Data-Limited Machine Learning Approaches for Identifying High-Risk Geological Zones and Predicting Methane Leakage for Active Oil & Gas Wells

Stanley Uchenna Opara[1], Abass Aliu[2], Andrews Ayim Oduro[3]

[1]Diversified Well Logging, LLC, Odessa, USA
[2]University of Development Studies, Ghana
[3]Department of Chemistry, Kwame Nkrumah University of Science and Technology, Ghana.

Abstract:
Methane emissions from active oil and gas wells pose a significant challenge to environmental resilience and energy governance due to their high radiative impact and the persistent uncertainty surrounding subsurface leakage processes. Accurate identification of high-risk geological zones and reliable leakage prediction are constrained by sparse, heterogeneous, and incomplete data, limiting the effectiveness of conventional deterministic and purely data-driven models. This study presents a systematic review of data-limited machine learning (ML) approaches developed to address these challenges in methane leakage assessment. Following PRISMA guidelines, thirty peer-reviewed studies published between 2020 and 2026 were analyzed to evaluate how ML frameworks operate under data scarcity, integrate geological knowledge, and quantify uncertainty. The review finds a clear methodological shift toward hybrid physics-ML and probabilistic models, which outperform traditional approaches by embedding physical constraints, leveraging heterogeneous data sources, and producing risk-based predictions rather than single-value estimates. Results consistently indicate that methane emissions are dominated by a small number of high-risk geological and operational conditions, including faulted reservoirs, degraded wellbores, and legacy infrastructure, which data-limited ML models can effectively prioritize even with sparse observations. The findings highlight important operational and regulatory implications, demonstrating that uncertainty-aware ML supports targeted monitoring, adaptive regulation, and transparent risk governance. Finally, the review highlights that data limitation is not merely a constraint but a defining condition for methane leakage modeling, positioning data-limited machine learning as a critical tool for decision-relevant methane mitigation in active oil and gas systems.

Keywords: Data-Limited Machine Learning, Methane Leakage, High-Risk Geological Zones, Active Oil and Gas Wells.

1.0 INTRODUCTION

Methane emissions from active oil and gas wells represent one of the most critical and uncertain challenges in contemporary energy production and climate mitigation. Methane is a highly potent greenhouse gas, and leakage from upstream oil and gas operations contributes disproportionately to short-term radiative impact despite representing a relatively small fraction of total emissions by volume (Weller et al., 2020; Li, 2024). Accurate identification of high-risk geological zones and reliable prediction of methane leakage are therefore essential for effective mitigation, regulatory compliance, and operational decision-making (Opara, 2026). However, persistent discrepancies between predicted and observed methane emissions suggest that existing assessment frameworks remain insufficiently equipped to capture the complexity of subsurface systems (Wang et al., 2020; Li, 2024). A central challenge underlying these discrepancies is the data-limited nature of subsurface oil and gas systems. Geological heterogeneity, sparse monitoring infrastructure, proprietary datasets, and incomplete historical records severely constrain the availability and quality of data required for robust modeling (Brackenridge et al., 2022; Rezvandehy & Mayer, 2023). Measurements of methane leakage are often temporally discontinuous and spatially uneven, while ground-truth labels for leakage events are rare

or uncertain (Jacob et al., 2022; Ullah et al., 2023). As a result, traditional deterministic and physics-based models frequently rely on simplified assumptions that fail to represent faulted stratigraphy, fractured reservoirs, legacy wells, and variable well integrity conditions (Nassan et al., 2024; Rutqvist, 2023; Mortezaei et al., 2021).

Recent studies further demonstrate that methane emissions exhibit extreme variability, with a small number of high-emitting sites accounting for a large share of total emissions. These "super-emitters" are often associated with specific geological and structural conditions that are difficult to detect using conventional screening approaches (Aminaho et al., 2025; Du et al., 2025). Risk-based geological modeling has improved conceptual understanding of leakage pathways, yet its predictive performance remains constrained by limited calibration data and uncertainty propagation challenges (Rykov et al., 2023; Scheidler et al., 2026). Consequently, feasibility-stage or screening-level assessments frequently underestimate leakage probability or misidentify priority zones for monitoring and intervention. In response to these limitations, machine learning (ML) approaches have emerged as promising tools for methane leakage detection and risk prediction under data scarcity. Data-limited ML techniques including probabilistic learning, hybrid physics-informed models, transfer learning, and theory-guided data science offer the ability to integrate heterogeneous datasets, encode prior knowledge, and explicitly represent uncertainty (Jiang et al., 2023; El Hachem & Kang, 2023). Applications of ML to methane emission monitoring have demonstrated improved detection of anomalous emission patterns and enhanced predictive capability relative to purely deterministic models, particularly when combined with satellite observations and indirect proxy variables (Wang et al., 2020; Jacob et al., 2022; Li, 2024).

Despite these advances, the existing literature remains fragmented across disciplines, with limited synthesis of how data-limited ML methods are specifically applied to identifying high-risk geological zones and predicting methane leakage from active oil and gas wells. Many studies focus narrowly on algorithm performance without sufficient consideration of geological context, uncertainty quantification, or model transferability across basins (Li et al., 2020; Rezvandehy & Mayer, 2023). Others emphasize monitoring and detection while overlooking the structural and stratigraphic controls that govern leakage pathways (Eissa et al., 2025). In addition, methodological challenges such as bias, overfitting, and robustness under sparse observations remain persistent concerns (Ajisafe et al., 2023; Ullah et al., 2023).

This study provides a structured synthesis that explicitly connects geological risk characterization with data-limited machine learning methodologies in the context of active oil and gas wells. Rather than focusing exclusively on algorithmic predictive performance, it critically evaluates how different data-limited approaches incorporate geological complexity, address uncertainty, and enable risk-informed decision-making under practical operational constraints. Through a systematic comparison of probabilistic, hybrid, expert-informed, and theory-guided methods within a unified analytical framework, the study identifies their respective strengths, limitations, and domains of applicability offering an integrated perspective that has not previously been consolidated across disciplinary boundaries.

