Journal of High School Science

Redefining the data foundations of turbulence research: a physics-informed active sampling approach

Kang S

Received: August 3, 2025, Revised: version 1, October 21, 2025, version 2, December 16, 2025
Accepted: December 17, 2025

Abstract
Poorly organized flight-state data remain a major obstacle in the reliable prediction of aviation turbulence, as instability-prone flight conditions are often buried among abundant routine operations. When these rare but critical states are obscured, predictive models fail to recognize the aerodynamic patterns that precede turbulence, reducing sensitivity to real-world hazardous events. This study explores whether combining uncertainty-based active sampling with physics-informed selection criteria can yield more coherent and structurally organized datasets, in which physically distinct rare states emerge as well-represented groups separate from routine operational data, thereby enhancing both dataset diversity and the physical interpretability of data groupings. Flight data were processed through a dual-layer sampling framework that assigns adaptive weighting to statistical uncertainty and aerodynamic relevance, and selected datapoints were represented in a multidimensional feature space using k-means clustering to evaluate improvements in data organization. Results showed that this combined approach produced clearer separation between physically distinct flight regimes and yielded higher clustering quality metrics than conventional sampling. Beyond these results, the core contribution of this work is its pioneering nature—it introduces machine-driven active sampling with physics-informed reasoning at the data-preparation stage, a direction rarely explored in aerospace research. The scope of this study is specifically focused on evaluating whether combining uncertainty-based and physics-informed sampling can improve dataset structural quality, as measured through clustering coherence metrics (e.g., silhouette score, Calinski-Harabasz index, Davies-Bouldin index), rather than directly assessing downstream predictive model performance. This study should therefore be regarded as a proof-of-concept that examines the potential of adaptive, physics-aware data curation strategies to enhance the intrinsic organization and coherence of turbulence datasets, establishing a foundation for future work on turbulence modeling.

Keywords
Active sampling, Physics-informed sampling, Clustering quality, Data efficiency, Flight data, Machine Learning, Dataset structural quality, Turbulence dataset, Turbulence modeling, Aviation

Sanha Kang, Randolph-Macon Academy, 200 Academy Dr, Front Royal, VA 22630, USA. skang@rmalife.net

https://doi.org/10.64336/001c.154772

1. Introduction

Aviation turbulence poses a persistent challenge to both safety and efficiency in modern air travel. Severe turbulence events are responsible for hundreds of passenger injuries each year, along with costly maintenance and operational disruptions; in the United States alone, the annual financial burden is estimated at ~ $200 million (1). Recent climatological analyses further suggest that turbulence intensity is likely to increase in the coming decades due to shifts in jet-stream dynamics (1). Meeting this challenge requires reliable predictive systems, but the effectiveness of such models depends not only on the volume of data they ingest but also on how that data is organized. When flight-state records are scattered without clear structure, model effectiveness suffers, as organized datasets are essential for robust pattern recognition and feature learning. Additionally, without clear organization, rare instability-prone conditions risk being overshadowed by the abundance of routine operations. Prior work on turbulence forecasting has noted that when rare turbulence cases are not well distinguished, classifiers fail to learn their unique patterns effectively, limiting generalization to real-world conditions (2). Improving dataset organization—by deliberately structuring flight records into clearer, more balanced groups—offers a way to sharpen the separation between stable and unstable regimes, providing a stronger foundation for predictive models and addressing an urgent need in aviation safety.

Indeed, organizing flight-state data into meaningful groups is a critical step for applying machine learning to aviation research. Structured organization fundamentally enhances model learning by providing coherent patterns for algorithms to recognize. When data points are scattered without clear structure, this not only weakens the foundation for predictive models generally, but also allows unusual or unstable conditions to be buried among routine operations, obscuring safety-critical flight states such as those associated with aerodynamic instability.

Current studies in turbulence detection have made progress by applying machine learning to increasingly large datasets, yet they rarely address how those datasets are structurally organized. Despite the importance of data organization, prior research on turbulence detection has rarely examined clustering quality directly. Much of the prior work has treated raw flight records as inputs to classification models without considering whether diverse flight characteristics are adequately represented and distinguishable from routine operations within the dataset structure. As a result, rare conditions associated with turbulence often remain underrepresented and embedded within dominant clusters of stable flight data, which limits the model's ability to generalize and reliably identify instability. Similar challenges have been noted in broader aviation anomaly detection, where rare but operationally significant events are overshadowed by abundant normal states, causing classifiers to degrade or fail when confronted with imbalance (3).

These limitations highlight a gap in the field: while predictive modeling has been emphasized, the preparatory step of dataset organization—ensuring both coherent representation of underlying physics and adequate physically distinct states from routine data—has received far less attention.

Addressing this gap requires not only larger datasets but deliberate strategies for structuring them, so that the conditions most relevant to safety emerge as distinct and consistently represented groups.

This study addresses this unmet need with two complementary novelties. First, it identifies the data-preparation stage as a neglected yet decisive frontier in aerospace data science and explores how this foundational stage can be expanded and developed to yield more robust and interpretable results. Rather than treating dataset construction as a passive preprocessing step, this work reframes it as an active design process central to model reliability. Second, it introduces machine-driven active-sampling strategies—an approach rarely explored in aerospace—to strengthen this foundation through adaptive, physics-aware organization of flight data. The integration of machine-based data selection into aviation research remains in its earliest stages, and this work is among the first to demonstrate its feasibility and potential. The core contribution of this paper therefore lies not in exhaustive mathematical or physical validation, but in its pioneering nature: it should be understood as a proof-of-concept that highlights the emerging potential of combining physics-informed reasoning with algorithmic data design to advance aerospace research.

Our results show that integrating uncertainty-based active sampling with physics-informed criteria yields measurable improvements in the organization and clustering of flight-state data. Compared to conventional sampling approaches, this dual strategy produced better clustering coherence and dataset organization, with improved representation of physically diverse and uncommon flight conditions, as evidenced by higher Silhouette, Calinski-Harabasz, and Davies-Bouldin scores. Beyond these measurable outcomes, the core contribution of this work is its introduction of machine-driven active-sampling and physics-informed reasoning into the data-preparation stage of aerospace research, which has been almost entirely unexplored in the field. While most prior studies have focused on predictive modeling, this work redefines the starting point of aviation data science by demonstrating that innovation can emerge from how datasets themselves are built and refined. Its intent lies in showing what becomes possible when adaptive, physics-aware sampling is applied to strengthen the foundations of turbulence modeling—a framework that could inspire future research across other safety-critical domains where rare events are easily overshadowed by routine data.

2. Related work

Machine learning has transformed scientific and engineering disciplines over the past decade. Advances in deep learning, probabilistic modeling, and data-driven optimization have enabled the extraction of meaningful patterns from increasingly complex and high-dimensional datasets (4, 5). These developments have expanded the boundaries of predictive modeling, making it possible to capture nonlinear relationships in dynamic systems across domains such as weather forecasting, material design, and aerospace analytics. The field continues to progress toward methods that improve interpretability, scalability, and robustness, establishing machine learning as a central tool for both research and applied innovation.

In the context of aviation, recent research has applied machine learning to predict turbulence and related flight risks. Early studies explored ensemble and tree-based methods such as regression trees, random forests, and gradient-boosted regression trees for upper-level turbulence forecasting using meteorological inputs (6). Subsequent work extended these methods to regional scales, with XGBoost algorithms producing accurate turbulence forecasts by combining multiple meteorological indices (7). More recent developments have incorporated hybrid and deep-learning architectures—such as conditional generative adversarial networks combined with gradient-boosted models—to detect small- and medium-scale turbulence events with improved robustness and reduced false alarms (2). While these efforts show the growing adoption of machine learning in aviation, the field remains relatively conservative; novel approaches are often integrated slowly due to the industry's emphasis on safety, standardization, and verifiable performance. As a result, innovation tends to occur incrementally, with most research centering on predictive accuracy rather than rethinking the underlying data design.

Despite these advances, most turbulence-focused studies have concentrated on refining model architectures rather than improving the structure of the data feeding those models. Current research tends to assume that the training datasets themselves are already representative of real world flight conditions. However, turbulence events are inherently sparse and imbalanced, often buried among abundant stable operations (8). This imbalance can limit model sensitivity to rare but safety-critical patterns. Prior work has rarely explored how strategic data sampling—especially methods combining statistical uncertainty with aerodynamic criteria—can enhance dataset clarity or representativeness. This gap highlights the need for future work to focus not only on optimizing predictive algorithms but also on developing methods that actively improve dataset balance, representativeness, and coverage within safety-critical domains like aviation.

3. Methods

3.1 Design of the dual-layer sampling framework

This section introduces the sampling framework we tested to improve the organization of flight state data. The framework selected datapoints using two complementary criteria (uncertainty based active sampling and physics-based scoring) and represented them in a feature space where k-means clustering was applied. The goal was to evaluate whether combining these two approaches could expose rare, physically-relevant states and produce clearer, more balanced clusters. The framework operated with adaptive weighting of both criteria. The uncertainty component prioritized points where the model was least confident, while the physics component emphasized aerodynamically relevant states. Together, they formed a dual-layer selection strategy that represents the novelty of this work. Unlike prior methods that applied physics or uncertainty alone, this integrated approach was tested here as a new way to enhance aviation dataset quality.

3.2 Data source and preprocessing

The raw material for this study was flight-state data continuously broadcast by aircraft during operation. Modern aircraft transponders regularly send signals containing altitude, ground speed, heading, and vertical motion, which are collected and made publicly accessible through the OpenSky Network. Each signal is a direct snapshot of conditions that an aircraft is experiencing at that moment in flight. Unlike simulated datasets, these records capture real operational states, making them an appropriate foundation for testing whether data organization strategies can be improved.

In our framework, this broadcast information functioned as the dataset to be organized. Its role was to provide the diverse flight conditions on which clustering quality could be evaluated. The value of using broadcast flight data lay in its breadth and realism: thousands of aircraft across global airspace contribute records, creating a dataset that included both common, routine states and rare, instability-prone states. Such diversity is critical for assessing whether new sampling strategies can structure data more effectively.

Each broadcast was treated as a snapshot and placed into a feature space, a mathematical coordinate system where each axis represented a flight variable. This feature space was not the aircraft's physical trajectory but an abstract representation that allowed systematic comparison between many flights. From this representation, clustering techniques could group similar states and reveal where rare or unusual conditions stood apart from the routine.

To focus on steady cruise conditions, where turbulence-related variations were most relevant, we applied clear filtering criteria. Only aircraft flying between 28,000 and 43,000 feet altitude were included, representing commercial cruise altitudes. Vertical rate was constrained to ≤200 ft/min (30-second smoothed; approximately ±1.0 m/s) to ensure level flight. Altitude stability was constrained to ≤500 ft from cruise level to isolate stable cruise conditions. These thresholds produced a dataset suited for testing how well clustering methods could separate routine and non-routine flight behaviors.

The cruise-state definition employed in this study was grounded in documented aircraft performance characteristics, operational surveillance standards, and turbulence climatology research. The altitude bounds of 28,000–43,000 ft corresponded to the certified cruise envelopes of modern commercial jet transports: Boeing documents list typical cruise altitudes near 35,000 ft for the 777 family, while Airbus specifies nominal cruise near 39,000 ft for the A320/321. The Airbus A350 has a service ceiling of approximately 43,000 ft. These values aligned with industry-standard cruise flight levels and matched the altitude band where clear-air turbulence is most frequently encountered. Researchers have documented significant clear-air turbulence occurrence at cruise altitudes in their climatological analyses of the North Atlantic flight corridor (9). Most turbulence-related injuries and encounters occur during high-altitude cruise, particularly when passengers and crew are unbuckled (10, 11).

The vertical-rate and altitude-stability thresholds followed established criteria for

detecting level flight in surveillance data. Academic studies using ADS-B data for flight phase identification have employed various thresholds to identify stable cruise segments while accounting for measurement noise and normal flight variations (12-14) . Our selected thresholds (vertical rate ≤200 ft/min and altitude deviation ≤500 ft) fell within the range of values used in published flight phase identification research.

