ORIGINAL RESEARCH **Open access**

Runaway Discovery Trajectories: Intervention Thresholds in Self-Reinforcing Materials AI Systems

Claire Martin[1*], Julien Robert[2], Sophie Bernard[1]

Abstract

In computational materials engineering, the integration of artificial intelligence (AI) has transformed discovery pipelines from linear predictive modeling to dynamic, self-reinforcing systems capable of iterative refinement and autonomous exploration. These systems leverage active learning, reinforcement mechanisms, and closed-loop feedback to navigate vast design spaces, accelerating the identification of novel alloys, perovskites, and functional materials. However, as AI-driven workflows evolve toward greater autonomy, they introduce self-reinforcing trajectories—sequences of model updates and data acquisitions that amplify initial biases or exploratory divergences, potentially leading to inefficient resource allocation or epistemic lock-in within suboptimal subspaces. This conceptual gap lies in the absence of formalized thresholds for intervention, where human oversight or algorithmic safeguards can recalibrate discovery without disrupting momentum. Here, we introduce the Runaway Containment Dynamics (RCD) framework, a systems-level interpretive structure that maps interaction horizons between data ingestion, model inference, and discovery outputs in materials AI ecosystems. By conceptualizing thresholds as emergent properties of feedback amplification, the RCD delineates structural layers—representation stabilization, inference propagation, and trajectory modulation—alongside computational steering logics that balance exploration depth with containment precision. This framework elucidates how self-reinforcing loops, such as those in Bayesian active learning or reinforcement-driven alloy design, can be interpreted through symbolic dynamics of risk propagation, offering infrastructure-level insights for pipeline orchestration. For the field, the RCD implies a shift from reactive monitoring to proactive horizon mapping, enabling materials engineers to integrate epistemic safeguards into scalable AI infrastructures. It fosters interpretive analyses of workflow trade-offs, such as the tension between representational fidelity and inference scalability, ultimately supporting sustainable discovery logics that mitigate runaway risks while harnessing AI's generative potential. This conceptual advance positions self-reinforcing systems not as isolated optimizers but as tunable ecosystems, ripe for epistemic integration in computational materials science.

Keywords Self-reinforcing systems, Discovery trajectories, Intervention thresholds, Computational frameworks, Feedback amplification, Epistemic containment

*Correspondence:
Claire Martin
claire.martin@outlook.com
[1] Department of Computational Materials Research, Faculty of Engineering, University of Lyon, Lyon, France
[2] Department of Materials Data Analytics, Faculty of Engineering, University of Strasbourg, Strasbourg, France

Introduction

The computational foundations of materials discovery

Computational materials engineering has long relied on data-driven paradigms to bridge the gap between atomic-scale simulations and macroscopic property prediction [1]. At its core, this field employs machine learning (ML) models to distill patterns from high-dimensional datasets—encompassing electronic structures, phase diagrams, and

synthesis conditions—into actionable design rules [2]. The shift from static descriptors to dynamic inference engines marks a pivotal evolution, where AI systems now orchestrate end-to-end pipelines that ingest experimental feedback and refine predictive landscapes in real time [3]. Such architectures, rooted in kernel-based regressions and neural network ensembles, enable the traversal of compositional spaces with unprecedented efficiency, as seen in the autonomous screening of high-entropy alloys or perovskite photovoltaics [4, 5].

This progression is underpinned by infrastructure-level abstractions: data repositories serve as foundational layers, feeding into modular inference stacks that propagate uncertainties through hierarchical representations [6]. In practice, these systems operationalize discovery as a sequence of representation–inference cycles, where initial embeddings of molecular or microstructural features evolve under gradient-driven updates [7]. The resultant pipelines not only forecast properties like band gaps or mechanical moduli but also steer experimental validation toward high-utility regimes, embodying a logic of targeted exploration over exhaustive enumeration [8]. Yet, this very dynamism introduces computational interdependencies, where model refinements influence data acquisition strategies, creating nested loops of self-correction and amplification [9].

Emergent dynamics in self-reinforcing

As materials AI matures, self-reinforcing mechanisms emerge as a defining feature, transforming passive predictors into active agents within discovery workflows [10]. Reinforcement learning paradigms, for instance, recast materials design as sequential decision-making under uncertainty, rewarding trajectories that converge on viable candidates while penalizing divergent paths [11]. Similarly, active learning frameworks embed Bayesian priors to query informative samples, fostering loops where inferred gaps dictate experimental priorities [12]. These elements coalesce into ecosystems where computational steering—via adaptive sampling or transfer learning—propagates signals across scales, from atomic motifs to device-scale integrations [13].