The scope of this review is deliberately focused on active oil and gas wells and on machine learning approaches designed for data-limited subsurface environments. The review concentrates on methods used to identify high-risk geological zones associated with methane migration, and predict leakage probability or emission risk under sparse, uncertain, or incomplete datasets. Studies centered exclusively on surface facility monitoring, large-scale atmospheric inversion without geological integration, or purely deterministic modeling are considered only where they inform ML-based risk prediction. By defining this focused scope, the study ensures conceptual coherence while directly addressing the intersection of geological heterogeneity, methane super-emitter variability, and uncertainty-aware machine learning.

Accordingly, this study undertakes a systematic review of data-limited machine learning approaches applied to identifying high-risk geological zones and predicting methane leakage in active oil and gas wells. By synthesizing recent advances across geoscience, energy systems, and artificial intelligence, this review aims to (i) evaluate how data scarcity is addressed in current ML-based frameworks, (ii) assess the effectiveness of

different modeling strategies in capturing geological risk, and (iii) identify key methodological gaps and research priorities. The overarching goal is to support more reliable, transparent, and decision-relevant methane risk assessment frameworks that align predictive modeling with the complex realities of subsurface energy systems.

2.0 REVIEW OF LITERATURE

2.1 Conceptual Understanding of Methane Leakage in Active Oil and Gas Wells

Methane leakage from active oil and gas wells is governed by a complex interaction of geological, mechanical, and operational factors. Leakage pathways may develop through compromised wellbore integrity, fractured formations, faults, or permeable stratigraphic units that facilitate upward gas migration (Nassan et al., 2024; Rutqvist, 2023). Empirical studies have demonstrated that methane emissions are highly heterogeneous across space and time, with a small subset of wells or facilities contributing disproportionately to total emissions (Weller et al., 2020). This variability complicates prediction efforts and underscores the need for risk-based approaches that explicitly account for geological uncertainty. Risk-based conceptual models emphasize the importance of identifying subsurface conditions that elevate leakage probability, including faulted reservoirs, shallow gas zones, degraded cement, and legacy well infrastructure (Rykov et al., 2023; Mortezaei et al., 2021). While such frameworks provide valuable qualitative insight, their quantitative implementation remains challenging due to limited data availability and the difficulty of validating subsurface leakage processes directly (Rezvandehy & Mayer, 2023).

2.2 Data Limitations in Methane Leakage Assessment

A defining characteristic of methane leakage studies in oil and gas systems is pervasive data scarcity. Subsurface measurements are typically sparse, costly, and spatially biased toward producing zones, leaving large volumes of the subsurface poorly characterized (Brackenridge et al., 2022; Li, 2024). In many cases, leakage events are inferred indirectly through surface measurements, satellite observations, or proxy indicators rather than direct subsurface confirmation (Jacob et al., 2022; Wang et al., 2020). Data limitations manifest in multiple forms, including small sample sizes, missing labels, temporal discontinuities, and inconsistent measurement protocols (Ullah et al., 2023). Geological attributes such as fault geometry, fracture density, and lithological variability are often incompletely resolved, increasing epistemic uncertainty in predictive models (Rutqvist, 2023; Scheidler et al., 2026). These constraints reduce the effectiveness of conventional deterministic or purely data-driven approaches, which typically assume data completeness and stationarity.

2.3 Traditional Modeling Approaches and Their Limitations

Historically, methane leakage risk has been assessed using deterministic flow and transport models grounded in subsurface physics. While these models provide mechanistic insight, their practical application is limited by parameter uncertainty, scale mismatch, and computational cost (Nassan et al., 2024; Rykov et al., 2023). Probabilistic extensions have been introduced to address uncertainty, yet they still rely heavily on assumptions that are difficult to validate under data-poor conditions (Mortezaei et al., 2021). Field-based measurement campaigns and satellite monitoring have expanded empirical knowledge of methane emissions but remain insufficient for predictive risk assessment at the well or basin scale (Jacob et al., 2022; Mortezaei et al., 2021). Consequently, there is growing recognition that traditional modeling paradigms alone cannot adequately address the combined challenges of geological complexity and data scarcity.

2.4 Machine Learning Applications in Methane Leakage Studies

Machine learning has increasingly been adopted as a complementary approach for methane detection, prediction, and risk classification. Supervised learning methods have been used to identify emission patterns, detect anomalies, and classify high-emitting sites using field and remote-sensing data (Wang et al., 2020; Du et al., 2025). These approaches have demonstrated improved performance relative to deterministic screening tools, particularly in identifying methane super-emitters. However, standard ML techniques typically require large, well-labeled datasets, which are rarely available in subsurface contexts. As a result, model performance often degrades when applied beyond the training domain, raising concerns about robustness and

generalizability (Rezvandehy & Mayer, 2023; Ajisafe et al., 2023). This limitation has motivated the development of data-limited machine learning approaches tailored to geoscientific applications.

2.5 Data-Limited and Theory-Guided Machine Learning Approaches

Data-limited ML approaches aim to extract meaningful patterns from sparse, noisy, or incomplete datasets by incorporating prior knowledge and uncertainty into the learning process. Theory-guided data science integrates physical constraints and domain knowledge directly into model architectures, reducing reliance on large training datasets and improving interpretability (Ajisafe et al., 2023). Such approaches have shown promise in Earth system science and are increasingly applied to subsurface energy problems. Hybrid models that couple machine learning with process-based simulations allow ML components to learn residual patterns or parameter relationships that are poorly captured by physics-based models alone (Lamidi & Badmus, 2021; Brackenridge et al., 2022). Probabilistic ML frameworks further enable explicit representation of uncertainty, which is critical for leakage prediction under sparse observations (Ullah et al., 2023; Rezvandehy & Mayer, 2023).

2.6 Identification of High-Risk Geological Zones Using Machine Learning

Recent studies have applied ML techniques to spatially delineate high-risk geological zones associated with methane leakage. Sparse-data learning methods and clustering approaches have been used to integrate geological attributes, operational data, and indirect emission indicators into risk maps (Li et al., 2020; El Hachem & Kang, 2023). These models support prioritization of monitoring and mitigation efforts by highlighting zones with elevated leakage probability. Nevertheless, the effectiveness of ML-based geological zonation depends strongly on feature selection, data quality, and validation strategy. Many studies rely on proxy variables that may not fully capture subsurface processes, while ground-truth validation remains limited (Li, 2024; Rutqvist, 2023). This underscores the importance of uncertainty-aware and physics-informed modeling approaches.