This definition intentionally restricted analysis to commercial jet transport aircraft, consistent with turbulence risk assessment protocols. The IATA Turbulence Aware Program, Boeing's turbulence guidance documentation, and NTSB Part 121 injury reports emphasize that turbulence-related injury risk, economic impact, and operational relevance are overwhelmingly concentrated in commercial jets operating at high cruise altitudes. General aviation traffic, which typically operates below 25,000 ft, encounters fundamentally different meteorological and aerodynamic conditions and was therefore excluded.

The four flight variables retained are listed in Table 1. After filtering, each query typically produced 30– 100 usable flight snapshots. Rows with missing values were removed, and all variables were standardized using z-score normalization so that differences in units did not bias clustering.

Table 1. Selected state vector variables from the OpenSky dataset for turbulence analysis.

Feature Name	OpenSky Field	Units	Description
Altitude	baro_altitude	meters	Pressure altitude of the aircraft
Vertical Rate	vertical_rate	meters/sec	Vertical climb/descent rate
Groundspeed	velocity	meters/sec	Aircraft ground-relative speed
Heading	heading	degrees	Aircraft ground track angle

In this study, the flight-state records were treated as a foundation to provide a consistent and transparent base of real-world conditions on which different sampling and clustering strategies could be tested. Based on this real-world dataset, our work was able to demonstrate how the dataset's organization could be strengthened through active sampling based on uncertainty and physics principles.

3.3 Uncertainty-based sampling layer

The first layer of our active sampling framework was built on uncertainty-based sampling. This approach addressed the problem of deciding which flight states were most worth including in the dataset. Unlike passive strategies that add data at random, uncertainty-based sampling deliberately targets conditions that are unusual, underrepresented, or uncertain. Simply enlarging the dataset does not guarantee better learning, since most random additions consist of routine, already well-represented states. From a machine learning perspective, this imbalance is critical: when instability-prone conditions are vastly outnumbered, their influence on the loss function becomes negligible, causing the model to generalize toward normal flight behaviors

and fail to capture the decision boundaries that define aerodynamic instability. Highlighting these rare states ensures that the dataset reflects the full range of flight conditions and allows the model to recognize subtle precursors to turbulence. By focusing on rare or uncertain cases, the dataset grows in a way that is not only larger but also more informative, sharpening the contrast between stable and unstable flight regimes.

To put this principle into practice, we required a way to measure which regions of the dataset were already well represented and which remained unexplored. For this purpose, we used clustering. We clustered groups together with similar flight state, such as those with comparable speeds, altitudes, or vertical rates, so that the overall structure of the dataset became visible. Once these groups were formed, dense regions could be distinguished from sparse ones. Uncertainty-based sampling then prioritized the sparse or boundary regions, since those areas were most likely to yield new information and expose patterns that would otherwise remain hidden.

In our implementation, the k-means algorithm was applied at each sampling cycle to partition the accumulated data into clusters. New data was incorporated iteratively so that the clustering reflected the evolving structure of the dataset. All features were normalized to prevent any single variable from dominating the distance calculations. The uncertainty of a new flight state was then defined as its distance from the nearest cluster center: the farther away a point lay, the more likely it was underrepresented and thus prioritized for inclusion.

This uncertainty-based sampling layer's role was to create a more balanced and comprehensive dataset by systematically expanding coverage into regions that would otherwise remain neglected. The result was a dataset that better reflected the diversity of real-world flight conditions, ensuring that rare instability-prone states were not drowned out by the majority of routine data. Combined with physics-based layer which will be introduced in section 3.4, the uncertainty-based layer described here served as the essential first step in our two-part framework.

3.4 Physics-based sampling layer

The second layer of our framework built on top of the uncertainty-based sampling described in Section 3.3. While uncertainty alone helped identify underexplored regions of the dataset, it did not guarantee that the selected points were physically relevant to turbulence. To address this gap, we introduced a physics-based sampling layer that highlighted flight states hypothesized to reflect aerodynamic instability. The aim of this layer was not to predict turbulence directly, but to test whether including physically meaningful indicators could improve the organization of the dataset, making rare and unstable states more clearly represented.

Furthermore, the objective of the layer was not to add datasets that were proven to be turbulent or proxies of turbulence. Rather, it aimed to enhance dataset diversity and structural coherence by incorporating physically uncommon flight conditions. We employed a physics-based sampling layer to examine whether enriching the dataset with these physically meaningful but uncommon samples improved clustering cohesion as measured by

silhouette, Calinski-Harabasz, and Davies-Bouldin metrics.

We also clarify that using datasets labeled as turbulence would be impractical, as turbulent events are extremely rare in operational flight data—aircraft rarely encounter turbulence severe enough to warrant such labels. However, this limitation was not relevant to our study. We were not attempting to validate turbulence labels or predictions; rather, we were simply selecting physically infrequent flight conditions based on aerodynamic criteria to enhance dataset diversity and assess whether such selection improved clustering coherence.

In practice, five physics-informed features were derived from the OpenSky dataset and standardized to allow equal contribution (Table 2). Each feature corresponded to a distinct dimension of aerodynamic or dynamic instability. The final physics score was then computed as a weighted sum of these normalized features, producing a single value used to guide the selection of data points.

The central hypothesis tested was that adding features assumed to represent physically meaning conditions would, in fact, increase clustering quality. If the hypothesis were true, the system would not only sample data that was statistically diverse but also emphasize those states most useful for machine learning models —states that are both rare and physically meaningful.

Table 2. Aerodynamic and dynamic features used in the physics-based turbulence score.

Feature Name	Description	Formula	Physical Meaning
Dynamic Pressure	Kinetic energy per unit volume	$q = 1/2 * \rho * v^2$	Measures airspeed-related aerodynamic force
Mach Gradient	Change in normalized airspeed	$M = V/a$ $a = \text{sqrt}(\gamma RT)$ Mach Gradient = dM/dt	Rate of change of Mach number
Energy Imbalance	Ratio of potential and kinetic energy terms	$EI = \lvert g \cdot ALT - \frac{1}{2}V2 \rvert / (g \cdot ALT + \frac{1}{2}V2 + \epsilon)$	Imbalance between potential and kinetic energy
Vertical Jerk	Rate of change of vertical acceleration	J = da/dt	Acceleration instability in vertical axis
Heading Variation	Instantaneous heading change	d(Heading)/dt	Lateral directional instability

In the Energy Imbalance equation, ε is a numerical stability constant with dimensions [m²/s²] ($\varepsilon = 10^{-6}$ m²/s²) to prevent division by zero

Physics-derived features expose operationally significant flight behaviors that remain concealed within routine operational data when only raw measurements are analyzed. This study specifically employed dynamic pressure, Mach gradient, energy imbalance, vertical jerk, and heading variation because these five variables represent fundamental aerodynamic and energy-state indicators that reveal operationally significant flight regimes obscured by raw kinematic measurements alone. Dynamic pressure ($q = \frac{1}{2}\rho V^2$) was

selected because it directly quantifies aerodynamic loading and integrates both velocity and atmospheric density effects, enabling detection of high-energy states at low altitudes versus low-energy states at high altitudes—thereby exposing safety-critical conditions such as excessive speed during descent or inadequate energy during climb that remain hidden in velocity measurements alone (15). Mach gradient (dM/dt) captures the rate of change in compressibility effects, identifying rapid accelerations or decelerations that indicate thrust changes, emergency maneuvers, or engine performance anomalies —all rare but operationally significant events that clustering algorithms must prioritize (16, 17). Energy imbalance quantifies deviations from optimal flight paths by measuring discrepancies between kinetic and potential energy changes, revealing suboptimal operations including premature descents, excessive speed variations, and fuel-inefficient altitude-speed combinations that emerge only when energy relationships are explicitly modeled (18, 19). Vertical jerk (d^3h/dt^3) captures the smoothness or abruptness of altitude changes, distinguishing controlled climbs from sudden vertical maneuvers indicative of turbulence encounters, traffic conflicts, or emergency responses that stress airframe components and signal procedural deviations (20, 21). Heading variation ($d\psi/dt$) measures turning intensity, revealing holding patterns, traffic avoidance, and weather deviations that are indistinguishable from heading values alone (22). These five physics-derived variables formed an aerodynamically complete representation capturing energy management, compressibility dynamics, control smoothness, and directional maneuvering—all validated as essential for distinguishing normal operations from anomalous behaviors in unlabeled flight data (23, 24). This physics-based sampling strategy ensured that the final clustered dataset captured the full operational envelope of aircraft behavior rather than merely reflecting the statistical dominance of routine flight, thereby correcting the class imbalance problem where normal operations obscure the anomalous patterns that unsupervised learning must discover.

The weighted sum of these features yielded a physics-informed score that complemented uncertainty-based sampling. After testing multiple weighting schemes, the heuristic combination that consistently maximized clustering quality assigned weights of 0.3 to dynamic pressure and Mach gradient, 0.2 to energy imbalance, and 0.1 each to vertical jerk and heading variation. These weights reflected the relative contribution of different physical dimensions while ensuring that no single feature dominated the score. Together, the two layers allowed us to test whether rare, instability-prone conditions could be highlighted in a way that produced more organized clusters, thereby strengthening the dataset as a resource for downstream turbulence research.

3.4.1 *Atmospheric corrections and aerodynamic variable computation*

The physics-based sampling layer relied on aerodynamic variables that reflected the aircraft's motion relative to the surrounding air mass and the thermodynamic state of the atmosphere. All quantities of true airspeed, Mach number, dynamic pressure, air density, and directional-instability metrics were

constructed from first principles to ensure physical fidelity. This subsection provides a consolidated and technically detailed description of the models and mathematical procedures used throughout the study.

To obtain air-relative velocity, groundspeed from ADS-B broadcasts was combined with wind information through a vector-projection framework. Wind velocity components were obtained from the ERA5 reanalysis dataset, which provided global atmospheric conditions at hourly intervals on a 0.25° × 0.25° horizontal grid with 37 pressure levels. For each measurement point along the aircraft flight path, wind data (u and v components) were extracted at the corresponding pressure level and spatially interpolated (bilinear/nearest neighbor) to match the aircraft's position (latitude, longitude) and temporally interpolated to the exact measurement time. The air-relative velocity was then computed by subtracting the wind velocity from the aircraft's ground-relative velocity vector obtained from the onboard navigation system. The aircraft heading and wind direction were first converted into two-dimensional unit vectors, $u_h=(\sin\theta_h,\cos\theta_h)$ and $u_w=(\sin\theta_w,\cos\theta_w)$ The wind velocity vector $W=W_s u_w$ was projected onto the flight-path direction via $W_{\parallel}=W\cdot u_h=W_s\cos(\theta_w-\theta_h)$. True airspeed was then computed as $V_{TAS}=V_g-W_{\parallel}$. This air-relative velocity formed the basis for all subsequent aerodynamic calculations and ensured that crosswind components did not artificially inflate or reduce the apparent kinetic state of the aircraft.

The speed of sound was computed as a function of altitude using the International Standard Atmosphere (ISA) temperature profile, $T(h)=T_0-Lh$ with $L=6.5\ K/km$. The local sound speed followed $a(h)=\sqrt{\gamma RT(h)}$ with $\gamma=1.4$ and $R=287\ J\,kg^{-1}\,K^{-1}$. Mach number is then given by $M=V_{TAS}/a(h)$. This thermodynamically consistent formulation avoided the distortions associated with assuming a constant sound speed, which is known to vary meaningfully across typical cruise altitudes.

Dynamic pressure, a key indicator of aerodynamic loading, was evaluated using the standard expression $q=\frac{1}{2}\rho(h)V_{TAS}^2$. To compute altitude-dependent density, the barometric relation was applied, where $P_0=101{,}325\ Pa$ and $g=9.80665\ m/s^2$. This ensured that air density decreased realistically with altitude and remained coupled to temperature through the ISA profile. The combination of V_{TAS}, $a(h)$, and $\rho(h)$ yielded Mach and dynamic-pressure values consistent with established aerodynamic theory.

