Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Based Stability Prediction for Magnet Packing in the Game Kluster

Sankalp Savarn
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11733631/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Machine Learning-Based Stability Prediction for Magnet Packing in the Game Kluster

Sankalp Savarn[1] and Mayank Goswami [1, 2, #]

[1] Department of Physics, IIT Roorkee
[2]*Mehta Family School of Data Science and Artificial Intelligence*, IIT Roorkee,
Roorkee, Uttarakhand-247667, India.
[#]mayank.goswami@ph.iitr.ac.in

Abstract

In the magnetic tabletop game Kluster, players sequentially place permanent magnets inside a cord-bounded arena without triggering inter-magnet collisions. Determining whether a candidate placement will relax into static equilibrium or snap into contact requires numerically integrating stiff, non-smooth equations of motion governed by steep $1/r^4$ dipole forces and dry Coulomb friction. This study investigates whether machine learning (ML) classifiers wrapped in size-conditional (Mondrian) Conformal Risk Control (CRC) can pre-screen candidate placements—discarding likely collisions prior to numerical integration—while bounding the rate of falsely discarded safe placements (the wrong-skip risk) below a user-specified tolerance α. We model identical cylindrical neodymium magnets as vertical point dipoles translating on a horizontal plane inside a rigid circular inelastic boundary. Using a Numba-compiled semi-implicit Euler simulator with step-halving convergence verification, we generate a dataset of 205,856 candidate placements across 1,500 sequential games with independently randomized friction coefficients and dipole force scales. Empirical evaluation across five experimental protocols reveals both the utility and the strict boundaries of surrogate screening: (1) Size-conditional CRC controls wrong-skip risk *on average only* across repeated game splits; for Histogram Gradient Boosting with force-ratio features at $\alpha = 0.05$ in board-size bin 5–8, the mean wrong-skip rate is 0.0457 (SE = 0.0012), yet 33% of individual calibration-test splits exceed α due to finite-sample variance. (2) Paired comparisons against an analytical Combined physics rule show no statistically significant ML gain on sparse boards (bin 2–4: +0.0015, 95% CI [−0.0010, +0.0056]), but significant gains on crowded boards where multi-body forces dominate (bin 5–8: +0.0581; bin 9–14: +0.0379). (3) Under unadapted physical parameter shifts ($q < 0.8$), empirical risk guarantees fail (0.0670 for gradient boosting vs. 0.1720 for static distance rules); explicit parameter conditioning restores overall risk control (0.0469) and central subgroup control (0.0459), though rare tail regimes remain above α due to small sample sizes. (4) A multi-layer perceptron neural network baseline requires more than twice the training time of gradient boosting (18.32 s vs. 8.61 s) without improving screening skip rates. On an AMD Ryzen 7 5800HS processor, hybrid ML screening reduces per-candidate latency from 6.78 ms (single run) and 20.075 ms (convergence-checked run) to 0.5709 ms for screened candidates, yielding net wall-clock speed-ups of 1.82× and 2.03×, respectively.

1 Introduction

Multi-body systems interacting through long-range magnetic forces and dry frictional contacts exhibit abrupt, history-dependent stability transitions. A practical macroscopic realization is the tabletop game Kluster, in which players take turns placing strongly magnetized stones inside a closed boundary cord. Whenever a newly introduced magnet disturbs the existing arrangement enough to cause two or more magnets to snap into physical contact, the placement fails and the active player must collect the clustered pieces. As the arena fills, the available free area shrinks and superposed dipole fields create intricate mechanical equilibria balanced solely by static friction and boundary normal forces.

Predicting whether a proposed placement will settle safely or trigger a cascade collision is a demanding mechanics problem. Because magnetic dipole-dipole forces scale inversely with the fourth power of separation distance [1, 2] and Coulomb friction introduces a non-smooth discontinuity at zero velocity [3], closed-form analytical stability criteria exist only for an isolated pair of magnets. Once three or more magnets occupy the arena, a magnet that is initially below its static friction limit can be destabilized mid-trajectory by the relaxation of a neighboring stone. Resolving these coupled trajectories requires numerical time-stepping at microsecond resolution ($\Delta t = 10^{-5}$ s) [4]. When evaluating dozens of candidate coordinates per turn across thousands of games in automated heuristic design, direct numerical integration becomes a severe computational bottleneck.

Machine learning (ML) surrogates offer an attractive route to bypass expensive integration by mapping static pre-placement board states directly to collision probabilities [5]. However, raw classifier probabilities lack formal statistical guarantees: thresholded heuristics can silently discard valid, stable placements or fail when surface friction and magnet strength vary. Here, we formulate candidate pre-screening as a distribution-free selective filtering problem using Mondrian Conformal Risk Control (CRC) [6–8], benchmarking physics-informed tree ensembles, linear classifiers, neural networks, and analytical rules across 1,500 simulated games.