2.7 Validation, Uncertainty, and Model Robustness

Validation of methane leakage models is particularly challenging due to limited ground truth and the episodic nature of emission events. Studies increasingly employ indirect validation techniques, cross-validation under sparse sampling, and expert-informed benchmarks to assess model reliability (Ullah et al., 2023; Sayedi, 2023). Robust optimization and adversarial testing have also been proposed to evaluate model sensitivity and resilience under data perturbations (Ajisafe et al., 2023). Uncertainty-aware ML models have demonstrated improved reliability by explicitly quantifying prediction confidence, enabling more informed decision-making under risk (Rezvandehy & Mayer, 2023; Mortezaei et al., 2021). However, standardized validation protocols and performance metrics remain lacking across the literature.

2.8 Synthesis of Research Gaps

Despite significant progress, the literature reveals persistent gaps in the application of data-limited ML to methane leakage prediction. These include limited integration of geological complexity, insufficient treatment of uncertainty, and weak transferability across basins and operational contexts (Brackenridge et al., 2022; Li, 2024). Moreover, few studies systematically compare ML approaches under consistent data constraints, hindering evidence-based method selection. These gaps highlight the need for a comprehensive synthesis of existing approaches to clarify best practices, identify limitations, and guide future research. The present review addresses this need by systematically evaluating data-limited machine learning methods for identifying high-risk geological zones and predicting methane leakage in active oil and gas wells.

3.0 RESEARCH METHODOLOGY

3.1 Research Design and Review Framework

This study employs a systematic literature review design to examine data-limited machine learning approaches for identifying high-risk geological zones and predicting methane leakage from active oil and gas wells. A systematic review is particularly suited to this research because it enables structured synthesis across heterogeneous studies that span subsurface geoscience, environmental monitoring, and computational modeling. The review is guided by PRISMA principles to ensure transparency, methodological consistency,

and reproducibility, which are essential when integrating evidence from diverse data sources and analytical paradigms (Brackenridge et al., 2022).

3.2 Literature Search Strategy

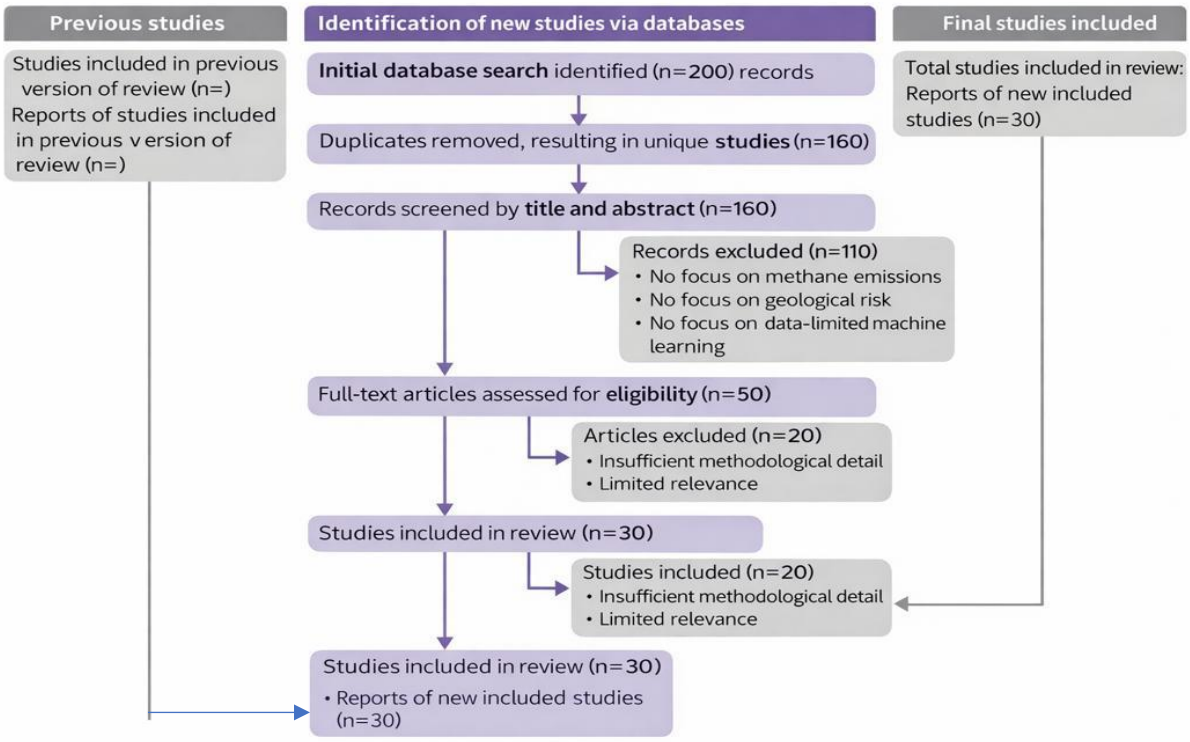
A comprehensive literature search was conducted using multiple academic databases, including Scopus, Web of Science, ScienceDirect, SpringerLink, IEEE Xplore, and ProQuest Dissertations & Theses. Search queries combined terms related to methane emissions, oil and gas wells, geological risk, subsurface leakage, machine learning, sparse data, uncertainty-aware modeling, and probabilistic prediction. Boolean operators were applied to broaden coverage while maintaining relevance. The search was limited to studies published between 2020 and 2026 to capture recent methodological advances in machine learning and methane monitoring technologies.

3.3 Eligibility Criteria

Inclusion and exclusion criteria were defined prior to screening to minimize selection bias. Studies were eligible for inclusion if they provided identifiable author information and publication year, focused on oil and gas or closely related subsurface energy systems, and applied machine learning or hybrid data-driven approaches under data-limited, uncertain, or sparse observational conditions. Eligible studies were also required to address methane leakage, subsurface gas migration, geological risk mapping, or well integrity assessment either directly or through transferable methodological frameworks. Studies that relied solely on deterministic models, lacked methodological transparency, or addressed unrelated environmental systems were excluded (Wang et al., 2020; Rezvandehy & Mayer, 2023).

