

Article

Toward Standardized Benchmarking of Search-Based Scenario Selection Methods in Autonomous System Validation

Daniel Autenrieth ^{1,*}, Daniel Paland ², Jan Mentjes ² and Adrian Zlocki ³¹ TKMS ATLAS ELEKTRONIK GmbH, Sebaldsbrücker Heerstraße 235, 28309 Bremen, Germany² German Aerospace Center (DLR), Institute of Systems Engineering for Future Mobility, Escherweg 2, 26121 Oldenburg, Germany³ fka GmbH, Steinbachstraße 7, 52074 Aachen, Germany

* Correspondence: daniel.autenrieth@tkmsgroup.com

Abstract

Autonomous systems must demonstrate reliable safety and robustness even under rare, safety-critical conditions. Because the vast range of possible operating situations cannot be exhaustively tested, scenario-based testing has emerged as a structured approach to expose systems to representative and challenging situations. Search-based scenario selection methods (SBSSMs) algorithmically explore the scenario space to identify critical and informative cases. Yet, existing implementations are often tightly coupled to specific domains and simulation environments, hindering reproducibility, comparability, and generalizability across application areas. This paper introduces a benchmarking framework that addresses these limitations by providing a domain-independent environment for the systematic evaluation of SBSSMs. Building on empirical observations from maritime simulation studies and the literature, the framework generates synthetic test instances that mimic the structural patterns found there while abstracting away application-specific semantics. It supplies standardized performance metrics and reference implementations of key method classes, enabling direct comparison of new approaches with established ones under controlled, repeatable conditions. Illustrative demonstrations show how the benchmark reveals characteristic strengths and weaknesses of different methods across diverse structural settings. As a reproducible and extensible baseline, the framework promotes standardized performance assessment and the transfer of methodological advances from maritime navigation to domains such as automotive safety.

Keywords: search-based scenario selection; benchmarking framework; autonomous systems validation



Academic Editor: Zhiguang Feng

Received: 5 June 2026

Revised: 20 July 2026

Accepted: 24 July 2026

Published: 29 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The increasing autonomy and operational complexity of systems incorporating artificial intelligence (AI), particularly in safety-critical applications such as maritime autonomous surface ships (MASS), have led to a substantial expansion of the parameter spaces relevant to system validation and the generation of meaningful test scenarios. This phenomenon, commonly referred to as the parameter space explosion [1], arises from the combinatorial growth in the number of influential factors, ranging from environmental conditions and system states to interaction dynamics between multiple agents. Consequently, exhaustive testing across all possible scenarios becomes infeasible, rendering traditional validation approaches insufficient for high-assurance systems. In the maritime domain,

this is exacerbated by the rarity of critical encounter situations relative to total voyage time and by the small size of available test fleets, which make statistical safety demonstration through real-world operational exposure impractical and shift the burden of evidence toward simulation-based testing. This challenge gains regulatory urgency with the International Maritime Organization's (IMO) goal-based MASS Code, effective July 2026, which emphasizes risk-based approval supported by simulation-based evidence [2].

To mitigate this challenge, the research community has embraced scenario-based testing (SBT) [3–5] as a principled alternative to indiscriminate test execution. Here, the objective is to systematically expose the system-under-test (SUT) to rare or safety-critical conditions. However, given the sheer size and complexity of the possible parameter combinations, manually identifying all such scenarios is neither scalable nor reliable. This has led to the emergence of search-based scenario selection methods (SBSSMs), which are algorithmic strategies that heuristically explore the scenario space using optimization, adaptive sampling, or surrogate modeling techniques to uncover interesting and representative cases with high failure potential.

SBSSMs provide a complementary means of identifying sparsely distributed, high-criticality regions within the parameter space, which may otherwise remain underrepresented in random or expert-curated test sets [6–8]. This risk of insufficient coverage is well documented in simulation-based safety validation: because critical events are rare, purely random sampling is unlikely to encounter the corresponding subregions of the parameter space without a prohibitive number of samples [9]. By systematically navigating the scenario space, SBSSMs can augment traditional validation approaches and improve overall coverage of critical areas. Moreover, their ability to target specific aspects such as performance boundaries [10–12] enables a more efficient allocation of validation resources, ultimately contributing to improved system robustness and operational safety.

Despite substantial progress in search-based scenario selection, methodological heterogeneity remains a key barrier to systematic comparison of existing approaches. Current methods vary widely in simulation environments [13–15], used models, criticality definitions, and evaluation metrics [16,17]. For instance, scenario selection strategies validated in the automotive domain, where highway traffic rules, sensor configurations, and interaction patterns dominate, cannot be directly transferred to the validation of a maritime autonomous surface vessel, which operates in fundamentally different hydrodynamic conditions, regulatory frameworks, and interaction topologies. A method optimized to explore corner cases in urban driving may fail to capture critical phenomena such as wave–current interactions, vessel maneuvering constraints, or collision avoidance under international maritime rules. In addition, many studies do not release source code or parameter settings and sometimes depend on proprietary or commercial tools with closed-source implementations [18,19], limiting reproducibility and generalizability. Comparisons are often restricted to random sampling baselines [20,21] or self-developed methods [11,12], hindering objective cross-method evaluations. As a result, robust performance comparisons are largely infeasible, obscuring whether new methods provide genuine advancements or are simply tailored to a specific setting. Consequently, assessing the applicability of existing approaches across domains remains challenging.

Addressing these obstacles demands more than isolated case studies or ad hoc comparisons. What is required is a common experimental foundation that enables different selection methods to be tested under identical and controllable conditions. Such a foundation must provide representative test cases, transparent evaluation protocols, and standardized metrics so that results become directly comparable and reproducible across research groups and application domains. In other words, a benchmarking framework provides the structural and methodological infrastructure that is currently missing. To

operationalize this need, the concrete benchmark requirements (BR1–BR7) listed next are derived from the limitations identified above, including lack of reproducibility, domain- and simulator-coupling, non-standardized protocols, and insufficient representativeness. These requirements define what a benchmark must provide in order to support meaningful and transferable evaluation of SBSSMs.

BR1: Reproducibility. It must ensure reproducibility by providing open implementations, transparent parameter settings, and clearly defined evaluation protocols.

BR2: Cross-domain comparability. It must support methodological comparability across domains by abstracting away from simulator internals, application semantics, and proprietary toolchains.

BR3: Structural realism. It must preserve the structural complexity and essential realism of real-world criticality spaces so that evaluations remain both challenging and representative of practical validation scenarios.

BR4: Controllability and determinism. It must guarantee controllability and determinism, ensuring that test cases and experimental procedures can be executed in a predictable and repeatable manner.

BR5: Structural diversity. It should expose structural diversity in the scenario space, allowing robust assessment across varied topologies, including multimodality and discontinuity.

BR6: Extensibility. It should support extensibility, enabling researchers to incorporate new selection methods, criticality-space findings, or evaluation metrics as the field evolves.

BR7: Task-specific metrics. Finally, it must accommodate task-specific evaluation metrics to reflect the distinct goals of different selection strategies.

These framework-level benchmark requirements are distinguished from the functional requirements of the synthetic test instances used within the benchmark. The latter are derived from empirical observations of realistic criticality landscapes and are formalized later in Section 3.3. This distinction separates how SBSSMs are evaluated (benchmark methodology and infrastructure) from what they are evaluated on (properties of the test landscapes).

In response to the above challenges and requirements, this paper introduces a reproducible, domain-agnostic benchmarking framework for the systematic evaluation of SBSSMs. The framework is motivated by and empirically grounded in the maritime domain: the structural properties of its synthetic landscapes are derived from simulation studies of COLREG-compliant collision avoidance and multi-vessel encounter scenarios (Section 3), so that evaluated methods face the same search challenges that arise in the validation of MASS. At the same time, the abstraction from domain-specific semantics preserves the transferability of the framework to other application areas. This enables controlled and repeatable performance evaluation across heterogeneous selection strategies. In addition, reference implementations of representative SBSSMs and a standardized metric suite are provided, establishing a transparent baseline that supports direct comparison of newly proposed methods under identical experimental conditions. The experiments reported in this work are intended as an illustrative demonstration of the benchmark rather than an exhaustive comparative study. Accordingly, this work is positioned as a step toward standardized benchmarking of SBSSMs.

The remainder of this paper is structured as follows. Section 2 provides theoretical background by introducing scenario-based testing, reviewing the construction of criticality spaces, and existing benchmarking initiatives, and further categorizing representative SBSSMs that motivate the benchmark design. Section 3 reports empirical observations from simulation environments, which inform both the structural design of synthetic criticality spaces and the benchmark requirements. Section 4 builds upon the previous sections by presenting the general methodology and the benchmark parts. It defines general as

well as method-specific evaluation metrics before concluding with a brief description of the benchmark structure. Section 5 demonstrates the applicability of the benchmark through selected illustrative examples. Section 6 discusses the methodological implications, limitations, and transferability of the proposed framework and outlines directions for future research. Finally, Section 7 concludes the paper.

2. Background and Related Work

This section consolidates the theoretical basis for the benchmark. It outlines key concepts of scenario-based testing, formalizes criticality spaces as scalar representations of safety risk, surveys representative search-based scenario selection methods, and reviews existing benchmarking efforts to highlight the methodological gaps that motivate the proposed framework.

2.1. Scenario-Based Testing

The validation of autonomous systems increasingly relies on SBT [3], which systematically evaluates system performance under parameterized representations of operational situations. In contrast to purely data-driven or exhaustive testing paradigms, SBT offers a structured approach to model diverse environmental conditions, agent interactions, and system configurations within the operational design domain (ODD) [3,22]. Recent surveys document the rapid maturation of SBT along the full validation chain, covering software-, hardware-, and vehicle-in-the-loop test methods, scenario databases, and evaluation frameworks [4,23] and confirm that the systematic construction and selection of test scenarios remains the central methodological bottleneck. Because each scenario constitutes a well-defined combination of initial conditions and contextual variables, it can be precisely reproduced, enables targeted stress tests of specific behaviors, and allows for the isolation of causal factors behind observed system responses.

Despite its conceptual advantages, SBT faces scalability challenges due to the combinatorial explosion of scenario permutations in high-dimensional environments [1]. Traditional scenario catalogs, whether expert-defined or derived from empirical data sources such as accident databases or driving logs [15,24], provide valuable coverage but are inherently limited in their ability to represent rare or unforeseen failure modes. Random sampling approaches, while scalable, risk undersampling critical regions, especially in high-dimensional or anisotropic landscapes where hazardous configurations are confined to narrow subspaces [25,26].

These challenges necessitate systematic scenario selection methods that can efficiently prioritize informative, high-risk, or underrepresented regions of the scenario space. In particular, SBSSMs have emerged as a promising approach to address these limitations. By leveraging optimization, adaptive sampling, and surrogate modeling techniques, SBSSMs aim to explore the scenario space strategically, identifying critical cases that are unlikely to be captured by naive sampling or conventional scenario libraries. The development of effective scenario selection techniques is therefore essential to ensure both the scalability and safety assurance of autonomous systems operating under complex and uncertain real-world conditions.

The need for systematic scenario selection is particularly pronounced for maritime autonomous surface ships, whose safety assessment increasingly follows scenario-based approaches [5]. Across the methodological spectrum of MASS path planning and collision avoidance, ranging from classical search and optimization over real-time reactive methods to learning-based and, most recently, large-language-model-based approaches such as Navigation-GPT, verifiable compliance with the COLREGs remains an unresolved challenge [27]. Since none of these method families can guarantee rule-compliant behavior by

construction, safe operation must instead be demonstrated through testing, which places the systematic selection of multi-vessel encounter scenarios at the core of MASS validation.

This demand is reinforced by recent regulatory developments: The International Maritime Organization adopted a non-mandatory, goal-based International Code of Safety for Maritime Autonomous Surface Ships by resolution MSC.595(111) [2], effective 1 July 2026, with a mandatory code expected to enter into force in 2032. Classification societies provide complementary guidance, including the DNV class guideline for autonomous and remotely operated ships [28] and the recommended practice on assurance of simulation models [29]. These instruments emphasize risk-based approval supported by simulation-based evidence, for which the reproducible identification of safety-critical scenarios, such as multi-vessel encounters under the COLREGs, is a central prerequisite. The benchmark proposed in this work addresses this need by enabling transparent and comparable evaluation of the selection methods employed in such validation processes, while remaining applicable beyond the maritime domain.

2.2. On the Creation of Criticality Spaces

This subsection first defines the notion of a criticality space and its formal basis, then reviews structurally distinct formulations reported in the literature, and finally describes the instantiation used in this work.

In this work, a criticality space denotes a scalar-valued function defined over a structured scenario parameterization, capturing the safety-relevant response of an autonomous system to specific configurations. Formally, it constitutes a response surface over a bounded subset of $\mathbb{R}^d$, where each point $\mathbf{x} \in \mathbb{R}^d$ represents a vector of initial or environmental parameters that define an instantiated concrete scenario [22]. The function value $C(\mathbf{x})$ encodes its criticality according to a predefined metric [16,30].

Schütt et al. [31] conceptualize criticality as a subdimension of general quality metrics, which implies that analogous spaces can be constructed based on broader quality-related criteria. Under this generalized perspective, such spaces may be referred to as score spaces. However, since our analysis of the simulated spaces exclusively focuses on safety-relevant metrics, the terms criticality metrics and criticality space are used to emphasize this specific analytical perspective.

Following the formalization introduced by Westhofen et al. [16] and the definitions of Schütt et al. [31], a scene-level criticality metric can be defined as a function

$$k : \mathcal{S} \times \mathbb{R}^+ \rightarrow \mathcal{O}, \quad (1)$$

which maps a scene $S \in \mathcal{S}$ at time $t \in \mathbb{R}^+$ to a value on a criticality scale $\mathcal{O} \subseteq \mathbb{R} \cup \{-\infty, +\infty\}$.

Scenario-level metrics extend this formulation by aggregating criticality [16], thus capturing retrospective or cumulative aspects of safety-relevant system behavior:

$$C(\mathbf{x}) = A(k(S, t)), \quad (2)$$

where $\mathbf{x} \in \mathbb{R}^d$ encodes parameters of a concrete scenario and A denotes a general aggregation operator, potentially acting over temporal, actor-related, or metric-specific dimensions.

The scenario parameters may include both system-intrinsic variables, such as initial velocity, maneuverability, or control latency, and exogenous factors, such as current, wind, visibility, obstacle position, or geometric constraints of the operational domain. Other relevant variables may include sensor placement configurations, communication delays, or topological features of the environment.

In practice, criticality metrics are often instantiated from measurable physical or perceptual safety quantities. For instance, Time-to-Collision (TTC) [32] can be used as a proxy and then transformed via a monotonic mapping onto a normalized scale (e.g., [0, 1]), where higher criticality corresponds to lower TTC values. While TTC is among the most widely used metrics, a broad range of alternative formulations exist, each reflecting different assumptions, input requirements, and levels of abstraction. A systematic review of these metrics, including their operational properties and suitability for different automated driving applications, is provided by Westhofen et al. [16].

Several recent works have proposed structurally distinct formulations of criticality or score spaces in the context of scenario-based validation. A brief comparison of these approaches provides context for the design decisions in the present study. Mullins et al. [12] define a parameterized state space comprising influential variables such as tidal current conditions, waypoint sequencing, and the spatial placement of structural elements. Over this space, they evaluate system behavior using scalar metrics including residual fuel and scenario performance, which jointly define what they refer to as the performance score space. In addition to simulation-based evaluations, they construct a suite of synthetic functions to assess the behavior of their search methods under controlled mathematical conditions. These include a composite benchmark composed of four independent 2D subfunctions, resulting in an eight-dimensional input space. Similarly, Lu et al. [11] evaluate binary performance outcomes (i.e., successful versus unsuccessful path planning) by varying the initial position of a static obstacle in both 2D and 3D environments. Their approach is further validated on nine custom-designed 2D synthetic spaces to examine the robustness and scalability of the proposed boundary search strategy.

Gladisch et al. [26] expand this spectrum of formulations by treating criticality as the real-valued robustness of Signal Temporal Logic (STL) specifications evaluated over closed-loop simulations of automated-driving controllers. Their study spans three control functions (adaptive-cruise, lane-keeping and low-level steering) whose input spaces range from two to six continuous parameters and whose robustness landscapes exhibit pronounced non-convex, multi-modal and potentially non-smooth structures even under comparatively simple STL predicates. In each case, the scalar robustness metric serves analogously to the criticality score described in the previous section. Gladisch et al. show that controller mode switches and physical feasibility constraints can imprint discontinuities and plateaus onto the landscape, thereby demonstrating that topological complexity arises not only from environmental variables but also from specification semantics and system-internal logic.

In their simulation-based testing framework for autonomous ships, Torben et al. [25] construct criticality spaces by evaluating the robustness of STL specifications over parameterized maritime scenarios. Each scenario is defined by a bounded subset of continuous input parameters, including system-internal variables (e.g., rudder angle, heading, velocity) and exogenous conditions (e.g., current direction, obstacle location). The criticality function is instantiated as the robustness score of the STL specification, mapped over the input space. The resulting landscapes reveal pronounced non-convexity, plateaus, and sharp transitions, reflecting the complex interplay between environmental parameters, system dynamics, and logical constraints. In particular, the study demonstrates that even relatively simple scenario configurations can induce highly structured robustness topologies, where regions of falsification (i.e., specification violation) are interspersed with large safe zones, enabling a nuanced analysis of system safety margins.

