapses to only two density-derived features (density_peak_height and density_peak_position) because RDF features and bulk_density parsing failed (Sec. II.4 vs Sec. III.6). With only 2D features and 91 samples, $k = 10$ risks over-partitioning a continuous trend into arbitrary bins; additionally, selecting $k = 10$ because it maximizes silhouette at the upper tested bound is not a robust model-selection justification (Sec. III.4.1, Fig. 9).**

   Recommendation: Make the feature set used for clustering explicit and consistent in Sec. II.4 and Sec. III.4.1 (move the parsing-failure disclosure earlier than Sec. III.6). Provide the full silhouette (and preferably Davies–Bouldin) curves over a wider k range and justify k based on a plateau/elbow and interpretability rather than the maximum tested endpoint. Add robustness checks: re-run clustering with fewer k

(e.g., **3–6**), show whether the key physical dichotomy (mobile/disordered vs trapped/ordered) persists, and/or compare with alternative clustering (GMM/hierarchical/DBSCAN). Ideally, fix RDF/bulk_density extraction and rerun; if not feasible, substantially temper "**10** distinct states/atlas" language and reframe as "density-profile-based regimes."

5. **Key control variables are not defined in transferable physical units, and this propagates into figure interpretability and generalizability: "coverage" is alternately described as a percentage and as an integer number of groups (Sec. II.2–II.3, Sec. III.2; Fig. 4–Fig. 6 captions), and salt is sometimes expressed as "NaCl pairs" without normalization by volume (Sec. II.1–II.3; multiple figures). This makes trends difficult to compare across geometries and undermines the "design principles" framing.**

   Recommendation: Define coverage canonically in one physical measure (preferred: groups per nm^2 or % of functionalizable sites) and provide an explicit mapping between "N groups" and "%" based on the graphene surface area and site count used. Standardize all axes/captions accordingly. Express salt in molarity or number density (ion pairs per nm^3) and optionally include the raw "NaCl pairs" in parentheses. Ensure all heatmaps and comparisons specify the fixed slice values with clear units (Sec. II.3, Sec. III.2.2).

6. **The interaction-effect analysis ("synergistic/antagonistic" deviations from an additive baseline) is not defined with sufficient mathematical precision, contains at least one arithmetic inconsistency, and is conceptually disconnected from the SHAP framework used elsewhere (Sec. III.5; Fig. 13). The additive baseline (what is averaged over, how categories are treated, whether it is a fitted model or marginal means) is unclear, and the interpretation sometimes aligns more with saturation/floor effects than "synergy."**

   Recommendation: In Sec. III.5, explicitly define the baseline used to compute interaction terms (equation, averaging sets, categorical handling, and uncertainty if any). Correct the numerical example where $\mathbf{0.8 - 1.2}$ is reported as $\mathbf{-0.3}$ instead of $\mathbf{-0.4}$ (Sec. III.5). Consider aligning interaction analysis with the ML model by reporting SHAP interaction values for the tree model, or alternatively fitting a simple linear/GLM model with and without interaction terms and comparing coefficients and fit; then describe effects as "positive/negative deviation from additivity" rather than "synergy" unless you define these terms rigorously.

7. **Claims of broad generality (e.g., "quantitative atlas," broadly applicable "design principles") are stronger than warranted by the study's restricted domain: single channel/pore geometry, limited functional group set, only**

NaCl, and classical non-polarizable MD (Abstract, Sec. I, Sec. IV). The limitations discussion focuses on parsing issues but does not sufficiently bound the physical domain of validity (Sec. III.6).

Recommendation: In Sec. III.6 and Sec. IV, clearly delimit the domain of validity: specify the channel geometry/size, functionalization chemistries, electrolyte type/range, and interaction model limitations. Rephrase "atlas"/"design principles" as applicable within this parameter space, and identify which qualitative trends you expect to be robust vs sensitive to geometry, electrolyte identity, or polarization/quantum effects. This will improve credibility without diminishing the paper's usefulness as a template workflow.

Minor issues

1. Methods/Results structure and cross-referencing are inconsistent (mixing Roman numerals, letters, and numeric subsection references; some incorrect references such as calling the regressor description "Section **2.2**" when it appears in Sec. II.4) (Sec. II, Sec. III).

   Recommendation: Standardize section/subsection numbering (e.g., Sec. II.1–II.4; Sec. III.1–III.6) and update all in-text references to match. Fix incorrect cross-references (e.g., in Sec. III.4.2/III.5 referencing the ML Methods).