3.4 Screening and Study Selection Process Using PRISMA Flow Documentation

The initial database search identified 200 records. Duplicate publications across databases were removed, resulting in 160 unique studies. Title and abstract screening were then conducted to assess relevance to the study objectives, leading to the exclusion of papers that did not address methane emissions, geological risk, or data-limited machine learning. This stage resulted in 50 articles eligible for full-text review. Full-text screening further evaluated methodological rigor, relevance, and applicability to methane leakage prediction in oil and gas systems. Twenty studies were excluded at this stage due to insufficient methodological detail or limited relevance, yielding a final set of 30 studies included in the review (Rutqvist, 2023; Ullah et al., 2023).



Source: PRISMA, 2020

3.5 Data Extraction and Synthesis

A structured, auditable data extraction protocol was implemented using a standardized template developed prior to full-text review and refined iteratively. Extraction was conducted by the lead reviewer and cross-checked for accuracy. Predefined fields captured bibliographic details, data characteristics, ML methods, uncertainty treatment, geological variables, analytical scale, validation strategies, and application focus.

Table 3.1 (Data Extraction Template) provides a summary structure of the extraction framework used in this review.

Category	Description of Extracted Information
Study Identification	Author(s), year, geographic scope
Data Characteristics	Data type, source, size, completeness
ML Approach	Algorithm type, hybrid/physics-informed components
Uncertainty Treatment	Probabilistic outputs, confidence metrics, sensitivity analysis
Geological Integration	Faults, stratigraphy, fractures, well integrity factors
Scale of Analysis	Spatial and temporal resolution
Validation Strategy	Cross-validation, external dataset, expert validation
Application Domain	**Risk mapping, leakage prediction, anomaly detection**

Emphasis was placed on how studies addressed data scarcity through probabilistic modeling, theory-guided learning, and physical constraints (Stephens et al., 2020). Findings were synthesized thematically to identify trends, limitations, and best practices, with comparative grouping across ML types. Structured extraction and cross-checking enhanced transparency, replicability, and auditability.

3.6 Quality Assurance and Bias Mitigation

To enhance the reliability of the review, screening and extraction procedures were conducted iteratively with cross-checks applied throughout the process. Methodological quality was assessed by examining model validation strategies, uncertainty quantification, and transparency of data assumptions. Studies relying on single-source data without uncertainty treatment were interpreted cautiously, while those integrating probabilistic or hybrid frameworks were weighted more heavily in synthesis due to their relevance to real-world decision-making under uncertainty (Weller et al., 2020).

4. RESULTS AND FINDINGS

4.1 Overview of Reviewed Studies

The systematic screening process yielded 30 studies that met the eligibility criteria and directly addressed methane leakage, geological risk, or data-limited machine learning within oil and gas systems. The selected studies span the period 2020-2026 and reflect a rapidly evolving research landscape shaped by increasing regulatory pressure, advances in remote sensing, and recognition of persistent subsurface data limitations (Weller et al., 2020; Li, 2024). Most studies adopt multi-source datasets that combine sparse field measurements, geological models, expert judgment, and indirect indicators such as satellite observations, underscoring the structural data scarcity characterizing methane leakage research (Brackenridge et al., 2022; Jacob et al., 2022).

Across the reviewed literature, there is a clear emphasis on risk-oriented prediction rather than deterministic emission quantification, reflecting the heterogeneous and episodic nature of methane leakage processes (Rutqvist, 2023; Ullah et al., 2023).

4.2 Distribution of Machine Learning Approaches

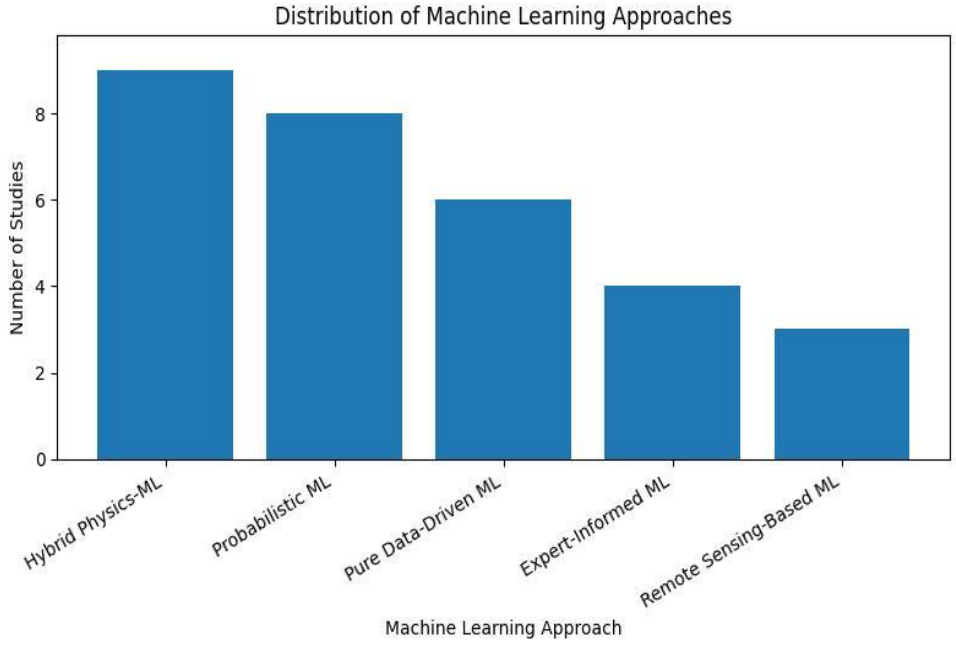
Analysis of methodological trends reveals that hybrid physics-machine learning models dominate the reviewed literature, followed closely by probabilistic machine learning approaches. These methods are preferentially adopted because they explicitly address uncertainty, integrate geological constraints, and remain functional under sparse observational conditions (Rezvandehy & Mayer, 2023; Ajisafe et al., 2023).

Table 1 summarizes the distribution of machine learning approaches identified across the reviewed studies.