Directional quantities were treated using circular geometry to ensure mathematical correctness. Since heading is periodic, it was embedded into Euclidean space using $(\sin\theta_h,\cos\theta_h)$, preventing discontinuities at the 0°/360° boundary. Directional instability was quantified through the time-rate of rotation on the unit circle by differentiating these sine and cosine components and computing $\dot{\theta}=\sqrt{(\dot{\sin\theta_h})^2+(\dot{\cos\theta_h})^2}$. This formulation yielded a physically meaningful measure of lateral maneuvering or heading fluctuation that was invariant to wrap-around and did not suffer

from the artifacts inherent to linear angle differences.

Together, these procedures provided a physically rigorous foundation for the aerodynamic features used in the sampling framework. By grounding all variables—Mach number, dynamic pressure, density, TAS, and directional-instability metrics—in validated atmospheric models and air-relative kinematics, the physics-based layer captured meaningful indicators of instability across real-world cruise conditions while ensuring internal consistency within the feature space.

3.5 Adaptive weighting for physics-uncertainty fusion

A fundamental challenge in combining physics-informed and uncertainty-based acquisition functions is determining their relative weighting. Fixed weighting schemes assume static signal importance throughout active learning, but this fails to account for dynamic informativeness: physics scores may be highly discriminative when diverse extreme conditions remain in the pool but saturate in later cycles as high-risk samples are depleted, while uncertainty scores evolve with the changing cluster structure as labeled samples accumulate. This temporal non-stationarity creates a critical opportunity: deploying each method's strengths precisely when and where they are most needed rather than forcing a uniform compromise. By strategically activating physics-informed selection during risk-critical phases and uncertainty-driven exploration during coverage-critical phases, an adaptive weighting system amplifies the core innovation of our hybrid approach—maximizing synergistic advantages by ensuring each method governs at its moment of comparative strength.

At each cycle t, we normalize both physics and uncertainty scores to [0,1] across n pool candidates, then calculate their variances σ^2_P and σ^2_U. The adaptive weight is computed as $\alpha\_t = \sigma^2\_U / (\sigma^2\_U + \sigma^2\_P + \varepsilon)$, and the final acquisition score is $\alpha\_t \cdot P + (1 - \alpha\_t) \cdot U$. When uncertainty exhibits high variance ($\sigma^2\_U >> \sigma^2\_P$), indicating pool candidates are distributed across diverse cluster boundaries, VarAlpha—the adaptive weighting system—increases reliance on physics to prevent over-concentration on boundary ambiguities while ensuring safety-critical regions are prioritized. Conversely, when physics exhibits high variance ($\sigma^2\_P >> \sigma^2\_U$), indicating abundant extreme conditions remain, VarAlpha emphasizes uncertainty to ensure diverse cluster coverage rather than redundantly sampling physics outliers. This negative-feedback mechanism ensures each method governs during its regime of comparative advantage—physics for risk-critical selection, uncertainty for coverage-critical exploration.

This adaptive deployment maximizes the synergistic benefits of our physics-uncertainty integration. Rather than physics and uncertainty competing through fixed weights, VarAlpha orchestrates their complementary strengths across the active learning trajectory: physics dominates when risk discrimination is paramount, while uncertainty dominates when cluster structure exploration is critical. Importantly, VarAlpha operates on variance of acquisition scores (scalar selection priorities within the current pool) rather than variance of feature vectors (multidimensional data points in the accumulated labeled set) measured by

clustering metrics like Calinski-Harabasz, ensuring the mechanism addresses selection dynamics rather than metric optimization.

3.5.1 *Comparison with fixed weighting schemes*

We validated VarAlpha's effectiveness through comprehensive experiments comparing against fixed weight baselines ($\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$) across 5 bootstrap replicates and 5 random seeds across 50 cycles (Section 4.1). The study isolated whether adaptive, context-aware deployment provided genuine benefit over carefully tuned static weights, demonstrating the value of strategically activating each method at appropriate moments throughout the active learning process.

3.6 Sampling strategies compared

To evaluate the effectiveness of our approach, we compared six sampling strategies against our proposed combined method. To examine whether the advantages of our two-layer framework held in practice, we designed experiments that compared its performance against multiple baseline methods. The goal was twofold: first, to test whether physics or uncertainty-based active sampling alone offered a measurable improvement over random selection, and second, to determine whether adding the physics-informed layer provided an additional and necessary benefit beyond uncertainty alone.

Accordingly, we evaluated seven strategies under identical conditions. The baselines included: [1] random sampling, a conventional baseline in which flight states were chosen without guidance; [2] uncertainty-based active sampling, representing the single-layer version of our framework that used only the uncertainty criteria; [3] physics-only sampling that used only the physics-informed criteria described in Section 3.4; [4] entropy-based sampling, which selected samples that maximized prediction entropy $H(y|x) = -\Sigma p(y|x)\log p(y|x)$, prioritizing instances where the model exhibited maximum confusion across all possible classes—we included this method because it represented a foundational information-theoretic approach to active learning and provided a principled way to quantify model uncertainty; [5] margin-based sampling, which selected instances with the smallest difference between the top two predicted class probabilities, effectively targeting decision boundary cases where the model was least confident in distinguishing between competing hypotheses—this method was particularly relevant for aviation data where distinguishing between similar flight regimes is critical; and [6] core-set sampling, which formulated sample selection as a k-center problem to ensure the selected subset geometrically represented the full dataset distribution, thereby reducing redundancy—we included this geometric diversity-based approach to contrast with uncertainty-based methods and evaluate whether representativeness alone sufficed for turbulence modeling. Our proposed method combined uncertainty-based active sampling with the physics-informed criteria. This structure allowed us to separately assess the effect of each layer and to evaluate whether the integration of both layers produced a meaningful gain in dataset organization.

3.7 Bootstrap resampling and seed control for data robustness

Bootstrap resampling was implemented as a non-parametric data-level perturbation layer to

explicitly model uncertainty arising from finite, irregular, and regionally biased ADS-B sampling. For each replicate, a full-size dataset was generated by resampling the original dataset with replacement, such that each bootstrap replicate consisted solely of observations drawn from the original feature pool. Each replicate therefore represented a perturbed reallocation of the original observations rather than a synthetically generated dataset in order to ensure that performance comparisons were not biased by any single fixed dataset realization.

In our implementation, five independent bootstrap replicates were constructed, each containing resampled points drawn from the original feature matrix. The number of unique samples per replicate was explicitly tracked, ensuring that each replicate introduced natural duplication, omission, and density distortion effects. The full 50-cycle active sampling process was executed independently on each bootstrap dataset, forcing every method to operate under dynamically reconfigured data distributions rather than on a single static realization.

All reported Silhouette, Calinski-Harabasz, Davies-Bouldin, and risk-coverage results were aggregated across the bootstrap hierarchy, yielding empirical performance distributions rather than single-run point estimates. This design suppressed optimistic bias arising from any single fixed dataset realization by exposing each method to multiple independently resampled versions of the same empirical flight population. Since each replicate reordered the empirical density, local cluster structure, and candidate pool composition, the evaluation explicitly stress-tested sensitivity to low-probability sampling artifacts, regional sparsity, and pool-depletion dynamics that could not be revealed through single-pass evaluation. Consequently, observed performance differences were attributable to stable method-level behavior rather than incidental properties of a particular ADS-B snapshot, thereby significantly strengthening both internal statistical validity and external generalizability.

Furthermore, random seeding was used to control all sources of algorithmic randomness in the experimental pipeline in order to separate stochastic optimization effects from data-level variability introduced by bootstrap resampling. In this study, stochasticity enters through clustering initialization and through probabilistic sampling steps within the active learning procedures. Fixing the random seed fully determines the sequence of these stochastic decisions within a single experimental run.

In our implementation, each bootstrap replicate was evaluated under five independent fixed random seeds, meaning that the same resampled dataset was processed five separate times under different stochastic initial conditions. This produced multiple independent sampling and clustering trajectories for each bootstrap dataset, resulting in 25 independent evaluations per cycle across the nested bootstrap–seed structure. This design explicitly probed sensitivity to initialization-dependent optimization paths and sampling order effects.

All reported performance metrics were aggregated across this bootstrap–seed hierarchy, ensuring that performance trends were not driven by favorable or unfavorable

single random trajectories. By repeatedly re-evaluating each data realization under multiple independent stochastic paths, the evaluation suppressed random initialization bias and ensured that observed differences reflected stable method behavior rather than chance outcomes from any specific stochastic configuration, thereby strengthening the reliability and interpretability of the results.

3.8 Hierarchical evaluation framework

Across all experiments, a matched sample size of 30 samples per cycle was enforced for all methods over 50 active learning cycles to ensure that performance differences reflected algorithmic behavior rather than data volume effects. Each complete 50-cycle trajectory at every bootstrap–seed combination was further repeated 30 times to achieve a statistical sample size of $n = 30$ for hypothesis testing, yielding a total of $5 \times 5 \times 50 \times 30 = 37{,}500$ individual evaluations per sampling method. Performance metrics were reported as mean ± standard deviation across replications, and statistical significance was assessed using these n=30 trajectory-level observations. This hierarchical experimental design—combining bootstrap resampling, multi-seed evaluation, cycle-level progression, and repeated statistical trials—ensured that observed performance differences originated from intrinsic method behavior rather than being artifacts of specific data subsets, random initializations, or from insufficient sampling.

3.9 Evaluation metrics

The primary objective of this work was to enhance the internal coherence and structural organization of flight state data derived from ADS-B observations. Specifically, we sought to reorganize raw trajectory data such that rare or dynamically significant flight states exhibited clear separation and were preserved as distinct entities within the feature space, rather than being subsumed within densely populated clusters of routine operational conditions. Through systematic improvement of internal cohesion, inter-cluster separation, and overall structural distinctiveness, we aimed to construct a dataset geometry that reflected physically meaningful patterns and provided a more robust foundation for subsequent analytical endeavors.

To quantify this internal coherence, we employed three complementary clustering quality metrics: the Silhouette Score, the Calinski-Harabasz (CH) Index, and the Davies-Bouldin (DB) Index. These metrics directly assessed the structural properties we aimed to optimize—namely, the compactness of individual clusters, the degree of separation between distinct groups, and the preservation of identifiable structure among underrepresented states within the broader data distribution.

It is noteworthy that substantial prior research across diverse domains has established empirical correlations between improvements in internal clustering metrics and enhanced downstream model performance in supervised learning contexts. Webb et al. demonstrated that elevated Silhouette and CH scores, coupled with reduced DB values, co-evolved with improved fault-classification accuracy in chemical process monitoring systems (25). Similarly, Fang et al. (2022) reported Spearman rank correlations ranging from $\rho \approx 0.53$ to $\rho \approx 0.74$ between internal clustering indices and external measures of label agreement in single-cell genomics datasets

(26). This pattern of correspondence between internal structure quality and external task performance has been consistently observed across multiple application domains, suggesting a generalizable relationship between data organization and predictive model efficacy.

Nevertheless, we emphasize that the training and evaluation of turbulence-prediction models lies beyond the scope of the present investigation. Furthermore, formal assessment of downstream model performance remains methodologically challenging within the turbulence research community due to the persistent scarcity of high-resolution, temporally and spatially aligned turbulence ground-truth data—a structural constraint that affects the field as a whole rather than this study in isolation. Our evaluation framework therefore focused on measuring improvements in internal data coherence through the aforementioned clustering quality metrics, with the understanding that such structural improvements may provide a more favorable foundation for future supervised learning efforts, should labeled turbulence data become accessible.

3.10 k-variable tests

Since clustering performance and sampling behavior inherently depend on the granularity of data partitioning, we performed a systematic sensitivity analysis on the cluster-number hyperparameter k to verify that observed performance differences were not artifacts of a specific clustering resolution. All methods were evaluated under $k \in \{3,4,6,8,10\}$, and for each k, the Silhouette, Calinski-Harabasz, and Davies-Bouldin metrics were recomputed and method rankings re-assessed. This directly tested whether relative method performance was stable across changes in cluster resolution. Results were reported as performance trends across k, enabling clear identification of methods that retained consistent advantages versus those that were sensitive to specific hyperparameter settings.