The PEGASUS (Project for the Establishment of Generally Accepted quality criteria, tools and methods as well as Scenarios and Situations for the release of highly automated driving functions) project documentation [33] presents demonstrative response

surfaces, heatmaps, and scatterplots for the Society of Automotive Engineers (SAE) Level 3 “Autobahn-Chauffeur” function, employing TTC and TTC combined with collision speed as safety metrics. These visualizations depict variations in ego- and target-vehicle speed and relative position, yielding continuous criticality landscapes characterized by plateaus. Several scenario selection strategies are demonstrated, such as methods targeting boundary regions of the parameter space or extrema of the metric surfaces.

In our experimental setup, each criticality value corresponds to the maximum of the chosen safety metric observed over the full simulation run. Alternative formulations may incorporate different aggregation strategies depending on the use case or specific system characteristics. For instance, a scalar indicator may encode the minimum value of a continuous-time distance metric, or the integral of a safety violation score over time. The resulting topology of the criticality surface is determined by the structural interactions among the parameterized inputs, the behavioral model of the system under test, and the operational semantics embedded in the metric definition. As a result, even identical system and parameter configurations may give rise to divergent criticality spaces, contingent upon the selected metric or the level of abstraction in the simulation model.

Extensions to the scenario analysis could involve incorporating full trajectories or other temporal components into criticality evaluation or parameter space construction. However, due to the significantly increased complexity and reduced interpretability associated with such extensions, this work deliberately omits these aspects.

2.3. Search-Based Scenario Selection Methods

Given the defined scope of this paper, subsequent sections concentrate exclusively on SBSSMs. These discussions do not constitute an exhaustive review. Instead, select representative methods are examined, serving as illustrative case studies that form the baseline for the benchmark. The methods are grouped according to their primary search objective.

2.3.1. Global Explorative Coverage

Sampling-based methods that aim for broad, uniform coverage of the scenario space are commonly used as baselines in the evaluation of search-based scenario selection techniques. Among these, Random Sampling, Latin Hypercube Sampling (LHS) [34], and Sobol Sequences [35] are frequently employed due to their simplicity and general applicability. LHS stratifies the sampling space along each dimension and ensures that each interval is sampled exactly once, promoting space-filling properties without full randomization. Sobol sequences generate low-discrepancy, quasi-random point sets that achieve uniformity over high-dimensional domains with fewer samples compared to purely random draws. While such methods do not actively target high-risk regions, they provide a valuable reference for assessing the effectiveness of more targeted or adaptive approaches. In surrogate-based settings, Uncertainty Sampling may also be used to guide exploration by prioritizing areas with high model uncertainty, although its performance depends on the underlying surrogate model.

2.3.2. Targeted Extremum Search

Methods in this category aim to identify high-criticality regions in the scenario space by explicitly optimizing criticality-related metrics. In contrast to global sampling techniques, these approaches focus computational resources on sparsely distributed optima that are unlikely to be captured by uniform designs.

Feng et al. [36] propose a multimodal critical-scenario search method based on an Improved Particle Swarm Optimization (IPSO) algorithm, aimed at comprehensively identifying critical scenarios in virtual testing of autonomous vehicles. Unlike conventional

PSO, IPSO incorporates LHS for better initial coverage, a neighborhood-based velocity update mechanism to encourage distributed exploration of multiple critical regions, and a restart strategy to escape premature convergence. The criticality of a scenario is evaluated using the minimum TTC, and the coverage of critical scenarios is quantified via a pixel-wise F1-score between discovered and ground-truth critical regions. While IPSO does not employ an explicit acquisition function, its hybrid mechanisms implicitly balance exploration and exploitation. As such, IPSO exemplifies a robust, heuristic-driven extremum search strategy particularly well-suited for discovering multimodal risk landscapes under limited evaluation budgets.

As a surrogate-assisted method, the Gaussian Process (GP)-based adaptive sampling approach introduced by Torben et al. [25] is included. Originally developed for simulation-based testing of autonomous ships, this method uses GP regression to construct a probabilistic surrogate model of an STL robustness function over a continuous input domain. The model is iteratively refined by selecting new input locations based on acquisition functions, such as Upper Confidence Bound (UCB), which balance exploration of uncertain regions with exploitation of high-predicted criticality. In the original work, robustness values are derived from STL specifications evaluated over simulation traces. In this benchmark, the approach is adapted to directly model analytically defined criticality functions. While surrogate-assisted methods offer efficient search strategies for rare or high-risk scenarios, they may become trapped in local optima if the acquisition landscape exhibits deceptive similarity structures or misleading uncertainty estimates, particularly in high-dimensional or non-smooth scenario spaces.

Song et al. [37] present an industry-oriented workflow for identifying critical scenarios using a three-stage process: scenario derivation from system requirements, closed-loop evaluation in Volvo's proprietary SPAS simulator, and multi-objective optimization via modeFrontier. The optimization is seeded through LHS and refined using either an internal algorithm (pilOPT) or Multi-Objective Simulated Annealing (MOSA), targeting objectives such as minimum TTC and jerk. Although the study demonstrates successful fault exposure in production-level systems, key components, such as the simulator interface, parameter mappings, and optimization setup, are not publicly available. Due to the absence of implementation details and the lack of reproducibility, the method is not included in the benchmark suite.

Beyond heuristic and surrogate-assisted optimization, learning-based adaptive testing has recently gained prominence. Feng et al. [38] train a dense-reinforcement-learning agent that edits a naturalistic driving environment so that safety-critical events are encountered orders of magnitude more frequently, and subsequent work combines such learned testing policies with adaptive importance sampling to correct for the induced distribution shift [39]. Because these approaches are tightly coupled to trained naturalistic environment models and to specific simulation stacks, they are not included as reference implementations here; the extensible interface of the proposed benchmark (cf. BR6), however, permits their future integration once suitable abstractions are defined.

2.3.3. Surrogate-Based Reconstruction

Surrogate-based methods approximate the criticality landscape using data-driven models and can use this approximation to guide further sampling. These approaches are particularly suited for high-cost evaluation settings, where efficient exploration of the parameter space is essential.

Zhang et al. [40] propose a surrogate-assisted safety testing framework aimed at efficiently identifying high-risk scenario areas, such as those with critically low TTC. Their approach evaluates six surrogate models, combining both interpolation- and regression-based

techniques, and employs a multiplicative acquisition function that balances exploitation (predicted risk) and exploration (distance to tested samples). Among these, the Inverse Distance Weighted (IDW) model consistently demonstrated the best performance across different ODDs, especially for rare and high-risk events. Due to its simplicity, runtime efficiency, and robustness across scenario types, IDW is included and implemented in this benchmark suite as a representative surrogate-assisted search method. It should be noted, however, that the method does not aim to reconstruct the entire criticality landscape perfectly, but rather tries to identify and delimit all distinct high-risk regions through targeted sampling.

Artificial Neural Networks (ANNs), here instantiated as feedforward Multi-Layer Perceptrons, are widely used surrogate models for approximating nonlinear and high-dimensional objective landscapes. Due to their universal function approximation capability, ANNs are particularly suited for modeling complex scenario-response relationships where the underlying structure may be non-smooth or highly multimodal [41]. However, their effectiveness strongly depends on the availability of sufficient training data, as they are prone to overfitting in sparse sampling regimes, which may occur in simulation-based testing [42]. Liu et al. [43] identify ANNs as a candidate model class for surrogate-assisted optimization in combinatorial spaces, though they emphasize the need for careful architecture selection and regularization to ensure generalization across discrete or hybrid domains.

In addition to standalone surrogate models, ensemble methods represent a powerful extension for surrogate-assisted safety testing. By combining multiple model types, ensembles can leverage complementary strengths and improve robustness in sparse or heterogeneous scenario spaces. Gorissen et al. [44] demonstrate a dynamic ensemble approach, where different surrogate models evolve cooperatively via specialized genetic algorithms and are recombined into hybrid predictors.

2.3.4. Boundary Region Detection

Performance-boundary identification within the high-dimensional scenario space used for autonomous-system validation identifies hypersurfaces where infinitesimal parameter changes induce large behavioral shifts. Concentrating evaluations on these regions elevates test efficiency, exposes critical failure modes for robustness assessment, and yields data that supports behavioral interpretation and systematic root-cause analysis.

Mullins et al. [12] investigated two adaptive sampling strategies within the Range Adversarial Planning Tool (RAPT) framework. The Gaussian Process Regression (GPR)-based approach fits a global surrogate model after each simulation batch and evaluates candidate scenarios using a score that combines the local gradient magnitude of the GPR mean with its predictive variance, thereby steering sampling toward regions likely to contain performance boundaries. In contrast, the Nearest Neighbor Density and Variance (NNDV) approach does not rely on a global surrogate model but instead estimates local variance and density directly from the k -nearest neighbors of each candidate point, favoring regions that are both variable in outcome and sparsely explored. Comparative evaluation on synthetic functions and an unmanned underwater vehicle case study demonstrated that NNDV achieved higher precision and coverage of boundary regions with lower computational cost and was therefore adopted as the preferred strategy.

Lu et al. [11] extend the concept with the Gradient Distance Nearest Neighbors Adaptive Sampling (GDNNAS) procedure. This model-free method selects candidate points through a joint gradient magnitude and Euclidean distance metric. Then the methods discard any scenario whose closest neighbor of a different mode lies beyond a dynamic distance threshold. While both RAPT and GDNNAS pursue boundary discovery, RAPT retains the entire sampling trace as the test set, whereas GDNNAS prunes off-boundary

points and delivers a highly condensed set of pure boundary cases. Evaluation metrics must therefore be aligned with these distinct output profiles. Both algorithms demonstrate nonetheless that adaptive sampling guided by local sensitivity metrics can efficiently target the sparsely distributed high-risk slices of the scenario space and therefore provide strong reference points for the benchmark proposed in this study.

Batsch et al. [45] propose the use of Gaussian Process (GP)-based surrogate models to reconstruct the criticality landscape of scenario spaces for automated driving. Their method leverages both Gaussian Process Classification (GPC) and Regression (GPR) to approximate the relationship between scenario parameters and a safety-relevant metric, specifically, the minimum time-to-collision (TTC). Unlike purely boundary-focused techniques, the aim is also to approximate the entire parameter space, including safe, unsafe, and intermediate regions. However, the method strategically concentrates sampling efforts near the performance boundary, where informative transitions between safe and unsafe behavior occur. The process begins with a space-filling design (LHS) to generate an initial surrogate. GPC is employed to filter infeasible or uninformative scenarios, while GPR estimates TTC values and associated predictive uncertainty. A straddle heuristic is then used to iteratively select new candidate scenarios that balance exploration and exploitation. The updated surrogate models are retrained after each simulation, enabling progressively refined reconstructions of the scenario landscape. Additionally, the GP posterior enables variance-based global sensitivity analysis, facilitating the identification of dominant scenario parameters and potential dimensionality reduction. Although excluded from the benchmark implementations, this approach is presented to illustrate the adaptability of similar surrogate modeling frameworks to distinct sampling strategies and optimization objectives.

2.4. Review of Available Benchmarks

Existing benchmarks relevant to scenario selection can be grouped into three categories: simulation-dependent frameworks, real-world data-driven benchmarks, and general-purpose optimization suites. As detailed below, none of them fully addresses the requirements of evaluating SBSSMs across multiple selection objectives.

Simulation-dependent benchmarks, such as SafeBench [46], constructed upon simulation platforms like CARLA [47], offer controlled, repeatable environments with various predefined scenarios. Their structured design facilitates the generation of highly realistic scenarios and allows for efficient modifications. However, their inherent domain specificity and limited structural variability constrain their diversity and restrict cross-domain generalization. OpenSBT [48,49] provides a modular, open-source framework that supports the integration of multiple simulation platforms, search algorithms, and fitness functions. While primarily designed for realistic system-level testing of automated driving systems, its flexible architecture allows for cross-simulator comparisons and configurable testing pipelines. The included metrics and example use cases primarily focus on multi-objective search and scenario coverage. As such, OpenSBT is well-suited for evaluating concrete scenario-based behaviors within automotive simulators or similar simulation environments but is less applicable to abstract problem classes that arise in broader scenario selection research.

Real-world data-driven benchmarks derived from empirical driving logs or accident databases [15] provide authentic scenario complexity and realistic data distributions. Nevertheless, they suffer from incomplete representation of a closely defined scenario space, difficulties in reproducing exact conditions, and limited extensibility beyond the original data domain. The CommonRoad Initiative [24] provides numerous real-traffic, hand-crafted, and automatically generated scenarios that can be used for benchmarking and

testing, particularly for motion planning algorithms. However, these scenarios are limited to this specific application domain.

In contrast, general-purpose black-box optimization suites such as Black-Box Optimization Benchmarking (BBOB)/Comparing Continuous Optimizers (COCO) [50] or Mixed-variable and Combinatorial Bayesian Optimization (MCBO) [51] provide standardized and reproducible test problems for evaluating optimization algorithms across diverse mathematical structures and variable types. These suites primarily assess an algorithm's ability to identify global or near-global optima under defined evaluation budgets. This objective is closely related to falsification, but it represents only one of several tasks relevant to search-based scenario selection. Although canonical optimization suites include functions with properties such as multimodality, separability, ill-conditioning, and varying degrees of regularity, their test instances are not specifically designed to reproduce the structural characteristics of safety-critical scenario spaces. In particular, properties such as sparsely distributed critical regions, extended plateaus, threshold-induced transitions, and empirically observed anisotropic interactions are not systematically incorporated as benchmark requirements. Benchopt [52] further supports reproducible comparisons of optimization methods across a broad range of problem formulations. However, its focus is primarily on algorithmic performance for predefined optimization problems rather than on the simultaneous evaluation of distinct scenario-selection objectives. In the cyber-physical-systems community, the annual Applied Verification for Continuous and Hybrid Systems Competition (ARCH-COMP) friendly competition [53,54] constitutes a long-running benchmarking initiative for the falsification of signal-temporal-logic specifications, providing standardized Simulink-based benchmark models, fixed evaluation budgets, and comparative tool reports. Closely related, Part-X [55] frames search-based test generation as a partitioning scheme with probabilistic guarantees on the identified falsifying regions. While these initiatives share the goal of reproducible cross-method comparison, they remain coupled to concrete system models and specification formalisms; they neither expose the structural properties of empirical criticality landscapes identified in the following section nor support scenario-selection objectives beyond falsification, such as landscape reconstruction or boundary coverage. Consequently, extending these numeric benchmarks to incorporate new empirical observations of scenarios proves challenging, highlighting the necessity for specialized frameworks tailored explicitly to scenario-based validation contexts. Furthermore, a comprehensive assessment of SBSSMs necessitates the incorporation of application-specific evaluation criteria that can extend beyond conventional optimization metrics.

3. Empirical Landscape Analysis

Building on the theoretical foundations and methodological review in the preceding section, this section empirically investigates the structural properties of criticality landscapes observed in realistic simulation environments. These empirical observations serve as both a structural template and a validation reference for the synthetic criticality landscapes developed in the subsequent benchmark design. In doing so, it is ensured that the abstracted test functions retain essential characteristics of real-world operational domains while allowing for controlled, repeatable evaluation of selection algorithms under diverse and challenging conditions.

Accordingly, the following subsections first describe the simulation setups and applied metrics, before illustrating key structural features through selected representative examples.

3.1. Scenarios and Systems Under Test

For the experimental evaluation, two algorithms were applied within their respective simulation environments. One of them, developed by TKMS and its segment ATLAS ELEKTRONIK [56], is an obstacle avoidance algorithm that relies on an optimization-based controller. The controller identifies situations that fall under the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs), the globally recognized set of navigational rules for maritime traffic, as well as encounters with static obstacles. It then continuously adjusts the vessel's heading and speed. These control decisions consider estimated energy expenditure, including detour penalties.

In the employed simulation environment, distinct criticality evaluation metrics were implemented to examine both the structural impact of metric selection on the resulting criticality landscape and to highlight the sensitivity of scenario selection performance to the choice of safety metric. This comparative design aims to underscore the methodological importance of aligning criticality definitions with domain-specific risk semantics.

The first metric, Time-to-Collision [32], was selected as a classical proxy for temporal proximity to hazardous interactions. TTC has been widely adopted in the validation of autonomous systems due to its interpretability, analytical tractability, and strong correlation with imminent collision risk in simplified kinematic settings. In the maritime context, TTC is particularly effective in identifying rapidly evolving, high-risk situations, especially when relative velocities are high and evasive maneuverability is limited.