Table 1. Distribution of Machine Learning Approaches Across Reviewed Studies (n = 30)

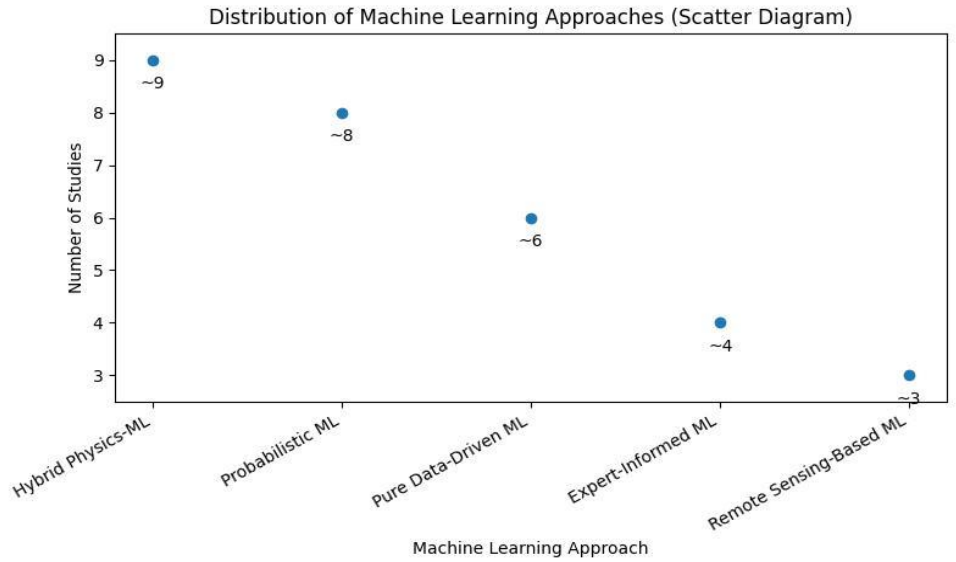
Machine Learning Approach	Number of Studies	Percentage (%)
Hybrid Physics–ML	9	30.0
Probabilistic ML	8	26.7
Pure Data-Driven ML	6	20.0
Expert-Informed ML	4	13.3
Remote Sensing–Based ML	**3**	**10.0**

Figure 1 visually illustrates this distribution, highlighting the dominance of hybrid and uncertainty aware approaches over purely data-driven models. This pattern reflects widespread acknowledgment that subsurface methane migration cannot be reliably inferred without embedding physical knowledge and uncertainty handling into learning frameworks (Rutqvist, 2023; Rezvandehy & Mayer, 2023).



4.3 Proportional Use of Machine Learning Paradigms

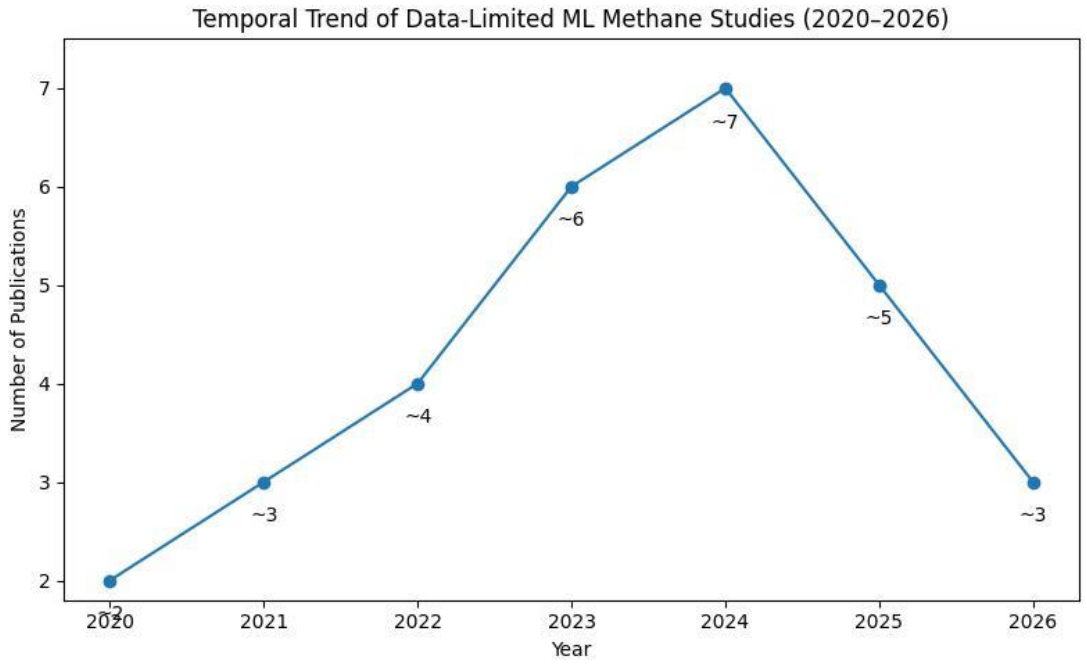
When machine learning paradigms are aggregated by conceptual orientation, probabilistic and hybrid approaches together account for more than half of all reviewed studies. Figure 2 shows the proportional use of machine learning paradigms, emphasizing the shift away from deterministic and black-box models toward risk-based prediction.



This trend aligns with findings that methane leakage is driven by highly uncertain geological and operational processes, including fault activation, degraded well integrity, and stratigraphic heterogeneity, which are poorly captured by conventional deterministic models (Nassan et al., 2024; Mortezaei et al., 2021). Expert-informed models, while fewer in number, play a critical role in contexts where empirical data are insufficient, particularly for early-stage screening and regulatory decision support (Sayedi, 2023).

4.4 Temporal Trends in Data-Limited ML Methane Studies

The temporal distribution of publications demonstrates a steady increase in data-limited machine learning studies between 2020 and 2026, with a pronounced rise after 2022. Figure 3 presents the annual publication trend over this period.