2 Problem Objectives and Why the Problem Needs Solving

When an automated planner or Monte Carlo search agent evaluates $K_{\mathbf{cand}}$ candidate placements on a board containing n magnets, it seeks to find a safe (settled) placement while invoking the numerical simulator as few times as possible. If a fast surrogate can reliably skip (discard) candidate placements that are almost certain to collide, only the remaining promising

candidates need to be routed to the full differential equation solver. However, if the filter is overly aggressive, it will erroneously discard truly safe placements, starving the search algorithm of valid moves in crowded arenas. We formalize the screening task around three concrete objectives:

1. Bounded Wrong-Skip Risk: Let $x \in \mathcal{X}$ denote the static pre-placement feature vector of a candidate configuration and $y \in \{0, 1\}$ denote its ground-truth simulation outcome, where $y = 1$ indicates a collision (`contact`) and $y = 0$ indicates a non-colliding outcome (`settled` or `timeout`). Given a contact risk score $s(x) \in \mathbb{R}$ (where higher values signify greater collision risk), the screening rule skips numerical simulation and discards the candidate whenever $s(x) \geq \tau_k$, and routes the candidate to the physics simulator when $s(x) < \tau_k$. A *wrong skip* occurs when a genuinely safe candidate ($y = 0$) is falsely discarded ($s(x) \geq \tau_k$). For a game g having candidate evaluations $\mathcal{J}_{g,k}$ within board-size bin k, let $\mathcal{S}_{g,k} = \{i \in \mathcal{J}_{g,k} : y_i = 0\}$ be the subset of truly safe candidates. The per-game wrong-skip loss is defined as

$$L_{g,k}(\tau_k) = \begin{cases} (1 / |\mathcal{S}_{g,k}|) \sum_{i \in \mathcal{S}_{g,k}} \mathbf{1}\{s(x_i) \geq \tau_k\} & \text{if } |\mathcal{S}_{g,k}| > 0, \text{ and} \\ 0 \quad \text{if } |\mathcal{S}_{g,k}| = 0. \end{cases} \tag{1}$$

Our primary objective is to calibrate τ_k such that the expected per-game loss on unseen test games satisfies $\mathbb{E}[L_{g,k}(\tau_k)] \leq \alpha$ for a user-chosen tolerance $\alpha \in (0, 1)$, while maximizing the fraction of skipped candidates $\Pr(s(x) \geq \tau_k)$.

2. Quantifying Surrogate Utility Against Analytical Physics Rules: Pairwise distance and static-friction heuristics require zero training data. We test whether ML surrogates genuinely outperform analytical physics rules across both sparse ($2 \leq n \leq 4$) and crowded ($5 \leq n \leq 14$) boards, reporting negative results where ML provides no benefit.

3. Robustness Under Physical Parameter Shift: Real magnet sets and tabletop surfaces vary in remanence and friction. We quantify how severely unadapted conformal thresholds degrade under shifted physical parameters, and whether explicit parameter conditioning restores valid risk control.

3 Theory

3.1 Planar Point-Dipole Magnetostatics

We model the game pieces as n identical cylindrical neodymium magnets of radius $R = 5.0$ mm ($D = 2R = 10.0$ mm), thickness $H = 3.0$ mm, volume $V = \pi R^2 H = 2.3562 \times 10^{-7}$ m^3, material density $\rho = 7500$ kg/m^3, and individual mass $m = \rho V = 1.7671$ g. Assuming uniform axial magnetization with nominal remanence $B_r = 1.3$ T and $\mu_0 = 4\pi \times 10^{-7}$ H/m, each disc carries a vertical dipole moment $m_{\mathbf{mag}} = B_r V / \mu_0 = 0.24375$ A m^2. Magnets are restricted to horizontal planar translation at coordinates $\boldsymbol{r}_i = (x_i, y_i) \in \mathbb{R}^2$ with fixed vertical orientation $s_i \in \{+1, -1\}$. Under the point-dipole approximation [1, 2], the interaction potential between magnets i and j separated by distance $r_{ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\|_2$ is

$$U_{ij}(r_{ij}) = s_i s_j \, k_{\mathbf{scale}} \, (\mu_0 / 4\pi) \, (m_{\mathbf{mag}}^2 / r_{ij}^3), \tag{2}$$

where $k_{\mathbf{scale}} > 0$ is a dimensionless force-scaling factor ($k_{\mathbf{scale}} = 1$ nominally). Differentiating U_{ij} with respect to $\boldsymbol{r}_i$ gives the pairwise central force exerted on magnet i by magnet j:

$$\begin{aligned} \boldsymbol{F}_{ij} = -\nabla_{r_i} U_{ij} = (3K \, k_{\mathbf{scale}} \, s_i s_j / r_{ij}^5) \, (\boldsymbol{r}_i - \boldsymbol{r}_j), \qquad K = (\mu_0 / 4\pi) \\ m_{\mathbf{mag}}^2 = 5.9414 \times 10^{-9} \text{ N m}^4. \end{aligned} \tag{3}$$

Parallel moments ($s_i s_j = +1$) repel radially ($\|\boldsymbol{F}_{ij}\|_2 = 3K \, k_{\mathbf{scale}} / r_{ij}^4$), whereas antiparallel moments ($s_i s_j = -1$) attract. The net magnetic force on magnet i is $\boldsymbol{F}_i = \sum_{j \neq i} \boldsymbol{F}_{ij}$, obeying $\sum_{i=1..n} \boldsymbol{F}_i = \boldsymbol{0}$.