Secondly, the Collision-Risk-Index (CRI), a compound metric based on Distance to Closest Point of Approach (DCPA) and Time to Closest Point of Approach (TCPA), was implemented, following established formulations from the domain of ship encounter analysis [57]. DCPA and TCPA are foundational metrics that independently capture spatial and temporal proximity in vessel interactions. DCPA measures the minimum distance between two vessels during an encounter, while TCPA quantifies the time remaining until this minimum distance is reached. When combined, these metrics offer a more comprehensive assessment of vessel interactions, reflecting both spatial and temporal dimensions. The CRI extends beyond the identification of direct collision courses by capturing scenarios where vessel trajectories may result in hazardous proximity.

Lastly, the Ship-Domain Overlap metric, based on the geometric safety buffer model originally proposed by Fujii, was included to capture violations of the so-called "ship domain", a vessel-specific area that must remain free of intrusions to ensure navigational safety [58]. Unlike point-based measures, this metric considers the spatial envelope of a vessel, accounting for its size and heading. It is particularly well-suited for evaluating passing situations and encounters involving low relative velocities, in which TTC-based methods may underestimate risk due to either small closing rates or parallel trajectories, despite critical spatial proximity.

The scenario (see Figure 1a) used for testing involves a situation in which the ego vessel must navigate around both a static obstacle and a crossing vessel exhibiting linear, collision-prone trajectories. Variations in the size and spatial positioning of the static obstacle were systematically introduced to elicit complex interaction dynamics between the ego vessel and the crossing vessel. To induce failure-prone behavior and thereby generate nontrivial criticality structures, the perception range of the algorithm was deliberately constrained. This controlled degradation in system capability, representative of consistently unfavorable environmental conditions, induces suboptimal decision-making under certain scenarios, thereby creating a diverse set of safety-relevant configurations within the criticality landscape.

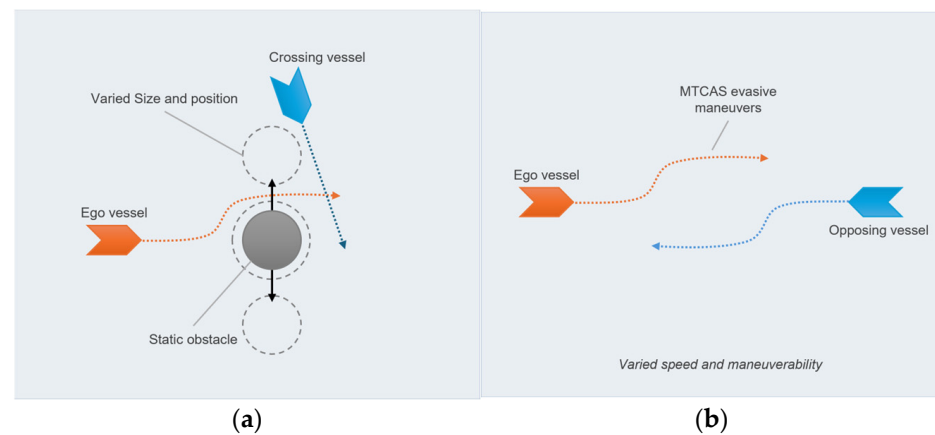


Figure 1. Schematic illustration of the simulation scenarios. (a) Crossing scenario used in the TKMS ATLAS ELEKTRONIK simulation: the ego vessel evades a static obstacle, whose size and position are varied between scenarios, while encountering a crossing vessel on a collision-prone course. (b) Head-on encounter scenario used in the DLR Maritime Traffic Simulation: both vessels approach on a collision course and follow the evasive maneuvers recommended by MTCAS.

The selected scenario was designed to reflect structurally representative navigation challenges that are characteristic of real-world operational domains. It embodies key elements such as static and dynamic obstacles, constrained maneuvering space, and COLREG-compliant decision logic, thereby offering a rich testbed for assessing the capacity of selection methods to uncover regions of elevated criticality.

The second scenario entails two vessels in a Head-On situation. It is executed using the Maritime Traffic Simulation (MTS). The MTS is the virtual side of the maritime testbed eMIR [59] developed by the German Aerospace Center (DLR) [60]. The MTS is capable of simulating interacting vessels with a configurable level of detail ranging from the collision avoidance behavior over sensors down to the physical dynamics of a ship. It has been employed during the evaluation of many heterogeneous maritime systems during research projects such as Maritime Traffic Collision Avoidance System (MTCAS) [61]. In this scenario, both ships are equipped with MTCAS and follow its recommendations for collision avoidance. MTCAS recommends evasive maneuvers based upon distance and time to the critical ship pose (CSP), a context-sensitive evolution of the closest point of approach (CPA). The schematic trajectories of the resulting evasive maneuver for each vessel are depicted in Figure 1b. Since the purpose of MTCAS is to reduce the risk of collision, the chosen risk metric is the maximum ship-domain overlap observed during the maneuver. The resulting metric is investigated under varied parameters, namely the speed and maneuverability of both vessels. The maneuverability in this scenario is defined as the maximum rate of turn a vessel can achieve. As the MTCAS recommendations depend on the time to CSP and thus, by extension, on the speed of both vessels as well as the distance to CSP, the resulting criticality space is non-linear and sometimes counterintuitive. For example, the risk of collision, as derived from the metric used, is relatively high because MTCAS is reacting late since the time to CSP is low (see Figure 2). The simulated criticality spaces serve as a structural point of reference for the generation of synthetic criticality landscapes and inform the qualitative assessment of their realism with respect to domain-specific safety behavior.

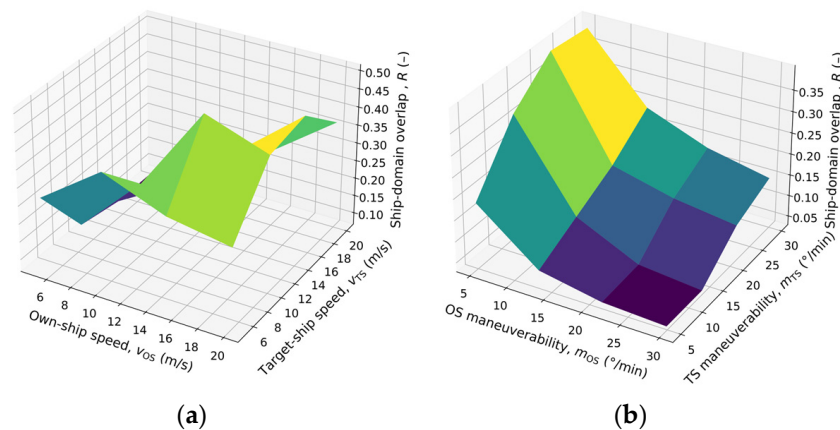


Figure 2. Criticality spaces from the DLR simulation. (a) Maximum ship-domain-overlap metric as a function of ego vessel speed and approaching vessel speed, showing multimodal and oscillatory behavior. (b) Maximum ship-domain-overlap metric as a function of the maneuverability of the ego vessel and the approaching vessel, showing a monotonic increase toward a single corner.

3.2. Observed Structural Patterns

The following figures present selected examples of criticality spaces generated from the simulation-based experimental setup described above. Each visualization corresponds to a two-dimensional slice of the underlying high-dimensional parameter space: two scenario parameters are varied to span the axes, while all remaining input parameters are held fixed at specific values. These visualizations illustrate a range of topological and mathematical structures that emerge from variations in scenario parameters. The primary objective of these figures is to exemplify the structural diversity inherent to realistic criticality landscapes, thereby highlighting the algorithmic challenges posed to scenario selection and search-based evaluation methods.

These structural patterns serve as the conceptual foundation for the construction of mathematical functions used in the synthetic benchmark framework. While the illustrated cases represent relatively simple interactions within low-dimensional subspaces, they are intended to inform the design of more abstract, generalizable function compositions. The goal is to extrapolate these localized features into complex, high-dimensional criticality functions that reflect the behavior of realistic, large-scale systems.

3.3. Mapping Observations to Benchmark Space Requirements

The surface plots above, combined with structural characteristics reported in the literature, reveal five recurring structural properties of empirical criticality landscapes. Each property poses a distinct algorithmic challenge to scenario selection methods and is therefore directly adopted as a functional requirement for the synthetic test instances used in this benchmark.

Non-linearity: As demonstrated by the oscillatory response surface in Figure 2, criticality functions must exhibit nonlinear dependencies on scenario parameters, reflecting the complex interaction dynamics of autonomous systems [25,26].

Non-convexity: The multimodal structures observed in Figure 2a and reported by Gladisch et al. [26] and Torben et al. [25] show that benchmark spaces must contain multiple local optima to capture multimodal risk distributions and prevent trivial optimization.

Anisotropy: The monotonic corner-directed gradient in Figure 2b, consistent with the directional parameter sensitivities reported in [12,25], shows that parameter dimensions must contribute unequally to system behavior. Directional scaling factors must therefore introduce non-axis-aligned gradients.

Sparsity: The isolated high-criticality islands visible in Figure 3b, and similarly reported by Mullins et al. [12] and Lu et al. [11], demonstrate that regions of elevated criticality are sparsely embedded in predominantly safe high-dimensional parameter spaces, reflecting realistic failure-mode distributions.

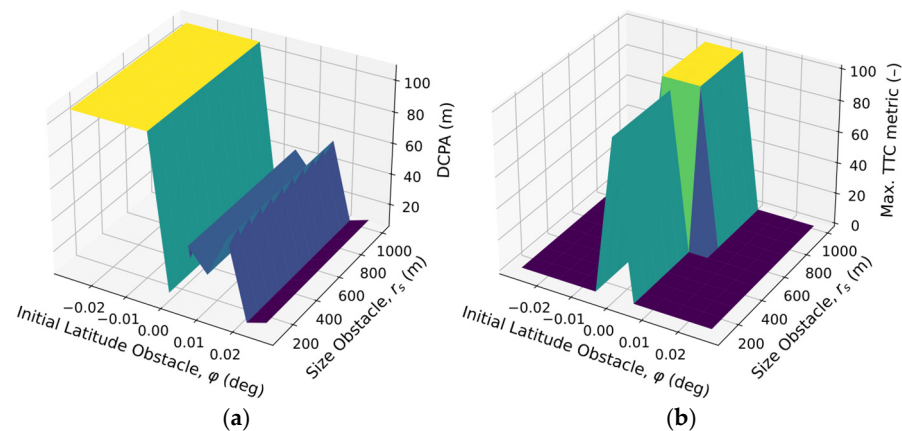


Figure 3. Criticality spaces from the TKMS ATLAS ELEKTRONIK simulation. (a) DCPA metric as a function of obstacle size and geographic position, showing discontinuities and multiple local maxima. (b) TTC metric as a function of obstacle size and geographic position, showing distinct discontinuities and plateaus.

Discontinuities: The sharp transitions and plateaus in Figure 3b arise from controller mode switches and specification thresholds. A phenomenon independently confirmed in STL-based maritime [25] and automotive [26] evaluations, as well as consistent with the threshold-induced plateau structures reported in the PEGASUS project documentation [33]. Sharp transitions and plateaus must therefore be included in synthetic landscapes to emulate such effects.

Together, these observations define the functional requirements for the synthetic benchmark landscapes. They ensure that evaluation scenarios are structurally representative of empirical criticality spaces and that tested algorithms are exposed to the same search challenges encountered in operational domains. Whether the generated instances satisfy these requirements is verified quantitatively in Section 4.5, where each of the five properties is operationalized as a scalar descriptor and compared between the empirical and synthetic landscapes.

4. Benchmark Design

As outlined in Section 3, representative criticality spaces were qualitatively identified using two distinct simulation setups, capturing complex interactions characteristic of maritime operational domains. These empirical observations serve as reference points, encapsulating salient structural features. Based on these insights and the spaces delineated in the literature, a synthetic criticality-space generation procedure was developed, described in detail in the following sections. Randomized compositions of mathematical primitives were strategically combined to reproduce observed structural properties and to construct challenging landscapes for search-based scenario selection methods.

Parameters of these primitives were sampled from probabilistic distributions explicitly designed to induce sparsity and anisotropy. To ensure representativeness and diversity, generated scenarios were filtered according to predefined structural criteria, enabling meaningful cross-method comparisons. Representative examples of resulting synthetic spaces are presented in Section 4.4 to illustrate the diversity of generated topologies.

Evaluation metrics were defined to include both general performance indicators and application-specific measures, enabling a comprehensive assessment of selection algorithms with respect to critical-region identification, boundary coverage, and sampling efficiency. This framework provides a reproducible basis for the comparative evaluation of SBSSMs.

Sections 4.1–4.7 describe the underlying functions, generation parameters, diversity assurance, resulting landscapes, their empirical validation, evaluation metrics, and the overall benchmark structure.

4.1. Functions and Models

The synthetic criticality spaces considered in this work are generated via randomized procedures based on various mathematical primitives. The selection of these functions was guided by two main objectives: first, to replicate structural patterns observed in exemplary spaces derived from the simulations, and second, to introduce structural complexities that pose significant challenges for SBSSMs. This section provides a brief overview of the utilized functions and the rationale behind their selection.

The non-normalized Gaussian function is employed due to its nonlinear characteristics. In the implementation, the function is truncated at five standard deviations to reduce computational complexity while preserving its essential shape. The function is defined as

$$f_{Gauss}(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top} \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad (3)$$

where $\mathbf{x}$ is the input vector, $\boldsymbol{\mu}$ the mean vector, and Σ the covariance matrix. The truncation is implemented by setting $f_{Gauss}(\mathbf{x}) = 0$ for $\|\mathbf{x} - \boldsymbol{\mu}\| > 5\sigma$.

A hypercube has a constant value within its boundaries. If this value is non-zero, it may arise from various causes, such as initial conditions, internal software constraints, or the structure of the criticality function itself. In cases where the constant value equals the domain maximum, this may indicate a system failure confined to that region. Due to discontinuities at the boundaries, such functions are particularly challenging for certain approximation methods.

Formally, the hypercube function is defined as

$$f_{Cube}(\mathbf{x}) = \begin{cases} c, & \text{if } \mathbf{x} \in [a_1, b_1] \times \cdots \times [a_d, b_d] \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

where $c \in \mathbb{R}$ is the constant value inside the hypercube and $[a_i, b_i] \subset \mathbb{R}$ are the interval bounds in each dimension $i = 1, \dots, d$.

The ramp function encodes a linear relationship among multiple parameters. Owing to the bounded nature of the domain, discontinuities may emerge at the edges. This behavior was also observed in simulated criticality spaces, particularly in scenarios involving challenging parameter combinations—such as imminent collision courses or intersecting trajectories of traffic participants. The ramp function can be written as:

$$f_{Ramp}(\mathbf{x}) = \mathbf{w}^{\top} \mathbf{x} + b, \quad (5)$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input vector, $\mathbf{w} \in \mathbb{R}^d$ is the weight vector (slope of the function), and $b \in \mathbb{R}$ is the intercept.

To evaluate the approximation capabilities for harmonic structures, various sine functions were incorporated into the generated spaces. Such patterns may emerge due to the structural positioning of sensors or environmental influences interacting with engineered surroundings. The identification of such harmonic structures, even when not directly

linked to critical behavior, may still yield insights into the improvement of autonomous systems. The Sine Function is defined as a superposition of multiple sinusoidal waves:

$$f_{\text{Sine}}(\mathbf{x}) = \text{offset} + \sum_{i=1}^N A_i \sin(f_i(\mathbf{w}_i^T \mathbf{x}) + \phi_i), \quad (6)$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input vector, N is the number of sinusoidal waves, A_i is the amplitude, f_i the frequency, $\mathbf{w}_i \in \mathbb{R}^d$ the weight vector determining the orientation, ϕ_i the phase of the i -th wave and offset is a constant value added to the sum of the waves. Optionally, the Sinus Function is constrained to a bounding box for each dimension.

In addition to structured components, random noise is incorporated into the synthetic criticality spaces to emulate stochastic or non-deterministic effects that may arise in real-world systems due to sensor uncertainties, environmental disturbances, or model imperfections. Furthermore, this noise can also originate from probabilistic AI modules, such as those employing stochastic policy approximations via deep reinforcement learning [62], which, despite introducing behavioral variability, tend to produce similar responses in comparably challenging scenarios. The noise function is constructed by sampling values independently from a standard normal distribution. This stochastic component introduces high-frequency, non-smooth variations in the response surface, thereby further challenging the approximation capabilities of surrogate models or SBSSMs. Mathematically, the noise function is defined as:

$$\varepsilon(\mathbf{x}) \sim N(0, \sigma^2), \forall \mathbf{x} \in \mathbb{R}^d, \quad (7)$$

where $N(0, \sigma^2)$ denotes a univariate normal distribution with mean 0 and variance σ^2 . The noise is applied pointwise and independently across the domain, such that:

$$\mathbf{x}_i \mapsto f_{\text{total}}(\mathbf{x}_i) = f_{\text{base}}(\mathbf{x}_i) + \varepsilon_i, \quad (8)$$

Here, f_{base} refers to the sum of the structured components and ε_i denotes the noise term at location $\mathbf{x}_i$.