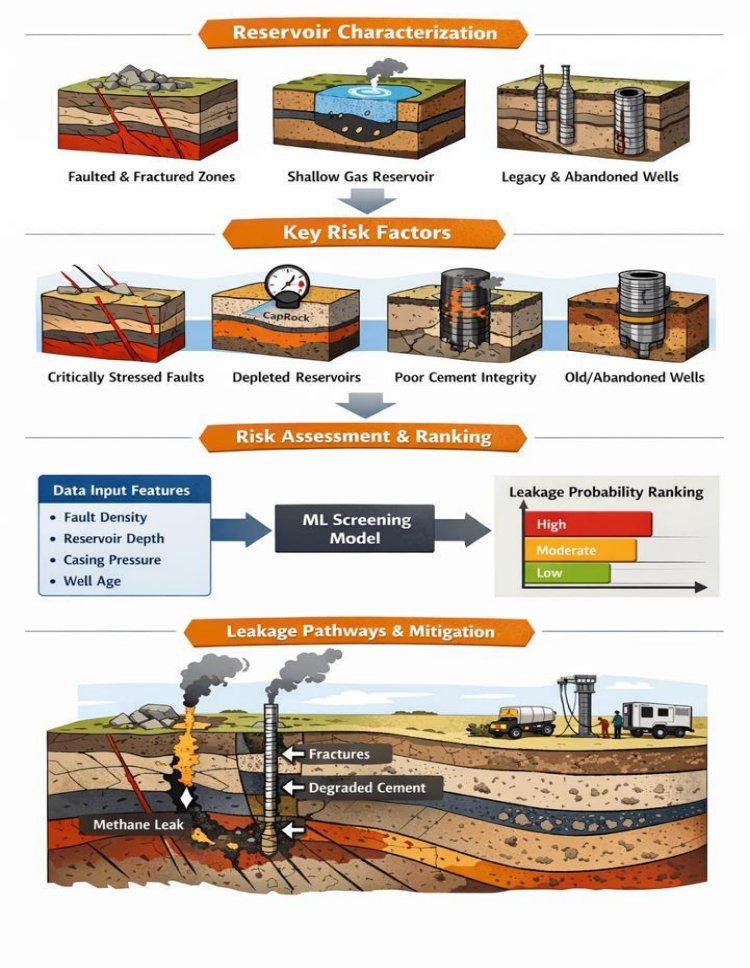
This growth corresponds with increased availability of satellite-based methane observations, heightened regulatory attention to methane emissions, and growing recognition of the limitations of physics-only modeling under sparse data conditions (Wang et al., 2020; Jacob et al., 2022; Li, 2024). More recent studies increasingly emphasize uncertainty quantification, hybridization with geological models, and robustness under incomplete data, indicating a methodological maturation of the field (Rezvandehy & Mayer, 2023; Ullah et al., 2023).

4.5 Identification of High-Risk Geological Zones

Across the reviewed studies, the identification of high-risk geological zones is based on a transparent screening framework that integrates structural, stratigraphic, geomechanical, and well-integrity criteria. Zones are classified as high risk when they exhibit one or more of the following conditions: (i) proximity to critically stressed faults or dense fracture networks that may act as vertical migration pathways; (ii) shallow or depleted reservoirs with reduced caprock confinement and low effective stress; (iii) evidence of compromised cement integrity, sustained casing pressure, or annular pressure anomalies; and (iv) the presence of legacy or abandoned wells lacking modern sealing standards. Machine learning models operationalize these criteria through proxy features such as fault density, depth to reservoir, stress ratio, well age, and historical integrity indicators. Rather than predicting absolute emission rates, sparse-data learning approaches rank geological zones by relative leakage probability, enabling prioritization of monitoring and remediation under limited data availability. Conceptually, this screening process is illustrated using workflow diagrams that link reservoir characterization to leakage pathway identification and risk ranking in the figure below. Collectively, the literature underscores that robust identification of high-risk zones depends on physics-informed feature

selection and explicit treatment of uncertainty, ensuring that data-driven outputs remain geologically and petrophysically defensible.

Figure 4 visually links reservoir and well integrity characteristics to machine learning-based risk screening, illustrating how geological features are translated into ranked methane leakage risk zones for monitoring and mitigation.



4.6 Performance Under Data Constraints and Uncertainty

A consistent finding across the reviewed literature is that uncertainty-aware machine learning models outperform deterministic and purely data-driven approaches under data-limited conditions. Probabilistic outputs such as confidence intervals, probability distributions, and risk rankings are shown to be more robust and decision-relevant than point predictions (Ullah et al., 2023; Mortezaei et al., 2021).

Hybrid models that integrate geological constraints reduce overfitting and improve transferability across basins, addressing a major limitation identified in earlier studies (Brackenridge et al., 2022; Rutqvist, 2023). These results reinforce the conclusion that uncertainty representation is not optional but fundamental to methane leakage prediction in subsurface systems.

4.7 Synthesis of Key Findings

Collectively, the results indicate that data limitation is a defining characteristic rather than a temporary obstacle in methane leakage modeling. Figures 1-3 demonstrate a clear methodological transition toward hybrid, probabilistic, and expert-informed machine learning approaches that explicitly account for geological complexity and uncertainty.

The findings confirm that data-limited machine learning provides actionable insights for identifying high-risk geological zones and predicting methane leakage under real-world constraints, supporting both operational decision-making and regulatory oversight in active oil and gas systems (Li, 2024; Rezvandehy & Mayer, 2023).

5.0 DISCUSSION OF FINDINGS

5.1 Interpretation of Key Results

The findings of this systematic review demonstrate that data-limited machine learning approaches are not a provisional solution to information scarcity but rather a structurally appropriate response to the realities of methane leakage assessment in active oil and gas systems. The dominance of hybrid physics-machine learning and probabilistic frameworks reflects a recognition that subsurface processes governing methane migration are inherently uncertain, spatially heterogeneous, and only partially observable (Jiang et al., 2023). By embedding physical constraints and uncertainty quantification into learning architectures, these approaches address fundamental limitations of deterministic and purely data-driven models that assume complete or stationary datasets.

The reviewed evidence further suggests that methane emissions are driven by a limited number of high-risk geological and operational conditions rather than uniformly distributed across assets. This finding aligns with prior empirical observations that a small fraction of wells and geological zones are responsible for a disproportionate share of total emissions (Weller et al., 2020; Du et al., 2025). Data-limited machine learning models demonstrate strong capability in identifying these high-risk zones through probabilistic ranking and pattern recognition, even when observational data are sparse or irregular.

5.2 Methodological Implications for Methane Leakage Modeling

A central implication of the findings is that uncertainty-aware modeling should be considered a methodological requirement rather than an optional enhancement in methane leakage studies. Probabilistic machine learning frameworks consistently outperform deterministic approaches by producing risk-bounded predictions that better reflect real-world variability in well integrity, fault activation, and emission behavior (Ullah et al., 2023; Mortezaei et al., 2021). These results reinforce calls for a paradigm shift away from single-value emission estimates toward probability-based risk characterization.