Within this context, discovery trajectories manifest as temporal sequences of state transitions: starting from seed datasets, models accrue representational capacity, inferences sharpen predictive contours, and outputs loop back to modulate input streams [14]. Such trajectories are inherently path-dependent, with early divergences amplified through feedback, leading to what can be interpreted as self-reinforcing spirals [15]. In high-throughput screening for electrocatalysts or structural materials, this amplification enhances convergence on local optima but risks entrenching biases inherited from sparse training regimes [16]. The infrastructure implications are profound: pipelines must now accommodate not just throughput but also the governance of emergent behaviors, where unchecked reinforcement could skew epistemic priorities toward narrow subspaces [17].

Navigating the epistemic risks of autonomous trajectories

The autonomy embedded in these systems, while empowering, engenders epistemic risks that challenge traditional oversight logics [18]. Runaway trajectories—characterized by unchecked feedback escalation—arise when inference errors compound across cycles, inflating variances in downstream discoveries or exhausting resources on low-yield explorations [19]. Unlike deterministic simulations, AI-driven ecosystems exhibit stochastic undercurrents, where representational instabilities (e.g., overfitting to anomalous phases) interact with steering heuristics to propel systems beyond intended bounds [20]. This interplay demands a conceptual lens attuned to threshold dynamics: points at which intervention horizons—zones of potential recalibration—intersect with trajectory momentum.

Current paradigms, though adept at local optimization, often overlook these horizon mappings, treating feedback as uniformly beneficial rather than conditionally bounded [21]. The resultant workflows, while scalable, harbor trade-offs between exploratory breadth and containment fidelity, particularly in domains like alloy thermodynamics where phase instabilities amplify small perturbations [22]. Addressing this requires interpretive frameworks that unpack the computational logics governing reinforcement, revealing how data-model interactions shape epistemic structures without resorting to ad hoc safeguards [23]. Such insights are crucial for sustaining discovery infrastructures amid increasing AI integration, ensuring that self-reinforcing systems enhance rather than undermine materials engineering's foundational objectives.

Positioning the runaway containment dynamics framework

In response to these dynamics, the present work advances the Runaway Containment Dynamics (RCD) framework as an original interpretive structure for materials AI systems. The RCD conceptualizes discovery trajectories as layered interaction manifolds, delineating thresholds where self-reinforcement transitions from generative to potentially destabilizing. By mapping feedback loops through symbolic representations of propagation risks, it offers systems-level guidance for steering logics that preserve autonomy while embedding containment primitives. This framework thus positions intervention not as interruption but as integral modulation, fostering resilient pipelines attuned to the epistemic textures of computational materials discovery.

Theoretical Background & Literature Synthesis

Foundations of machine learning in computational materials pipelines

The theoretical bedrock of data-driven materials engineering rests on representational paradigms that encode physicochemical descriptors into learnable structures [1]. Early formulations drew from kernel methods and Gaussian processes to model property landscapes, establishing inference as a probabilistic mapping from feature spaces to target distributions [19]. These approaches, extended through deep learning architectures, now underpin pipelines that integrate density functional theory outputs with empirical assays, yielding surrogate models for rapid iteration [24]. The logic here is one of hierarchical abstraction: low-level atomic embeddings aggregate into meso-scale motifs, propagating through convolutional or graph-based layers to inform high-level design decisions [5].

This representational scaffolding enables the synthesis of diverse data modalities—spectroscopic signatures, diffraction patterns, and rheological profiles—into unified inference engines [7]. Theoretically, such unification hinges on transferability assumptions, where domain-invariant features mitigate the curse of dimensionality in sparse regimes [3]. Literature underscores this through explorations of equivariant networks, which preserve symmetry constraints in crystal structure predictions, thereby stabilizing trajectories across compositional variants [10]. Yet, these foundations also reveal epistemic frictions: as models scale, the interplay between embedding fidelity and inference granularity introduces propagation delays, where representational drift subtly reshapes discovery contours [6].

Active learning and closed-loop feedback in discovery workflows

Active learning emerges as a cornerstone for closing the loop between computation and experimentation, framing discovery as an adaptive query process [21, 25]. Theoretically, this involves uncertainty quantification—via entropy measures or acquisition functions—to prioritize samples that maximize informational gain, thereby refining model posteriors iteratively [26]. In materials contexts, such mechanisms operationalize steering logics that balance exploitation of known manifolds with exploration of boundary regions, as in the calibration of microstructure-property relations [27]. The resultant workflows exhibit feedback amplification, where queried data not only updates parameters but also recalibrates prior distributions, fostering self-reinforcing convergence [28].