2. Figure labeling and caption clarity need tightening: several figures are referenced as "Figure ??" (Sec. III.1) and multiple plots/captions do not consistently state units/scaling for D, salt, and coverage or the meaning of error bars.

   Recommendation: Resolve all figure-number placeholders, standardize diffusion-unit formatting (cm^2/s with explicit $\times 10^{-5}$ scaling if used), and specify in each caption: (i) n (number of systems in each aggregate/bin), (ii) what error bars represent (SD/SE/CI), and (iii) the exact fixed parameters for any sliced heatmap (Sec. II.3, Sec. III.2.2).

3. XGBoost/SHAP implementation details are incomplete (hyperparameters, early stopping, encoding scheme/reference category, software versions), limiting reproducibility and interpretability (Sec. II.4, Sec. III.4.2).

   Recommendation: Add a concise hyperparameter/configuration table in Sec. II.4 (max_depth, n_estimators, learning_rate, subsample, colsample_bytree, regularization, random seed) and specify categorical encoding and reference categories for functionalization. Provide library versions for xgboost and shap.

4. Clustering results are described in depth for only a small subset of clusters, making it hard to evaluate whether the **10** clusters add scientific value beyond "high mobility vs low mobility" (Sec. III.4).

Recommendation: Add a compact table (main text or SI) listing for each cluster: number of systems, centroid (peak height/position), mean±SD diffusion, and dominant ranges of functionalization/coverage/salt. Optionally include one representative density profile per cluster (or per major regime).

5. Mechanistic language (e.g., "ice-like," "premelting-like," "trapped-immobile") is sometimes stronger than supported by the presented observables, especially given the reliance on density peaks after RDF/bulk-density parsing failures (Sec. III.2–III.3, Sec. III.5).

   Recommendation: Either support these interpretations with additional structural/dynamical metrics available from trajectories (e.g., tetrahedrality/order parameters, hydrogen-bond statistics, residence times, ion distributions), or temper the wording to clearly indicate qualitative analogy rather than demonstrated phase-like ordering.

6. Choice of fixed slices in heatmaps (e.g., coverage fixed at **24%**, functionalization fixed to CH_3, salt fixed at **18** pairs) is not justified and may bias perception of trends (Sec. II.3, Sec. III.2.2).

   Recommendation: Briefly justify these choices as mid-range/representative or add a supplementary panel showing that alternative slice values give qualitatively similar conclusions (or explicitly note where they differ).

7. Nomenclature inconsistencies for functionalization labels (e.g., UNFUNC vs 0UNFUNC; CO vs carbonyl; feature-name variants) can confuse mapping between simulations, plots, and one-hot features (Sec. II.1, Sec. III.2, Sec. III.4.2).

   Recommendation: Provide a single naming table in Sec. II.1 mapping (i) simulation labels, (ii) figure labels, and (iii) ML feature names for each functional group, and standardize usage throughout.

8. The limitations section notes parsing failures but does not quantify the scope (how many systems affected) nor discuss how missing bulk_density/RDF may bias conclusions about "interfacial" vs "bulk-like" behavior (Sec. III.6).

   Recommendation: State explicitly whether the parsing issue affected all **91** systems and describe the likely impact (e.g., inability to distinguish similar first-layer peaks with different mid-channel densities). If feasible, fix and rerun; otherwise state how this constrains interpretation.

Very minor issues

1. Typographical/LaTeX inconsistencies: unit formatting (cm^2/s vs $cm^2\ s^{-1}$), mixed quote styles, Å formatting artifacts, inconsistent capitalization in system labels (nacl vs NaCl), and minor spacing/punctuation issues (various sections).

Recommendation: Perform a formatting pass to standardize units, typography, Å notation, label capitalization, and spacing; ensure consistent significant figures aligned with reported uncertainty.

2. Acronyms and state labels are sometimes introduced without definition or consistent styling (e.g., SHAP in the Abstract; inconsistent capitalization/quoting of state names) (Abstract, Sec. I, Sec. III.4, Sec. IV).

   Recommendation: Define acronyms on first use (Abstract and main text) and adopt a consistent style for named states (e.g., Title Case without quotes) while reserving quotes for qualitative descriptors (e.g., "ice-like").

3. Citation and section-reference formatting is inconsistent (e.g., "[1; 2]" vs "[1,2]"; "Sec. II.B" vs "Sec. II.2"), which slightly reduces polish (Sec. I–II; References).