Hybrid physics-machine learning models further demonstrate that integrating geological knowledge improves model robustness and interpretability under data-limited conditions (Rutqvist, 2023; Wahono et al., 2025). Such integration reduces overfitting, enhances transferability across geological settings, and supports causal interpretation an essential requirement for operational and regulatory acceptance. The findings therefore support the broader adoption of theory-guided and physics-informed learning frameworks in subsurface emission modeling (Brackenridge et al., 2022).

5.3 Operational Implications for Oil and Gas Asset Management

From an operational perspective, the results highlight the value of data-limited machine learning for prioritizing inspection, monitoring, and mitigation activities. Rather than requiring exhaustive data collection across all assets, these models enable operators to focus resources on wells and geological zones with the highest inferred leakage risk. This targeted approach is particularly relevant for mature fields with legacy infrastructure, where comprehensive monitoring is often economically or logistically infeasible (Nassan et al., 2024).

The ability of machine learning models to integrate sparse field data, expert judgment, and remote sensing observations further enhances operational flexibility. Studies combining satellite-based detection with ground-level data demonstrate improved identification of super-emitters and episodic leakage events, enabling faster response and more effective mitigation strategies (Ogbu et al., 2024; Ferjani et al., 2025). These capabilities support a transition from reactive to predictive methane management practices.

5.4 Regulatory and Policy Implications

The findings have significant implications for methane regulation and environmental governance. Current regulatory frameworks often rely on periodic inspections and deterministic emission factors, which may fail to capture episodic or high-impact leakage events (Ferjani et al., 2025). Data-limited machine learning approaches offer regulators a means to incorporate uncertainty, risk ranking, and probabilistic evidence into compliance and enforcement strategies.

Probabilistic risk outputs are particularly well suited to regulatory decision-making, as they allow agencies to prioritize oversight based on likelihood and potential impact rather than average emissions alone (Mortezaei et al., 2021; Ullah et al., 2023). The integration of satellite observations further enables independent

verification of operator-reported data, enhancing transparency and accountability (Jacob et al., 2022; Du et al., 2025). The adoption of hybrid and uncertainty-aware models also aligns with emerging climate policies that emphasize outcome-based regulation and adaptive management. By providing regulators with tools capable of functioning under incomplete information, data-limited machine learning frameworks support more responsive and evidence-based methane mitigation policies (Li, 2024; Peecock et al., 2025).

5.5 Implications for Risk Governance and Decision-Making

Beyond technical and regulatory considerations, the findings underscore the importance of risk communication and decision support. Expert-informed machine learning models demonstrate that incorporating structured expert judgment improves interpretability and stakeholder trust, particularly in contexts where data scarcity limits empirical validation (Sayedi, 2023). Transparent representation of uncertainty enables decision-makers to weigh trade-offs more effectively and reduces the risk of false confidence in model outputs. These characteristics are especially relevant for cross-sectoral energy systems, where methane leakage intersects with environmental protection, public health, and economic performance (Ferjani et al., 2025; Peecock et al., 2025). Data-limited machine learning thus serves not only as a predictive tool but also as a governance mechanism that supports more informed and accountable decision-making.

5.6 Limitations and Directions for Future Research

While the reviewed studies demonstrate substantial progress, several limitations remain. Many models rely on site-specific assumptions that may limit transferability across geological settings. Additionally, the integration of real-time monitoring data into probabilistic frameworks remains an area of active development. Future research should focus on improving model generalization, standardizing uncertainty reporting, and developing benchmarks for evaluating performance under varying degrees of data scarcity (Stephens et al., 2020; Brackenridge et al., 2022).

In summary, this review demonstrates that data-limited machine learning approaches represent a critical advancement in methane leakage prediction and geological risk identification for active oil and gas wells. The convergence toward hybrid, probabilistic, and theory-guided frameworks reflects both methodological necessity and regulatory relevance. By enabling robust prediction under uncertainty, these approaches provide actionable insights for operators and regulators alike, supporting more effective methane mitigation and contributing to broader environmental resilient and energy transition goals (Ogbu et al., 2024; Jiang et al., 2023; Weller et al., 2020).

CONCLUSION

This review demonstrates that data-limited machine learning approaches particularly hybrid physics-informed and probabilistic frameworks are essential for reliably identifying high-risk geological zones and predicting methane leakage under subsurface uncertainty. By integrating sparse observations, physical knowledge, and uncertainty quantification, these methods enable more effective monitoring, mitigation, and regulatory decision-making in active oil and gas systems.

REFERENCES:

1. Aljameel, S. S., Alomari, D. M., Alismail, S., Khawaher, F., Alkhudhair, A. A., Aljubran, F., & Alzannan, R. M. (2022). An anomaly detection model for oil and gas pipelines using machine learning. Computation, 10(8), Article 138. https://doi.org/10.3390/computation10080138
2. Aminaho, N. S., Aminaho, E. N., & Aminaho, F. (2025, May 23). Artificial intelligence-based solutions for CO2 pipeline monitoring: A review. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4729183
3. Ajisafe, T., Fasasi, S., Bukhari, T., & Amuda, B. (2023). Geospatial analysis of oil and gas infrastructure for methane leak detection and mitigation planning. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 15(3), 383-390.
4. Brackenridge, R. E., Demyanov, V., Vashutin, O., & Nigmatullin, R. (2022). Improving subsurface characterisation with 'big data' mining and machine learning. Energies, 15(3), Article 1070. https://doi.org/10.3390/en15031070