3.2 Coulomb Friction, Wall Normal Complementarity, and Critical Distance

Planar motion is resisted by dry Coulomb friction [3] with static coefficient μ_s (nominal 0.30) and kinetic coefficient μ_k (nominal 0.20). With $g = 9.81$ m/s^2, the static and kinetic friction magnitudes are $F_s = \mu_s mg$ (nominal 5.2007 mN) and $F_k = \mu_k mg$ (nominal 3.4671 mN). The arena boundary is modeled as a rigid, frictionless, perfectly inelastic circular wall of radius $R_{\mathbf{wall}}$, restricting centers to $\|\boldsymbol{r}_i\|_2 \leq r_{\mathbf{max}} \equiv R_{\mathbf{wall}} - R$. When a magnet touches the wall ($\|\boldsymbol{r}_i\|_2 \geq r_{\mathbf{max}} - \epsilon_{\mathbf{wall}}$, $\epsilon_{\mathbf{wall}} = 10^{-9}$ m) and experiences an outward radial force $F_{i,n} = \boldsymbol{F}_i \cdot \hat{\boldsymbol{n}}_i > 0$ ($\hat{\boldsymbol{n}}_i = \boldsymbol{r}_i / \|\boldsymbol{r}_i\|_2$), the wall normal reaction balances $F_{i,n}$, leaving static friction to resist only the tangential component:

$$\boldsymbol{F}_i^{\text{eff}} = \begin{cases} \boldsymbol{F}_i - (\boldsymbol{F}_i \cdot \hat{\boldsymbol{n}}_i)\hat{\boldsymbol{n}}_i & \text{if } \|\boldsymbol{r}_i\|_2 \geq r_{\mathbf{max}} - \epsilon_{\mathbf{wall}} \text{ and } \boldsymbol{F}_i \cdot \hat{\boldsymbol{n}}_i > 0; \\ \boldsymbol{F}_i. & \text{else} \end{cases} \tag{4}$$

For an isolated antiparallel pair ($n = 2$, $s_1 s_2 = -1$) away from the wall, attraction overcomes static friction if and only if $3K \, k_{\mathbf{scale}} / r_{12}^4 > F_s$, defining the analytical critical interaction distance r_c and dimensionless shift ratio q:

$$\begin{aligned} r_c(\mu_s, k_{\mathbf{scale}}) = (3K \, k_{\mathbf{scale}} / (\mu_s mg))^{1/4}, \qquad q \equiv r_c(\mu_s, k_{\mathbf{scale}}) / \\ r_{\mathbf{c,default}} = [(k_{\mathbf{scale}} / \mu_s) / (k_{\mathbf{scale},0} / \mu_{s,0})]^{1/4}, \end{aligned} \tag{5}$$

where $r_{\mathbf{c,default}} = 43.033$ mm under nominal parameters ($\mu_{s,0} = 0.30$, $k_{\mathbf{scale},0} = 1.0$).

3.3 Mondrian Conformal Risk Control

Let $\{(\boldsymbol{X}_g, \boldsymbol{Y}_g)\}_{g=1..n_cal}$ be exchangeable calibration games disjoint from the training set. In Mondrian Conformal Risk Control [6–8], candidates are partitioned into board-size bins $k \in \{2\text{-}4, 5\text{-}8, 9\text{-}14\}$ by total magnet count n_{total}. Because $L_{g,k}(\tau) \in [0, 1]$ is monotonically non-increasing in τ, the threshold $\hat{\tau}_k$ for bin k at target risk α is calibrated as

$$\hat{\tau}_k = \inf\{\, \tau \in \mathbb{R} : [n_{\text{bin,k}} / (n_{\text{bin,k}} + 1)]\, \hat{R}_k(\tau) + 1 / (n_{\text{bin,k}} + 1) \leq \alpha \,\}, \quad \hat{R}_k(\tau) = (1 / n_{\text{bin,k}}) \sum_{g=1..n_bin,k} L_{g,k}(\tau), \tag{6}$$

where $n_{\text{bin,k}}$ is the number of calibration games with evaluations in bin k. For an exchangeable test game g_{test}, $\mathbb{E}[L_{g_test,k}(\hat{\tau}_k)] \leq \alpha$, with theoretical finite-sample slack bounded by $1 / (n_{\text{bin,k}} + 1)$. Whenever $\alpha < 1 / (n_{\text{bin,k}} + 1)$, no finite threshold satisfies Eq. (6), and the rule strictly abstains ($\hat{\tau}_k = +\infty$, skip rate = 0, wrong-skip risk = 0).

4 Previous Work by Others

Closed-form expressions for dipole-dipole forces and torques in permanent magnet assemblies were established by Yung et al. [2] and Griffiths [1]. In computational contact mechanics, Stewart [3] formalized complementarity conditions for dry Coulomb stick-slip transitions and inelastic impacts, while Hairer et al. [4] analyzed symplectic and semi-implicit Euler integrators for stiff dynamical systems. To accelerate expensive physical trajectories, physics-informed machine learning (PIML) surrogates have been widely deployed across molecular and continuum dynamics [5]. On structured tabular physical representations, gradient-boosted decision trees [9] implemented in scikit-learn [10] provide strong predictive baselines with low inference latency. To wrap black-box predictors in rigorous statistical guarantees, Vovk et al. [8] introduced conformal prediction and Mondrian taxonomy partitioning, which Bates et al. [7] and Angelopoulos et al. [6] extended to monotonic risk functions via Conformal Risk Control. Recently, Tibshirani et al. [11] and Barber et al. [12] showed that conformal coverage degrades when exchangeability is violated by covariate or parameter shift, motivating our parameter-shift experiments.