   Recommendation: Normalize citation formatting to the target venue's style and use a single consistent convention for section references throughout.

Key statements and references

- ✔ **Graphene's surface can be chemically functionalized with various chemical groups to modify surface wettability and interaction potentials, thereby altering the behavior of confined water molecules in nano-confined environments such as graphene-based materials [3; 4].**
- *Reference(s):* **3**, **4**
- *Justification:* Reference **4** shows that graphene oxide, bearing oxygen functional groups, is highly hydrophilic and exhibits strong water–surface interactions; MD and experiments demonstrate water molecules cluster near functional groups, form H-bonds, and that varying oxidation level/defects changes adsorption capacity and dynamics—i.e., chemical functionalization tunes wettability and interaction potentials, altering confined water behavior. Reference **3** demonstrates that water behavior in graphene-based nanoconfinement (adsorption/depletion, orientation, PMF) is sensitive to surface properties (here, surface charge), supporting that surface modifications alter confined water behavior. Together, they directly support the statement.

- ✖ **The present study leverages a large dataset of 91 pre-computed molecular dynamics simulations of water on functionalized graphene, previously generated to explore how variations in functionalization type, surface coverage, and salt concentration affect water structuring and dynamics [5; 6; 7].**

- *Reference(s):* **5**, **6**, **7**
- *Justification:* None of the cited papers describe or provide a dataset of **91** pre-computed MD simulations. Ref. **5** studies water permeation in pristine vs. graphene oxide multilayers (no salt-variation dataset). Ref. **6** examines interfacial heat transfer with different functional groups and O/C ratios (no salt, no **91**-simulation dataset). Ref. **7**

models thermal-driven flow in pristine graphene channels with NaCl (not functionalized graphene, no dataset). Therefore the claim about leveraging a **91**-simulation dataset spanning functionalization type, surface coverage, and salt concentration is not supported by **5**, **6**, or **7**.

- **✖ The analysis of diffusion coefficients across all 91 systems shows that the maximum diffusion coefficient (3.84×10^{-5} cm^2/s) is over six times larger than the minimum (0.61×10^{-5} cm^2/s), quantitatively confirming that functionalization type, coverage, and salt concentration can profoundly and tunably impact interfacial water dynamics in line with prior work on confined and interfacial diffusion [12, 13].**
- *Reference(s):* **12**, **13**
- *Justification:* Neither **12** nor **13** analyzes **91** systems or reports diffusion coefficients with the stated extremes ($\mathbf{3.84 \times 10^{-5}}$ vs $\mathbf{0.61 \times 10^{-5}}$ cm^2/s). Ref. **12** develops methods and reports interfacial/bulk diffusion near a water–vapor interface, not across varied functionalization, coverage, or salt conditions. Ref. **13** studies premelting at an ice–rubber vs ice–vapor interface and reports diffusion coefficients $\sim \mathbf{10^{-10}}$ m^2/s for a few temperatures, with no exploration of functionalization type, coverage, or salt concentration. Thus the quantitative claim and its scope are not supported by **12**, **13**.
- **✖ A systematic parametric analysis framework, similar in spirit to prior studies of water and ion transport in low-hydration or confined media, was used to deconstruct the multi-dimensional design space and isolate the individual and pairwise influences of functionalization type, coverage, and salt concentration on water diffusion [14, 15].**
- *Reference(s):* **14**, **15**
- *Justification:* **15** studies SWCNT diffusion and quantum friction under variations such as excitation power, solvent, and functionalization type, but does not present a systematic multi-dimensional parametric framework for water diffusion, nor analyze coverage or salt concentration or their pairwise effects. Thus the claim about deconstructing design space for functionalization type, coverage, and salt concentration on water diffusion is not supported.
- **✖ To move beyond simple correlations, a two-part machine learning methodology inspired by prior physics-informed and nanoconfinement-focused ML studies was implemented: unsupervised clustering to identify interfacial water states and an interpretable Gradient Boosting Regressor with SHAP analysis to extract quantitative design principles for water transport [16, 17].**
- *Reference(s):* **16**, **17**

- *Justification:* Reference **16** presents a physics-informed reinforcement learning framework for interfacial area transport in two-phase flow; it does not use unsupervised clustering, gradient boosting, or SHAP. Reference **17** studies nanoconfined water density profiles using supervised models (random forest, XGBoost, MLP) without unsupervised clustering or SHAP interpretability, and it does not extract design principles for water transport. Thus, the stated two-part methodology is not supported by **16** or **17**.

