

Machine-Learned Hydraulic Conductivity Beyond Random Splits

Abstract

What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions? This review answers by treating extrapolation-aware geotechnical prediction as a property of a sociotechnical workflow rather than a feature that can be read from average accuracy. The focal setting is sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, where a simplified or learned representation can look accurate while omitting the physical regime that determines transfer. Evidence from the assigned publications is synthesized with foundational studies of calibration, distribution shift, causal structure, and responsible deployment. Four requirements follow: preserve the lineage of instrument observations, constitutive parameters, geometry, loading history, simulation output, and validation measurements; measure stability across relevant perturbations; connect confidence to a specific action; and maintain a route for human challenge and correction. The framework distinguishes descriptive performance from decision utility and separates uncertainty about the world from uncertainty created by the model and its evaluator. It also shows why faster inference or richer reasoning is valuable only when it improves a defined decision under a transparent resource budget. The article is a literature review and research agenda, not a report of a newly completed trial.

Introduction

What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions? The difficulty begins with a basic observation: prediction is only one stage of a deployed workflow. Before an output appears, designers have already decided what counts as evidence, which representations are admissible, how alternatives are compared, and how uncertainty changes action. In sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, these choices interact with institutions and physical processes that the model only partially represents. A high score can coexist with fragile inputs, an evaluator that rewards the wrong surface feature, or an action rule that turns modest uncertainty into a costly decision. The resulting argument frames extrapolation-aware geotechnical prediction as an evidence problem. The inquiry focuses on the minimum evidence and comparison set required for a credible claim. This requirement gives practical content to the article's central question: What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions?

The argument proceeds from an operational sequence perspective. Its unit is the chain connecting boundary conditions, material state, experiment or simulation, reduced representation, and engineering decision. The benefit is that it avoids attributing every problem to the predictor: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. This framing places the focal studies on comparable evidential terms. Although their application settings differ, they each make some part of an inference chain more explicit - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The synthesis therefore asks whether that explicit component remains meaningful when instrument observations,

constitutive parameters, geometry, loading history, simulation output, and validation measurements change, and whether the proposed safeguards lead to bounded extrapolation, targeted experiment, model revision, and conservative design screening rather than merely producing another metric. This requirement gives practical content to the article's central question: What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions?

Evidence From the Focal Studies

A useful test case is Prediction of Hydraulic Conductivity of Sodium Bentonite GCLs by Machine Learning Approaches. Rather than treating its output as an isolated score, the study frames a specific problem: how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions. Its central mechanism is comparison of six machine-learning approaches on compiled literature data using features such as bentonite mass per area, cations, ionic strength, relative abundance, swell behavior, and stress, and its evidential basis is that an 80/20 split identifies XGBoost as the strongest reported approach, including an R-squared of 0.826 and classification accuracy of 0.887 for one feature subset. The work demonstrates the value of multivariate nonlinear prediction for a property governed by coupled chemical and mechanical factors. For sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, the transferable lesson is methodological: a system should preserve the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. Random splitting of heterogeneous literature data can place closely related experiments on both sides of the split, and extrapolation to new bentonites, solutions, stresses, or laboratories remains unproven (Li et al. 2025).

Evidence for extrapolation-aware geotechnical prediction becomes concrete in Research on Simplified Evaluation Method for Soil-Rock Mixed Slope Stability under Dam-Break Flood Impact. The paper addresses a defined problem—how a simplified stability method can retain the dominant hydraulic and mechanical effects of a high-Froude-number flood—through a dam-break impact-pressure expression supported by flume experiments and numerical simulation, coupled with a simplified pore-pressure and stability evaluation. Its evaluation is informative because application to the Wangzongdian landslide case produces a safety factor reported to fall to 0.98 under impact. The study offers a screening route for a hazard whose transient loading is often poorly represented by static slope calculations. This does not make the method a universal component; it identifies a design hypothesis that must be re-tested in the article's focal setting. In particular, one principal case, simplified material behavior, uncertain hydraulic parameters, and limited representation of erosion and progressive failure restrict direct transfer. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Huang et al. 2025).

A Broader Evidence Base

Several established literatures clarify the design space. Combustion simulation requires consistent transport, chemistry, numerics, and boundary conditions (Poinsot and Veynante 2012). Surface forces provide the mechanistic basis for selective colloidal attraction, repulsion, and adsorption (Israelachvili 2011). Colloidal transport couples surface interactions with hydrodynamics and concentration (Russel, Saville, and Schowalter 1989). Classical soil mechanics links pore pressure, effective stress, and stability (Terzaghi, Peck, and Mesri 1996). Together these findings imply that extrapolation-aware geotechnical prediction

should be evaluated against explicit alternatives and failure costs. In *Machine-Learned Hydraulic Conductivity Beyond Random Splits*, this literature supplies a specific test of extrapolation-aware geotechnical prediction within sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions.