5 Method

5.1 Numerical Simulator and Progressive Dataset Generation

The simulator (`src/kluster/simulator.py`) integrates the planar equations of motion using a Numba JIT-compiled [13] semi-implicit Euler scheme with step $\Delta t = 10^{-5}$ s up to $T = 1.0$ s. At each step: (1) Pairwise forces $\boldsymbol{F}_i$ and $r_{\text{min}} = \min_{i<j} r_{ij}$ are accumulated; if $r_{\text{min}} \leq D(1 + \text{TOL}) = 10.5$ mm (TOL = 0.05) at $t > 0$ with $r_{\text{min}} \geq 0.5D$, the trajectory terminates as `contact`. (2) Wall normal projection (Eq. (4)) is applied at r_{max}; if all magnets satisfy $\|\boldsymbol{v}_i\|_2 < 10^{-9}$ m/s and $\|\boldsymbol{F}_i^{\text{eff}}\|_2 \leq F_s$, the trajectory terminates in static equilibrium as `settled`. (3) Moving magnets experience kinetic friction $\boldsymbol{F}_i^{\text{net}} = \boldsymbol{F}_i^{\text{eff}} - F_k(\boldsymbol{v}_i / \|\boldsymbol{v}_i\|_2)$, with exact zero-velocity clamping on stick transitions ($\boldsymbol{v}_i^{\text{new}} \cdot \boldsymbol{v}_i < 0$ and $\|\boldsymbol{F}_i^{\text{eff}}\|_2 \leq F_s$) and inelastic radial clamping at r_{max}. Every candidate is integrated at both $\Delta t = 10^{-5}$ s and $\Delta t/2 = 5 \times 10^{-6}$ s (`label_checked`); disagreements or deep overshoots ($r_{\text{min}} < 0.5D$) are labeled `unresolved`. All 20 analytical unit checks in `tests/test_physics.py` pass.

Using `src/kluster/generate.py`, we generated a fixed-parameter dataset (`kluster_dataset.csv`, 300 games, seed 2026) and our main randomized-parameter dataset (`kluster_dataset_v2.csv`, 1,500 games, 205,856 rows, seed 2027). In `v2`, each game g draws independent parameters via `np.random.default_rng([2027, game_id])`:

$$R_{\text{wall}} \in \{60, 70, 80, 90, 100\}\ \text{mm}, \quad \mu_s \sim \text{U}(0.15, 0.60), \quad \mu_k \sim \mu_s\, \text{U}(0.50, 0.80), \quad k_{\text{scale}} \sim \exp(\text{U}(\ln 0.6, \ln 1.6)). \tag{7}$$

Games start with 1 magnet and progress up to $n_{\text{total}} = 14$. At each turn, $K_{\text{cand}} = 16$ non-overlapping candidates are sampled uniformly inside r_{max} with random polarity $s_{\text{new}} \in \{-1, +1\}$. One `settled` candidate is selected at random to continue the game from its *final relaxed resting coordinates*. Across 205,856 rows in `v2`, 110,303 (53.58%) are `contact`, 95,036 (46.17%) `settled`, 513 (0.25%) `timeout`, and 4 (0.002%) `unresolved` (dropped).

5.2 Feature Engineering, Analytical Baselines, and Partitioning

Features (`src/kluster/features.py`) are extracted strictly at $t = 0$ without trajectory leakage across three nested sets:

- **GEOM (15 features):** `n_total`, `rw_mm`, `s_new`, nearest-neighbor distance `d_nn`, same/opposite-pole distances `d_same` and `d_opp`, local neighbor counts within 40 mm and 60 mm (`n_same_40`, `n_opp_40`, `n_same_60`, `n_opp_60`), `wall_gap`, global minimum pair distance `r_min_all`, `mean_r_centre`, `net_pol`, and `frac_up`.
- **PHYS (18 features):** Adds wall-corrected force ratios `f_ratio_max` = $\max_i \|\boldsymbol{F}_i^{\text{eff}}\|_2 / F_s$, `f_ratio_new` = $\|\boldsymbol{F}_{\text{new}}^{\text{eff}}\|_2 / F_s$, and count `n_over_fs` = $\sum_{i=1..n} \mathbf{1}\{\|\boldsymbol{F}_i^{\text{eff}}\|_2 > F_s\}$.
- **PHYS_PARAMS (20 features):** Adds `rc_mm` = 1000 $r_c(\mu_s, k_{\text{scale}})$ and normalized distance `d_opp_over_rc` = $d_{\text{opp}} / r_c(\mu_s, k_{\text{scale}})$.

Five analytical rules are scored for CRC calibration: (1) **Static-friction rule** (s = `f_ratio_max`); (2) **Distance rule** (s = −`d_opp`); (3) **Critical-distance rule** (s = −`d_opp`, identical rank to Distance rule); (4) **Combined rule** (s = −`d_opp` if `n_over_fs` > 0, else −1.0); and (5) **Parameter-aware critical-distance rule** (s = −`d_opp_over_rc`). We compare these

against `HistGradientBoostingClassifier` (lr = 0.05, depth 8, L_2 = 1.0, 200 trees), `StandardScaler` + `LogisticRegression` (C = 10.0), and `StandardScaler` + `MLPClassifier` (layers (32, 16), early stopping, 60 epochs). Splits are grouped strictly by `game_id`: 50% Train (750 games) / 15% Val (225 games) / 20% Cal (300 games) / 15% Test (225 games) for Tables 1, 2, and 3, and 750 Train / 525 Cal Pool / 225 Test for Table 3.

6 Results

6.1 Mondrian Risk Control and Split Variability (Table T1)

Across 100 repeated game splits on `kluster_dataset_v2.csv`, size-conditional Mondrian CRC bounds the mean wrong-skip rate below α across all bins (Table 1 and Fig. 1). At α = 0.05, Gradient Boosting (+force ratios) achieves a mean wrong-skip rate of 0.0460 (SE = 0.0006) in bin 2–4, 0.0457 (SE = 0.0012) in bin 5–8, and 0.0417 (SE = 0.0024) in bin 9–14. However, **risk control holds on average only**: at α = 0.05, 29% (bin 2–4), 33% (bin 5–8), and 28% (bin 9–14) of individual calibration-test splits exceed α due to finite-sample calibration variance.