- ⚠ **The observed five-fold tunability of water diffusion in functionalized graphene channels, together with the identified hierarchy of control parameters (salt concentration > functional group chemistry > surface coverage), is consistent with and extends previous reports of breakdown or enhancement of fast water transport in graphene-based nanochannels [9; 10].**

- *Reference(s):* **9**, **10**
- *Justification:* Refs. **9** and **10** collectively document both breakdown and enhancement of fast water transport in graphene-based nanochannels: **9** shows strong suppression of slip and flow with hydroxyl functionalization (breakdown), while **10** reports ultrafast water permeation and only slight reduction in the presence of **0.1** M $MgCl_2$ (enhancement). However, neither paper reports a five-fold tunability of water diffusion nor establishes the stated hierarchy of control parameters (salt concentration > functional group chemistry > surface coverage). In fact, **9** indicates a large impact of surface coverage, and **10** shows only modest salt effects. Thus the statement is only partially supported.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains little explicit derivation-level mathematics; it primarily defines extracted features (peak heights/positions), reports diffusion-coefficient summaries, and describes ML procedures (K-means clustering, silhouette-score selection, SHAP-based feature attribution). The main audit-relevant issues are internal consistency of reported quantities (especially diffusion statistics and units/exponents) and consistency of variable/feature definitions across Methods, Results, and Limitations.

Checked items

1. ✔ **Diffusion coefficient unit specification** (Sec. II.A.1 (p.**2**); Table I (p.**3**); multiple figure captions (pp.**6**–**8**))
    - **Claim:** The target diffusion coefficient D is reported in cm^2/s (often displayed as $\times 10^{-5}$ cm^2/s).
    - **Checks:** dimensional/units consistency, notation consistency

- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** diffusion_cm2s is the same quantity across sections unless stated otherwise
- **Notes:** Units for D are repeatedly stated as cm^2/s; formatting varies but is interpretable.

2. ✖ **Diffusion summary statistics consistency (Table I vs Results)** (Table I / Sec. II.B (p.3) vs Sec. III.A (p.5) vs Conclusions (p.14))

   - **Claim:** The paper reports a single distribution/range of diffusion coefficients across 91 systems.
   - **Checks:** internal consistency across sections, definition consistency
   - **Verdict:** FAIL; confidence: high; impact: critical
   - **Assumptions/inputs:** Table I and Sec. III.A describe the same dataset of 91 systems
   - **Notes:** Table I reports mean=2.15×10^{-5}, min=0.61×10^{-5}, max=3.84×10^{-5}, whereas Sec. III.A reports mean=1.17×10^{-5}, min=0.40×10^{-5}, max=1.98×10^{-5} and Conclusions repeat that. These are incompatible without an explicit explanation (different subset/definition).

3. ✖ **Feature set used for clustering (Methods vs Results/Limitations)** (Sec. II.D.1 (p.4) vs Sec. III.D.1 (p.10) and Sec. III.F (p.13))

   - **Claim:** K-means clustering is performed on a structural feature set including density peaks, bulk density, and RDF peak height.
   - **Checks:** definition consistency, method/result alignment
   - **Verdict:** FAIL; confidence: high; impact: critical
   - **Assumptions/inputs:** The clustering described in Methods is the same clustering whose results appear in Sec. III.D.1
   - **Notes:** Methods specify 4 structural features (including bulk_density and rdf_peak_height). Later text states RDF feature extraction failed and bulk_density parsed as zero, and clustering was done only with density_peak_height and density_peak_position.

4. ✖ **Coverage variable definition (percent vs group-count)** (Sec. II.C.1 (p.3); Sec. III.B.1 and Figs. 3–6 (pp.6–8))

   - **Claim:** Coverage is a single well-defined input variable used consistently across analysis.
   - **Checks:** notation/definition consistency, dimensional/units consistency
   - **Verdict:** FAIL; confidence: high; impact: moderate
   - **Assumptions/inputs:** Coverage is used as a numeric feature in regression and in heatmap axes

- **Notes:** Coverage is described as a percentage (e.g., **24%** functionalized) and as counts of functional groups (**8, 16, 24** groups). Without a mapping, heatmaps and averages are ambiguous.