4.2. Parameters and Assumptions

The generation of synthetic criticality spaces in this paper is guided by a set of design assumptions, which are informed by observations from simulation-based validation environments and theoretical considerations regarding the structure of high-dimensional parameter spaces in autonomous systems. To evaluate the effectiveness of scenario selection methods under challenging conditions, the synthetic criticality spaces are deliberately constructed to contain sparsely distributed regions of elevated criticality embedded in otherwise low-criticality domains. This structural design enforces the need for targeted search strategies and serves to rigorously test the capability of the methods to identify isolated, high-risk areas within a high-dimensional space. To achieve this, Beta-distributed sampling is employed for determining function amplitudes during the generation phase. This probabilistic shaping introduces a natural bias towards low criticality values, while preserving the possibility of sharply peaked regions. This approach is both analytically tractable and flexible enough to emulate observed sparsity patterns in empirical response surfaces, as illustrated in Section 3.2.

Recognizing the impracticality of manually constructing diverse high-dimensional criticality spaces, all test instances are generated through randomized composition of analytic function primitives, including Gaussian, ramp, hypercube, sine functions, and stochastic noise components (cf. Section 4.1). Each criticality space is synthesized by combining a randomly selected subset of these functions, with parameter values (e.g., location, scale) sampled from predefined distributions. This ensures the generation of heterogeneous topologies across different test instances.

In real-world systems, not all parameters contribute equally to safety-relevant outcomes. Moreover, interactions between parameters often lead to critical behavior that cannot be attributed to any single axis [63,64]. To model this, anisotropy is explicitly embedded into the synthetic criticality spaces via dimension-specific scaling factors and directional weighting schemes. These weights are drawn randomly and applied to both linear and nonlinear components, leading to non-axis-aligned criticality gradients.

Optionally, one or more parameters per scenario can be defined as structurally irrelevant, meaning that variations along these dimensions have no effect on the criticality function. Their inclusion enables the evaluation of whether selection methods can effectively differentiate between informative and non-informative input dimensions, an essential requirement for scalability and robustness in high-dimensional real-world applications. This option was not activated in the benchmark instances used in this study, as the illustrative experiments and evaluation metrics in Sections 4.6 and 5 do not target this capability; the generator configuration, however, supports the inclusion of irrelevant dimensions for evaluations that require it.

4.3. Diversity Assurance and Instance Selection

As motivated in the previous section, the generation of synthetic criticality spaces in this work relies on randomized construction techniques. However, due to inherent limitations in the size and diversity of any finite test set, it becomes essential to select a subset of generated samples in a way that maximizes structural variability. To this end, a systematic filtering procedure is applied following the initial random generation phase.

To quantify the structural characteristics of each synthetic space, a discrete Fourier transform (DFT) [65,66] is applied to the spatial representation of the criticality function. The resulting frequency spectra are compared across all pairwise combinations using correlation coefficients. By evaluating the correlations between all permutations of the parameter axes, the method is able to identify structurally similar spaces, even if their critical structures are rotated or reflected within the parameter space. While this approach is not strictly invariant to spatial translations, it effectively captures global structural similarities in the spectral domain, which is well-suited for the characterization of function topologies in high-dimensional spaces.

The resulting pairwise similarity matrix is further enriched with statistical descriptors such as the mean and variance of the criticality values. Each descriptor is transformed into a distance metric, and the resulting metrics are aggregated into a single dissimilarity matrix. This matrix serves as input to a k-Medoids clustering algorithm [67,68], where the number of clusters corresponds to the desired number of benchmark instances. By selecting the medoids from each cluster, the final set of synthetic spaces is constructed to span a wide range of structural and statistical characteristics. This selection procedure ensures that the benchmark suite includes both regular and irregular criticality structures, while maintaining statistical rigor and diversity across the test cases.

4.4. Example Synthetic Criticality Landscapes

To complement the structural and filtering descriptions given above, Figure 4 presents three examples of generated criticality functions used in the benchmark. Each surface plot illustrates distinct topological properties that are representative of the benchmark's function pool.

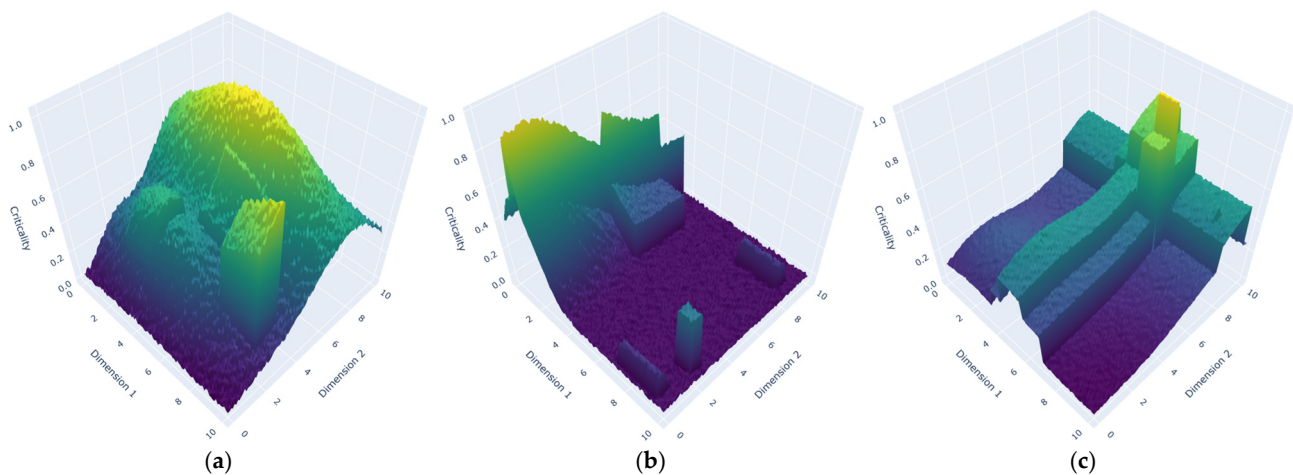


Figure 4. Examples of synthetic criticality landscapes illustrating structural diversity across the benchmark suite. (a) Noisy surface with a dominant peak and two critical areas. (b) Stepwise hypercubic structure with a directional ramp and isolated high-risk blocks. (c) Piecewise-plateau regions with varying criticality levels and abrupt boundary transitions.

Figure 4 presents three representative examples of synthetic criticality functions used in the benchmark, each illustrating distinct topological properties. Panel (a) shows a surface with a dominant central peak overlaid by significant additive noise, creating multiple spurious local maxima and challenging methods that rely on smooth or noise-free gradient estimates. Panel (b) exhibits sharp discontinuities and isolated high-criticality blocks, making reliable interpolation difficult for surrogate models and causing gradient-free optimizers to miss critical regions entirely. Panel (c) presents a piecewise-constant structure with abrupt level transitions, representative of systems governed by discrete thresholds, where the absence of gradient information within flat regions complicates boundary localization.

4.5. Quantitative Structural Comparison with Empirical Landscapes

To substantiate the qualitative correspondence between the synthetic landscapes and the empirical observations of Section 3 quantitatively, five scalar descriptors were computed for both, each operationalizing one structural requirement of Section 3.3: non-linearity as the residual variance of a least-squares hyperplane fit ($1 - R^2$), multimodality as the number of local maxima with a topographic prominence of at least 10% of the value range, anisotropy as the ratio of the largest to the smallest eigenvalue of the gradient covariance matrix, the critical-region fraction as the share of grid cells exceeding the criticality threshold, and discontinuity as a combined measure of the extent and severity of grid cells whose gradient magnitude exceeds five times the median.

The empirical references must be interpreted in light of how they were obtained. The three TKMS surfaces stem from a setup in which the perception range was deliberately constrained to provoke failures (Section 3.1); they therefore represent a failure-rich excerpt of the scenario space, characterized by extended critical regions, plateaus, and sharp transitions. The DLR surfaces, in contrast, capture regular collision-avoidance behavior under MTCAS and are correspondingly smooth, with small critical regions and no discontinuities. The two sources thus mark opposite ends of a structural spectrum rather than a representative sample of operational criticality spaces. In addition, all empirical surfaces are reconstructed from factorial grids with eight levels per axis, and the interpolation onto the descriptor resolution attenuates sharp transitions, biasing the empirical discontinuity values downward. The individual descriptor values should therefore be read as indicative anchors, not as a ground-truth distribution.

Figures 5 and 6 compare the descriptor distributions of the synthetic instance pools with the empirical references for the two- and four-dimensional case, respectively. Two observations follow. First, in the two-dimensional comparison (Figure 5), the synthetic pool spans both empirical regimes: the smooth, weakly critical DLR surfaces and the discontinuous, failure-dominated TKMS surfaces lie within or adjacent to the synthetic range for multimodality, anisotropy, and critical-region fraction, while non-linearity and discontinuity of the individual references bracket the synthetic distribution from opposite sides. The anisotropy value of the TKMS DCPA surface constitutes an exception: it reaches the numerical cap of the descriptor because its plateau-dominated topology yields a nearly degenerate gradient covariance and marks a boundary case rather than a measurable ratio. This coverage of heterogeneous structural regimes, rather than agreement with any single reference, is the property required of a benchmark intended to expose selection methods to diverse conditions (BR5). Second, the most reliable comparison is the four-dimensional case (Figure 6), which covers the full parameter space of the head-on scenario rather than a low-dimensional excerpt and matches the dimensionality of the benchmark application in Section 5. Here the empirical space falls within or at the boundary of the synthetic distribution for all five descriptors, indicating that the generated landscapes reproduce the global structural characteristics of a complete empirical criticality space.

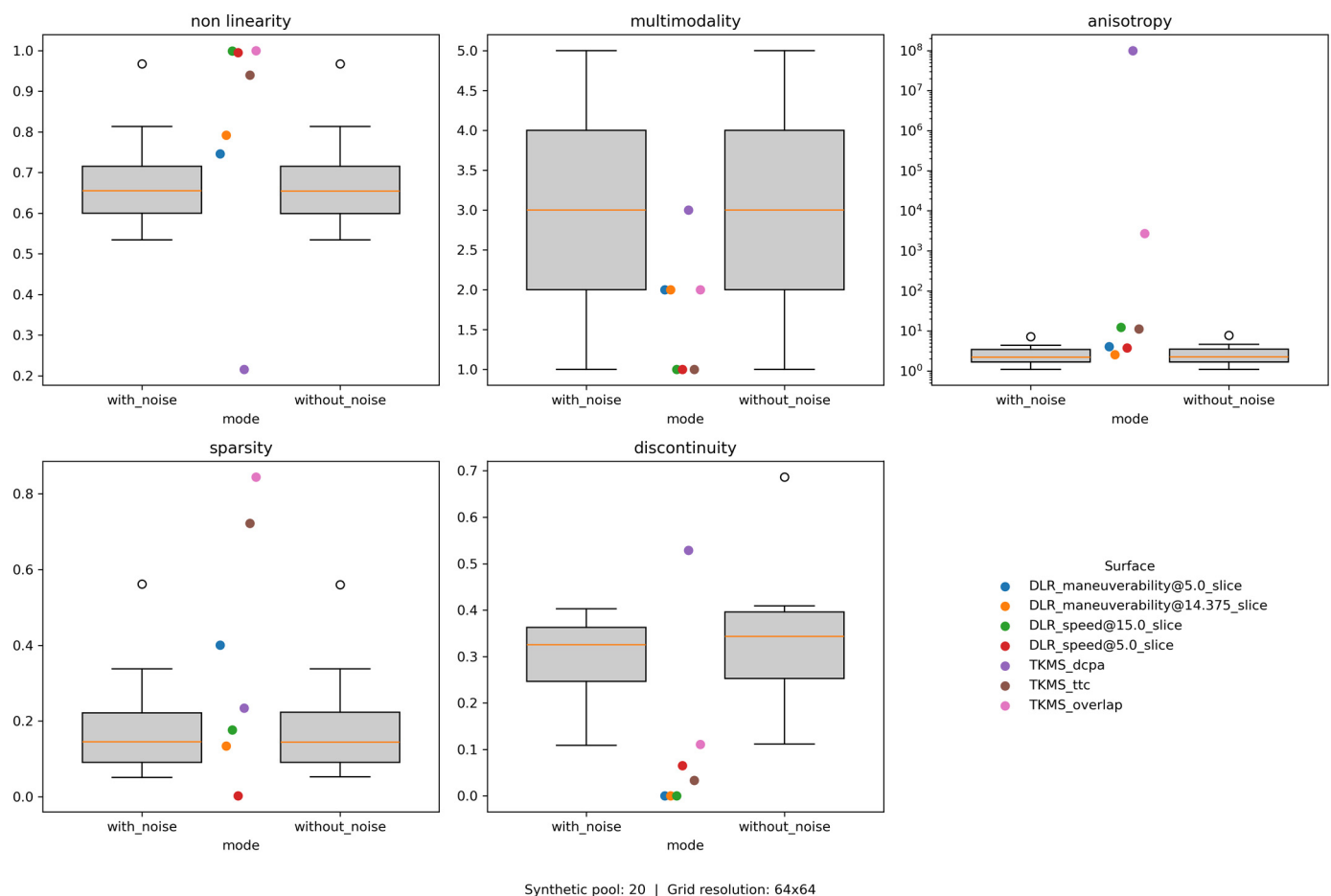


Figure 5. Structural descriptor distributions of the twenty two-dimensional synthetic benchmark instances (boxplots, grouped into instances with and without noise components) with the seven empirical reference surfaces overlaid as individual points: three TKMS surfaces (DCPA, TTC, ship-domain overlap) and four two-dimensional slices of the DLR head-on scenario. Anisotropy is shown on a logarithmic scale; the TKMS DCPA value coincides with the numerical cap of the descriptor due to its plateau-dominated topology.

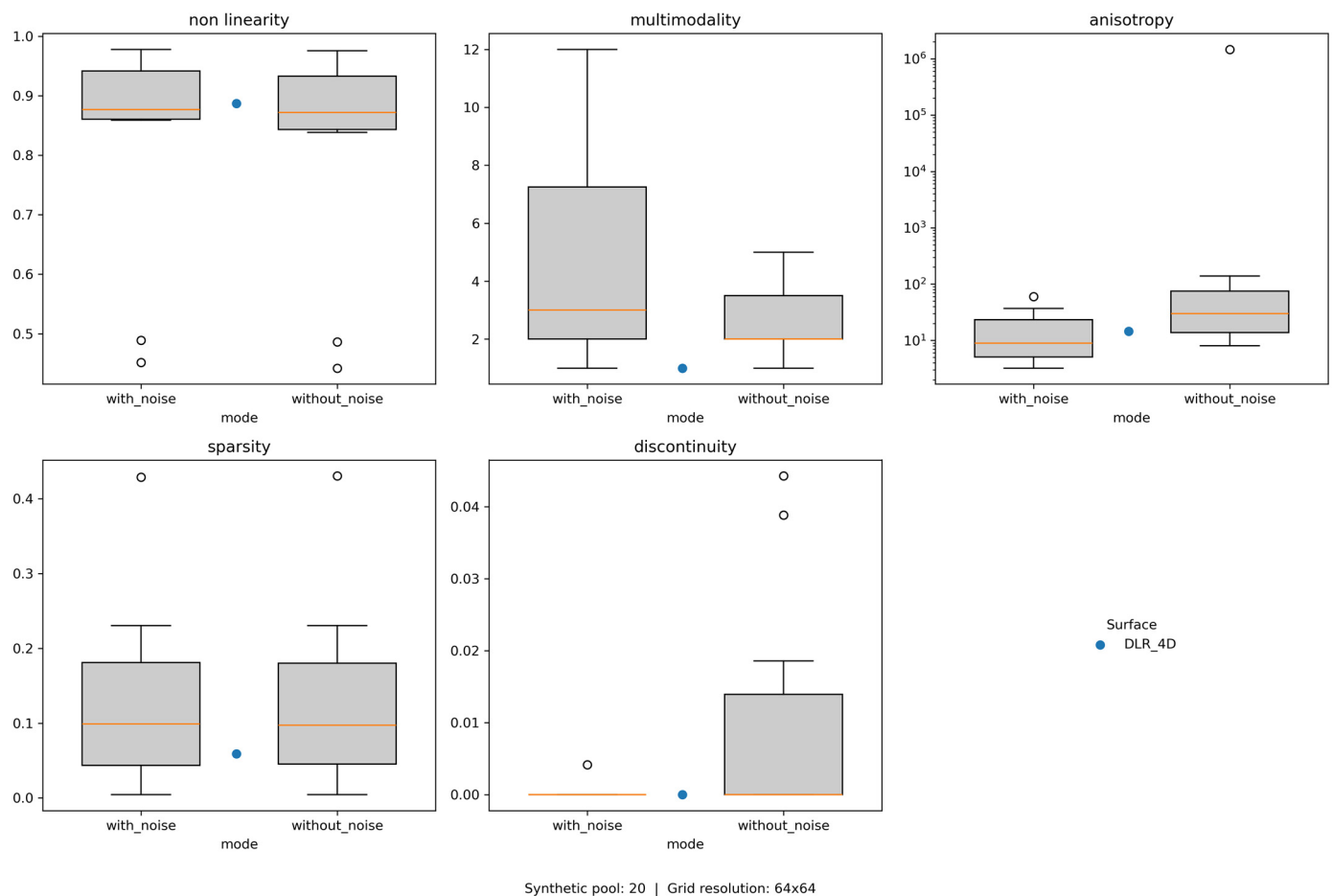


Figure 6. Structural descriptor distributions of the twenty four-dimensional synthetic benchmark instances (boxplots, grouped into instances with and without noise components) with the full four-dimensional DLR criticality space overlaid as a single point. Anisotropy is shown on a logarithmic scale.