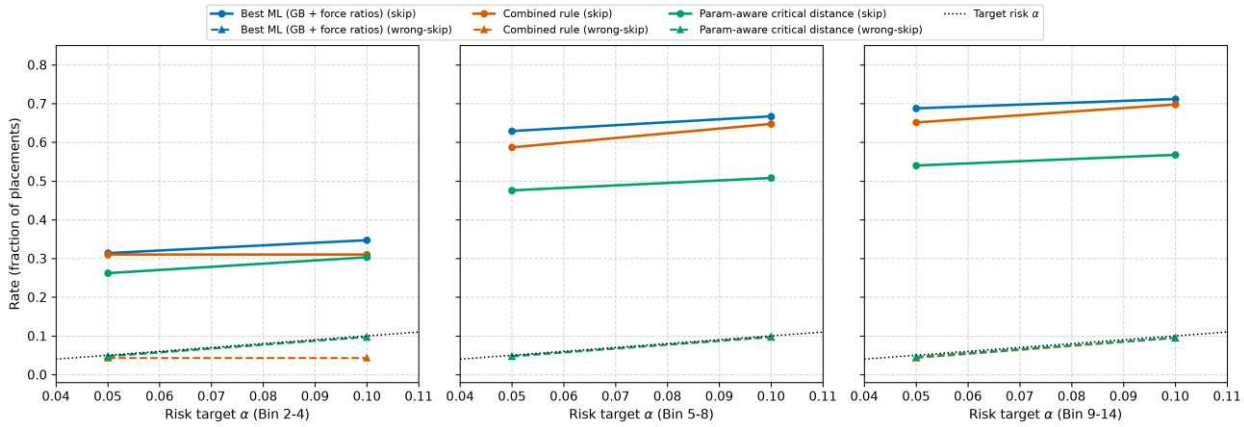


Figure 1: Size-conditional Mondrian CRC performance across board-size bins (`Bin 2-4`, `Bin 5-8`, `Bin 9-14`) at n_{cal} = 320. Solid lines (circles) show skip rates; dashed lines (triangles) show realized within-bin wrong-skip rates vs. target α (dotted line).

Table 1: Confirmatory Mondrian CRC screening risk (T1) and paired skip-rate gains vs. the analytical Combined rule (T2) at α = 0.05 across 100 repeated game splits.

Board Bin	Best ML Model	Realized Risk (SE)	Exceed Frac	ML Skip Rate	Combined Skip Rate	Paired Gain [95% CI]
Bin 2–4	GB (+force ratios)	0.0460 (0.0006)	0.29	0.3254	0.3239	+0.0015 [−0.0010, +0.0056]
Bin 5–8	GB (+force ratios)	0.0457 (0.0012)	0.33	0.6573	0.5992	+0.0581 [+0.0409, +0.0752]
Bin 9–14	GB (+force ratios)	0.0417 (0.0024)	0.28	0.6674	0.6296	+0.0379 [+0.0127, +0.0614]

6.2 Negative Result: Sparse Board Inefficiency vs. Crowded Gains (Table T2)

Table 1 shows that on sparse boards (`Bin 2-4`) at α = 0.05, the paired skip-rate difference between Gradient Boosting (+force ratios) and the Combined rule is +0.0015 (95% CI [−0.0010, +0.0056]). Because the confidence interval spans zero, **there is no statistically significant ML gain on sparse boards**, where two-body attraction within r_c and immediate static friction violation suffice. Conversely, on crowded boards where multi-body superpositions dominate, ML achieves significant paired skip-rate gains of +0.0581 (95% CI [+0.0409, +0.0752]) in bin 5–8 and +0.0379 (95% CI [+0.0127, +0.0614]) in bin 9–14.

6.3 Parameter Shift Breakdown and Subgroup Audit (Table T3)

We evaluate generalization across physical regimes defined by $q = r_c / r_{c,\text{default}}$: the central confirmatory subgroup $0.8 \le q \le 1.25$ (1,377 games, mean 207.35 test games/split) and two exploratory tail subgroups, $q < 0.8$ (91 games, mean 13.35 test games/split) and $q > 1.25$ (32 games, mean 4.30 test games/split). In **Experiment A** (models trained/calibrated on fixed nominal parameters and tested on shifted `v2` subgroups), risk control collapses in the shifted tails (Table 2 and Fig. 2): at α = 0.05 in $q < 0.8$, unadapted distance rules fail at 0.1720 (SE = 0.0017) and unadapted Gradient Boosting (+force ratios) reaches 0.0670 (SE = 0.0032), while in $q > 1.25$ the unadapted `Static-friction rule` reaches ≈ 0.120.