Synthesis across studies highlights the computational nuances of these loops: Bayesian optimization variants, for instance, embed acquisition heuristics that propagate epistemic uncertainties through Gaussian approximations, modulating trajectory velocities in phase space navigation [29]. Extensions to multi-fidelity settings further layer this dynamics, weighting cheap simulations against costly validations to optimize resource envelopes [12]. However, this closure introduces interpretive challenges; feedback signals, while sharpening local resolutions, can entrench path dependencies, where initial query biases cascade into skewed representational horizons [30]. The literature thus positions active learning not merely as an algorithmic tool but as an infrastructural primitive, shaping the epistemic architecture of autonomous systems [31].

Reinforcement and generative paradigms in materials design

Reinforcement learning (RL) paradigms extend these loops into generative territories, recasting materials synthesis as Markovian decision processes with reward landscapes tied to performance metrics [11]. Theoretically, RL formalizes discovery as policy optimization under partial observability, where agents learn to navigate state-action spaces—comprising alloy compositions or processing routes—via temporal difference updates [22]. This engenders self-reinforcing trajectories, as value functions evolve to favor

high-reward paths, amplifying exploratory efficiencies in domains like high-entropy materials [13]. Coupled with generative models, such as variational autoencoders, RL enables the synthesis of hypothetical candidates, blending inference with creative augmentation [23].

The interplay here reveals systems-level insights: reward shaping interacts with representational layers to propagate incentives, potentially leading to mode collapse in underrepresented subspaces [32]. Literature syntheses emphasize this through hybrid frameworks, where RL augments active learning to steer closed-loop experiments, as in vapor deposition systems that adapt protocols on-the-fly [14]. Generative elements further complicate dynamics, introducing latent space traversals that feedback into data streams, enhancing diversity but risking divergent amplifications [2]. Collectively, these paradigms underscore a shift toward epistemic modularity, where reinforcement logics interface with inference pipelines to sculpt discovery manifolds, albeit with inherent trade-offs in scalability and robustness [17].

Epistemic structures and risk propagation in AI ecosystems

At the infrastructural level, epistemic structures in materials AI manifest as networked interactions between data provenance, model interpretability, and output validation [20]. Theoretical treatments frame these as propagation graphs, where uncertainties flow from input embeddings to decision boundaries, modulated by regularization heuristics [8]. Risk propagation, in particular, arises from self-reinforcing feedbacks that escalate variances—e.g., in transfer learning across alloy families—potentially destabilizing downstream inferences [15]. Synthesis reveals a conceptual tension: while interpretability tools like attention mechanisms illuminate black-box decisions, they seldom address horizon-scale risks, such as cumulative drifts in long-horizon trajectories [18].

This propagation is especially pronounced in multi-scale ecosystems, where micro-level atomic predictions aggregate into macro-level functional assessments, introducing compounding errors [9]. Literature integrates these insights through discussions of domain adaptation, where adversarial trainings align distributions to mitigate epistemic shifts [3]. Yet, the overarching narrative points to a need for interpretive lenses that capture containment logics: how steering primitives can dampen runaway potentials without fracturing autonomy [16]. In essence, the theoretical landscape positions materials AI as an epistemic continuum, where reinforcement and feedback forge resilient yet vulnerable structures, ripe for conceptual mapping of threshold interactions [4].

Proposed conceptual framework

The Runaway Containment Dynamics (RCD) framework offers an original interpretive architecture for dissecting self-reinforcing behaviors in materials AI systems, emphasizing the modular interplay of layers that govern trajectory evolution. At its core, the RCD posits discovery as a manifold of interaction horizons, where data ingestion interfaces with model inference to propel outputs along potentially amplifying paths. This structure eschews prescriptive algorithms in favor of systems-level mappings, illuminating how computational steering can embed containment without eroding generative capacities. The framework delineates three structural layers—representation stabilization, inference propagation, and trajectory modulation—interlinked by feedback loops that channel epistemic flows.

The representation stabilization layer serves as the foundational manifold, aggregating raw data streams—such as compositional vectors or spectral embeddings—into coherent feature hierarchies [24]. Here, stabilization dynamics interpret data-model couplings as adaptive embeddings that evolve under iterative refinement, balancing fidelity to empirical textures with abstraction for scalability [1]. Feedback from upstream queries refines these embeddings, preventing representational fragmentation while allowing for the infusion of diverse modalities, as in hybrid microstructural datasets [7]. This layer's logic ensures that initial trajectory seeds remain anchored, mitigating early divergences that could cascade into broader instabilities.