5. ✖ **Cluster 9 diffusion magnitude/exponent consistency** (Sec. III.D.1, Cluster 9 description (p.10))
   - **Claim:** Cluster-average diffusion values are reported on the same scale/units as other diffusion values.
   - **Checks:** units/exponent consistency, internal consistency
   - **Verdict:** FAIL; confidence: high; impact: moderate
   - **Assumptions/inputs:** Cluster diffusion coefficient refers to the same diffusion metric D as elsewhere
   - **Notes:** Cluster 9 reports 0.55×10^{-6} cm^2/s, conflicting with the paper's consistent $\times 10^{-5}$ scale and the stated global minima. Likely exponent typo or mismatched definition.

6. ✔ **Interaction term definition** (Sec. III.E and Fig. 13 caption (pp.12–13))
   - **Claim:** Interaction effect is defined as deviation from additive prediction: (actual diffusion − additive-model predicted diffusion).
   - **Checks:** algebraic form check, units consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Additive-model prediction is in the same units as actual diffusion
   - **Notes:** The expression 'actual − predicted' is algebraically coherent and yields diffusion units.

7. ✖ **Interaction example arithmetic** (Sec. III.E, antagonistic example for 0UNFUNC_45nacl (p.13))
   - **Claim:** Given actual=0.8×10^{-5} and predicted=1.2×10^{-5}, the interaction term is -0.3×10^{-5} cm^2/s.
   - **Checks:** algebra/arithmetic consistency
   - **Verdict:** FAIL; confidence: high; impact: minor
   - **Assumptions/inputs:** Interaction term equals actual − predicted as stated
   - **Notes:** By the stated definition, $0.8 - 1.2 = -0.4$ (in $\times 10^{-5}$ units), not -0.3.

8. ⚠ **SHAP value dimensionality** (Sec. III.D.2 (p.11–12))
   - **Claim:** Mean |SHAP value| magnitudes (e.g., 3.07×10^{-6}) quantify contributions to diffusion predictions.
   - **Checks:** dimensional/units consistency, definition consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor

- **Assumptions/inputs:** SHAP values are in the same units as the model output (diffusion)
- **Notes:** The text reports SHAP magnitudes without explicitly attaching diffusion units; this is plausibly fine, but the paper does not explicitly confirm the SHAP values' unit interpretation in-text.

9. ⚠ **Bulk density definition vs later parsing failure** (Sec. II.A.2 (p.2) vs Sec. III.F (p.13))
    - **Claim:** bulk_density is computed as average density in the middle 10 Å and used as a meaningful feature.
    - **Checks:** definition consistency, pipeline consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** The feature engineering described in Methods was successfully applied to produce the dataset
    - **Notes:** Methods define bulk_density, but Limitations state bulk_density was consistently parsed as zero due to a scripting/definition issue. The paper does not clearly state whether bulk_density was excluded from all downstream analyses (except clustering) or retained as a degenerate feature.

Limitations

- The provided PDF content contains almost no explicit equations/derivations (e.g., no MSD→diffusion Einstein relation is written), so algebraic step-by-step verification of derivations is largely not possible.
- Many quantitative claims are presented as reported values (means/ranges/examples) without formula definitions; the audit therefore focuses on internal consistency and dimensional/notation coherence rather than derivation correctness.
- Some figure references are placeholders (e.g., "Figure ??"), limiting precise cross-location verification of some claims.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Eight numerical/logic checks were run. Four passed (range ratio > 6 in Table I; quartile ordering; mean within min–max in Table I; and "five-fold" range consistency for 0.40 to 1.98×10^{-5}). Four failed, driven by cross-section inconsistencies between Table I and Results/Conclusions (min/max and mean/SD), a unit/scale contradiction for Cluster 9 versus the stated global minimum, and an arithmetic mismatch in an interaction-term example.

Checked items

1. ✔ **C1_range_ratio_tableI** (Page 3, Section II.B + Table I)

- **Claim:** “maximum diffusion coefficient (3.84×10^{-5} cm^2/s) being over six times larger than the minimum (0.61×10^{-5} cm^2/s)”
- **Checks:** ratio_check
- **Verdict:** PASS
- **Notes:** Computed ratio $3.84/0.61 = 6.295081967...$, which satisfies “over six times”.

2. ✔ **C2_tableI_quartile_ordering** (Page 3, Table I)
   - **Claim:** Table I summary statistics should be ordered consistently: Min $\leq$ 25 th $\leq$ Median $\leq$ 75th $\leq$ Max.
   - **Checks:** monotonic_order_check
   - **Verdict:** PASS
   - **Notes:** Sequence is nondecreasing: $0.61 \leq 1.59 \leq 2.11 \leq 2.76 \leq 3.84$.