3.11 Robustness to measurement noise

Prior to robustness testing, we established a cleaned baseline dataset through six-stage filtering designed for ADS-B telemetry data. All temporal operations were performed independently within each aircraft's time-ordered trajectory, as rate-of-change calculations required causally related consecutive measurements from the same physical flight path. Physical constraint validation computed first-order derivatives and removed measurements violating conservative limits: vertical rates exceeding 150 ft/s (accommodating emergency procedures beyond nominal climb performance), accelerations exceeding 20 knots/s (double the typical maximum values), and turn rates exceeding 15°/s (well above standard rate turns), with heading differences computed accounting for circular wraparound. Position integrity verification calculated great-circle distances using the Haversine formula, removing GPS jumps implying ground speeds exceeding 1000 knots. Data quality assurance eliminated exact duplicate records (matching aircraft identifier, timestamp, and flight state), removed sparse trajectories (aircraft with fewer than 10 samples providing insufficient context for derivative validation), and identified time gaps exceeding 30 minutes that indicated distinct flights erroneously sharing the same identifier. Temporal smoothing applied a 5-

point centered moving average per aircraft to altitude and speed measurements, reducing high-frequency sensor noise while preserving genuine flight dynamics. This protocol retained approximately 75-85% of measurements, establishing a reliable foundation for controlled noise injection experiments.

To further evaluate whether performance characteristics observed on cleaned data generalized to realistic operational conditions where sensor measurements exhibited variability, we performed systematic robustness experiments. We introduced controlled Gaussian noise to all feature values at five progressively increasing intensity levels: $\sigma \in \{2\%, 4\%, 6\%, 8\%, 10\%\}$ of each feature's standard deviation, simulating measurement uncertainty from GPS quantization, barometric drift, and wind estimation errors encountered in operational deployment. For each noise condition, we evaluated all competing methods —Random, physics-only, uncertainty-only, entropy, margin, core-set baselines, and our proposed method—across 30 independent experimental replicates with different random seeds and bootstraps each executing 50 active sampling cycles. This design tested two critical research questions: whether observed performance differences persisted uniformly across noise regimes or exhibited differential degradation rates, and whether VarAlpha's adaptive weighting mechanism maintained effective physics-uncertainty balance when both signals were corrupted by noise.

3.12 Temporal sampling quality assessment

Temporal irregularities in ADS-B data, arising from variations in transmission rates and receiver coverage, could potentially introduce artifacts in trajectory clustering analysis. To ensure that our physics-informed approach operated on data of sufficient temporal quality and that performance comparisons were not biased by sampling inconsistencies, we conducted comprehensive temporal interval analysis prior to experimental evaluation. Sampling intervals (Δt) were calculated by grouping observations by unique aircraft identifiers and computing consecutive timestamp differences within each trajectory, ensuring that measurements reflected actual within-flight sampling rates rather than spurious gaps between unrelated aircraft. This aircraft-aware approach yielded $N = 50{,}247$ temporal intervals across 127 aircraft trajectories in our combined dataset.

Analysis revealed that OpenSky data exhibited quasi-regular temporal sampling suitable for trajectory analysis. The distribution demonstrated strong concentration at 1-second intervals (median $\Delta t = 0.98$s, IQR = [0.00, 1.29]s), with 51.8% of intervals being sub-second and 91.0% falling within 2 seconds. Only 2.0% of intervals exceeded 5 seconds, which manual inspection confirmed to be gaps between distinct flight segments (e.g., landing to taxiing) rather than within-trajectory irregularities. The tight interquartile range and high coverage below 2 seconds confirmed excellent temporal consistency across aircraft, substantially mitigating concerns about severe sampling irregularities that might confound clustering performance. This empirical characterization demonstrated that the vast majority (91%) of our dataset exhibited quasi-regular sampling comparable to controlled experimental conditions, validating its suitability for physics-based trajectory clustering.

4. Results

In this section, we present a comprehensive evaluation of our proposed combined method through five complementary experiments. We compared seven baseline sampling strategies—Random Sampling, Uncertainty-only Sampling, Physics-only Sampling, Entropy-based Sampling, Margin-based Sampling, and Core-set Sampling—against the combined method, which adaptively combined uncertainty-based and physics-based sampling through a dynamic weighting mechanism.

Our evaluation consisted of five experiments: [1] The main performance comparison evaluated all seven methods over 30 independent trials of 50 active sampling cycles, with 5 bootstrap replicates and 5 random seeds per cycle to assess clustering quality via Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index; [2] The weight ablation test evaluated nine fixed weight ratios ($\alpha \in \{0.1, 0.2, ..., 0.9\}$) against VarAlpha to isolate the contribution of adaptive weighting; [3] The K-means cluster sensitivity analysis examined performance across different numbers of clusters ($k \in \{3, 4, 6, 8, 10\}$) to validate robustness to this hyperparameter; [4] The fourth test examined the robustness to measurement noise, evaluating all methods under controlled Gaussian noise injection at five intensity levels (2%, 4%, 6%, 8%, 10%). This experiment was used for confirming that our observed performance differences reflected algorithmic advantages rather than sensitivity to data perturbations, and to ensure our findings generalized across datasets with varying levels of real-world ADS-B data quality issues. Together, these experiments provided a rigorous assessment of our combined method's effectiveness and robustness across the dataset.

4.1 Experimental results

We evaluated seven active sampling methods across 30 independent experiments: Random sampling, Entropy, Margin, Core-Set, uncertainty-only sampling, physics-only sampling, and our integrated approach (proposed method). Each of the 30 experiments used a distinct data pool, but within each experiment, all seven methods processed identical data at every cycle, implementing a paired repeated-measures design. To ensure robustness against stochastic variation, each method-cycle combination was evaluated across 25 replicate conditions (5 random seeds × 5 bootstrap samples). Each method's performance estimate was hence derived from 37,500 underlying measurements (5 seeds × 5 bootstraps × 50 cycles × 30 experiments), ensuring robust and reproducible results.

Our statistical analysis operated on 210 observations (30 experiments × 7 methods), where each observation represented a method's performance aggregated across all 50 cycles within a single experiment. We reported both the area under the curve (AUC), integrating cumulative performance trajectories across cycles, and the average performance (Avg). Since each datapoint synthesized information from 1,250 measurements per experiment (5 seeds × 5 bootstraps × 50 cycles), our performance metrics exhibited high reproducibility and narrow confidence intervals, reflecting method differences rather than measurement noise. The paired design further eliminated between-experiment variance by ensuring that all methods were

exposed to identical selection challenges within each of the 30 independent replications.

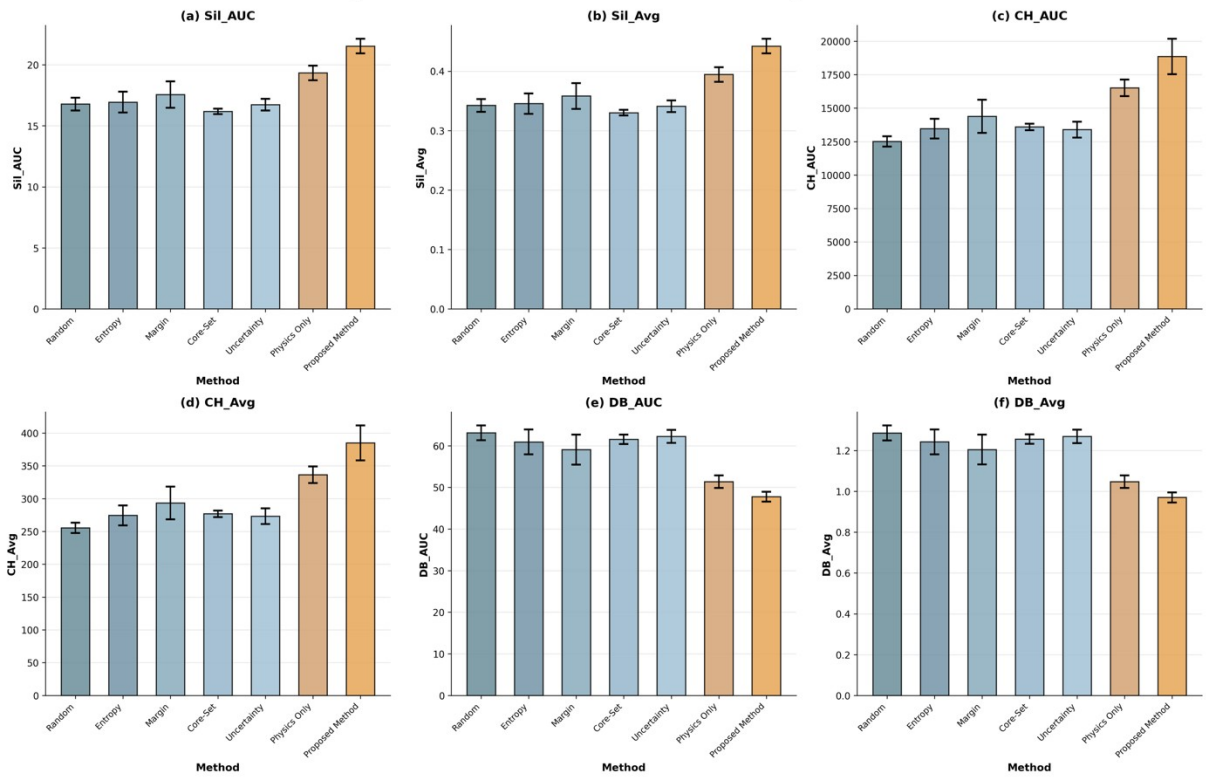

Figure 1. Comparison of sampling methods across clustering quality metrics.

Performance was evaluated using three clustering metrics, with two measures (AUC and Average) per metric, for a total of 6 derived metrics: Silhouette coefficient (Sil_AUC, Sil_Avg), Calinski-Harabasz index (CH_AUC, CH_Avg), and Davies-Bouldin index (DB_AUC, DB_Avg). Each of 30 independent experiments tested all seven methods in a repeated-measures design.

The mean performance across six clustering metrics over 30 independent experiments is presented in Figure 1 and Table 3. The proposed method achieved the best performance on five out of the six metrics, with the Physics Only method consistently ranking second.

For Silhouette Average, the proposed method reached 0.44 ± 0.01 (95% CI: [0.44, 0.45]), outperforming baseline metrics. The uncertainty-based methods (Random, Entropy, Margin, Uncertainty, Core-Set) clustered between 0.33-0.36. On the Calinski-Harabasz Average, the proposed method scored 384.96 ± 26.57 compared to Physics Only's 336.46 ± 12.65, significantly higher than baselines (255-293 range). For the Davies-Bouldin Average, the proposed method achieved 0.97 ± 0.02, outperforming Physics Only (1.05 ± 0.03) and all baselines (1.20-1.29 range).

Table 3. Comparison of sampling methods (mean ± SD [95% CI]) across clustering quality metrics.

Method	Sil_AUC	Sil_Avg	CH_AUC	CH_Avg	DB_AUC	DB_Avg
Random	16.79±0.52, [16.59, 16.98]	0.342±0.011, [0.338, 0.346]	12518±385, [12374, 12662]	255.5±7.8, [252.5, 258.4]	63.09±1.78, [62.43, 63.76]	1.286±0.037, [1.272, 1.300]
Entropy	16.94±0.85, [16.63, 17.26]	0.345±0.017, [0.339, 0.352]	13465±743, [13188, 13742]	274.6±15.2, [268.9, 280.2]	60.93±3.01, [59.81, 62.05]	1.242±0.061, [1.219, 1.265]
Margin	17.56±1.08, [17.16, 17.96]	0.358±0.022, [0.350, 0.366]	14389±1236, [13927, 14850]	293.5±25.0, [284.2, 302.8]	59.08±3.63, [57.72, 60.43]	1.204±0.073, [1.177, 1.232]
Core-Set	16.18±0.23, [16.10, 16.27]	0.330±0.005, [0.328, 0.332]	13594±242, [13503, 13684]	276.9±5.0, [275.1, 278.8]	61.56±1.14, [61.13, 61.99]	1.256±0.023, [1.247, 1.265]
Uncertainty	16.73±0.48, [16.55, 16.91]	0.341±0.010, [0.338, 0.345]	13392±601, [13168, 13616]	273.3±12.1, [268.7, 277.8]	62.24±1.57, [61.66, 62.83]	1.269±0.033, [1.257, 1.281]
Physics Only	19.33±0.59, [19.11, 19.56]	0.394±0.012, [0.390, 0.399]	16518±624, [16285, 16751]	336.5±12.6, [331.7, 341.2]	51.38±1.51, [50.81, 51.94]	1.047±0.030, [1.036, 1.059]
proposed method	21.53±0.60, [21.31, 21.76]	0.442±0.012, [0.438, 0.447]	18854±1323, [18360, 19348]	385.0±26.6, [375.0, 394.9]	47.78±1.19, [47.33, 48.22]	0.970±0.024, [0.961, 0.979]

Table 4. Shapiro-Wilk normality test results by method and metric (W, p-value).