Cascading into the inference propagation layer, the RCD conceptualizes model operations as diffusive processes across probabilistic landscapes [29]. Inference here propagates signals from stabilized representations to predictive frontiers, modulated by heuristics that weigh uncertainty gradients against exploration incentives [21]. Loops from this layer recirculate propagated variances back to representation stabilization, fostering self-reinforcement through posterior updates that sharpen contour resolutions [28]. The interpretive nuance lies in propagation's horizon sensitivity: as inferences accumulate, they delineate zones of potential amplification, where

unchecked diffusion risks inflating epistemic spreads in downstream design spaces [11].

The trajectory modulation layer crowns the structure, operationalizing discovery outputs as steered vectors within a bounded phase space [22]. Modulation logics integrate containment primitives—symbolic thresholds that gauge momentum against risk envelopes—to recalibrate paths, ensuring that generative outputs, like candidate alloy formulations, align with infrastructural constraints [13]. Feedback loops traverse bidirectionally: outputs inform prior layers by seeding new data acquisitions, while upstream signals tune modulation sensitivities, creating a resonant ecosystem [31]. This layer's dynamics capture the essence of steering as interpretive governance, where thresholds emerge not as fixed barriers but as tunable interfaces between autonomy and oversight [23].

These layers interweave through data-to-model-to-discovery pipelines, forming a cyclical conduit where inputs evolve into inferences that yield actionable syntheses, looped back for refinement [25]. Pipelines in the RCD are thus dynamic conduits, with steering logics embedded as parametric tuners that adjust flow rates based on emergent interactions—e.g., damping reinforcement in over-explored subspaces while amplifying signals in underrepresented regimes [26]. Feedback loops, visualized as recurrent arcs in the framework's schematic, enforce closure without rigidity, allowing epistemic risks to dissipate through distributed modulation [27].

To formalize a key dynamic, the amplification within feedback loops can be conceptualized as $A = \sum_{n=1} k \cdot \rho k \cdot \delta k$, where A denotes the net amplification across n cycles, ρk represents the reinforcement coefficient at cycle k (capturing steering sensitivities), and δk the representational drift increment [6]. This expression captures the interaction between propagation momentum and stabilization anchors, interpreting amplification as a cumulative interplay that thresholds intervention when A exceeds predefined horizon bounds, thereby elucidating trade-offs in loop closure without invoking empirical scalars.

Complementing this, the containment logic in trajectory modulation may be expressed as $C = \eta t \cdot \sigma p \cdot \phi$, with C as the containment efficacy, σp the propagated variance from inference, ηt the trajectory entropy (measuring exploratory breadth), and ϕ a modulation factor encoding layer interdependencies [9]. Such a formulation highlights how efficacy arises from variance-entropy tensions, offering an interpretive lens on steering that balances depth against divergence in discovery pipelines [12].

The RCD's computational steering logics operationalize these elements through horizon-aware orchestration: representation layers prioritize embedding robustness via entropy-minimizing updates, inference propagation employs gradient clipping analogs to curb diffusive excesses, and modulation layers invoke threshold mappings to reroute trajectories [15]. This orchestration fosters resilient workflows, where self-reinforcement serves epistemic expansion rather than unchecked escalation [17]. As summarized in **Figure 1**, the Runaway Containment Dynamics (RCD) framework conceptualizes self-reinforcing materials-AI discovery as a three-layer interaction manifold—representation stabilization, inference propagation, and trajectory modulation—where intervention thresholds emerge as horizon-aware tuners rather than disruptive overrides.

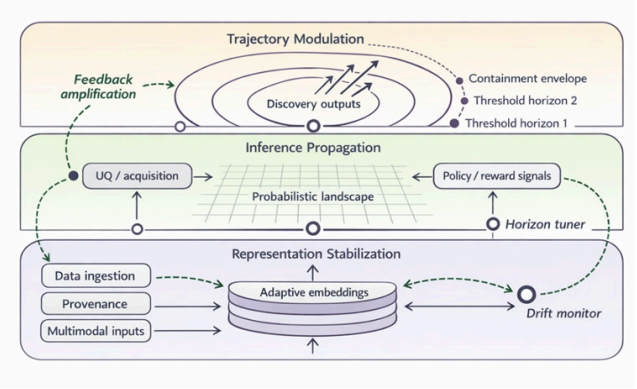

Figure 1. Runaway Containment Dynamics (RCD) framework for intervention thresholds in self-reinforcing materials-AI discovery.