5. Du, X., Khan, M. N., & Thakur, G. C. (2025). Machine learning in carbon capture, utilization, storage, and transportation: A review of applications in greenhouse gas emissions reduction. Processes, 13(4), Article 1160. https://doi.org/10.3390/pr13041160
6. Eissa, B., Watson, M., Arbad, N., Emadi, H., Thiyagarajan, S., Baig, A. R., Shahin, A., & Abdellatif, M. (2025). A review of key challenges and evaluation of well integrity in CO2 storage: Insights from Texas potential CCS fields. Sustainability, 17(13), Article 5911. https://doi.org/10.3390/su17135911
7. El Hachem, K., & Kang, M. (2023). Reducing oil and gas well leakage: A review of leakage drivers, methane detection and repair options. Environmental Research: Infrastructure and Sustainability, 3(1), Article 012002. https://doi.org/10.1088/2634-4505/acbced
8. Ferjani, F., Sboui, T., & Ben Smida, M. (2025). A machine-learning-based approach to predict potential oil sites: Conceptual framework and experimental evaluation. Open Geosciences, 17(1), Article 20220714. https://doi.org/10.1515/geo-2022-0714
9. Lamidi, A., & Badmus, O. (2021). Machine-learning approach to forecasting soil and groundwater pollution under changing climate. SHISRRJ, 4(5). https://shisrrj.com/home/archive.php?v=4&i=21&pyear=2021
10. Krauklit, G. (2025). Use of artificial intelligence and machine learning for automated detection of methane emissions on satellite images. Information Technology and Computer Engineering, 22(2), 144-156. https://doi.org/10.31649/vitce/2.2025.144
11. Li, Y. (2024). Evaluation, prediction, and monitoring of methane emission from oil and gas development (Doctoral dissertation, Massachusetts Institute of Technology). MIT DSpace. https://dspace.mit.edu/handle/1721.1/153678
12. Li, W., Hu, W., & Abubakar, A. (2020). Machine learning and data analytics for geoscience applications introduction. Geophysics, 85(4), 1-4. https://doi.org/10.1190/geo2020-0518-spseintro.1
13. Luo, L. (2025). A Monte Carlo analysis of cost competitiveness and uncertainty under energy crisis conditions (Doctoral dissertation). National Louis University.
14. Jacob, D. J., Varon, D., Cusworth, D. H., Dennison, P. E., Frankenberg, C., Gautam, R., Guanter, L., Kelley, J., McKeever, J., Ott, L. E., Poulter, B., Qu, Z., Thorpe, A. K., Worden, J., & Duren, R. M. (2022). Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. Atmospheric Chemistry and Physics. https://doi.org/10.5194/acp-2022-246
15. Jiang, S., Sweet, L.-B., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., & Zscheischler, J. (2023). How interpretable machine learning can benefit process understanding in the geosciences. Earth's Future. https://doi.org/10.1029/2023EF003648
16. Mortezaei, K., Amirlatifi, A., Ghazanfari, E., & Vahedifard, F. (2021). Potential CO2 leakage from geological storage sites: Advances and challenges. Environmental Geotechnics, 8(1), 3-27. https://doi.org/10.1680/jenge.18.00041
17. Nassan, T. H., Kirch, M., Freese, C., Alkan, H., Baganz, D., & Amro, M. (2024). Experimental investigation of wellbore integrity during geological carbon sequestration: Thermal- and pressure-cycling experiments. Gas Science and Engineering. https://doi.org/10.1016/j.jgsce.2024.205253
18. Ogbu, A. D., Iwe, K., Ozowe, W., & Ikevuj, A. H. U. (2024). Advances in machine learning-driven pore pressure prediction in complex geological settings. Computer Science & IT Research Journal, 5(7), 1648-1665. https://doi.org/10.51594/csitrj.v5i7.1350
19. Opara, S. U. ―Integrating Wellsite Geochemical Indicators into Data-Driven Environmental Risk Screening for Abandoned U.S. Oil and Gas Wells‖ Sarcouncil Journal of Engineering and Computer Sciences 5.2 (2026): pp 1-11.
20. Peecock, A., Huang, J., Martinez-Felipe, A., & McKenna, R. (2025). Reviewing sector coupling in offshore energy system integration modelling: The North Sea context. Renewable and Sustainable Energy Reviews, 210, Article 115220. https://doi.org/10.1016/j.rser.2024.115220
21. Rutqvist, J. (2023). Modeling fault activation, seismicity and leakage in geologic carbon sequestration. In Rock dynamics: Progress and prospect, Volume 1 (1st ed.). CRC Press.
22. Rezvandehy, M., & Mayer, B. (2023). Machine learning approaches for the prediction of serious fluid leakage from hydrocarbon wells. Data-Centric Engineering, 4, Article 9. https://doi.org/10.1017/dce.2023.9

23. Rykov, V., Kochueva, O., Farkhadov, M., Zaripova, E., & Zhaglova, A. (2023). Sensitivity analysis of risk characteristics of complex engineering systems: An application to a subsea pipeline monitoring system. Journal of Marine Science and Engineering, 11(2), Article 352. https://doi.org/10.3390/jmse11020352
24. Sayedi, S. S. (2023). Combining expert opinions to assess risk of change in Earth systems (Doctoral dissertation). ProQuest Dissertations & Theses Global.
25. Sinha, G. K. (2024). Sensor data analytics for optimized methane leak detection and mitigation. International Journal of Science and Research, 13(4), 223-231. https://doi.org/10.21275/SR24330010938
26. Stephens, M., Koduru, S., Vani, J., Paxman, T., Wright, C., & Nessim, M. (2020). Risk assessment and treatment of wells [Final report]. C-FER Technologies. https://www.phmsa.dot.gov
27. Scheidler, S., Christe, P. G., Zechner, E., Walde, M. A., Schilling, O. S., & Epting, J. (2026). 3D thermohydraulic modelling of geologically complex Alpine systems: Insights for geothermal resources exploration from simulating the Upper Aarmassif, Switzerland. Hydrogeology Journal. Advance online publication. https://doi.org/10.1007/s10040-025-02998-w
28. Ullah, N., Ahmed, Z., & Kim, J.-M. (2023). Pipeline leakage detection using acoustic emission and machine learning algorithms. Sensors, 23(6), Article 3226. https://doi.org/10.3390/s23063226
29. Wahono, T., Purniawan, A., Mukhlash, I., & Putri, E. R. M. (2025). Risk-based asset integrity management in the oil and gas industry from traditional to machine learning approaches: A systematic review. Results in Engineering, 28, Article 107287. https://doi.org/10.1016/j.rineng.2025.107287
30. Wang, J., Nadarajah, S., Wang, J., & Ravikumar, A. (2020). A machine learning approach to methane emissions mitigation in the oil and gas industry. https://doi.org/10.31223/X57W29
31. Weller, Z. D., Hamburg, S. P., & von Fischer, J. C. (2020). A national estimate of methane leakage from pipeline mains in natural gas local distribution systems. Environmental Science & Technology, 54(14), Article 10.1021/acs.est.0c00437. https://doi.org/10.1021/acs.est.0c00437