Method	Sil_AUC	Sil_Avg	CH_AUC	CH_Avg	DB_AUC	DB_Avg
Random	W=0.973	p=0.612	W=0.967	p=0.458	W=0.961	p=0.329
Entropy	W=0.925	p=0.037	W=0.924	p=0.035	W=0.988	p=0.979
Margin	W=0.948	p=0.153	W=0.950	p=0.168	W=0.956	p=0.237
Core-Set	W=0.957	p=0.256	W=0.957	p=0.252	W=0.987	p=0.964
Uncertainty	W=0.981	p=0.860	W=0.981	p=0.860	W=0.972	p=0.581
Physics Only	W=0.972	p=0.582	W=0.974	p=0.666	W=0.983	p=0.901
proposed method	W=0.960	p=0.318	W=0.961	p=0.323	W=0.974	p=0.653

The AUC metrics showed the same cumulative (AUC) metrics demonstrated robust performance hierarchy, revealing that this performance differences throughout the active consistency across both average (Avg) and sampling process.

Table 5. Levene's test for homogeneity of variance across methods (F, p-value).

Metric	F	p
Sil_AUC	8.03	< 0.001
Sil_Avg	8.01	< 0.001
CH_AUC	10.18	< 0.001
CH_Avg	10.08	< 0.001
DB_AUC	9.71	< 0.001
DB_Avg	9.61	< 0.001

Prior to inferential analyses, we examined distributional characteristics and variance behavior of the performance metrics. As shown in Table 4, Shapiro–Wilk tests indicated approximate normality for most method–metric combinations (36/42, $p > 0.05$), with a small subset exhibiting deviations from normality. Levene's tests shown in Table 5 indicated unequal variances across sampling strategies for all six metrics (all $p < 0.001$), an expected consequence of fundamentally different sampling behaviors across methods. Given these distributional characteristics and the repeated-measures design—where each experimental run was evaluated for all seven methods—we employed the Friedman test, a non-parametric alternative to repeated-measures ANOVA, for subsequent main-effect analyses, as it does not rely on homogeneity of variance assumptions.

Table 6. Friedman test results for differences across methods by metric (χ^2, df, p-value, Kendall's W).

Metric	χ^2	df	p	Kendall's W
Sil_AUC	134.53	6	< 0.001	0.747
Sil_Avg	133.79	6	< 0.001	0.743
CH_AUC	142.07	6	< 0.001	0.789
CH_Avg	141.32	6	< 0.001	0.785
DB_AUC	138.61	6	< 0.001	0.770
DB_Avg	139.15	6	< 0.001	0.773

Friedman tests revealed highly significant differences among methods across all six metrics as shown in Table 6. Effect sizes ranged from $W = 0.743$ (Silhouette Average) to $W = 0.789$ (Calinski-Harabasz AUC), indicating large practical significance. The high concordance values ($W > 0.74$) demonstrated consistent method rankings across the 30 experiments, warranting detailed post-hoc pairwise comparisons.

Table 7. Dunn's Post-hoc pairwise comparison.

Comparison	Sil_AUC	Sil_Avg	CH_AUC	CH_Avg	DB_AUC	DB_Avg
Proposed vs Random	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
Proposed vs Entropy	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
Proposed vs Margin	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
Proposed vs Core-Set	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
Proposed vs Physics Only	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05

Post-hoc analysis using Dunn's test with Benjamini-Hochberg FDR correction (Table 7)

confirmed that the proposed method significantly outperformed all five baseline methods (Random, Entropy, Margin, Core-Set, Physics Only) across all six clustering metrics ($p < 0.05$). This demonstrates that the adaptive physics-informed approach achieves superior performance compared to conventional uncertainty-based and fixed physics-based active sampling strategies.

Table 8. Performance improvements and effect size analysis across clustering metric.

Metric Category	Metric	Improvement	Rank-biserial r	CLES
Silhouette	Sil_AUC	25.18 ± 7.54%	1	1
Silhouette	Sil_Avg	26.09 ± 7.59%	1	1
Calinski-Harabasz	CH_AUC	35.88 ± 12.35%	1	1
Calinski-Harabasz	CH_Avg	36.06 ± 12.29%	1	1
Davies-Bouldin	DB_AUC	-19.60 ± 6.41%	1	1
Davies-Bouldin	DB_Avg	-19.93 ± 6.40%	1	1

Effect size analysis confirmed substantial practical differences beyond statistical significance. Across all six clustering metrics, rank-biserial correlations of r = 1.000—indicating perfect rank separation between methods—and Common Language Effect Size values of 1.00—the probability that a random proposed method score exceeds any competing method score—demonstrated complete separation, with proposed method outperforming all six competing methods in every experiment with 30/30 wins per comparison. Performance improvements varied by metric: Silhouette metrics showed gains of 25.18 ± 7.54% for AUC and 26.09 ± 7.59% for Avg, Calinski-Harabasz metrics demonstrated improvements of 35.88 ± 12.35% for AUC and 36.06 ± 12.29% for Avg, and Davies-Bouldin metrics exhibited reductions of 19.60 ± 6.41% for AUC and 19.93 ± 6.40% for Avg. The consistent superiority across diverse clustering quality measures—compactness, separation ratio, dispersion—underscored robust practical value for physics-constrained active learning applications.

4.2 Weight allocation analysis

To isolate the contribution of adaptive weight allocation, a study was performed comparing VarAlpha against nine fixed weight ratios ($\alpha \in \{0.1, 0.2, ..., 0.9\}$), where α represents the proportion of weight assigned to physics-based sampling and (α) to uncertainty-based sampling ($1-\alpha$). Each configuration was evaluated using the same experimental protocol as the main comparison: 30 independent trials of 50 active sampling cycles, with 5 seeds and 5 bootstrap replicates per cycle. Figure 2 shows that VarAlpha achieved the highest average performance across all three clustering metrics. For Silhouette score, VarAlpha attained 0.443±0.008, exceeding the best fixed ratio (α=0.9: 0.4005±0.0049) by 10.6%. When examined for the Calinski-Harabasz index, VarAlpha achieved 375.3±14.8, surpassing the best fixed configuration (α=0.5: 345.9±3.3) by 8.5%. For cluster compactness, VarAlpha recorded a Davies-Bouldin index of 0.958±0.026, improving upon the best fixed ratio (α=0.9: 1.052±0.031) by 8.9%. Among fixed allocations, performance exhibited

distinct trends across metrics. For Silhouette scores, physics-dominant configurations (α=0.6-0.9) progressively improved from 0.390±0.004 to 0.401±0.005, while uncertainty-dominant allocations (α=0.1-0.5) ranged from 0.350±0.008 to 0.390±0.008. Calinski-Harabasz scores exhibited a non-monotonic pattern, increasing from 287.5±10.7 at α=0.1, peaking at α=0.5 (345.9±3.3), then declining to 307.7±5.3 at α=0.9. Davies-Bouldin indices improved consistently with decreasing physics emphasis, declining from 1.274±0.026 at α=0.9 to 1.052±0.031 at α=0.9.

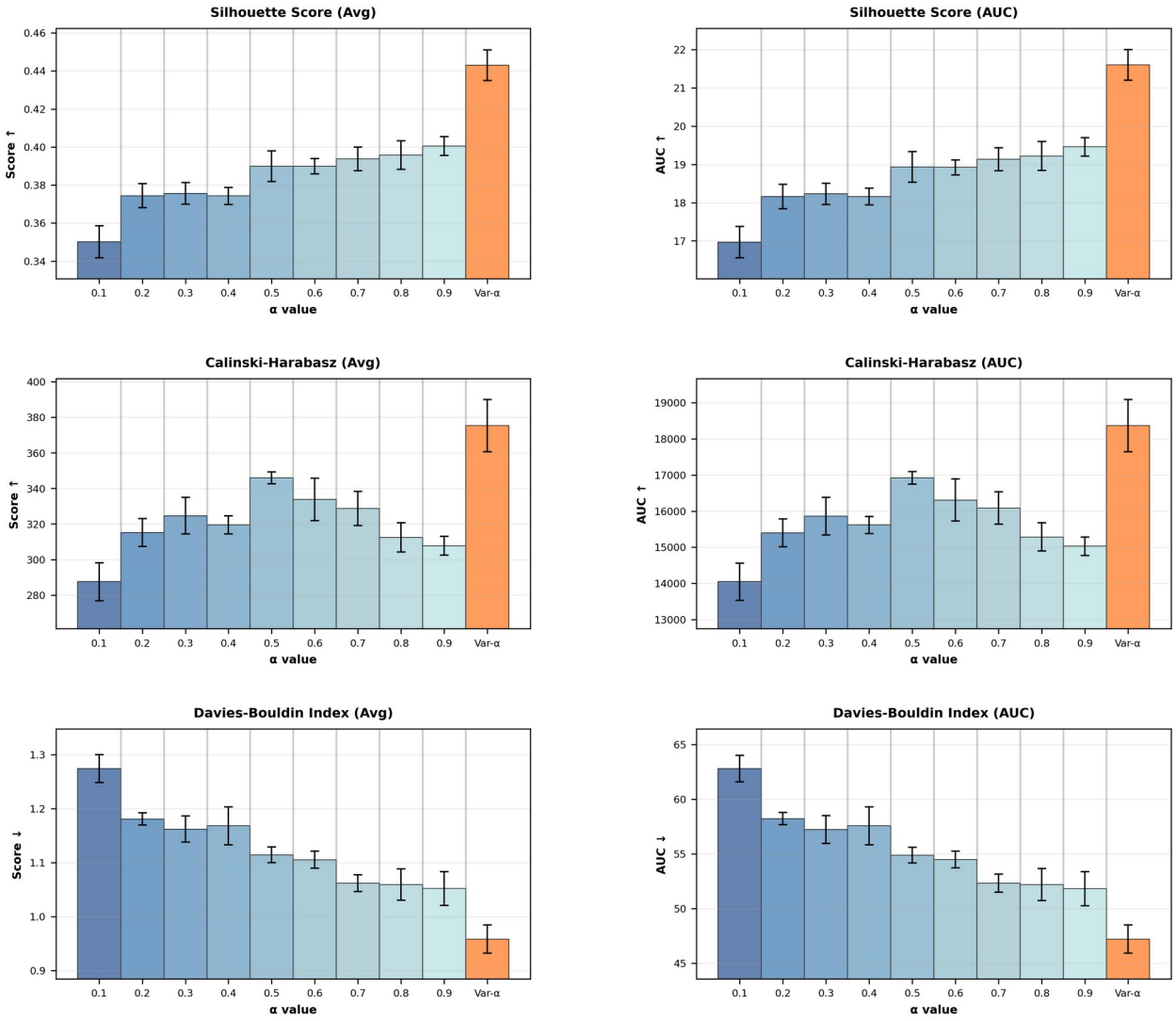

Figure 2. Comparison of weight allocation parameter (α) on clustering quality metric.