The evidence base offers complementary tests rather than one universal metric. Geosynthetic design connects material characterization to field boundary conditions and durability (Koerner 2012). Random forests provide a nonlinear baseline for mixed physical and chemical predictors (Breiman 2001). Gradient boosting is a strong comparator for structured engineering data (Friedman 2001). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In *Machine-Learned Hydraulic Conductivity Beyond Random Splits*, this literature supplies a specific test of extrapolation-aware geotechnical prediction within sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions.

A credible assurance case can be built from findings that operate at different levels. Attribution can diagnose learned associations but cannot replace mechanism (Lundberg and Lee 2017). FAIR data practice supports reproducible physical-model comparisons (Wilkinson et al. 2016). Calibration theory separates observation error, parameter uncertainty, and structural discrepancy (Kennedy and O'Hagan 2001). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In *Machine-Learned Hydraulic Conductivity Beyond Random Splits*, this literature supplies a specific test of extrapolation-aware geotechnical prediction within sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions.

A Layered Model of Reliability

The analysis distinguishes evidence, representation, inference, and decision as separate layers. The evidence layer records observations and their provenance. The representation layer turns those observations into features, graphs, tokens, or simulated states. The inference layer produces predictions, explanations, translations, anomaly scores, or candidate actions. The decision layer maps those outputs to bounded extrapolation, targeted experiment, model revision, and conservative design screening. Reliability can fail at any boundary. Better inference cannot repair evidence that omitted the relevant condition, and a calibrated probability cannot rescue an action rule whose costs were never specified. Conversely, a conservative decision policy may reduce harm even when the underlying model remains imperfect. This layered view makes extrapolation-aware geotechnical prediction testable because each claim can be assigned to a component and a measurable transition. The implication is especially consequential in sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, where a small modeling choice can redirect attention and resources.

A further distinction separates aleatory variation, epistemic uncertainty, and strategic adaptation. Aleatory variation concerns irreducible differences among cases. Epistemic uncertainty concerns missing knowledge, limited samples, or unsupported regions of the input space. Strategic adaptation occurs when users, adversaries, markets, or institutions respond to the deployed workflow itself. These sources demand different controls. Repeated measurement can characterize variation; new evidence and conservative extrapolation can reduce epistemic uncertainty; red teaming and feedback-aware assessment address adaptation. Combining them into one confidence score hides which intervention is appropriate. A defensible system therefore reports uncertainty in a form tied to an available response rather than presenting confidence as a decorative number. The implication is especially consequential in sodium-bentonite geosynthetic clay

liners exposed to saline or leachate solutions, where a small modeling choice can redirect attention and resources.

An Evaluation Protocol

The action and its costs should determine the reported metrics. Discrimination measures whether cases are ranked correctly; calibration asks whether confidence corresponds to observed frequency; selective risk asks what error remains after deferral; robustness measures performance across defined perturbations; and utility assigns costs to actions. These are not interchangeable. For extrapolation-aware geotechnical prediction, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to instrument observations, constitutive parameters, geometry, loading history, simulation output, and validation measurements. If the operational tool updates incrementally, the evaluation must also test forgetting, stale evidence, and rollback. If an automatic judge supplies labels, a sample needs independent human review and content-preserving perturbations that reveal whether the judge is grading substance. Seen through extrapolation-aware geotechnical prediction, the issue is not merely technical; it determines which claims remain open to challenge.

Evaluation begins by writing each claim in testable form. Each intended claim - such as improved robustness, safer abstention, faster updating, or better structural localization - needs a target population, comparator, outcome, and boundary condition. Random splits are insufficient when related samples share a patient, repository, instrument run, season, organization, or simulated design family. Grouped and temporal splits better approximate the novelty encountered after deployment. Stress tests should vary one factor at a time when attribution is important, then combine factors to measure interaction. Repeated runs are necessary whenever decoding, optimization, retrieval, or numerical kernels can vary. Reporting only the best seed or a pooled mean makes operational instability disappear. This requirement gives practical content to the article's central question: What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions?

Institutional Conditions for Reliability

Human intervention works only when the review task is feasible and consequential. Reviewers need enough context to challenge the output, but not so much generated rationale that weak evidence is obscured by volume. Escalation queues require prioritization and workload limits. A reject option that sends nearly everything to experts is safe only on paper; a reject option that rarely fires may be equally ineffective. For sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, oversight should therefore be evaluated with realistic staffing, time pressure, and disagreement. The system should record the final action and its reason without turning that record into surveillance unrelated to the stated purpose. Contestability, privacy, and security are properties of this institutional process as much as of the predictive component. Seen through extrapolation-aware geotechnical prediction, the issue is not merely technical; it determines which claims remain open to challenge.

Measurements become useful only when they alter deployment limits. A deployment specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the system. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed process, or

new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the predictive component and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. This requirement gives practical content to the article's central question: What evidence is needed to determine how machine learning can predict hydraulic conductivity without hiding literature heterogeneity or unsupported chemical conditions?