3. ✔ **C3_tableI_mean_within_range** (Page 3, Table I)
   - **Claim:** Mean diffusion coefficient (2.15×10^{-5} cm^2/s) should fall within [min,max].
   - **Checks:** range_inclusion_check
   - **Verdict:** PASS
   - **Notes:** Mean 2.15 lies within $[0.61, 3.84]$ in $\times 10^{-5}$ units.

4. ✔ **C4_results_range_ratio_fivefold** (Page 5, Section III.A; Page 14, Conclusions)
   - **Claim:** Diffusion range “from 0.40×10^{-5} cm^2/s to 1.98×10^{-5} cm^2/s” is described as “approximately five times greater / five-fold range.”
   - **Checks:** ratio_check
   - **Verdict:** PASS
   - **Notes:** Computed ratio $1.98/0.40 = 4.95$; relative difference from 5.0 is 1%, within the 10% tolerance for “approximately five-fold”.

5. ✖ **C5_inconsistency_tableI_vs_results_range_minmax** (Page 3 Table I vs Page 5 Section III.A vs Page 14 Conclusions)
   - **Claim:** Table I reports min/max 0.61 and 3.84 ($\times 10^{-5}$ cm^2/s), while Results/Conclusions report min/max 0.40 and 1.98 ($\times 10^{-5}$ cm^2/s). These should be consistent if they describe the same dataset of 91 systems.
   - **Checks:** cross_section_consistency_check
   - **Verdict:** FAIL
   - **Notes:** Min mismatch: $0.61 - 0.40 = 0.21$ ($\times 10^{-5}$). Max mismatch: $3.84 - 1.98 = 1.86$ ($\times 10^{-5}$).

6. ✖ **C6_inconsistency_tableI_vs_results_mean_sd** (Page 3 Table I vs Page 5 Section III.A)

- **Claim:** Mean/SD differ: Table I mean 2.15 and SD 0.78 ($\times 10^{-5}$) vs Results mean 1.17 and SD 0.41 ($\times 10^{-5}$). If both refer to "all 91 systems" diffusion coefficients, they should match.
- **Checks:** cross_section_consistency_check
- **Verdict:** FAIL
- **Notes:** Mean mismatch: $2.15 - 1.17 = 0.98$ ($\times 10^{-5}$). SD mismatch: $0.78 - 0.41 = 0.37$ ($\times 10^{-5}$).

7. ✖ **C7_cluster9_unit_consistency** (Page 10, Section III.D.1 (Cluster 9 description))
   - **Claim:** Cluster 9 average diffusion: "0.55×10^{-6} cm^2/s" should be on the same unit scale as other diffusion values (typically $\times 10^{-5}$ cm^2/s). Convert and compare plausibility relative to reported minima (0.40×10^{-5}).
   - **Checks:** unit_conversion_and_comparison
   - **Verdict:** FAIL
   - **Notes:** 0.55×10^{-6} converts to 0.055×10^{-5}, which is below the reported minimum 0.40×10^{-5} (difference 0.345×10^{-5}).
8. ✖ **C8_interaction_term_arithmetic_UNFUNC_45nacl** (Page 13, Section III.E (antagonistic interactions example))
   - **Claim:** For 0UNFUNC_45nacl: actual 0.8×10^{-5}, predicted 1.2×10^{-5}, interaction term reported -0.3×10^{-5}; but actual−predicted equals -0.4×10^{-5}.
   - **Checks:** difference_check
   - **Verdict:** FAIL
   - **Notes:** Computed actual−predicted $= -0.4 \times 10^{-5}$; reported interaction is -0.3×10^{-5}, outside the $\pm 0.05 \times 10^{-5}$ tolerance.

Limitations

- Only the provided PDF text/images were available; no underlying dataset tables (all 91 diffusion coefficients, per-parameter group means) are included for recomputation.
- Checks avoid extracting numbers from plots/heatmaps because pixel-based value extraction is out of scope.
- Several internal consistency checks can only flag mismatches between reported numbers; they cannot determine which section is correct without raw data.
- Some claims depend on values shown in heatmaps/plots and cannot be numerically verified without image-based extraction.
- Some global claims (e.g., monotonic trends across all systems, cluster totals, silhouette-score optimal k, SHAP importances) cannot be verified without full per-system data or model outputs.