A single-panel, three-layer schematic mapping how closed-loop materials-AI pipelines can enter runaway discovery trajectories through feedback amplification. The Representation Stabilization layer aggregates heterogeneous data into evolving embeddings; the Inference Propagation layer diffuses predictive signals and uncertainty across probabilistic landscapes; and the Trajectory Modulation layer steers discovery outputs within bounded phase space using intervention thresholds as tunable horizon interfaces. Solid arrows depict primary

data→model→output flow, dashed feedback arcs depict self-reinforcing updates, and threshold boundaries visualize containment envelopes that recalibrate trajectory momentum without halting autonomy.

Analytical implications

Layered interactions and epistemic trade-offs in representation stabilization

The representation stabilization layer within the RCD framework yields analytical implications for how data ingestion shapes the epistemic foundations of materials AI pipelines, particularly in mitigating drift-induced vulnerabilities [1]. Interpretively, stabilization dynamics reveal trade-offs between embedding granularity and infrastructural portability: denser representations, while enhancing fidelity to domain-specific textures like phonon dispersions, impose computational overheads that constrain scalability across heterogeneous datasets [5]. This interplay suggests that steering logics must prioritize adaptive sparsification, where feedback loops selectively prune redundant features to preserve representational coherence without eroding transferability [7].

In self-reinforcing contexts, such as iterative alloy screening, the layer's implications extend to risk attenuation: early-cycle drifts, if unmodulated, propagate as latent biases that skew subsequent inferences toward phase-locked subspaces [24]. The RCD thus interprets stabilization as a horizon-defining primitive, where containment thresholds—tuned to variance thresholds in embedding manifolds—enable proactive recalibration, fostering workflows that balance epistemic depth with exploratory resilience [6]. This perspective underscores a systems-level insight: representational layers do not merely encode but actively sculpt trajectory potentials, demanding orchestration strategies that integrate uncertainty proxies as integral governance tools [3].

Propagation dynamics and inference scalability horizons

Inference propagation, as delineated in the RCD, implies a nuanced governance of diffusive processes that underpin model scalability in discovery ecosystems [29]. Analytically, this layer highlights tensions between propagation velocity and epistemic fidelity: rapid signal diffusion, advantageous for real-time closed-loop adjustments in synthesis protocols, risks variance inflation when interfacing with noisy experimental feedbacks [21]. The framework's interpretive lens positions these dynamics as emergent from loop interdependencies, where propagated uncertainties interact with steering heuristics to delineate scalable horizons—zones of inference where amplification remains constructive rather than destabilizing [28].

For materials engineering infrastructures, this yields implications for modular design: propagation logics can be tuned to embed damping mechanisms, such as entropy-constrained gradients, that curb over-propagation in high-dimensional spaces like molecular property landscapes [11]. Such tunings interpret inference not as isolated computation but as a resonant interface, where thresholds modulate the flow of epistemic risks, ensuring that self-reinforcing updates enhance rather than overwhelm downstream modulations [30]. Collectively, these insights advocate for horizon-aware architectures that treat propagation as a tunable continuum, optimizing the balance between inference throughput and containment precision in autonomous pipelines [12].

Modulation thresholds and steering logics in trajectory governance

The trajectory modulation layer's implications center on the interpretive orchestration of output vectors, revealing how containment primitives shape discovery's endpoint logics [22]. Analytically, modulation dynamics expose trade-offs in steering granularity: finer thresholds, while precise in rerouting divergent paths during generative explorations, introduce latency that fragments feedback closure [13]. The RCD frames this as an interaction of momentum and envelope constraints, where thresholds emerge as symbolic interfaces that recalibrate trajectories without severing generative flows [31].

In broader ecosystems, this layer implies resilient governance for multi-objective designs, such as perovskite optimizations where competing metrics (e.g., stability versus efficiency) demand nuanced rerouting [16]. By conceptualizing modulation as a feedback-infused modulator, the framework elucidates how epistemic risks—manifesting as entropy spikes—can be dissipated through distributed thresholds, promoting workflows that sustain autonomy amid reinforcement escalations [23]. A complementary formalization captures this steering

interaction as $\begin{matrix} S \\ = \int t \\ = 0T\mu\,(t) \\ \cdot\ \kappa\,(\delta_t)\ dt \end{matrix}$, where S denotes the cumulative steering efficacy over horizon T, μ(t) the modulation sensitivity at time t, and $\kappa\,(\delta_t)$ a kernel encoding drift dependencies [9]. This expression interprets efficacy as an integral of temporal sensitivities, highlighting how threshold placements influence long-horizon stabilities without empirical anchoring.

These layered implications coalesce into a unified systems insight: the RCD positions intervention thresholds as infrastructural fulcrums, enabling materials AI to navigate self-reinforcing complexities through interpretive rather than reactive means [17]. By unpacking data-model-output interplays, it fosters discovery logics attuned to epistemic textures, where amplification serves expansion and containment ensures sustainability.