4.6. Evaluation Metrics

The evaluation of scenario selection methods requires tailored metrics that not only quantify overall sampling performance but also reflect the specific objectives of the employed search strategy. In the context of high-dimensional and structurally diverse criticality landscapes, such metrics must be sensitive to both the spatial characteristics of the sample distribution and the functional relevance of the identified scenarios. Accordingly, this section distinguishes between general-purpose metrics, which are method-agnostic and capture geometric or statistical aspects of the sample set, and strategy-specific metrics, which target core objectives such as extremum identification, model approximation, or boundary localization.

The metrics presented herein serve as a foundational set for comparative analysis. However, they are not intended to be exhaustive. Since evaluation goals can vary substantially depending on application-specific safety arguments or optimization criteria, the framework explicitly allows for the integration of user-defined metrics that better reflect task-dependent performance requirements.

4.6.1. General Metrics

Independent of the employed search strategy, scenario selection performance can be characterized along two fundamental dimensions: sampling efficiency and spatial coverage.

Sampling efficiency is quantified via the mean criticality score μ_C and the maximum observed criticality C_{max} of the selected scenario set:

$$\mu_C = \frac{1}{|X_{sel}|} \sum_{x \in X_{sel}} C(x), \quad C_{max} = \max_{x \in X_{sel}} C(x), \quad (9)$$

with X_{sel} as the selected scenarios and $C(x)$ Criticality value at point x .

To assess the spatial distribution quality of selected samples, two complementary metrics are employed: the discrepancy and the spatial entropy. Both aim to characterize how uniformly and extensively the sample set populates the parameter space, albeit from fundamentally different perspectives.

The discrepancy metric quantifies the maximum deviation between the empirical distribution of the samples and the ideal uniform distribution. Formally, it is defined as:

$$D = \sup_{u \in [0,1]^d} \left| \frac{1}{N} \sum_{i=1}^N \mathbb{I}(x_i \leq u) - \prod_{j=1}^d u_j \right|, \quad (10)$$

where $\mathbb{I}$ is the indicator function. This “worst-case” metric is sensitive to clustering, gaps, and boundary effects, and is particularly suited for evaluating quasi-uniform sampling strategies. In practice, numerical approximations of discrepancy (e.g., centered or star discrepancy) are computed using established libraries such as SciPy’s qmc module [69].

In contrast, the spatial entropy measures the distributional diversity of sample locations by quantifying how evenly the samples are spread across a partitioned parameter space. Given a decomposition into K disjoint subregions $\{R_k\}$, the spatial entropy is defined as:

$$H_{spatial} = -\sum_{k=1}^K p_k \log p_k, \quad (11)$$

where p_k is the proportion of samples falling within region R_k . This Shannon entropy [70] quantifies the diversity of the spatial sampling pattern, with higher values indicating more uniform coverage.

Together, discrepancy and spatial entropy offer a comprehensive view of sampling behavior: discrepancy detects local irregularities and deviations from uniformity, while entropy captures the global balance and diversity of sample allocation. Both metrics are agnostic to the criticality values of the samples and remain interpretable across different scenario selection strategies. However, their expected values and practical relevance may vary depending on whether the method prioritizes global exploration or targeted search, as in extremum-driven approaches.

The time cost metric approximates the practical computational expenses associated with the selection method, including the sampling process, algorithmic overhead, and possible training or fitting times. It reflects the computational feasibility of the sampling method for practical deployment. Depending on the application, time cost can be reported either as an absolute runtime for initialized benchmark instances or as an asymptotic complexity expression (Big- $\mathcal{O}$ notation) to characterize scalability behavior with respect to sample size or input dimensionality.

Beyond these general indicators, distinct search strategies introduce specific evaluation objectives that require tailored metrics. The following subsections specify evaluation criteria aligned with the methodological design of the corresponding search classes.

4.6.2. Global Explorative Coverage

For global exploration methods, such as LHS or Sobol sequences, performance is primarily assessed through the spatial entropy and the discrepancy metric, both of which were introduced and formally defined in Section 4.6.1. While spatial entropy captures the

global dispersion of samples, discrepancy quantifies their local uniformity. In this context, discrepancy serves as the main indicator of sampling quality, as uniform coverage is the primary objective of these methods.

4.6.3. Targeted Extremum Search

For extremum-oriented algorithms, including simulated annealing and evolutionary strategies, the primary evaluation focus lies on the identification of high-criticality regions. Performance is assessed via two key metrics. Convergence rate, defined as the number of samples required to exceed a predefined criticality threshold τ :

$$N_{\tau} = \min\{n \in \mathbb{N} \mid C(\mathbf{x}_n) \geq \tau\}, \quad (12)$$

Extremum gap, which quantifies the deviation between the highest observed criticality value and a reference optimum obtained through offline evaluations:

$$\Delta C = \max_{\mathbf{x} \in X_{\text{sel}}} C(\mathbf{x}) - C_{\text{opt}}^{\text{ref}}, \quad (13)$$

Additionally, across-run mean criticality and variance can quantify central tendency and dispersion under a fixed evaluation budget. Region detection and coverage metrics provide strong diagnostic value for extremum search. Scalability to higher dimensional spaces remains limited due to high sample complexity for reliable region masks and unfavorable computational complexity of clustering algorithms. Evaluation cost increases sharply with dimension, which restricts practical use to moderate dimensionality.

4.6.4. Surrogate-Based Reconstruction

Surrogate-based selection methods are assessed through the model approximation error, computed as the mean squared error (MSE) between surrogate predictions and ground-truth criticality values across a dense evaluation grid. In addition, the coefficient of determination (R^2) may be reported as a complementary metric to the MSE. While the MSE quantifies absolute approximation error, R^2 expresses the proportion of variance in criticality values that is explained by the surrogate model. This scale-independent measure facilitates the interpretation of model fit and supports comparisons across criticality landscapes with differing statistical properties.

To assess the ability of search-based methods to comprehensively discover multiple critical regions, Feng et al. [36] propose an F1-based coverage metric. Both the reference set (obtained via dense grid sampling) and the set of sampled scenarios are thresholded to generate binary masks of critical areas. A pixel-wise comparison yields true positives, false positives, and false negatives, from which the F1 score is computed as the harmonic mean of precision and recall. Since this metric reflects how completely critical regions are discovered, it is particularly suited for multimodal extremum search. However, it is important to note that the F1 score is computed on a per-point basis and does not explicitly verify whether all distinct critical regions have been discovered. Small, high-risk regions may be underrepresented or entirely missed without significantly affecting the global F1 score if larger regions dominate the sample distribution. While region-aware clustering could address this limitation, such methods are known to scale exponentially with dimensionality and sample size, limiting their practical applicability in high-dimensional benchmarks. As such, the F1 score remains a useful but inherently localized measure of search effectiveness.

Moreover, the spatial entropy of the selected samples serves as a proxy for the exploration behavior of the selection method, quantifying the distributional spread of sample locations within the parameter space. To assess the exploration–exploitation balance more comprehensively, additional composite metrics may be introduced. For example, by com-

binning and weighing spatial entropy with the average criticality of the selected samples, the method can be adapted to the proposed safety argumentation of the user.

4.6.5. Boundary Region Detection

The evaluation of boundary detection methods requires dedicated metrics that reflect the precision, efficiency, and completeness of the identified boundary regions. In contrast to conventional classification-based metrics, which rely on binary labels for individual samples, the following set of metrics provides a continuous and global assessment of sampling performance in high-dimensional boundary search tasks. The metrics are adapted from recent work on adaptive sampling for performance boundaries [11].

The precision metric $M_{\text{precision}}$ quantifies the proportion of selected samples that are correctly located within the true boundary regions. It reflects the sampling specificity and utilization efficiency:

$$M_{\text{prec}} = \frac{N_p}{N_{\text{all}}}, \quad (14)$$

where N_p is the number of samples located inside the true boundary region, and N_{all} denotes the total number of generated samples.

The effectiveness metric M_{eff} captures the sampling efficiency with respect to the total number of system queries. Since system evaluations are often costly, this metric provides a normalized indicator of query efficiency:

$$M_{\text{eff}} = \frac{N_p}{N_{\text{query}}}, \quad (15)$$

where N_{query} represents the total number of evaluations of the SUT. It is important to note that this metric is particularly meaningful for methods that select a reduced subset of critical scenarios intended for downstream use, e.g., in physical or high-fidelity simulations, where system evaluations are costly. In other approaches, the total number of model or simulation calls should be considered, and M_{acc} may serve as the more appropriate indicator of sampling efficiency.

The coverage metric M_{cov} assesses the extent to which the true boundary region has been covered by the sampled points. It reflects the completeness of boundary exploration:

$$M_{\text{cov}} = \frac{N_{t,\text{covered}}}{N_{t,\text{all}}}, \quad (16)$$

where $N_{t,\text{covered}}$ is the number of true boundary elements covered by the sampling, and $N_{t,\text{all}}$ denotes the total number of boundary reference points. It should be noted that the coverage percentage depends on both the resolution of the evaluation grid and the chosen proximity criterion for associating sampled points with true boundary elements (in cases where a fully grid-based selection is not applied).

4.7. Benchmark Structure

The proposed benchmark comprises the described generated ground-truth spaces that define the complete set of test cases. SBSSMs under evaluation can query this ground truth to obtain criticality information. Each query is recorded by the benchmark to enable transparent logging of the sampling process. These logged interactions serve as the basis for computing a metric suite tailored to the evaluation objectives of the respective SBSSM. If desired, a reference to the model can be supplied to permit the calculation of additional model-specific error terms, thereby facilitating direct assessment of prediction accuracy or model bias.

All benchmark modules are openly accessible [71] and designed for extensibility, allowing researchers to incorporate new criticality landscapes, evaluation metrics, or application domains as the field evolves.

5. Benchmark Application

This section demonstrates the application of the proposed benchmark using a small set of representative search-based scenario selection methods. The purpose of these experiments is not to conduct a comprehensive comparative study, but to illustrate the benchmark's functionality, the defined evaluation metrics, and the implemented baseline algorithms. To this end, results are reported for selected criticality spaces and parameter dimensions, illustrating the operation of the logging interface, the computation of performance metrics, and the benchmark's ability to discriminate between fundamentally different search behaviors. All source code and extended experimental results are provided in the accompanying repository [71] to facilitate independent reproduction and further analysis. The data-flow architecture of the framework, including the interface between the benchmark and the method under evaluation, is shown in Figure 7.

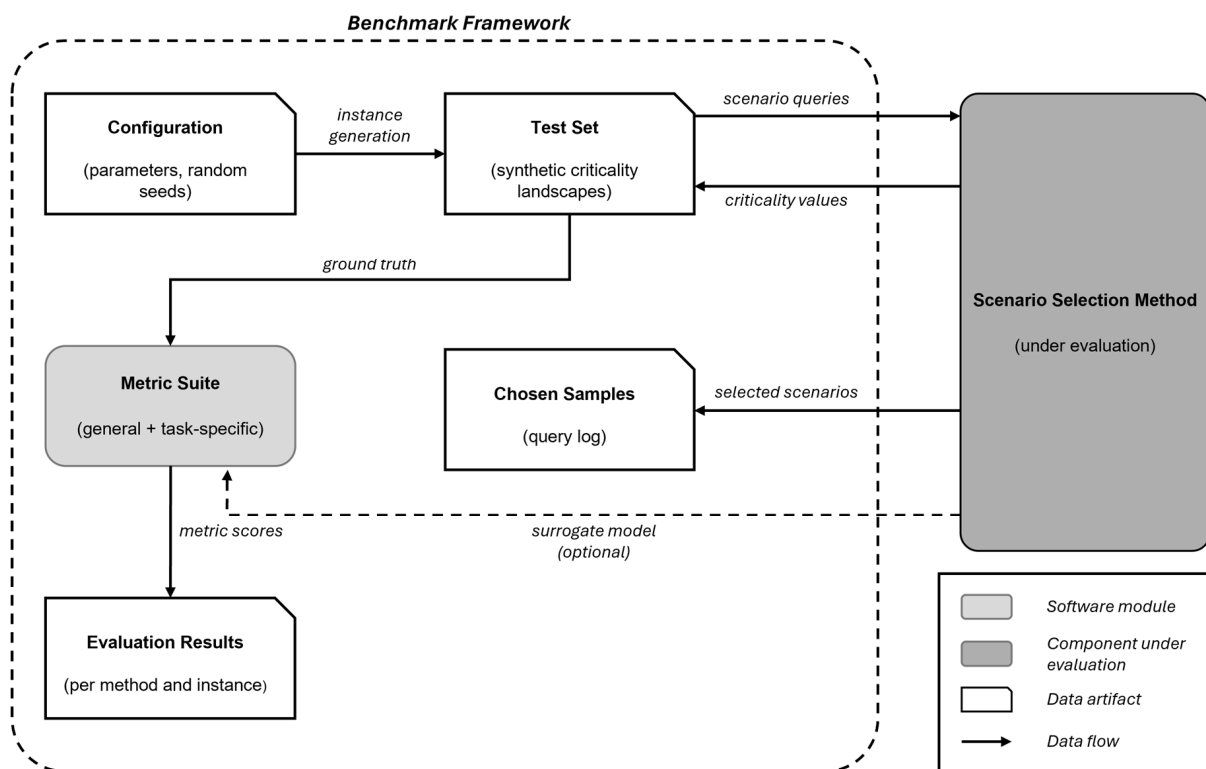


Figure 7. Data-flow architecture of the benchmark framework.

5.1. Evaluation Protocol

All algorithms are evaluated on ten distinct synthetic criticality spaces designed to represent heterogeneous boundary and density structures. Each method receives an identical evaluation budget, defined by the number of scenario queries, to ensure comparability. For every method-space combination, ten independent replications with varying random seeds are conducted to account for stochastic variability [71]. Performance values are reported as mean and standard deviation across replications. Evaluation metrics follow the definitions in Section 4.6, with subsets selected according to the methodological category (boundary identification, coverage, diversity). Initial points are generated uniformly across methods using LHS to avoid initialization bias.

The reference methods are selected purposively to represent the principal scenario-selection objectives introduced in Section 2.3 rather than to provide an exhaustive comparison of all available algorithms. Selection is guided by four criteria: methodological relevance to scenario-based validation, representation of distinct search objectives, sufficient implementation and parameter information to support reproducibility, and the possibility of conducting meaningful comparisons between related approaches. Random Sampling and Sobol sequences are included as established non-adaptive baselines for global coverage. For extremum search, IPSO is selected as a representative population-based heuristic that has been frequently referenced in critical-scenario search research and for which a sufficiently reproducible formulation is available. GPR-UCB is included as an established surrogate-assisted approach that explicitly balances exploration and exploitation. For boundary identification, NNDV and GDNNAS are selected because they address the same task using related but distinct sampling principles. GDNNAS was proposed as an alternative to NNDV and was previously compared with it, making their evaluation on the proposed benchmark particularly informative. For landscape reconstruction, IDW is chosen as a simple and widely applied interpolation-based surrogate, while the ANN represents a standard nonlinear data-driven model. Their inclusion enables a comparison between a transparent local interpolation method and a flexible learned approximation. Overall, the selected methods provide reproducible and methodologically diverse reference cases for demonstrating the benchmark's evaluation capabilities.

For the demonstration, the methods described in Section 2.3 are applied using the parameters in Table 1. The configuration follows a uniform rationale. For all methods introduced in prior work, parameter values are adopted directly from the respective original publications. Where the original publications do not report complete settings, the remaining parameters are calibrated such that the reference implementations reproduce the qualitative behavior and performance reported in those publications. For the ANN surrogate, for which no reference configuration exists, a small-scale hyperparameter and architecture search over layer width, depth, and learning rate was conducted, and the most robust configuration across preliminary runs was selected. We emphasize that none of the resulting settings is claimed to be optimal for the benchmark instances. Per-instance tuning is deliberately avoided for two reasons. First, the reference implementations are intended as reproducible baselines under a fixed, documented configuration rather than as best-effort competitors. Second, identical settings across all instances ensure that observed performance differences reflect method characteristics and landscape structure rather than differences in tuning effort. Consequently, benchmark users seeking a fair comparison should optimize the hyperparameters of all methods under consideration with respect to the selected evaluation metric, subject to an identical tuning budget.

Table 1. Overview of the SBSSMs considered in this study together with their core parameters.