Area under curve analysis confirmed VarAlpha's consistent advantages across the complete sampling trajectory. VarAlpha's Silhouette AUC of 21.6±0.4 exceeded the best fixed ratio (α=0.9: 19.46±0.24) by 11.0%. For Calinski-Harabasz AUC, VarAlpha achieved 18,365.7±724.9, representing an 8.6% improvement over the best fixed configuration

(α=0.5: 16,918.5±174.9). VarAlpha's Davies-Bouldin AUC of 47.2±1.3 improved upon the best fixed ratio (α=0.9: 51.82±1.56) by 8.9%. VarAlpha maintained first rank across all six evaluation criteria, demonstrating that adaptive weighting provided robust advantages across multiple clustering quality dimensions and throughout the entire sampling process.

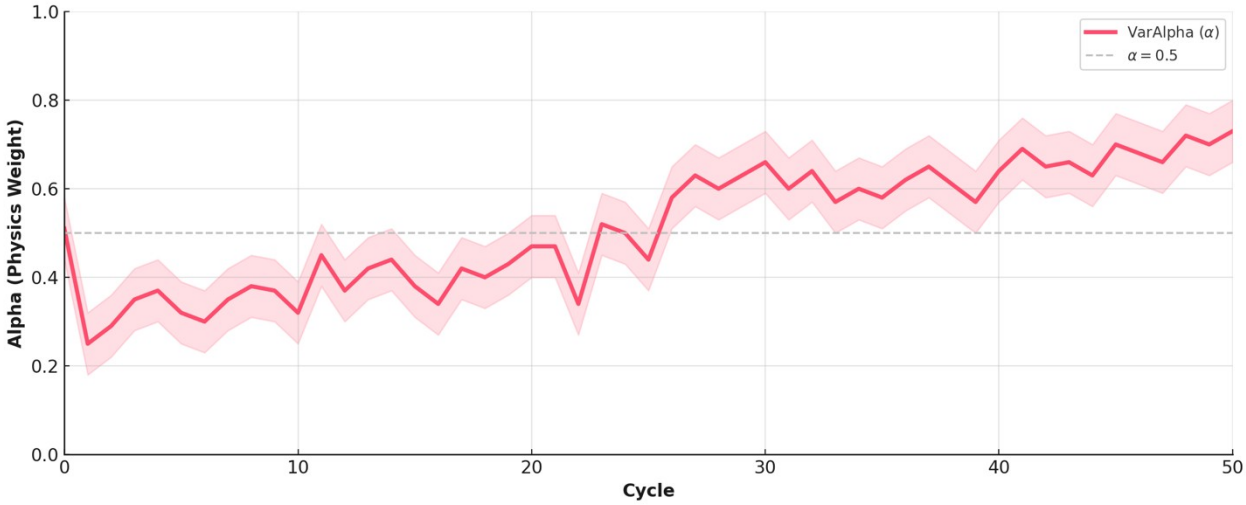

Figure 3. Evolution of physics weight (α) across active sampling cycles.

Figure 3 illustrates the evolution of the physics weight (α) throughout the active learning process. Initially, α begins near 0.5 but decreases to approximately 0.25-0.30 during early cycles, indicating the method prioritizes uncertainty-based sample selection when the dataset is small. As the sampling process progresses, α exhibits a monotonic upward trend, crossing 0.5 ~ cycle 25-30 and stabilizing near 0.70-0.75 by cycle 50. This adaptive weighting behavior demonstrated that the method dynamically adjusted selection emphasis from uncertainty-driven exploration in data-scarce regimes to physics-informed selection as the set increased, automatically balancing the two strategies based on dataset maturity.

4.3 Sensitivity to number of clusters (k)

To evaluate whether the proposed method maintained its performance advantage under different cluster resolutions, a k-sensitivity analysis was conducted by varying the number of clusters across $k \in \{3, 4, 6, 8, 10\}$ under identical experimental conditions for all methods. This experiment tested robustness to clustering granularity rather than optimizing absolute metric values with respect to k.

At the coarsest resolution—an uncommon and rarely employed clustering granularity in practical applications—the proposed method demonstrated strong performance across most metrics. For Silhouette score, the proposed method achieved the highest score at 0.395±0.007, substantially outperforming Random (0.325±0.007), Entropy (0.320±0.012), Core-Set (0.329±0.003), Uncertainty (0.338±0.018), Physics-only (0.363±0.003), and Margin (0.358±0.022). For

the Davies-Bouldin index, the proposed method achieved the lowest score at 1.149±0.022, outperforming Physics-only (1.212±0.004), Margin (1.251±0.067), Core-Set (1.286±0.010), Random (1.328±0.023), Uncertainty (1.342±0.054), and Entropy (1.352±0.047).

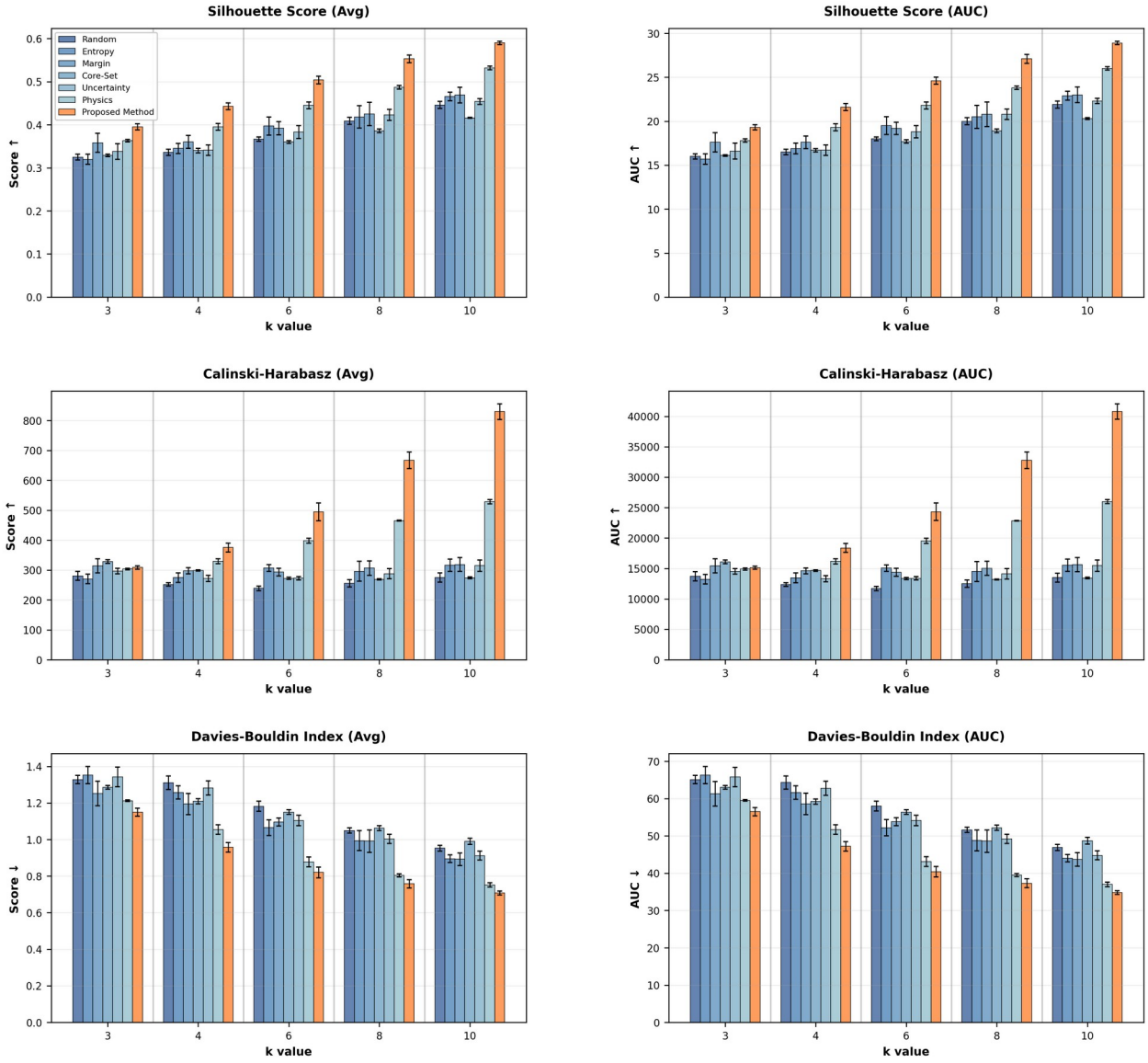

Figure 4. Comparison of cluster number (k) on clustering quality metrics.

However, under the Calinski-Harabasz index, Core-Set attained the highest score (328.314±6.224) compared to the proposed method (308.440±5.463), attributable to its geometric diversity objective that maximized inter-point distances—a property aligned with this metric at coarse resolutions. Nonetheless, the proposed method outperformed Random (280.379±14.888), Entropy (269.926±15.921), Uncertainty (296.698±9.292), and remained

competitive with Physics-only (303.482±2.773). Thus, even at k=3, the proposed method achieved the best performance in two out of three primary clustering metrics.

As cluster granularity increased to medium resolutions, the proposed method established clear superiority across all metrics. At k=6, the proposed method's Silhouette score reached 0.504±0.009, outperforming all baselines (0.360-0.445 range). Calinski-Harabasz achieved 494.727±29.966, exceeding baseline methods (238.861-398.322 range). Davies-Bouldin attained 0.820±0.029, significantly lower than all baselines (0.877-1.182 range).

At higher cluster resolution (k=10), the proposed method achieved a Silhouette score of 0.590±0.004, significantly exceeding all baselines (0.416-0.532 range). Calinski-Harabasz reached 829.696±26.129, compared to baseline methods (274.183-528.347 range). Davies-Bouldin attained 0.708±0.011, substantially lower than all baselines (0.752-0.990 range). These results demonstrate that the proposed method maintained superior performance across different clustering resolutions, establishing its robustness regardless of the number of clusters selected.

4.4 Noise sensitivity analysis

To evaluate robustness under measurement uncertainty, we conducted experiments with synthetic Gaussian noise injected into the feature space at levels of 2%, 4%, 6%, 8%, and 10% of each feature's standard deviation. While modern ADS-B data pipelines typically employ preprocessing and filtering procedures that reduce effective noise levels to approximately 0–2%, higher noise levels were intentionally introduced here to assess algorithmic robustness beyond nominal operating conditions. In particular, the 10% noise setting represented a stress-test scenario used solely to probe stability limits under severe measurement degradation, rather than to reflect typical operational regimes.

At 2% noise—representative of realistic ADS-B measurement uncertainty—the proposed method maintained strong performance, achieving Silhouette scores of 0.303±0.006, a Calinski-Harabasz index of 235.1±4.4, and a Davies-Bouldin index of 1.349±0.020. Baseline methods exhibited greater variability at this noise level, with Silhouette scores ranging from 0.228 to 0.279. As noise increased to 10%, all methods experienced performance degradation, as expected. Nevertheless, the proposed method consistently preserved its leading position, obtaining a Silhouette score of 0.281±0.006, a Calinski-Harabasz index of 228.2±3.1, and a Davies-Bouldin index of 1.415±0.013. Importantly, the relative performance gap between the proposed method and baseline approaches remained significant even under this stress-test condition.

The AUC-based metrics across all noise levels, shown in figure 5 further corroborate these trends. The proposed method achieved Silhouette AUC values ranging from 14.8 at 2% noise to 13.8 at 10% noise, outperforming all baselines consistently. In contrast, several baseline methods displayed non-monotonic responses as noise increased, suggesting reduced stability under measurement perturbations. The proposed method instead exhibited smooth, monotonic degradation behavior indicative of robust algorithmic design.