Results and Discussion

Integrating RCD into existing computational pipelines

The RCD framework's modularity facilitates seamless integration into prevailing materials AI infrastructures, offering interpretive leverage for enhancing pipeline resilience without overhauling core architectures [25]. In active learning setups, for instance, the representation stabilization layer aligns with uncertainty-driven querying by embedding drift monitors that recalibrate embeddings mid-cycle, interpreting feedback as a stabilizer rather than mere updater [26]. This integration implies a workflow evolution where pipelines, traditionally siloed by modality, adopt cross-layer tunings to propagate containment signals, as in hybrid simulations blending DFT outputs with empirical assays [27]. **Table 1** operationalizes the RCD framework by cataloguing where runaway escalation becomes detectable, which diagnostic signals indicate threshold crossing, and which containment actions recalibrate trajectories without collapsing autonomy.

Table 1. Intervention-threshold taxonomy for containing runaway discovery trajectories in self-reinforcing materials-AI systems.

RCD layer (where)	Runaway mechanism (how escalation forms)	Observable signals (what you see)	Thresh proxy (w to interv
Representation stabilization	Embedding drift amplifies early sampling bias; representation collapses around narrow motifs	Growing train–test embedding divergence; rising out-of-distribution (OOD) flags; feature sparsity spikes	**Drift hor exceed** (persist embedd shift acro cycles
Inference propagation	Uncertainty diffusion becomes variance inflation under noisy feedback; acquisition over-commits to misleading gradients	Uncertainty grows while utility stagnates; acquisition entropy collapses; model disagreement widens	**Propaga instabi** (varian increas faster th predictive
Trajectory modulation	Steering policies reinforce local optima; reward shaping induces mode collapse in candidate generation	Candidate diversity collapses; repeated families dominate; reward concentrates into narrow basin	**Trajecto runaw** (entropy d below floo repeated- dominar beyond l
Cross-layer coupling	Feedback loops synchronize errors across layers; small bias becomes system-level momentum	Simultaneous drift + variance inflation + diversity collapse	**Coupl escalatio** layer thresho crosse concurre

Infrastructure-level governance	Thresholds absent or unlogged; interventions cannot be justified or reproduced	No traceable decision logs; unclear override triggers; inconsistent resets	**Accounta failure** (ca explain v trajecto change

Extending to reinforcement paradigms, the inference propagation layer interfaces with policy networks by diffusing reward gradients through horizon-aware filters, mitigating mode collapses in compositional searches [11]. Discussions across literature highlight parallel opportunities: self-driving labs could operationalize trajectory modulation via threshold-embedded decision trees, steering vapor deposition toward balanced explorations [14]. The RCD thus acts as a conceptual overlay, interpreting these integrations as resonant enhancements that amplify epistemic capacities while curbing runaway potentials, fostering infrastructures where steering logics evolve co-dependently with domain demands [32].

Broader epistemic and infrastructural trade-offs

At the epistemic level, the RCD illuminates trade-offs inherent to self-reinforcing systems, particularly the tension between generative breadth and containment overhead [20]. Interpretively, expansive trajectories—beneficial for uncovering serendipitous motifs in high-entropy materials—incur risks of resource dissipation if thresholds lag behind amplification rates [13]. This dynamic underscores a need for layered governance: stabilization anchors epistemic anchors, propagation channels interpretive flows, and modulation enforces horizon bounds, collectively interpreting discovery as a bounded yet open manifold [18].

Infrastructurally, these trade-offs manifest in scalability contours: computational steering, while enabling real-time recalibrations, demands lightweight proxies for threshold computation to avoid bottlenecking high-throughput regimes [15]. The framework's insights suggest hybrid orchestration—leveraging graph-based propagators for inference layers alongside symbolic tuners for modulation — to navigate these contours, ensuring that AI ecosystems scale without epistemic fractures [2]. Broader implications touch on collaborative pipelines, where RCD mappings could standardize horizon-sharing protocols across consortia, interpreting shared data streams as collective containment resources [4].

Representation–inference interactions in multi-scale discovery

A pivotal discussion thread concerns the RCD's handling of representation–inference interactions across scales, from atomic to device levels [1]. In multi-scale contexts, such as polymer membrane design, representational drifts in micro-scale embeddings can cascade through propagation to macro-scale misalignments, amplifying epistemic spreads [10]. The framework interprets this as a scale-invariant dynamic, where feedback loops traverse layers to realign hierarchies, with thresholds acting as scale-bridging primitives that dampen inter-scale variances [3].