Category	Method	Key Parameters	Parameter Origin
Global Coverage	Random	parameter-free	not applicable
	Sobol Sequence	parameter-free	not applicable
	Latin Hypercube	parameter-free	not applicable
Extremum Search	IPSO	particles $n_p = 10$; max. iterations $T_{\max} = 100$; inertia weight $w = 0.8$; cognitive coefficient $c_1 = 0.075$; social coefficient $c_2 = 0.075$; restart threshold = 1	adopted from [36]
	GPR-UCB	initial samples $n_{\text{init}} = n_{\text{select}}/5$; candidate pool size = 100 batch size = 5; confidence level = 0.997; exploration weight $\kappa = 2.0$; kernel: Matérn ($\nu = 2.5$) + white noise	adopted from [25], κ tuned to reproduce reported behavior

Table 1. *Cont.*

Category	Method	Key Parameters	Parameter Origin
Landscape Reconstruction	IDW	initial samples $n_{init} = n_{select}/5$; power $p = 5.0$; neighbors $k = 5$; candidates per iteration = 250; batch size = 5; exploitation weight $\beta = 1.0$	adopted from [40]
	ANN	MLP, 2 hidden layers with 64 units each; learning rate $lr = 0.001$; max. epochs = 20,000	small-scale hyperparameter and architecture search
Boundary Identification	NNDV	initial samples $n_{init} = n_{select}/5$; candidate pool size = 5000; batch size = 1; variance weight $g = 1.0$; density weight $v = 1.0$	adopted from [12]
	GDNNAS	initial samples $n_{init} = n_{select}/5$; gradient weight $g = 2$; distance weight $v = 1$; initial elite fraction $p_0 = 0.2$; sampling-density parameters $a = 240$, $b = 72$; elimination parameters $\omega_0 = 5$, $\omega_{min} = 1$; initial neighborhood radius $r_0 = R/10$	adopted from [11], tuned to reproduce published results

5.2. Global Explorative Coverage

As outlined in Section 2.3.1, global sampling strategies such as Random Sampling, LHS and Sobol sequences represent widely adopted baselines in scenario selection research. While these methods do not explicitly target interesting regions, they provide valuable reference baselines for assessing the geometric coverage quality of the parameter space. In this sense, LHS and Sobol approximate the best-case uniformity achievable without incorporating knowledge. However, their sampling efficiency in terms of capturing sparse, high-criticality regions is inherently limited, making them an important but insufficient benchmark when evaluating the advantages of adaptive search strategies.

Table 2 reports results for a 6D space with 1000 samples across discrepancy and spatial entropy. Consistent with theoretical expectations, Sobol sequences achieved the lowest discrepancy, thereby exhibiting the most uniform coverage of the parameter space. LHS demonstrated similarly favorable uniformity, albeit with slightly higher local clustering effects, while purely random draws displayed both reduced spatial entropy and higher variability between runs.

Table 2. Discrepancy and spatial entropy for 1000 samples in a 6D space. Lower discrepancy and higher spatial entropy indicate more uniform coverage.

	Spatial Entropy	Discrepancy
Random	2.7676 ± 0.0026	0.002594 ± 0.000495
LHS	2.7651 ± 0.0028	0.000733 ± 0.000061
Sobol	2.7725 ± 0.0	0.000075 ± 0.000001

Therefore, global coverage methods exhibit the expected behavior and serve as baseline references for spatial representativeness. Adaptive strategies focusing on high-criticality regions are not expected to exceed these baselines in uniformity metrics. Instead, coverage measures should be viewed as complementary and can be combined with extremum- and boundary-oriented metrics to form more comprehensive performance indicators.

5.3. Targeted Extremum Search

Extremum search targets the most critical points under a fixed evaluation budget. Gaussian process-based and improved particle swarm optimizations use multiple random or space filling initializations to increase the probability of reaching the global maximum. Single start methods such as simulated annealing depend strongly on the initial state. In sparse and high dimensional landscapes with fragmented basins of attraction dependence on the initial state reduces the likelihood of escaping local optima within the available budget and reduces the probability of reaching the global maximum. Multistart strategies

do not guarantee global optimality, and effectiveness depends on the restart policy the exploration-exploitation balance and the fidelity of the model specification.

Tables 3 and 4 show that GPR-UCB achieves the highest mean criticality in both the two-dimensional and four-dimensional benchmark settings. In the two-dimensional setting, GPR-UCB also requires fewer evaluations to reach the predefined convergence threshold than Random Sampling and IPSO. IPSO attains a maximum criticality comparable to GPR-UCB but produces a lower mean criticality, reflecting its broader exploration of the parameter space. In the four-dimensional setting, GPR-UCB is the only method that reaches the convergence threshold within the available budget, doing so in two of ten runs. Nevertheless, the reduced convergence frequency across all methods demonstrates the increased difficulty of identifying sparse critical regions as dimensionality increases.

Table 3. Criticality and Convergence Rate for 100 samples in a 2D space.

	Average Criticality	Average Maximum Criticality	Average Convergence Rate
Random	0.534 ± 0.0123	0.988 ± 0.0097	35.5
GPR	0.864 ± 0.0039	0.999 ± 0.0001	27.0
IPSO	0.757 ± 0.0878	0.998 ± 0.0020	32.1

Table 4. Criticality and Convergence Rate for 300 samples in a 4D space. Converged denotes the proportion of runs reaching the predefined threshold τ .

	Average Criticality	Average Maximum Criticality	Converged
Random	0.248 ± 0.0084	0.688 ± 0.0510	0/10
GPR	0.433 ± 0.0078	0.849 ± 0.0645	2/10
IPSO	0.374 ± 0.0601	0.712 ± 0.1067	0/10

Figures 8 and 9 illustrate the corresponding sampling behavior in a two-dimensional benchmark space. GPR-UCB progressively concentrates evaluations in regions of high predicted criticality, whereas IPSO distributes samples more broadly along particle trajectories. This difference reflects the stronger exploitation behavior of GPR-UCB and the comparatively broader exploration performed by IPSO.

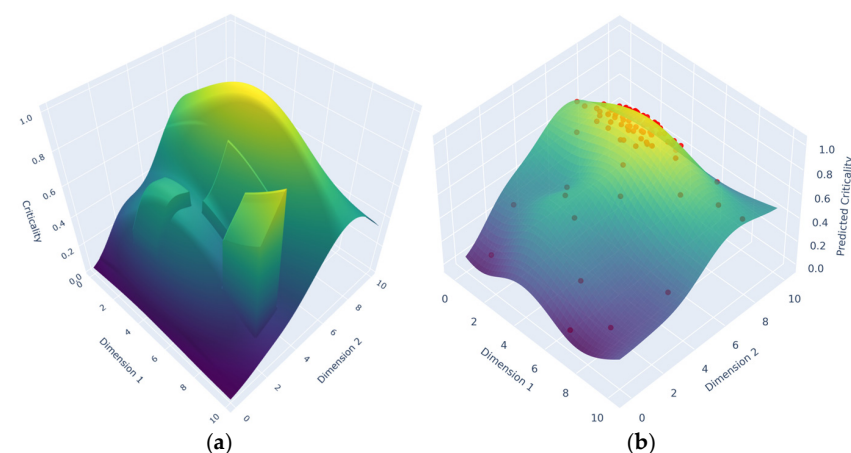


Figure 8. Ground-truth criticality landscape and observed sampling behavior of the Bayesian optimization method. (a) Three-dimensional surface representation of the ground-truth criticality landscape. (b) Observed sampling locations (red points) overlaid on the predicted criticality surface in the two-dimensional search space, demonstrating progressive accumulation in regions of high predicted criticality and convergence toward the most critical area.

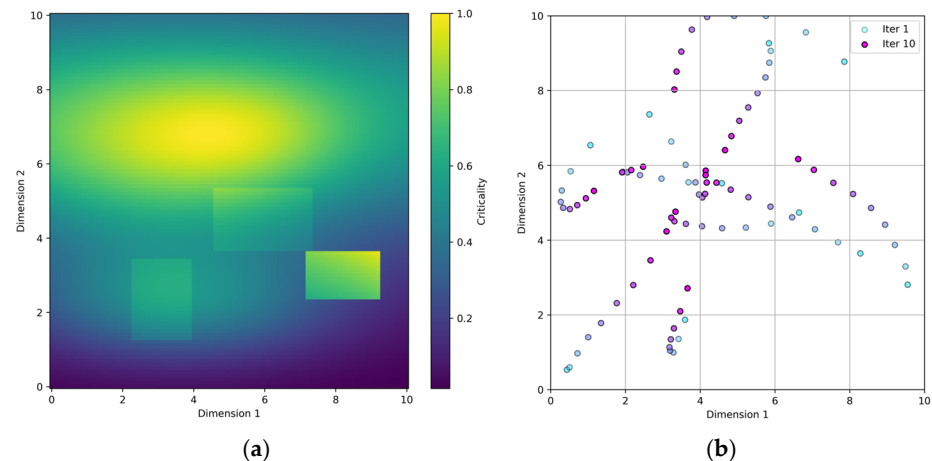


Figure 9. Ground-truth criticality landscape and observed sampling behavior of the IPSO method. (a) Top-down view of the ground-truth criticality landscape in the two-dimensional search space, with criticality values ranging from low (purple) to high (yellow). (b) IPSO sampling locations in the two-dimensional search space across iterations 1 to 10, showing progressive convergence toward regions associated with the best-identified neighbors and gradients.

5.4. Surrogate-Based Reconstruction

Surrogate modeling seeks to approximate the underlying criticality landscape based on a limited set of sampled data. Depending on the application objective, such models may prioritize the accurate reconstruction of high-risk regions or aim to capture the overall structure of the space. Performance is typically evaluated using regression metrics, such as MSE and R^2 , as well as classification-oriented measures that assess the correct identification of critical regions, for example F1 coverage at predefined thresholds. In this study, IDW and ANNs are employed as representative regression-based surrogates. More generally, classification models or hybrid approaches may also be applied. Surrogate methods are particularly advantageous in settings where system evaluations are computationally expensive and only a limited number of ground-truth samples can be obtained.

As reported in Tables 5 and 6, the relative performance of ANN and IDW depends on both the benchmark space and the dimensionality. In the two-dimensional experiments, neither method uniformly dominates: the ANN achieves better values across all three metrics on Space 1, whereas IDW performs better on Space 5. In the three-dimensional experiments, IDW outperforms the ANN on both instances and across all metrics. MSE and R^2 quantify reconstruction quality across the full continuous landscape, whereas F1 coverage evaluates the identification of critical regions after applying the predefined threshold of 0.6. These metrics can therefore diverge: sparse critical regions contribute relatively little to global regression metrics, while small prediction deviations near the threshold can substantially affect F1 coverage. This effect is most pronounced for the ANN on Space 1 in the three-dimensional setting, where a moderate global fit coincides with an F1 coverage close to zero, indicating that the critical regions are almost entirely missed despite reasonable overall regression accuracy. The threshold was fixed across methods and spaces to ensure comparability and was not optimized for either surrogate. Since only a limited hyperparameter and architecture search was conducted for the ANN, its results should be interpreted as illustrative rather than conclusive; the pronounced performance drop from the two- to the three-dimensional setting under a fixed configuration may reflect the sparser sampling relative to the volume of the space. IDW additionally offers lower computational complexity in the evaluated configuration.

Table 5. MSE, R^2 and F1 Coverage at a 0.6 threshold for 100 samples in a 2D space. Results are reported for IDW and ANN surrogates on two representative instances (Space 1 and Space 5).

		Average MSE	Average R^2	Average F1 Coverage @ 0.6
Space 1	IDW	0.0102 ± 0.0009	0.835 ± 0.014	0.895 ± 0.006
	ANN	0.0057 ± 0.0018	0.906 ± 0.029	0.924 ± 0.014
Space 5	IDW	0.0049 ± 0.0005	0.860 ± 0.015	0.849 ± 0.020
	ANN	0.0105 ± 0.0036	0.697 ± 0.104	0.688 ± 0.044

Table 6. MSE, R^2 and F1 Coverage at the 0.6 threshold for 200 samples in a 3D space. Results are reported for IDW and ANN surrogates on two representative instances (Space 1 and Space 5).

		Average MSE	Average R^2	Average F1 Coverage @ 0.6
Space 1	IDW	0.0052 ± 0.0003	0.743 ± 0.012	0.608 ± 0.063
	ANN	0.0086 ± 0.0010	0.579 ± 0.052	0.001 ± 0.003
Space 5	IDW	0.0060 ± 0.0005	0.830 ± 0.012	0.795 ± 0.010
	ANN	0.0240 ± 0.0086	0.320 ± 0.243	0.381 ± 0.104

5.5. Boundary Region Detection

Boundary region detection balances exploration of the parameter space with exploitation of detected boundary regions. Coverage quantifies the proportion of detected boundaries. Precision and effectiveness quantify the proportion of points located in the vicinity of boundary regions relative to all selected and evaluated samples.

Methods by Mullins et al. [12] (NNDV) and Lu et al. [11] (GDNNAS) were reimplemented from the published descriptions. Figure 10 illustrates the operational differences. Figure 10b displays only GDNNAS-selected points. Because GDNNAS filters sampled points further away from the boundaries after evaluation, the number of system queries exceeds the number of displayed points. This characteristic is reflected in the effectiveness metric.

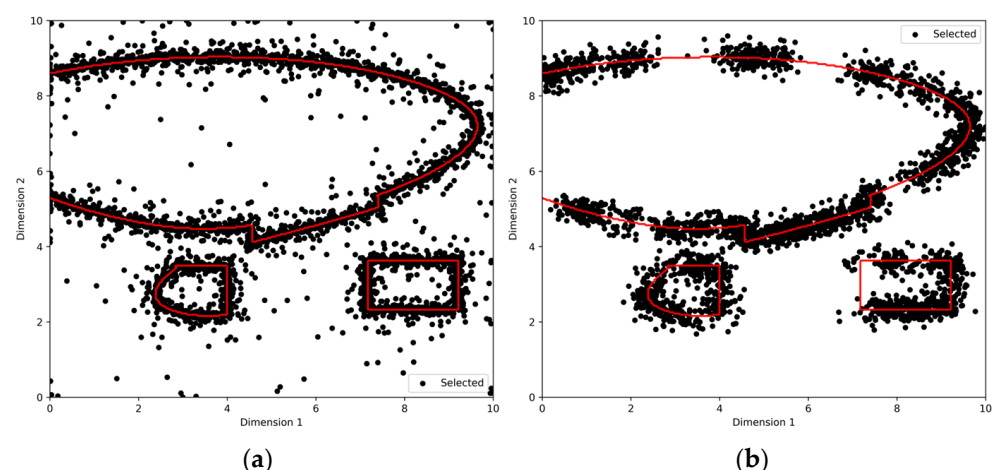


Figure 10. Comparison of sampling behavior between NNDV and GDNNAS in a two-dimensional search space. (a) NNDV sampling locations (black points) with boundary (red lines), demonstrating broad domain coverage while concentrating samples along boundary regions. (b) GDNNAS locations (black points) with boundary (red lines), producing compact sets of boundary samples through iterative refinement and exclusion of points beyond the threshold.

Results indicate that NNDV surpasses the reference values reported by Lu et al. [11] on the corresponding test cases. GDNNAS achieves high precision through filtering, while effectiveness and coverage depend strongly on the number of initial samples. Larger initial sets increase coverage but reduce effectiveness drastically due to higher discard rates. NNDV is less sensitive to initialization due to its exploratory breadth, which improves coverage but may still fail to detect regions without initial samples. Performance differences in two- and three-dimensional spaces are shown in Figures 11 and 12. For both dimensions the number of selected samples was fixed at 3000. The number of initial samples was aligned with the metric ratio reported in Lu et al.'s publication and retained for the higher-dimensional case.

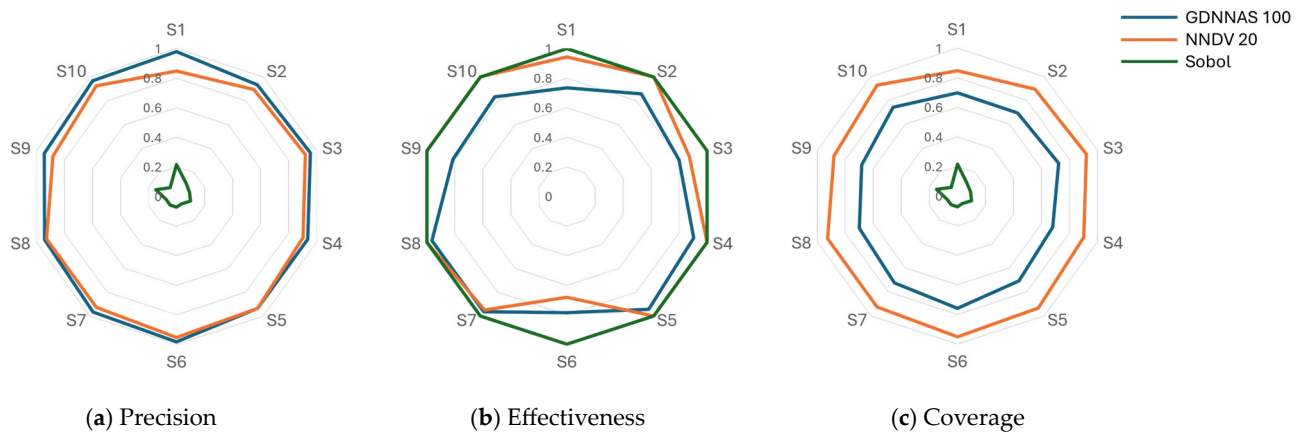


Figure 11. Evaluation metrics comparing NNDV and GDNNAS for 20 and 100 initial samples with 3000 selected samples across 10 two-dimensional spaces. Axes S1–S10 denote the ten evaluated criticality spaces. Sobol achieves broad boundary coverage but lower precision and effectiveness, while GDNNAS yields the highest precision and NNDV balances effectiveness with boundary coverage.