From Evidence Capture to Escalation

The next design choice concerns where additional computation is worth its cost. Easy, well-supported cases can take a fast path. Shifted, ambiguous, or high-cost cases should activate additional precision, retrieval, alternative drafts, simulation, or expert review. This conditional design is preferable to applying maximum computation everywhere because resource cost is part of deployment quality. It is also preferable to an unconditional fast path because efficiency can amplify a systematic error at scale. The trigger must be evaluated as carefully as the expensive model: coverage, false reassurance, subgroup effects, latency, and the consequences of abstention all matter. The output is not simply a prediction but a package containing evidence links, uncertainty, the action taken, and the conditions that caused escalation. The implication is especially consequential in sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, where a small modeling choice can redirect attention and resources.

The evidence architecture for sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions begins with typed evidence. Each record should identify its source, time, collection conditions, transformations, and relation to other records. Graphs are useful when relations carry meaning, but graph construction is itself a hypothesis. An edge may denote temporal adjacency, physical causation, code dependency, shared infrastructure, or analyst assertion; treating these as interchangeable creates false certainty. The architecture should therefore preserve edge types and confidence, retain alternative links when evidence is ambiguous, and version both the data schema and the rules that construct the graph. A model may aggregate this structure, but the original evidence must remain recoverable for audit. The implication is especially consequential in sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, where a small modeling choice can redirect attention and resources.

Next Steps for Evidence Building

Beyond individual benchmarks, research must address how findings are retained and updated. Benchmarks should retain negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting should include data lineage, versioned testing code, confidence intervals, and enough error material to reproduce major conclusions. Prospective studies should predefine monitoring and stopping rules, then measure how people adapt to the application. Finally, researchers should compare simple baselines with complex architectures under equivalent information. Complexity is justified when it adds a measurable capability - for example, structural localization, calibrated deferral, or incremental updating - not merely when it creates a richer narrative around the same errors. In sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, this distinction should be tested before it changes an operational decision.

Future testing sets should be organized around plausible mechanisms of failure. Cases should represent unsupported regions, ambiguous evidence, conflicting modalities, rare but costly conditions, and realistic

adversarial transformations. Each case needs provenance and a reason for inclusion. The second priority is intervention testing: when uncertainty is detected, compare additional computation, retrieval, alternative reasoning paths, model ensembles, and human escalation under the same resource budget. This reveals whether a proposed safeguard actually changes the decision or only changes the explanation. The third priority is transfer. A method should be re-evaluated across sites, devices, languages, organizations, or physical regimes before broad claims are made. The implication is especially consequential in sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, where a small modeling choice can redirect attention and resources.

Conclusion

Extrapolation-aware geotechnical prediction is credible only when it is attached to an documented evidence chain and decision policy. The focal literature contributes several ways to improve the chain: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated evaluation. A reported benchmark gain still leaves the boundary of safe transfer to be established. For sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions, the practical standard should be a traceable workflow that measures shift and instability, supports bounded extrapolation, targeted experiment, model revision, and conservative design screening, and preserves enough evidence for an independent reviewer to reconstruct failure. Future work should compare safeguards under realistic resource and staffing limits, publish negative and subgroup results, and treat monitors and automated judges as components requiring their own validation. The final standard is not uninterrupted automation but evidence-linked action with documented deferral and independent verifiability. This point narrows the research task for sodium-bentonite geosynthetic clay liners exposed to saline or leachate solutions: compare alternatives under the same information and resource constraints.

References

1. Li, Dong, et al. “Prediction of Hydraulic Conductivity of Sodium Bentonite GCLs by Machine Learning Approaches.” *Environmental Geotechnics*, vol. 12, no. 2, 2025, pp. 154–73, <https://doi.org/10.1680/jenge.22.00181>.
2. Huang, Shuai, et al. “Research on Simplified Evaluation Method for Soil-Rock Mixed Slope Stability under Dam-Break Flood Impact.” *Bulletin of Engineering Geology and the Environment*, vol. 84, no. 1, 2025, <https://doi.org/10.1007/s10064-024-04055-4>.
3. Poinso, Thierry, and Denis Veynante. *Theoretical and Numerical Combustion*. 3rd ed., R. T. Edwards, 2012.
4. Israelachvili, Jacob N. *Intermolecular and Surface Forces*. 3rd ed., Academic Press, 2011.
5. Russel, William B., Dudley A. Saville, and William R. Schowalter. *Colloidal Dispersions*. Cambridge University Press, 1989.
6. Terzaghi, Karl, Ralph B. Peck, and Gholamreza Mesri. *Soil Mechanics in Engineering Practice*. 3rd ed., Wiley, 1996.
7. Koerner, Robert M. *Designing with Geosynthetics*. 6th ed., Xlibris, 2012.
8. Breiman, Leo. “Random Forests.” *Machine Learning*, vol. 45, 2001, pp. 5–32.
9. Friedman, Jerome H. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189–1232.
10. Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
11. Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*, vol. 3, 2016, article 160018.
12. Kennedy, Marc C., and Anthony O’Hagan. “Bayesian Calibration of Computer Models.” *Journal of the Royal Statistical Society: Series B*, vol. 63, no. 3, 2001, pp. 425–64.