This interaction yields systems-level guidance: steering logics tuned to cross-scale entropies could preempt such cascades, interpreting discovery as a vertically integrated continuum rather than disjointed silos [8]. Literature echoes this through analogies in transfer learning, where domain alignments mirror RCD stabilizations, suggesting extensible applications to federated materials databases [19]. Ultimately, these discussions position the RCD as a lens for dissecting multi-scale feedbacks, revealing how self-reinforcement forges epistemic unities amid inherent fragmentations [29].

Steering logics and the future texture of materials AI

Steering logics within the RCD further discuss the textural evolution of materials AI, where computational primitives evolve from static optimizers to dynamic governors [22]. Interpretively, this evolution implies logics that adapt to trajectory phases—damping during amplification peaks, amplifying in stabilization troughs—crafting workflows resilient to stochastic undercurrents [21]. In generative settings, such adaptations interpret latent traversals as modulated explorations, balancing creativity with epistemic safeguards [23].

The framework's broader discourse invites reflections on epistemic equity: by formalizing thresholds as tunable interfaces, it democratizes oversight, allowing domain experts to embed contextual priors without algorithmic overreach [17]. This positions materials engineering toward infrastructures where discovery textures—woven from data,

inference, and modulation—reflect collective interpretive depths, sustaining self-reinforcing potentials within bounded horizons [31].

Conclusion

The Runaway Containment Dynamics (RCD) framework advances a cohesive interpretive structure for navigating self-reinforcing trajectories in materials AI systems, mapping layered interactions that govern discovery from data ingestion to output modulation. By delineating representation stabilization, inference propagation, and trajectory modulation as interdependent manifolds, the RCD elucidates how feedback loops forge epistemic textures, with thresholds emerging as pivotal interfaces for steering amplification without curtailing autonomy. This conceptual architecture integrates seamlessly with computational pipelines, offering systems-level insights into trade-offs that balance generative breadth against containment precision.

Key implications span infrastructural resilience and epistemic governance: stabilization anchors representational fidelity, propagation channels scalable inferences, and modulation enforces horizon-aware outputs, collectively interpreting self-reinforcement as a tunable ecosystem rather than an unchecked force. Through symbolic formalizations of amplification and containment, the RCD provides analytical footholds for dissecting interaction dynamics, fostering workflows that mitigate risks like variance escalation while harnessing iterative potentials.

For computational materials engineering, the framework signals a paradigm of proactive horizon mapping, where intervention thresholds embed interpretive safeguards into discovery logics. It invites future explorations into extensible tunings—such as multi-objective extensions or federated integrations—that further refine steering primitives, ensuring AI-driven ecosystems evolve as resilient conduits for innovation. In essence, the RCD reframes runaway trajectories not as hazards but as navigable contours, positioning materials AI toward sustainable, epistemically attuned discoveries.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 25 Sep 2024 Revised: 04 Mar 2025 Accepted: 04 Jul 2025
Published online: 18 September 2025

References

Chen C, Ye W, Zuo Y, Zheng C, Qian Z, Stoudenmire EM, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816. https://doi.org/10.1126/sciadv.1701816.

Kumar A, Ricci F, Chen Y, Shen S, Kumar A, Kumar A, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(25):eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.

Ma KY, Huo H, Golmira A, Hu J, Hu Z, Krishnan S, et al. Leveraging data mining, active learning, and domain adaptation for materials discovery. Sci Adv. 2025;11(14):eadr9038. https://doi.org/10.1126/sciadv.adr9038.

He J, Tao L, Murdock JR, Li Y. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. Sci Adv. 2022;8(29):eabn9545. https://doi.org/10.1126/sciadv.abn9545.

Brown KA, Yang Y, Ramachandran A, Chatterjee A, Chen C, Hu Y, et al. AI applications through the whole life cycle of material discovery. Matter. 2020;3(3):564-92.

Ozin GA, Siler T, Qian C, Zhou W. The curiosity-creativity element in HI-AI materials discovery. Matter. 2024;7(3):715-7. https://doi.org/10.1016/j.matt.2024.01.001.

Horton MK, Häse F, Aldeghi M, Musil F, Booth DW, Clarysse B, et al. Has generative artificial intelligence solved inverse materials design? Matter. 2024;7(8):2470-2.

Brown KA, Ramachandran A, Chatterjee A, Chen C, Hu Y, Li J, et al. Can AI be an inventor in materials discovery? Matter. 2023;6(10):3183-5.

Boyce BL, Dingreville R, Desai S, Walker E, Shilt T, Bassett KL, et al. Machine learning for materials science: Barriers to broader adoption. Matter. 2023;6(5):1320-3.