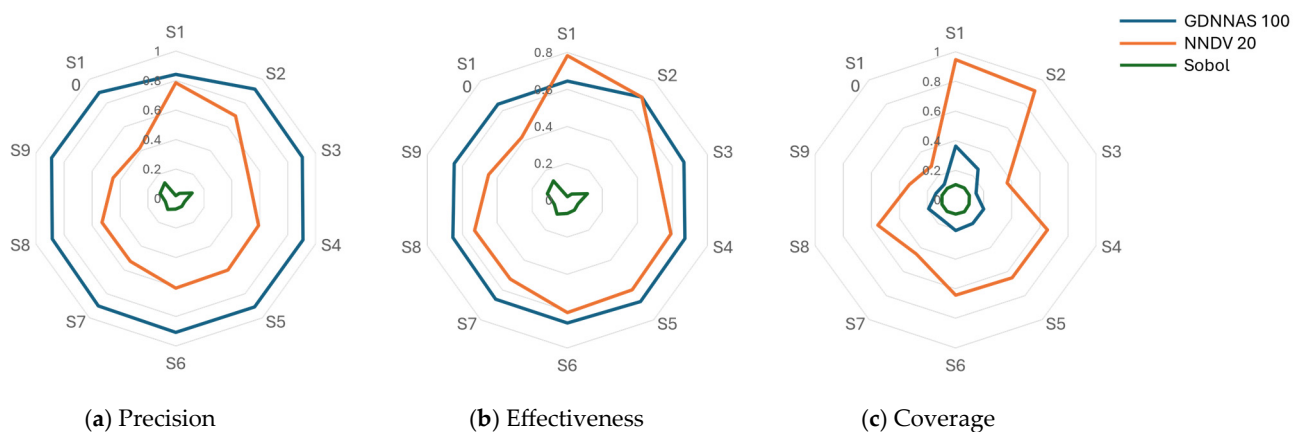


Figure 12. Evaluation metrics comparing NNDV and GDNNAS for 20 and 100 initial samples with 3000 selected samples across 10 three-dimensional spaces. Axes S1–S10 denote the ten evaluated criticality spaces. GDNNAS tends to achieve higher precision and effectiveness, whereas NNDV provides stronger boundary coverage. Sobol performs consistently lower across all metrics.

5.6. External-Validity Check: Real vs. Synthetic Method Rankings

The preceding sections demonstrated the benchmark on synthetic landscapes only. To assess whether conclusions drawn from these landscapes transfer to empirical criticality spaces, the head-on scenario from the DLR simulation (Section 3.1) was evaluated as an additional ground-truth instance. The full factorial simulation grid over the four parameters own-ship speed, target-ship speed, and the maneuverability of both vessels was embedded

as a queryable criticality space via multilinear interpolation and rescaled to the nominal benchmark domain, so that all selectors and metrics operate on it unmodified. All nine reference methods were then executed on the real space and on ten synthetic 4D instances under the protocol of Section 5.1, and the resulting method rankings per metric were compared using Kendall's rank correlation. Rank correlation is appropriate here because metric values are not comparable in magnitude across landscapes; only the induced method ordering is.

Figure 13 shows the agreement between the ranking obtained on the real space and the ranking obtained on each individual synthetic space. For the discrepancy metric, computed over all nine methods, the rankings agree strongly (mean $\tau = 0.85 \pm 0.07$). The residual disagreement corresponds to occasional swaps between adjacent methods with nearly identical discrepancy values, while the overall ordering of coverage-oriented and adaptive methods is preserved on every instance. The category-specific metrics compare two methods each, so τ is restricted to ± 1 per space; agreement across the ten spaces is therefore assessed jointly. The model reconstruction metrics (MSE, R^2 , F1 coverage for IDW vs. ANN) and the average maximum criticality (GPR vs. IPSO) reproduce the real-space ordering on 10 of 10 synthetic instances (sign test, $p \approx 0.002$), and the boundary detection metrics (effectiveness and precision for NNDV vs. GDNNAS) on 9 of 10 instances ($p \approx 0.021$). The average selected criticality is the only metric whose ordering does not transfer consistently ($\tau = +1$ on six, $\tau = -1$ on four instances). This instability is present among the synthetic instances themselves and reflects the landscape dependence of mean criticality already observed in Section 5.3 rather than a systematic real–synthetic divergence.

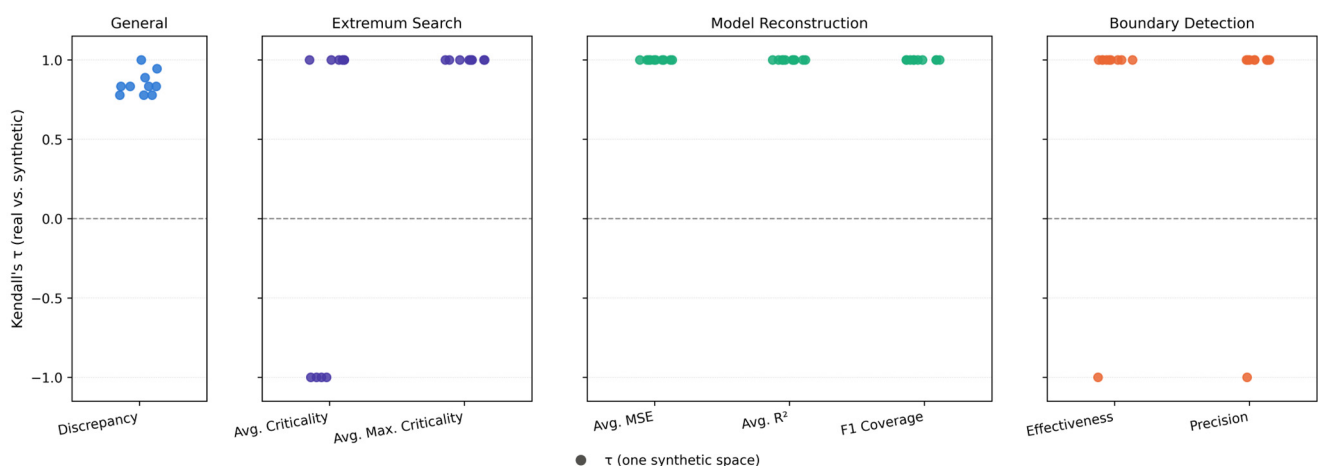


Figure 13. Kendall's τ between the method ranking on the real DLR criticality space and the ranking on each of the ten synthetic 4D spaces, grouped by metric category. Dots denote individual synthetic spaces, diamonds the mean $\pm$ standard deviation across spaces.

These results provide quantitative evidence complementing the qualitative structural argument of Section 3: for the metrics that discriminate robustly between methods, rankings obtained on the synthetic benchmark are predictive of rankings on an empirical simulation-derived criticality space. The scope of this validation is discussed in Section 6.3.

6. Discussion

This section provides a comprehensive discussion of the proposed benchmark. It evaluates the framework against the functional and methodological requirements established in the preceding sections, examines its current limitations, and outlines avenues for future development. Rather than comparing the performance of individual SBSSMs,

the analysis focuses on the benchmark itself as the basis for systematic evaluation of scenario-selection methods.

6.1. Alignment with Benchmark Requirements

This section summarizes how the proposed benchmark addresses the requirements defined in Section 1.

BR1 (Reproducibility). Reproducibility is ensured through open-source reference implementations of all benchmark components, including the criticality-function generators, configuration files, evaluation scripts, and a getting-started guide [71]. Transparent parameter settings are provided via the published configuration files, and the evaluation protocol is specified by the accompanying execution and scoring scripts. In addition, deterministic random-seed handling is enforced to support exact replication of experiments.

BR2 (Cross-domain comparability). Cross-domain comparability is achieved by employing analytically defined criticality landscapes that are independent of simulator internals, proprietary toolchains, and application semantics. This abstraction enables evaluation of search-based scenario selection methods under identical conditions without embedding domain-specific constraints, thereby facilitating transfer of methodological insights across operational domains.

BR3 (Structural realism). Structural realism is maintained by designing synthetic landscapes to reproduce phenomena reported in empirical validation studies and simulation-based testing practice. As a result, the benchmark reflects the topological complexity characteristic of real criticality spaces, rather than relying on overly smooth or simplistic objective functions.

In addition, Sections 4.5 and 5.6 provide quantitative support: the structural descriptors of the synthetic instances cover the range observed in the empirical reference surfaces, and method rankings obtained on the synthetic instances agree with rankings obtained on an empirical simulation-derived criticality space for nearly all evaluation metrics, indicating that the synthetic landscapes preserve the performance-relevant structure of real criticality spaces.

BR4 (Controllability and determinism). The framework guarantees controllability and determinism by making benchmark-instance generation fully deterministic given the published configuration files and random seeds. This permits repeated execution on identical instances. For inherently stochastic scenario-selection methods, repeated runs enable statistically meaningful estimates of average performance and variability.

BR5 (Structural diversity). Structural diversity is supported through a two-stage procedure consisting of randomized generation of candidate landscapes followed by a filtering step. Diversity is quantified using the proposed formulation based on DFT-derived features together with complementary metrics, and benchmark instances are selected to represent the identified clusters within the candidate pool. This procedure increases coverage of varied topologies and reduces the risk of over-specialization to a narrow landscape class.

BR6 (Extensibility). Extensibility is enabled by a modular software architecture that exposes explicit interfaces for integrating new criticality functions, sampling schemes, and evaluation metrics. This design allows the benchmark to evolve with advances in scenario-selection research and to incorporate emerging empirical findings on criticality-space structure.

BR7 (Task-specific metrics). Task-specific evaluation is supported by allowing the inclusion of application- and objective-oriented performance measures in addition to general-purpose metrics. Examples include surrogate-model prediction error for model-

based methods or coverage measures for exploration-oriented strategies, thereby reflecting the distinct goals of heterogeneous selection approaches.

6.2. Practical Implications

The benchmark can support practitioners in selecting scenario selection methods before integrating them into computationally expensive simulation environments. By comparing candidate methods under standardized evaluation budgets and structurally diverse conditions, developers can identify whether a method is suitable for critical-region detection, boundary identification, or landscape reconstruction. This preliminary assessment can reduce implementation effort and simulation cost while improving the transparency of method selection. The benchmark does not replace domain-specific validation, but provides a reproducible screening and qualification stage before deployment in high-fidelity simulation, software-in-the-loop, or hardware-in-the-loop testing.

6.3. Limitations

The approach remains subject to several important constraints. First, an abstraction gap persists between the analytic benchmark and high-fidelity simulators or real-world trials: physical sensor noise, closed-loop controller dynamics, and hardware latencies are modeled only through simple noise schemes, so performance observed here may not extrapolate linearly to operational environments. The predictive-validity analysis in Section 5.6 partially narrows this gap but is itself limited: it rests on a single empirical criticality space from one domain, resolved at eight factorial levels per axis with linear interpolation in between, and the category-specific rank comparisons involve only two methods each, restricting their individual statistical power. Extending this validation to additional empirical spaces from other domains remains future work. The structural comparison in Section 4.5 rests on a small and heterogeneous empirical reference set, resolved at eight factorial levels per axis; higher-resolution empirical criticality spaces from additional domains would be required to turn the descriptor comparison into a statistically grounded validation. Second, empirical knowledge of large, highly coupled real-world scenario spaces is still sparse. The synthetic landscapes draw on limited in-house simulations and a small body of literature, which may not capture all behavioral patterns emerging in more complex, large-scale systems. Third, deterministic enrichment of the criticality landscapes is practical only in low-to-moderate dimensions. Under the multidimensional Nyquist–Shannon sampling criterion, the number of samples required to reproduce a given spatial frequency grows exponentially with dimensionality. In higher dimensions, the benchmark can therefore introduce fine-grained patterns only through stochastic function components, which limits the precision with which specific topological features can be prescribed. Finally, despite deliberate diversity, a residual structural bias remains because the library of function primitives cannot span the full topology that future autonomous-system domains might reveal. In addition, the experiments in Section 5 are deliberately illustrative in scope, and broader validation across additional benchmark instances, dimensionalities, algorithms, and evaluation budgets remains necessary. Moreover, the benchmark characterizes methods on synthetic structures; when selecting or comparing algorithms for a dedicated application, its results should therefore be complemented by domain-specific evidence, such as evaluations on representative application scenarios.

6.4. Future Work

Further development could advance along several complementary axes. Benchmark evolution could transition the present static, single-shot design toward dynamic, multi-stage scenarios and introduce heteroscedastic noise as well as low-rank pattern generators, enabling fine-grained structures in dimensions beyond ten without prohibitive sampling

requirements. Regular benchmark updates could incorporate newly observed criticality motifs from additional domains, thereby maintaining representative coverage. Moreover, foundation-model-based scenario generation, including large-language-model- and diffusion-model-driven synthesis of safety-critical scenarios [72,73], constitutes a rapidly evolving method class whose selection behavior could be evaluated within the proposed framework once suitable query interfaces are defined. On the search-based method side, hybrid human-in-the-loop pipelines and realistic uncertainty-aware acquisition functions could reconcile expert priors with probabilistic risk assessment, whereas statistically grounded stopping criteria could align search effort with tight evaluation budgets. Regarding evaluation, the current scalar criticality score could be generalized to multi-objective formulations that illuminate trade-offs, such as safety versus efficiency, and to metrics that penalize undue exploration of demonstrably irrelevant dimensions. Furthermore, adaptive stopping criteria and optimizations across multiple spaces with a limited test budget could be tested on the framework. Finally, external validation could couple the analytic framework to high-fidelity simulators in order to quantify discrepancies in scenario criticality rankings arising from differences in model fidelity, and selected algorithms could be deployed in hardware- or software-in-the-loop environments to verify failure-exposure rates under realistic latencies. Establishing a public leaderboard and continuous-integration pipeline could support these initiatives, promoting transparent comparison and facilitating broader adoption within the research community.

7. Conclusions

This work introduces a reproducible and domain-agnostic benchmark for the evaluation of SBSSMs, addressing a methodological gap in the validation of autonomous systems where cross-domain comparability has remained limited. It responds to a concrete need in the validation of MASS, where simulation-based safety evidence depends on the systematic and comparable selection of critical encounter scenarios. By abstracting from application-specific semantics and focusing instead on structural characteristics, its synthetic criticality landscapes emulate the complexity of real operational environments while remaining analytically tractable and extensible. The standardized evaluation metrics and logging facilities enable a systematic assessment of heterogeneous algorithmic strategies under controlled and repeatable conditions. Quantitative analyses indicate that the synthetic instances span the structural range of the empirical reference surfaces and that method rankings obtained on the benchmark largely agree with those on an empirical simulation-derived criticality space. Taken together, these contributions address the benchmark requirements BR1–BR7 established in Section 1, as substantiated by the alignment analysis in Section 6.1.

Importantly, the benchmark is not intended as an isolated academic artifact but as a bidirectional bridge between concrete use cases and methodological development. Insights gained from a specific application domain, for example, maritime autonomy or automotive safety, can inform the design of new benchmark instances by contributing empirically observed characteristic criticality structures, selection methods, or aggregation schemes. Conversely, algorithms refined and validated within the benchmark can be transferred back to the original domain with evidence-based expectations regarding their relative suitability. This interplay allows practitioners to iterate between real-world observations and synthetic testing, using the benchmark as a proving ground for new methods and as a controlled environment for exploring operating conditions beyond those directly encountered in their specific use case. In practical validation workflows, the benchmark can thus serve developers and validation authorities as an initial method-selection stage before costly domain-specific testing. In the maritime domain specifically, this supports the validation of MASS by enabling reproducible comparison of scenario selection methods

for COLREG-compliant collision avoidance and multi-vessel navigation, in line with the simulation-based evidence emphasized by the recently adopted IMO MASS Code and classification society guidance [2,28,29]. While transferability beyond the maritime domain is structurally motivated by criticality landscapes reported in automotive studies [26,33], predictive validity has so far been demonstrated only for a maritime criticality space.

By enabling systematic, cross-method comparisons under controllable conditions, the benchmark lays a foundation for evidence-based selection and adaptation of SBSSMs across diverse autonomous system validation applications and constitutes a step toward standardized benchmarking in this field.

Author Contributions: Conceptualization, D.A.; Methodology, D.A.; Investigation, D.A.; Visualization D.A.; Software, D.A. and D.P.; Writing—Original Draft, D.A. and D.P.; Writing—Review and Editing, D.A., D.P., J.M. and A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted within the project “Validation of Maritime AI” and partially funded by the Federal Ministry of Economic Affairs and Energy under the grant 03SX587.

Data Availability Statement: The source code and benchmark data supporting this study are openly available on Zenodo at <https://doi.org/10.5281/zenodo.21441489> [71].