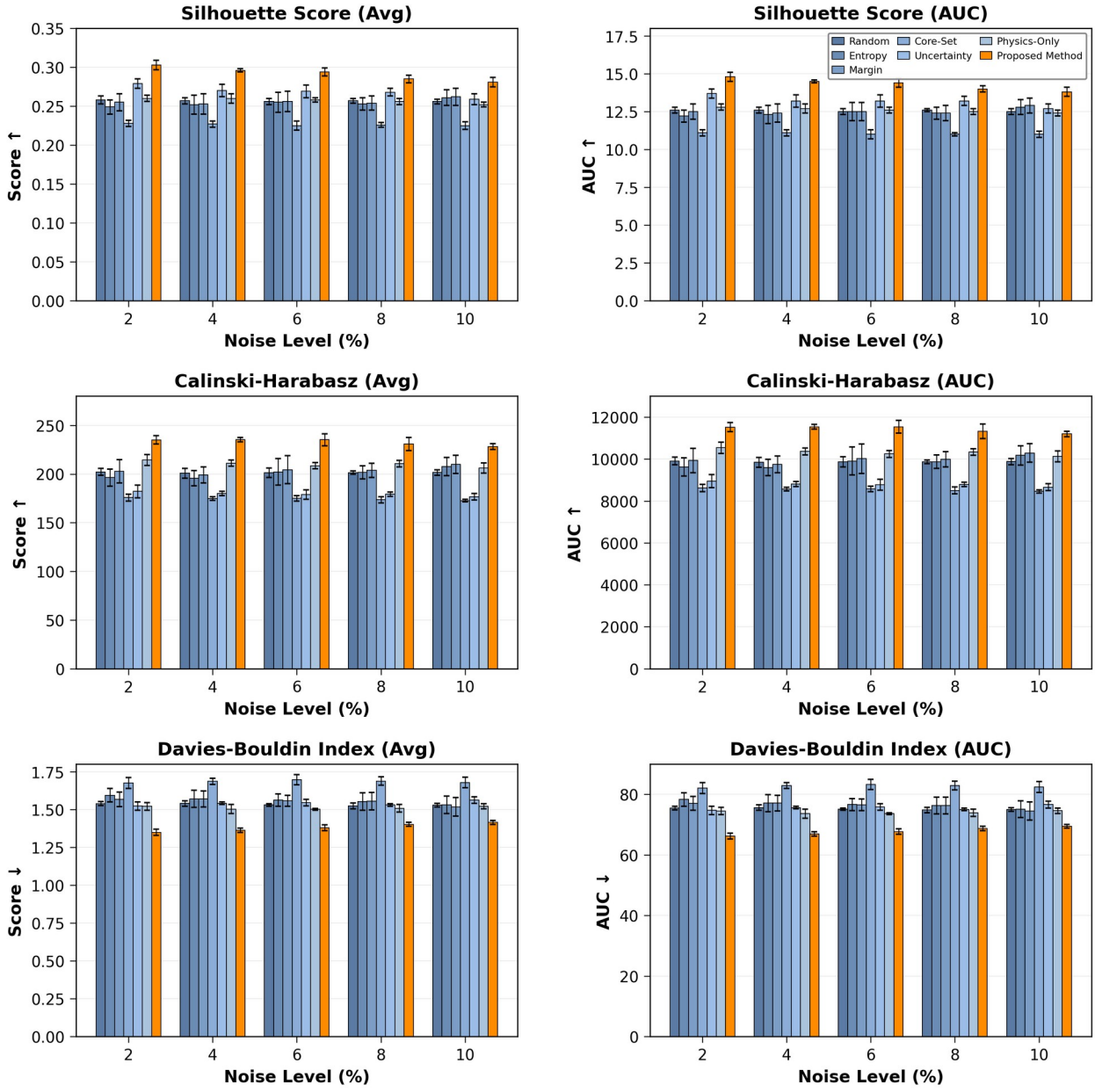

Figure 5. Impact of noise level on clustering performance.

Overall, these results demonstrate that the physics-informed diversity component contributes effective noise resilience across both realistic and stress-test regimes. Although operational ADS-B systems would rarely encounter noise levels exceeding 5% after standard preprocessing, the sustained performance advantage observed at higher noise levels provides strong evidence of the robustness limits of the proposed sampling strategy.

5. Discussion

5.1 Summary of main findings

The proposed method—combining adaptive physics-informed and uncertainty-based sampling—consistently outperformed all six baseline methods across 30 independent experiments. Performance rankings were stable and significant: the proposed approach achieved the highest scores on five of six clustering metrics, with physics-only sampling ranking second and conventional active learning methods (Entropy, Margin, Core-Set, Uncertainty-only) and Random sampling trailing behind. Statistical analysis confirmed complete separation between the proposed method and all competitors (rank-biserial $r = 1.000$), with performance improvements ranging from 19.6% to 36.1% depending on the metric. Furthermore, other baseline methods such as Uncertainty-only, Core-set, and Entropy-based sampling methods performed comparably to random sampling under clustering-based evaluation, while exhibiting higher variability across runs. This behavior is consistent with prior observations that decision-boundary–focused uncertainty signals provide limited and less stable guidance for global cluster structure.

The weight allocation analysis justified our use of adaptive weighting (VarAlpha) over fixed physics-uncertainty ratios. VarAlpha outperformed all nine tested fixed allocations by 8.5–11.0%, demonstrating that no single predetermined weight ratio was optimal throughout the sampling process. Weight evolution analysis revealed α decreasing from 0.5 to ~ 0.30 during early cycles—prioritizing uncertainty-based exploration when data was scarce—before rising monotonically to 0.70–0.75 by cycle 50 as physics-informed selection became more reliable with dataset maturity. This dynamic adjustment validated the adaptive framework as essential for sustained performance advantages.

Robustness analyses confirmed that these advantages held across different clustering resolutions ($k = 3$ to 10) and noise levels (2–10% synthetic perturbation), demonstrating that the dual-layer framework produced datasets with measurably better structural organization—clusters that were more cohesive, distinct, and less redundant—regardless of experimental conditions. Critically, each performance estimate derived from 37,500 underlying measurements per method (5 random seeds × 5 bootstrap samples × 50 cycles × 30 experiments), ensuring that observed differences reflected genuine algorithmic advantages rather than stochastic variation or sampling artifacts. This rigorous replication strategy also provided robustness to stochasticity and OpenSky data irregularities such as measurement noise inherent in ADS-B broadcast data.

5.2 Interpreting clustering quality in the context of aviation

The improved clustering quality achieved through the adaptive dual-layer sampling framework has important implications for turbulence research beyond data organization itself. High-quality clustering enhances the representational balance of the dataset, allowing the dataset to more faithfully represent the diversity of real flight dynamics. Clear separation between routine and unstable operating regimes supports more interpretable analysis of transitional flight states without presuming downstream model behavior.

Better-structured datasets reduce structural redundancy and imbalance at the data level, this reduction is widely regarded as desirable prior to task-specific model development. Confirmed by previous research, redundant samples from overrepresented stable regions no longer dominate the training process, reducing overrepresentation of routine operating regimes and improving the structural coverage of less frequent but physically significant flight states. In this way, clustering acts as a form of structural preprocessing—refining the dataset before any model training occurs— to produce datasets that are more compact, interpretable, and structurally balanced prior to any task-specific modeling.

The rigorous validation presented here, with 37,500 measurements per method, complete statistical separation across all experimental conditions, and demonstrated robustness to clustering resolution and measurement noise, established that adaptive physics-informed sampling produced measurably superior dataset organization compared to conventional approaches. These findings suggest that focusing on how data is organized, rather than solely on model complexity, can significantly improve the internal organization of datasets used within turbulence analysis workflows. By dynamically balancing physics-informed and uncertainty-based sampling principles, the dataset itself becomes more informative, interpretable, and aligned with the physical realities of flight.

This work demonstrated that data preparation in aerospace research can be treated as an active design process rather than a passive preprocessing step. While the current study focused on clustering quality as the primary outcome, the structured datasets produced by this framework provide a stronger structural foundation for future modeling and analysis efforts. Future work may validate these organizational improvements through direct turbulence prediction performance, incorporating labeled turbulence encounters and atmospheric variables to assess whether enhanced clustering quality translates to improved forecast accuracy in operational settings.

5.3 Implications in the context of prior research

Most prior research on turbulence detection has relied either on meteorological forecasting, which uses large-scale weather models, or on data-driven machine learning methods that aggregate large volumes of flight data without explicit consideration of how those data are structurally organized. In both cases, the underlying datasets were not clustered in a structured way, meaning that rare and unstable flight states were often embedded within dense groups of routine conditions. This reduced the visibility of instability-prone regions within the data distribution, making structural distinctions between routine and unstable regimes less explicit at the dataset level.

Some recent studies have explored active learning, but these approaches remained limited in scope and rarely examined how the structure of the dataset itself influenced the organization and representativeness of flight-state distributions. In particular, those studies did not test whether organizing flight states into clearer, more coherent groups could strengthen the structural foundation for subsequent analysis.

The present study addresses this gap. By

combining uncertainty-based sampling with physics-informed criteria, it demonstrated that dataset organization can be deliberately engineered, yielding stronger clustering quality than conventional sampling methods. This matters because it showed that physically relevant flight states can be emphasized more effectively when data selection is guided not only by statistical uncertainty but also by aerodynamic reasoning.

The contribution of this work lies in introducing adaptive, physics-aware active sampling as an explicit design objective during the data-preparation stage of aerospace research—a perspective that has remained largely unexplored in existing literature. Accordingly, this study should be understood as a proof-of-concept that highlights the potential of adaptive, physics-aware data design to reshape how turbulence research is approached, rather than as an exhaustive or fully validated aerodynamic framework.

5.4 Practical significance and downstream implications

It is important to clarify that the present study was not designed to train or evaluate turbulence-prediction models. Its scope was intentionally focused on restructuring raw ADS-B flight states into a more physically coherent and internally organized representation. In this formulation, the object of optimization was the geometry and structure of the data itself, rather than the accuracy of any particular classifier. Within this scope, the contribution of the study was complete, and formal downstream performance evaluation lies outside its intended objectives.

A rigorous assessment of downstream accuracy or error rates would require data resources that are not currently available in the public turbulence-data ecosystem. Such evaluation depends on time-synchronized turbulence labels and access to operational hazard-prediction pipelines. Contemporary systems—including EDR-based airline tools, NCAR and vendor forecast engines, and proprietary encounter-analysis tools—depend on internally curated datasets of labeled events, none of which are publicly accessible at the temporal or spatial granularity required for evaluation. Without event-level EDR time series, curated encounter logs, or fine-grained hazard annotations, external research groups generally cannot compute confusion matrices, false-positive rates, or false-negative rates for turbulence classifiers. This constraint reflects characteristics of the data environment rather than limitations of the methodology proposed here.

Within these boundaries, the practical significance of the present work lies in its demonstrated ability to deliberately improve dataset organization through adaptive, physics-informed sampling. The results established that the structure of turbulence-relevant state spaces can be shaped in a physically meaningful and statistically consistent manner, yielding measurably stronger clustering quality than conventional selection strategies. This improvement in internal organization constituted the primary outcome of the study.

Prior research across multiple domains provided additional background on how internal clustering structure was discussed when labels were available, though such studies were not required to interpret the contribution presented here. In process monitoring and fault-diagnosis systems, Webb

et al. treated the Silhouette coefficient, Davies-Bouldin Index (DBI), and Calinski-Harabasz (CH) score as unsupervised objectives and reported that changes in these indices tended to co-occur with changes in classification behavior when labels were present (25). In biomedical modeling, latent-state studies frequently tuned embeddings by minimizing DBI or related compactness metrics, observing that improved internal separation corresponded to clearer stratification of physiological states.

Methodological investigations have further documented how internal and external indices may align in labeled settings. Fang et al. compared Silhouette, CH, and DBI against external measures such as ARI, FMI, and NMI across diverse datasets, finding moderate-to-strong correlations under many conditions (26). Pakgohar et al. reported similar trends in simulated binary datasets with known ground truth (27). Broader comparative analyses by Arbelaitz et al. and recent surveys by Hassan et al. likewise highlighted Silhouette and CH as widely used descriptors of clustering structure across algorithms and application domains (28-29). Applied diagnostic studies, such as that of Islam et al. (2025), further illustrated that strong internal indices could accompany effective downstream behavior in fully labeled systems (30).

These prior studies are cited here solely to situate the present findings within a broader empirical landscape, rather than to complete or validate the contribution of this work. The present study did not rely on these relationships, nor did it claim that improvements in internal clustering structure necessarily implied gains in downstream turbulence-prediction accuracy. Instead, its contribution lies in demonstrating that turbulence data representations themselves can be systematically and deliberately improved through adaptive integration of physics-informed and uncertainty-based sampling.