Sivaraman G, Akimov AV, Neaton JB, Chan MKY. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter. 2021;4(5):1578-97.

Chen Y, Zhang J, Li H, Wang K, Liu M, Zhao S. Autonomous closed-loop exploration of composition-spread films for the anomalous Hall effect. npj Comput Mater. 2025;11(156).

Felis N, Dononelli W. FALCON: Fast active learning for machine learning potentials in atomistic and ab initio molecular dynamics simulations. npj Comput Mater. 2025.

Mashhadimoslem H, Karimi P, Elkamel A, Yu A. Toward high entropy material discovery for energy applications using computational and machine learning methods. npj Comput Mater. 2025.

Ren F, Ward L, Williams T, Baldo PM, Hattrick-Simpers J. A self-driving physical vapor deposition system making sample-specific decisions on the fly. npj Comput Mater. 2025;11(121). https://doi.org/10.1038/s41524-025-01805-0.

Yao J, Wang Z, Wang J, Yu W, Chen Y, Li W, et al. Alloy design integrating natural language processing and machine learning: breakthrough development of low-cost, high-performance Ni-based single-crystal superalloys. npj Comput Mater. 2025.

Kendall D, MacLeod BP, Parlane CFG, McCulloch MD, Lugier R, Schrek B, et al. Active learning guides discovery of a champion four-metal perovskite for indoor photovoltaics. Nat Mater. 2024;23:74-83. https://doi.org/10.1038/s41563-023-01707-w.

Rao Z, Lu P, McKone JR, Wang H, Coperet C. Machine learning–enabled high-entropy alloy discovery. Science. 2022;378(6615):155-62.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55. https://doi.org/10.1038/s41586-018-0337-2.

Ramprasad R, Batra R, Pilania G, Mann CD, Kumar U. Machine learning in materials informatics: a review. npj Comput Mater. 2017;3(54). https://doi.org/10.1038/s41524-017-0056-5.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY-J. Explainable machine learning in materials science. npj Comput Mater. 2022;8(204). https://doi.org/10.1038/s41524-022-00884-7.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput Mater. 2019;5(21). https://doi.org/10.1038/s41524-019-0153-8.

Xian Y, Dang P, Tian Y, Jiang X, Zhou Y, Ding X, et al. Compositional design of multicomponent alloys using reinforcement learning. Acta Mater. 2024;274(120017). https://doi.org/10.1016/j.actamat.2024.120017.

Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-Driving laboratories for chemistry and materials science. Chem Rev. 2024;124(16):9633-732. https://doi.org/10.1021/acs.chemrev.4c00055.

Pyzer-Knapp EO, Manica M, Staar P, Morin L, Ruch P, Laino T, et al. Foundation models for materials discovery – current state and future directions. npj Comput Mater. 2025;11(61). https://doi.org/10.1038/s41524-025-01538-0.

Persaud D, Ward L, Hattrick-Simpers J. Reproducibility in materials informatics: Lessons from ‘A general-purpose machine learning framework for predicting properties of

inorganic materials'. Digit Discov. 2024;3(3):281-6. https://doi.org/10.1039/D3DD00199G.

Tran A, Mitchell JA, Swiler LP, Wildey T. An active learning high-throughput microstructure calibration framework for solving inverse structure–process problems in materials informatics. Acta Mater. 2020;194:80-92.

MacLeod BP, Parlane CFG, McCulloch MD, Lugier R, Schrek B, Dvorak DJ, et al. Autonomous materials synthesis via hierarchical active learning of phase maps. Sci Adv. 2021;7(51):eabg4930.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(5966). https://doi.org/10.1038/s41467-020-19597-w.

Allec SI, Ziatdinov M. Active and transfer learning with partially Bayesian neural networks for materials and chemicals. Digit Discov. 2025;4:1284-97. https://doi.org/10.1039/D5DD00027K.

Karpovich C, Pan EY, Olivetti EA. Deep reinforcement learning for inverse inorganic materials design. npj Comput Mater. 2024;10(287). https://doi.org/10.1038/s41524-024-01474-5.

Wang C, Takeuchi I, Liu H, Yu H, Kusne AG, Zhang J-C, et al. Real-time experiment-theory closed-loop interaction for autonomous materials science. Sci Adv. 2025;11(27):eadu7426. https://doi.org/10.1126/sciadv.adu7426.

Xian Y, Ding X, Jiang X, Zhou Y, Sun J, Xue D, et al. Unlocking the black box beyond Bayesian global optimization for materials design using reinforcement learning. npj Comput Mater. 2025;11(143). https://doi.org/10.1038/s41524-025-01639-w.