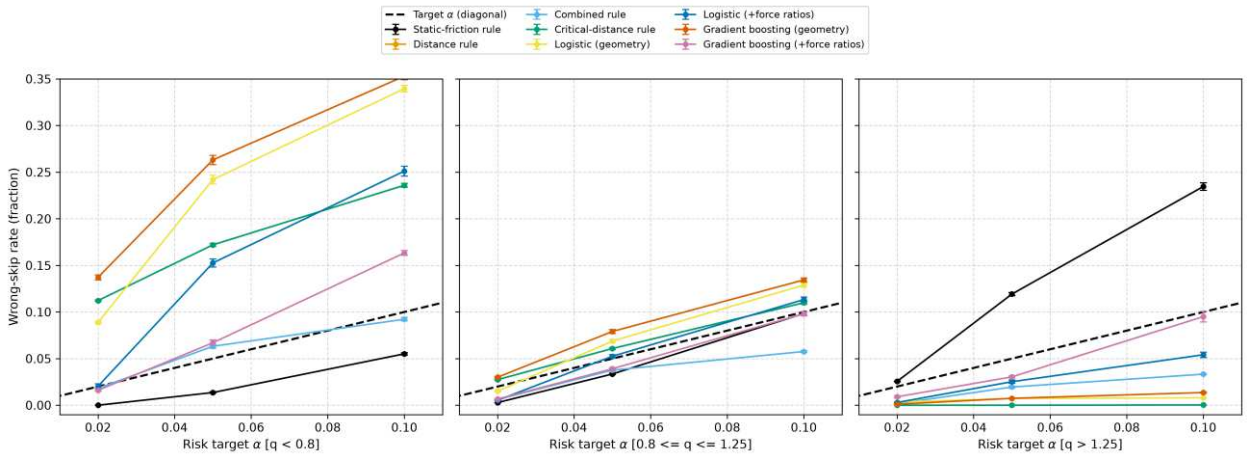


Figure 2: Breakdown of conformal risk guarantees under unadapted parameter shift (Experiment A) across $q < 0.8$ (left), $0.8 \le q \le 1.25$ (center), and $q > 1.25$ (right). Error bars denote ±1 SE; the dashed diagonal marks target α.

Table 2: Parameter-shift wrong-skip risk at $\alpha = 0.05$ across 20 splits (Table T3). Asterisks (*) denote exploratory subgroups (< 30 test games/split) exceeding α.

Experiment	Model / Screening Rule	Overall Risk (SE)	q < 0.8 Risk (SE)	0.8 ≤ q ≤ 1.25 Risk (SE)	q > 1.25 Risk (SE)
Exp A (Fixed → Shift)	Distance / Critical-Distance Rule	0.0894 (0.0008)	0.1720 (0.0017)	0.0469 (0.0008)	0.0041 (0.0006)
Exp A (Fixed → Shift)	Gradient Boosting (+force ratios)	0.0526 (0.0008)	0.0670 (0.0032)	0.0467 (0.0009)	0.0162 (0.0022)
Exp B (Random Trained)	Gradient Boosting (+parameters)	0.0469 (0.0012)	0.0548 (0.0038)*	0.0459 (0.0010)	0.0741 (0.0115)*
Exp C (Param Heuristic)	Param-Aware Critical-Distance Rule	0.0474 (0.0008)	0.0502 (0.0021)*	0.0469 (0.0008)	0.0605 (0.0052)*

In **Experiment B** (models trained/calibrated on `v2` with `PHYS_PARAMS`) and **Experiment C** (the Parameter-aware critical-distance rule using $-d_{opp} / r_c$), overall risk control (0.0469 and 0.0474) and central subgroup control (0.0459 and 0.0469) are restored. However, because calibration is conditioned on board size rather than q, rare exploratory tails still exceed α (0.0548 in $q < 0.8$; 0.0741 in $q > 1.25$). A 100-split audit (`scripts/audit_static_paramshift.py`) confirmed that the `Static-friction rule` has 100% distinct monotonic scores and controls risk within every bin (≤ 0.05) and under pooled calibration (≤ 0.05).

6.4 Calibration Sample Size Scaling and Neural Baselines (Tables T4 & T5)

Table 3 and Fig. 3 evaluate calibration scaling across $n_{cal} \in \{20, 40, 80, 160, 320\}$. Continuous-score models closely track the theoretical slack curve $1 / (n_{cal} + 1)$, whereas two negative results stand out: (1) at $n_{bin,k} = 20$ and $\alpha = 0.02$, $1 / (20 + 1) = 0.0476 > 0.02$, forcing strict abstention ($\hat{\tau}_k = +\infty$, skip rate = 0); and (2) at $\alpha = 0.10$ in bin 2–4, the discrete `Combined rule` plateaus at a constant risk gap of ≈ 0.0567 (realized risk ≈ 0.0433) because its `n_over_fs` > 0 floor prevents generating higher risk. In architectural comparisons (Table 3 and Fig. 4), the MLP (32, 16) neural network takes over twice as long to train as Gradient Boosting (18.32 s vs. 8.61 s) without improving skip rates in any bin (0.6242 vs. 0.6319 in bin 5–8 at $\alpha = 0.05$).