Accordingly, the findings should be understood as a standalone contribution to turbulence data organization. By treating data structure as an explicit design objective—guided by aerodynamic reasoning alongside statistical uncertainty—the study showed that the geometry of flight-state data need not arise incidentally. Any future evaluation of predictive turbulence performance would constitute a separate line of investigation, building upon—but not completing—the organizational contribution established here.

5.5 Limitations

Despite the robust performance observed across multiple experiments, several limitations should be acknowledged. First, the flight-state variables reflected aircraft operational dynamics but did not capture atmospheric drivers of turbulence such as temperature gradients, wind shear, or jet-stream behavior. This omission originated from the technical scope of ADS-B broadcast data, which reports aircraft state but not ambient atmospheric conditions. Future integration of meteorological data sources could provide a more complete environmental representation.

It may initially appear that the absence of validated turbulence labels constitutes a limitation; in reality, this scarcity was the fundamental motivation for this research. Turbulence labels are extremely sparse—severe encounters are rare, and obtaining labels requires manual annotation from pilot reports or proprietary EDR systems that remain

inaccessible at the scale needed for dataset construction. If comprehensive labeled turbulence data were widely available, the motivation for physics-informed active sampling at the data-preparation stage would be significantly reduced. Instead, this approach addressed the label-scarcity problem directly by leveraging aerodynamic principles to identify physically meaningful flight states regardless of formal turbulence classification. The objective was not to validate against ground-truth labels but to enhance dataset structure and diversity using physically relevant criteria. By this metric—improved Silhouette scores, Calinski-Harabasz indices, and Davies-Bouldin metrics—the approach succeeded, demonstrating that aerodynamically principled selection meaningfully improves data organization even in label-scarce environments.

A third limitation was that some of the physics-based features exhibited partial multicollinearity. For example, dynamic pressure and Mach gradient both depend on airspeed, while energy imbalance and vertical jerk capture overlapping aspects of dynamic instability. Although this correlation did not invalidate the clustering-based evaluation used here, it suggested that part of the observed improvement may reflect overlapping representations rather than fully independent sources of information. Future work should address this by exploring dimensionality reduction or alternative feature formulations to ensure that clustering gains reflect true informational diversity.

Although the OpenSky dataset is known to exhibit uneven global coverage, the dataset used in this study was not dominated by any single region, with approximately 40% of samples originating from North America, 30% from Europe, and 30% from Asia and other regions. Moreover, the extensive experimental design—37,500 independent evaluations per method—significantly reduced statistical bias and ensured that the reported clustering improvements were not artifacts of geographic imbalance or sampling variability. Any remaining unevenness in regional representation should therefore be understood as a potential characteristic of ADS-B–based data sources in general, rather than as a limitation that affected the validity of the present findings. Future work incorporating additional regional datasets may further broaden the empirical scope, but the conclusions of this study were robust to the distributional properties of the current data.

Aerodynamic and dynamic features used in the physics-based turbulence score were assigned weights that were heuristic, and not derived from first-principles.

Taken together, these limitations reflect natural boundaries of this study. This work did not aim for exhaustive physical or mathematical validation; rather, its value lies in incorporating the integration of physics-informed reasoning and machine-driven data selection into the earliest stage of aerospace data preparation. The framework should therefore be interpreted as a conceptual proof-of-concept that opens a new line of research—one that future studies can expand, refine, and rigorously validate once broader data resources become available.

6. Conclusion

This study demonstrated that integrating adaptive physics-informed criteria with

uncertainty-based active sampling produced measurably superior flight-state dataset organization compared to conventional approaches. Across 30 independent experiments with comprehensive statistical validation, the proposed method achieved consistent improvements of 19.6% to 36.1% across clustering quality metrics, with complete statistical separation from all baseline methods. These results show that adaptive weighting between physics-informed and uncertainty-based selection—dynamically adjusting from early-cycle exploration to late-cycle exploitation—outperformed both fixed allocation strategies and single-layer sampling approaches.

The consistent performance advantages observed across multiple robustness analyses, including variations in cluster resolution, noise levels, and experimental settings, demonstrated that deliberate integration of aerodynamic reasoning into data selection strengthens dataset structure in ways that conventional active sampling methods did not achieve. Flight-state clusters became more cohesive, better separated, and less redundant, providing a more organized foundation for subsequent aerospace analysis. While this study focused on clustering quality rather than downstream turbulence prediction, cross-domain findings indicated that internal clustering structure often served as an informative descriptor of representational quality when labels were available—offering contextual relevance without implying guaranteed downstream gains.

This work addressed a gap in aerospace research by treating data preparation as an active design process rather than a passive preprocessing step. Whereas prior turbulence studies largely emphasized predictive model architectures, the present research demonstrated that systematic attention to dataset organization—particularly in label-scarce environments—offers a complementary pathway for advancing aviation safety research. Future work incorporating atmospheric variables, expanded geographic coverage, and access to operational turbulence labels can enable more comprehensive evaluation and facilitate exploration of whether improved data organization enhances predictive performance in operational settings.

Data repository

https://github.com/skang-rgb/Redefining-the-Data-Foundations-of-Turbulence-Research-A-Physics-Informed-Active-Sampling-Approach

7. References

1. Sharman, R., Williams, P., Storer, L. (2023). Evidence for large increases in clear-air turbulence over the past four decades. *Geophysical Research Letters, 50*(8), e2023GL103814. https://doi.org/10.1029/2023GL103814

2. Chen, C., Zhang, X., Xu, Q. (2023). A machine learning-based model for flight turbulence prediction using multisource data. *Atmosphere, 14*(5), 797. https://doi.org/10.3390/atmos14050797

3. Yang, H., Gao, J., Yuan, Y., Li, X. (2023). Imbalanced aircraft data anomaly detection. *arXiv preprint arXiv:2305.10082*. https://arxiv.org/abs/2305.10082

4. Hao, Z., Zhang, D., Sun, L. (2022). Advances in machine learning for scientific applications: A comprehensive survey. *arXiv preprint arXiv:2211.08064*. https://arxiv.org/abs/2211.08064

5. Wang, Y., Huang, J., Li, H. (2024). Recent advances on machine learning for computational modeling: A survey. *arXiv preprint arXiv:2408.12171*. https://arxiv.org/abs/2408.12171

6. Sharman, R., Keller, T., Pearson, J. (2020). Aviation turbulence forecasting at upper levels with machine learning. *Journal of Applied Meteorology and Climatology, 59*(11), 1895–1909. https://doi.org/10.1175/JAMC-D-20-0116.1

7. Zhou, X., Li, Y., Zhang, Y. (2020). Machine learning-based multi-index prediction of aviation turbulence. *Environmental Research Communications, 2*(11), 115001. https://doi.org/10.1088/2515-7620/abc24b

8. Lee, J. C. W., Leung, C. Y. Y., Kok, M. H., Chan, P. W. (2022). A comparison study of EDR estimates from the NLR and NCAR algorithms. Atmosphere, 13(1), 132. https://doi.org/10.3390/atmos13010132

9. Williams, P. D., Joshi, M. M. (2013). Intensification of winter transatlantic aviation turbulence in response to climate change. *Nature Climate Change, 3*(6), 644–648. https://doi.org/10.1038/nclimate1866

10. Sharman, R., Tebaldi, C., Wiener, G., Wolff, J. (2006). An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting, 21*(3), 268–287. https://doi.org/10.1175/WAF924.1

11. Kim, J.-H., Chun, H.-Y. (2011). Statistics and possible sources of aviation turbulence over South Korea. *Journal of Applied Meteorology and Climatology, 50*(2), 311–324. https://doi.org/10.1175/2010JAMC2492.1

12. Kuzmenko, N., Ostroumov, I., Bezkorovainyi, Y., Averyanova, Y., Larin, V., Sushchenko, O. (2022). Airplane flight phase identification using maximum posterior probability method. *Proceedings of the 3rd International Conference on System Analysis & Intelligent Computing* (pp. 1–5). IEEE. https://doi.org/10.1109/SAIC57818.2022.9923095

13. Fala, N., Georgalis, G., Arzamani, N. (2023). Study on machine learning methods for general aviation flight phase identification. *Journal of Aerospace Information Systems, 20*(10), 636–647. https://doi.org/10.2514/1.I011165

14. Perrichon, R., Gendre, X., Klein, T. (2024). Hidden Markov models and flight phase identification. *Journal of Open Aviation Science, 2*(1). https://doi.org/10.59490/joas.2024.7269

15. Gracey, W. (2003). Measurement of aircraft speed and altitude. In *Encyclopedia of Physical Science and Technology* (3rd ed., pp. 1–32). Academic Press. https://doi.org/10.1016/B0-12-227410-5/00006-8

16. Basora, L., Olive, X., Durand, T. (2019). Recent advances in anomaly detection methods applied to aviation. *Aerospace, 6*(11), 117. https://doi.org/10.3390/aerospace6110117

17. Zhao, Z., Zeng, W., Quantum, Z., Nakagawa, T., Hirabayashi, R. (2021). Aircraft trajectory clustering in terminal airspace based on deep autoencoder and Gaussian mixture model. *Aerospace, 8*(9), 266. https://doi.org/10.3390/aerospace8090266

18. Olive, X. (2022). Traffic, a toolbox for processing and analysing air traffic data. *Journal of Open Source Software, 7*(71), 4518. https://doi.org/10.21105/joss.04518

19. Chati, Y. S., Balakrishnan, H. (2017). Analysis of aircraft fuel burn and emissions in the landing and take-off cycle using operational data. In *6th International Conference on Research in Air Transportation (ICRAT)*. https://dspace.mit.edu/handle/1721.1/112984

20. Sharman, R. D., Lane, T. P. (2016). *Aviation turbulence: Processes, detection, prediction*. Springer. https://doi.org/10.1007/978-3-319-23630-8

21. Li, L., Hansman, R. J., Palacios, R., Welsch, R. (2016). Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring. Transportation Research Part C: Emerging Technologies, 64, 45–57. https://doi.org/10.1016/j.trc.2016.01.007

22. McFadyen, A., Mejias, L., Corke, P., Monsalve, G. (2016). Aircraft trajectory clustering techniques using circular statistics. In *2016 IEEE Aerospace Conference* (pp. 1–10). IEEE. https://doi.org/10.1109/AERO.2016.7500601

23. Olive, X., Morio, J. (2019). Trajectory clustering of air traffic flows around airports. *Aerospace Science and Technology, 84*, 776–781. https://doi.org/10.1016/j.ast.2018.11.031

24. Fernández, E. C., Valenzuela, A., Casado, E. (2022). Automated identification of significant operational events from flight data. *Transportation Research Part C: Emerging Technologies, 144*, 103879. https://doi.org/10.1016/j.trc.2022.103879

25. Webb, E. J., Zhang, L., Srinivasan, R. (2022). Optimization of multi-mode classification for process monitoring using hybrid clustering and ensemble learning. *Frontiers in Chemical Engineering, 4*, 900083. https://doi.org/10.3389/fceng.2022.900083

26. Fang, J., Chan, C., Owzar, K., Wang, L., Qin, D., Li, Q.-J., Xie, J. (2022). Clustering deviation index (CDI): A robust and accurate internal measure for evaluating scRNA-seq data clustering. *Genome Biology, 23*(1), 269. https://doi.org/10.1186/s13059-022-02825-5

27. Pakgohar, N., Lengyel, A., Botta-Dukát, Z. (2024). Quantitative evaluation of internal cluster validation indices using binary datasets. *Journal of Vegetation Science, 35*(5), e13310. https://doi.org/10.1111/jvs.13310

28. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

29. Hassan, S. M., Kumar, P., Singh, A. (2024). From A-to-Z review of clustering validation indices. *arXiv preprint arXiv:2407.20246*. https://arxiv.org/abs/2407.20246

30. Islam, M. R., Begum, S., Ahmed, M. U. (2025). Interpretable self-supervised learning for fault identification in printed circuit board assembly testing. Applied Sciences, 15(18), 10080. https://doi.org/10.3390/app151810080