Acknowledgments: The authors gratefully acknowledge the administrative support provided by Project Management Jülich (PTJ).

Conflicts of Interest: Author Daniel Autenrieth is employed by TKMS ATLAS ELEKTRONIK GmbH and Author Adrian Zlocki is employed by fka GmbH. Daniel Paland and Jan Mentjes are employed at the german aerospace center (Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)). Jan Mentjes is a founder of the Marivation GmbH.

References

1. Amersbach, C.; Winner, H. Defining Required and Feasible Test Coverage for Scenario-Based Validation of Highly Automated Vehicles. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*; IEEE: Piscataway, NJ, USA, 2019; pp. 425–430. [CrossRef]
2. International Maritime Organization. *International Code of Safety for Maritime Autonomous Surface Ships (MASS Code): Resolution MSC.595(111)*; IMO: London, UK, 2026.
3. Riedmaier, S.; Ponn, T.; Ludwig, D.; Schick, B.; Diermeyer, F. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access* **2020**, *8*, 87456–87477. [CrossRef]
4. Tang, S.; Zhang, Z.; Zhang, Y.; Zhou, J.; Guo, Y.; Liu, S.; Guo, S.; Li, Y.-F.; Ma, L.; Xue, Y.; et al. A Survey on Automated Driving System Testing: Landscapes and Trends. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 124. [CrossRef]
5. Hake, G.; Austel, A.; Becker, J.S.; Putze, L.; Wetzig, N. Safety Assessment of Maritime Autonomous Surface Ships: A Scenario-Based Approach. *J. Phys. Conf. Ser.* **2025**, *3123*, 012022. [CrossRef]
6. Kramer, B.; Neurohr, C.; Büker, M.; Böde, E.; Fränzle, M.; Damm, W. Identification and Quantification of Hazardous Scenarios for Automated Driving. In *Model-Based Safety and Assessment, Proceedings of the 7th International Symposium, IMBSA 2020, Lisbon, Portugal, 14–16 September 2020*; Zeller, M., Höfig, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 163–178, ISBN 978-3-030-58920-2. [CrossRef] [PubMed]
7. Schütt, B.; Ransiek, J.; Braun, T.; Sax, E. 1001 Ways of Scenario Generation for Testing of Self-driving Cars: A Survey. In *2023 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8. [CrossRef]
8. Gangopadhyay, B.; Khastgir, S.; Dey, S.; Dasgupta, P.; Montana, G.; Jennings, P. Identification of Test Cases for Automated Driving Systems Using Bayesian Optimization. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1961–1967. [CrossRef]
9. Corso, A.; Moss, R.; Koren, M.; Lee, R.; Kochenderfer, M. A Survey of Algorithms for Black-Box Safety Validation of Cyber-Physical Systems. *J. Artif. Intell. Res.* **2021**, *72*, 377–428. [CrossRef]
10. Cai, J.; Deng, W.; Wang, Y.; Guang, H.; Li, J.; Ding, J. Boundary Scenario Generation for HAVs Based on Classification and Local Sampling. *Machines* **2023**, *11*, 426. [CrossRef]
11. Lu, H.; Wang, S.; Zhang, Y.; Cheng, S. Sampling-based test scenarios searching method for autonomous system robustness evaluation. *Appl. Soft Comput.* **2024**, *162*, 111808. [CrossRef]

12. Mullins, G.E.; Stankiewicz, P.G.; Hawthorne, R.C.; Gupta, S.K. Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles. *J. Syst. Softw.* **2018**, *137*, 197–215. [\[CrossRef\]](#)
13. Brinkmann, M.; Bode, E.; Lamm, A.; Maelen, S.V.; Hahn, A. Learning from automotive: Testing maritime assistance systems up to autonomous vessels. In *OCEANS 2017—Aberdeen*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8. [\[CrossRef\]](#)
14. Perabo, F.; Park, D.; Zadeh, M.K.; Smogeli, O.; Jamt, L. Digital Twin Modelling of Ship Power and Propulsion Systems: Application of the Open Simulation Platform (OSP). In *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*; IEEE: Piscataway, NJ, USA, 2020. [\[CrossRef\]](#)
15. Ding, W.; Xu, C.; Arief, M.; Lin, H.; Li, B.; Zhao, D. A Survey on Safety-Critical Driving Scenario Generation—A Methodological Perspective. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 6971–6988. [\[CrossRef\]](#)
16. Westhofen, L.; Neurohr, C.; Koopmann, T.; Butz, M.; Schütt, B.; Utesch, F.; Neurohr, B.; Gutenkunst, C.; Böde, E. Criticality Metrics for Automated Driving: A Review and Suitability Analysis of the State of the Art. *Arch. Comput. Methods Eng.* **2023**, *30*, 1–35. [\[CrossRef\]](#)
17. Cai, J.; Deng, W.; Guang, H.; Wang, Y.; Li, J.; Ding, J. A Survey on Data-Driven Scenario Generation for Automated Vehicle Testing. *Machines* **2022**, *10*, 1101. [\[CrossRef\]](#)
18. Best, A.; Narang, S.; Pasqualin, L.; Barber, D.; Manocha, D. AutonoVi-Sim: Autonomous Vehicle Simulation Platform with Weather, Sensing, and Traffic Control. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2018; pp. 1048–1056. [\[CrossRef\]](#)
19. Araujo, H.; Mousavi, M.R.; Varshosaz, M. Testing, Validation, and Verification of Robotic and Autonomous Systems: A Systematic Review. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 51. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Porres, I.; Azimi, S.; Lilius, J. Scenario-based Testing of a Ship Collision Avoidance System. In *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*; IEEE: Piscataway, NJ, USA, 2020; pp. 545–552. [\[CrossRef\]](#)
21. Goss, Q.; Akbas, M.I. Eagle Strategy with Local Search for Scenario Based Validation of Autonomous Vehicles. In *2022 International Conference on Connected Vehicle and Expo (ICCVE)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [\[CrossRef\]](#)
22. Menzel, T.; Bagschik, G.; Maurer, M. Scenarios for Development, Test and Validation of Automated Vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: Piscataway, NJ, USA, 2018. [\[CrossRef\]](#)
23. Sun, J.; Zhang, H.; Zhou, H.; Yu, R.; Tian, Y. Scenario-Based Test Automation for Highly Automated Vehicles: A Review and Paving the Way for Systematic Safety Assurance. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 14088–14103. [\[CrossRef\]](#)
24. Althoff, M.; Koschi, M.; Manzing, S. CommonRoad: Composible benchmarks for motion planning on roads. In *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 719–726, ISBN 978-1-5090-4804-5. [\[CrossRef\]](#)
25. Torben, T.R.; Glomsrud, J.A.; Pedersen, T.A.; Utne, I.B.; Sørensen, A.J. Automatic simulation-based testing of autonomous ships using Gaussian processes and temporal logic. *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **2023**, *237*, 293–313. [\[CrossRef\]](#)
26. Gladisch, C.; Heinz, T.; Heinzemann, C.; Oehlerking, J.; von Vietinghoff, A.; Pfitzer, T. Experience Paper: Search-Based Testing in Automated Driving Control Applications. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*; IEEE: Piscataway, NJ, USA, 2019; pp. 26–37. [\[CrossRef\]](#)
27. Namgung, H.; Kim, J.-S.; Jang, D.-U. Path planning and collision avoidance technologies for maritime autonomous surface ships: A review of COLREGs compliance, algorithmic trends and the navigation-GPT framework. *J. Navig.* **2026**, *78*, 430–450. [\[CrossRef\]](#)
28. DNV. *Autonomous and Remotely Operated Ships*; Class Guideline DNV-CG-0264; DNV: Høvik, Norway, 2021.
29. DNV. *Assurance of Simulation Models*; Recommended Practice DNV-RP-0513; DNV: Høvik, Norway, 2021.
30. Junietz, P.M. *Microscopic and Macroscopic Risk Metrics for the Safety Validation of Automated Driving*; Technische Universität Darmstadt: Darmstadt, Germany, 2019.
31. Schutt, B.; Steimle, M.; Kramer, B.; Behnecke, D.; Sax, E. A Taxonomy for Quality in Simulation-Based Development and Testing of Automated Driving Systems. *IEEE Access* **2022**, *10*, 18631–18644. [\[CrossRef\]](#)
32. Hayward, J.C. *Near Miss Determination Through Use of a Scale of Danger*; Pennsylvania Transportation and Traffic Safety Center: University Park, PA, USA, 1972.
33. Pegasus Method Documentation. Available online: <https://www.pegasusprojekt.de/en/pegasus-method> (accessed on 13 August 2025).
34. McKay, M.D.; Beckman, R.J.; Conover, W.J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **2000**, *42*, 55–61. [\[CrossRef\]](#)
35. Sobol', I.M.; Asotsky, D.; Kreinin, A.; Kucherenko, S. Construction and Comparison of High-Dimensional Sobol' Generators. *Wilmott* **2011**, *2011*, 64–79. [\[CrossRef\]](#)
36. Feng, T.; Liu, L.; Xing, X.; Chen, J. Multimodal critical-scenarios search method for test of autonomous vehicles. *J. Intell. Connect. Veh.* **2022**, *5*, 167–176. [\[CrossRef\]](#)
37. Song, Q.; Tan, K.; Runeson, P.; Persson, S. Critical scenario identification for realistic testing of autonomous driving systems. *Softw. Qual. J.* **2023**, *31*, 441–469. [\[CrossRef\]](#)

38. Feng, S.; Sun, H.; Yan, X.; Zhu, H.; Zou, Z.; Shen, S.; Liu, H.X. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **2023**, *615*, 620–627. [[CrossRef](#)] [[PubMed](#)]
39. Yang, J.; Wang, Z.; Wang, D.; Zhang, Y.; Lu, Q.; Feng, S. Adaptive Safety Performance Testing for Autonomous Vehicles with Adaptive Importance Sampling. *Transp. Res. Part C Emerg. Technol.* **2024**, *179*, 105256. [[CrossRef](#)]
40. Zhang, H.; Sun, J.; Tian, Y. Accelerated Safety Testing for Highly Automated Vehicles: Application and Capability Comparison of Surrogate Models. *IEEE Trans. Intell. Veh.* **2024**, *9*, 2409–2418. [[CrossRef](#)]
41. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [[CrossRef](#)]
42. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. Available online: <https://jmlr.org/papers/v15/srivastava14a.html> (accessed on 19 July 2026).
43. Liu, S.; Wang, H.; Peng, W.; Yao, W. Surrogate-assisted evolutionary algorithms for expensive combinatorial optimization: A survey. *Complex Intell. Syst.* **2024**, *10*, 5933–5949. [[CrossRef](#)]
44. Gorissen, D.; Dhaene, T.; de Turck, F. Evolutionary model type selection for global surrogate modeling. *J. Mach. Learn. Res.* **2009**, *10*, 2039–2078.
45. Batsch, F.; Daneshkhah, A.; Palade, V.; Cheah, M. Scenario Optimisation and Sensitivity Analysis for Safe Automated Driving Using Gaussian Processes. *Appl. Sci.* **2021**, *11*, 775. [[CrossRef](#)]
46. Xu, C.; Ding, W.; Lyu, W.; Liu, Z.; Wang, S.; He, Y.; Hu, H.; Zhao, D.; Li, B. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25667–25682. [[CrossRef](#)]
47. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16.
48. Sorokin, L.; Munaro, T.; Safin, D.; Liao, B.H.-C.; Molin, A. OpenSBT: A Modular Framework for Search-based Testing of Automated Driving Systems. In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, New York, NY, USA; ACM: New York, NY, USA, 2024; pp. 94–98. [[CrossRef](#)]
49. Sorokin, L.; Safin, D.; Nejati, S. Can search-based testing with pareto optimization effectively cover failure-revealing test inputs? *Empir. Softw. Eng.* **2025**, *30*, 26. [[CrossRef](#)]
50. Hansen, N.; Auger, A.; Ros, R.; Mersmann, O.; Tušar, T.; Brockhoff, D. COCO: A platform for comparing continuous optimizers in a black-box setting. *Optim. Methods Softw.* **2021**, *36*, 114–144. [[CrossRef](#)]
51. Dreczkowski, K.; Grosnit, A.; Ammar, H.B. Framework and benchmarks for combinatorial and mixed-variable Bayesian optimization. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 69464–69489. [[CrossRef](#)]
52. Moreau, T.; Massias, M.; Gramfort, A.; Ablin, P.; Bannier, P.-A.; Charlier, B.; Dagréou, M.; La Tour, T.D.; Durif, G.; Dantas, C.F.; et al. Benchopt: Reproducible, efficient and collaborative optimization benchmarks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25404–25421. [[CrossRef](#)]
53. Khandait, T.; Formica, F.; Arcaini, P.; Chotaliya, S.; Fainekos, G.; Hekal, A.; Kundu, A.; Lew, E.; Loreti, M.; Menghi, C.; et al. Arch-comp 2024 category report: Falsification. In *Proceedings of the 11th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH24)*; EasyChair: Stockport, UK, 2024; pp. 122–144.
54. Khandait, T.; Lyu, D.; Arcaini, P.; Fainekos, G.; Formica, F.; Gon, S.; Hekal, A.; Kundu, A.; Menghi, C.; Pedrielli, G.; et al. ARCH-COMP 2025 Category Report: Falsification. In *Proceedings of 12th Int. Workshop on Applied Verification for Continuous and Hybrid Systems*; EasyChair: Stockport, UK, 2025; pp. 169–189.
55. Pedrielli, G.; Khandait, T.; Cao, Y.; Thibeault, Q.; Huang, H.; Castillo-Effen, M.; Fainekos, G. Part-X: A Family of Stochastic Algorithms for Search-Based Test Generation with Probabilistic Guarantees. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 4504–4525. [[CrossRef](#)]
56. TKMS AG & Co. KGaA. ATLAS ELEKTRONIK. Available online: <https://www.tkmsgroup.com/atlas-elektronik> (accessed on 2 February 2026).
57. Lee, D.-H.; Seo, M.-Y.; Lee, H.-G.; Yoo, W.-J.; Kim, S.; Paik, K.-J. A Study on Multi-ship Avoidance System for Unmanned Surface Vehicles Using the Quaternion Ship Domain and Collision Risk Index. *J. Ocean Eng. Technol.* **2025**, *39*, 103–121. [[CrossRef](#)]
58. Steidel, M.; Mentjes, J.; Hahn, A. Context-Sensitive Prediction of Vessel Behavior. *J. Mar. Sci. Eng.* **2020**, *8*, 987. [[CrossRef](#)]
59. Deutsches Zentrum für Luft- und Raumfahrt. eMIR—e-Maritime Integrated Reference Platform. Available online: <https://www.dlr.de/en/se/research-transfer/projects/emir-research-projects> (accessed on 20 July 2026).
60. Hahn, A.; Thoralf, N. *EMaritime Integrated Reference Platform*; Deutsche Gesellschaft für Luft-und Raumfahrt-Lilienthal-Oberth eV: Bonn, Germany, 2016.
61. Steidel, M.; Hahn, A. MTCAS—An Assistance System for Collision Avoidance at Sea. In Proceedings of the 18th International Conference on Computer and IT Applications in the Maritime Industries, Tullamore, Ireland, 25–27 March 2019.
62. Chen, Y.; Ji, C.; Cai, Y.; Yan, T.; Su, B. Deep Reinforcement Learning in Autonomous Car Path Planning and Control: A Survey. *arXiv* **2024**, arXiv:2404.00340.

63. Kuhn, D.R.; Wallace, D.R.; Gallo, A.M. Software fault interactions and implications for software testing. *IEEE Trans. Softw. Eng.* **2004**, *30*, 418–421. [[CrossRef](#)]
64. Gründl, M. *Fehler und Fehlverhalten als Ursache von Verkehrsunfällen und Konsequenzen für das Unfallvermeidungspotenzial und die Gestaltung von Fahrerassistenzsystemen*; Universität Regensburg: Regensburg, Germany, 2005. [[CrossRef](#)]
65. Cooley, J.W.; Tukey, J.W. An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comput.* **1965**, *19*, 297–301. [[CrossRef](#)]
66. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: Cambridge, UK, 2007.
67. Maranzana, F.E. On the location of supply points to minimize transportation costs. *IBM Syst. J.* **1963**, *2*, 129–135. [[CrossRef](#)]
68. Park, H.-S.; Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **2009**, *36*, 3336–3341. [[CrossRef](#)]
69. Discrepancy—SciPy v1.16.0 Manual. Available online: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.qmc.discrepancy.html> (accessed on 2 July 2025).
70. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
71. Autenrieth, D. SBSSM-Benchmark v1.1.1. Available online: <https://zenodo.org/records/21441489> (accessed on 20 July 2026).
72. Zhang, J.; Xu, C.; Li, B. ChatScene: Knowledge-Enabled Safety-Critical Scenario Generation for Autonomous Vehicles. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 15459–15469. [[CrossRef](#)]
73. Gao, Y.; Piccinini, M.; Brusnicki, R.; Zhang, Y.; Betz, J. NuRisk: A Visual Question Answering Dataset for Agent-Level Risk Assessment in Autonomous Driving. *arXiv* **2026**, arXiv:2509.25944.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.