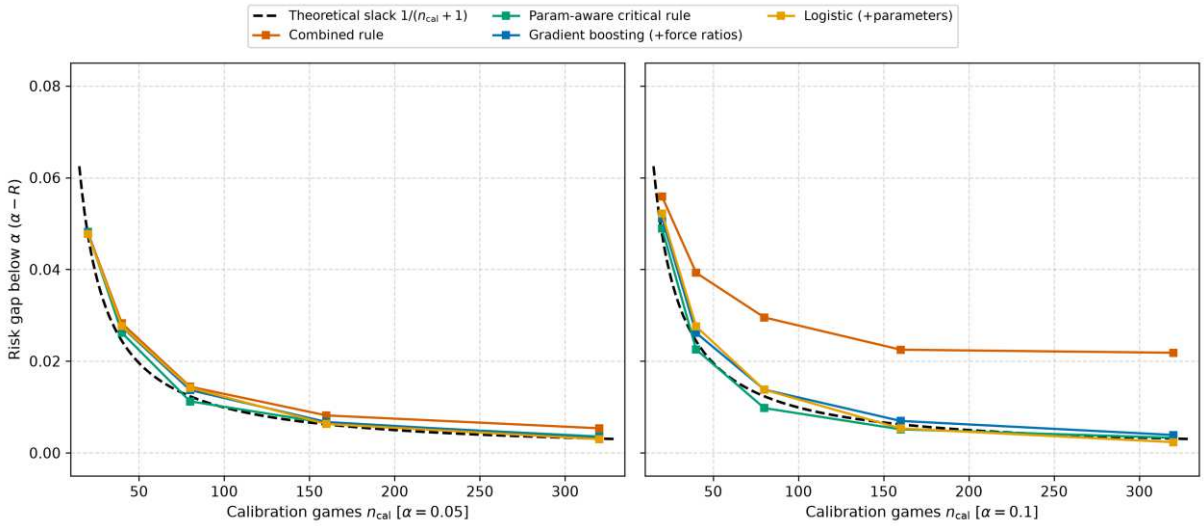


Figure 3: Empirical risk gap below α ($\alpha - R$) vs. calibration games $n_{\text{cal}} \in \{20, 40, 80, 160, 320\}$ at $\alpha = 0.05$ (left) and $\alpha = 0.10$ (right), alongside theoretical slack $1 / (n_{\text{cal}} + 1)$ (dashed).

Table 3: Calibration size scaling in Bin 5–8 at $\alpha = 0.05$ (T4, left) and Neural Network vs. GBDT/Logistic comparison at $\alpha = 0.05$ (T5, right).

n_cal (T4: Bin 5–8, GB+force)	Risk (SE)	Gap / Theory	Skip Rate	Model (T5: α = 0.05)	Bin	Skip (SE)	Bin Risk	Train (s)
20 (α = 0.02)	0.0000	0.0200 / 0.0476	0.0000 (Abstain)	GB (+params)	2–4	0.3126 (0.0018)	0.0456	8.61
20 (α = 0.05)	0.0020 (0.0006)	0.0480 / 0.0477	0.3712	GB (+params)	5–8	0.6319 (0.0028)	0.0459	8.61
40 (α = 0.05)	0.0240 (0.0018)	0.0260 / 0.0244	0.5665	GB (+params)	9–14	0.6834 (0.0041)	0.0442	8.61
80 (α = 0.05)	0.0365 (0.0019)	0.0135 / 0.0124	0.6085	MLP (32, 16)	2–4	0.3128 (0.0019)	0.0458	18.32
160 (α = 0.05)	0.0434 (0.0016)	0.0066 / 0.0062	0.6227	MLP (32, 16)	5–8	0.6242 (0.0031)	0.0462	18.32
320 (α = 0.05)	0.0465 (0.0012)	0.0035 / 0.0031	0.6290	MLP (32, 16)	9–14	0.6790 (0.0044)	0.0448	18.32
320 (Combined)	0.0447 (0.0013)	0.0053 / 0.0031	0.5872	Logistic (+p)	5–8	0.6001 (0.0026)	0.0461	0.647

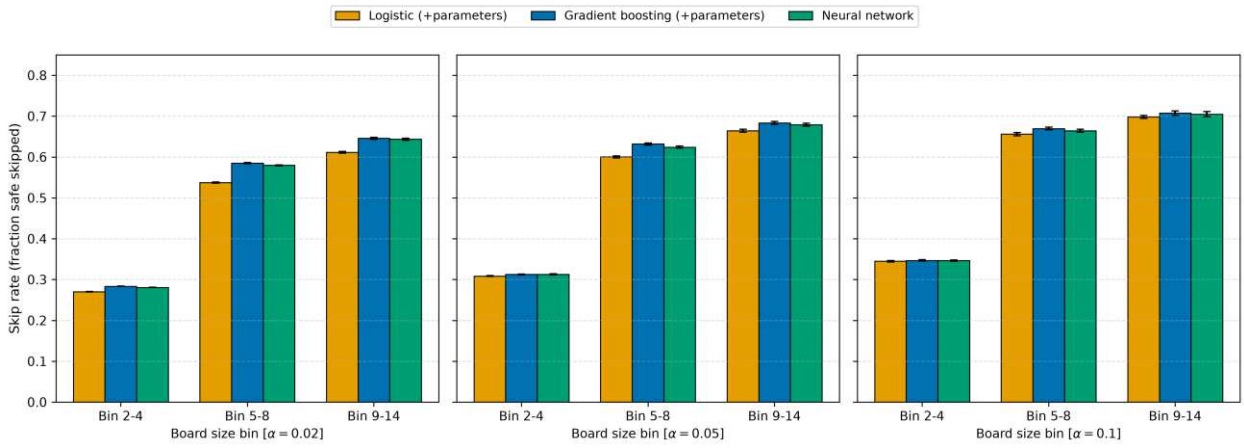


Figure 4: Safe-calibrated skip rates for Logistic (+parameters), Gradient Boosting (+parameters), and Neural Network across board-size bins at $\alpha \in \{0.02, 0.05, 0.10\}$ (Table T5).

6.5 Computational Latency and Hybrid Speed-Up Benchmarks (Figure F5)

Single-threaded runtime profiling on an AMD Ryzen 7 5800HS CPU (8C/16T, 15.4 GB RAM) shows that a single simulation run requires $t_{\text{sim,single}} = 6.780$ ms (6.78 ms) and a convergence-checked run requires $t_{\text{sim,checked}} = 20.075$ ms (Fig. 5). In contrast, feature extraction takes $t_{\text{feat}} = 0.562$ ms and Gradient Boosting inference takes $t_{\text{infer}} = 0.0089$ ms ($t_{\text{eval}} = 0.5709$ ms total). With expected latency $t_{\text{screened}} = t_{\text{eval}} + (1 - s_{\text{frac}})t_{\text{sim}}$ ($s_{\text{frac}} \approx 0.535$ at $\alpha = 0.05$), hybrid screening yields net wall-clock speed-ups of 1.82× (single-run) and 2.03× (checked-run).

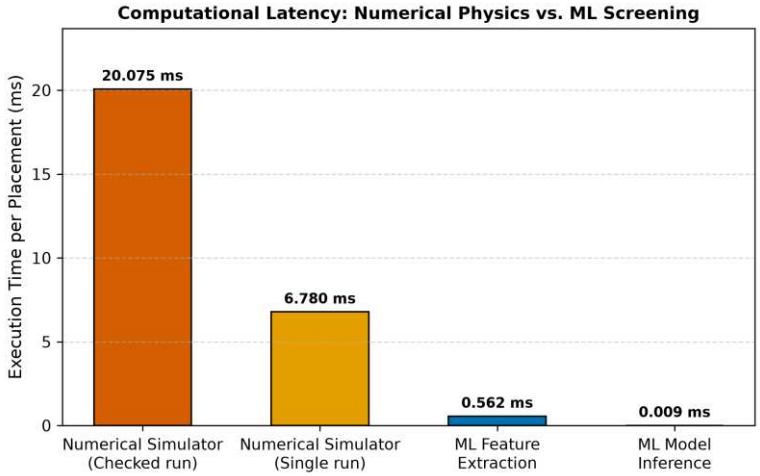


Figure 5: Per-placement execution latency comparing checked (20.075 ms) and single-run (6.780 ms) physics simulation against ML feature extraction (0.562 ms) and inference (0.009 ms).

7 Conclusion

Mondrian Conformal Risk Control paired with physics-informed Gradient Boosting *partially solves* the magnet stability screening problem in Kluster: (1) On crowded boards ($5 \leq n_{\text{total}} \leq 14$), ML outperforms analytical heuristics by +3.79 to +5.81 percentage points in skip rate, delivering a 1.82× to 2.03× wall-clock speed-up while bounding mean wrong-skip risk below $\alpha = 0.05$ (0.0457 in bin 5–8) and training 2.1× faster than an MLP neural network. (2) Conversely, ML offers no significant gain on sparse boards ($n_{\text{total}} \leq 4$, +0.0015), where the analytical Combined rule suffices; furthermore, CRC bounds risk in expectation only (33% of splits exceed α), abstains when $n_{\text{bin,k}} < \alpha^{-1} - 1$, and requires explicit parameter conditioning (d_{opp} / r_c) under environmental parameter shifts.

8 Limitations

Ten explicit physical and statistical boundaries govern this study: **(1) Point-Dipole Idealization:** Finite-size multipole corrections near contact ($r_{ij} \to D$) are omitted. **(2) Fixed-Vertical Moments:** Dipole moments are locked vertically ($s_i = \pm 1$), excluding 3D magnetic torques, tipping, or pole-flipping. **(3) Planar 2D Translation:** Out-of-plane liftoff and rotational tumbling are excluded. **(4) Rigid Circular Boundary:** The boundary is a rigid, frictionless, inelastic circle rather than a deformable fabric cord. **(5) Uniform Cylindrical Geometry:** Commercial Kluster stones are irregular pebbles rather than identical discs. **(6) Simulation-Derived Labels:** Ground truth comes from numerical ODE integration rather than physical experiments. **(7) Finite Horizon ($T = 1.0$ s):** `timeout` (0.25%) indicates motion at 1.0 s, not infinite-time instability. **(8) Illustrative Constants:** Nominal parameters ($B_r = 1.3$ T, $\rho = 7500$ kg/m^3, $\mu_s = 0.30$, $\mu_k = 0.20$) are uncalibrated to specific hardware. **(9) Uniform Sampling:** Candidates are sampled uniformly rather than by human tactics. **(10) Expectation-Only Guarantees:** CRC bounds expected risk across exchangeable splits, not per-split or rare-tail ($q < 0.8$, $q > 1.25$) realizations.

Declarations

Public Review and Online Upload Declaration: The author (Sankalp Savarn, Enrolment No. 24123038) declares no objection to the online upload and public review of this preprint article, the VideoPPT presentation, and the source code repository on public platforms.

Data and Code Availability: All code, 20-check unit tests (`tests/test_physics.py`), environment pins (`requirements-tested.txt`), and split datasets (`train_games_v2.csv`, `test_games_v2.csv`) are provided in the submission archive. https://github.com/mggm1982/ML-Based-Stability-Prediction-for-Magnet-Packing-in-the-Game-Kluster.git

Video: https://youtu.be/rYwu2FHvjfQ

References

[1] David J. Griffiths. Introduction to Electrodynamics. Cambridge University Press, Cambridge, UK, 4th edition, 2017.

[2] K. W. Yung, P. B. Landecker, and D. D. Villani. An analytic solution for the force between two magnetic dipoles. Magnetic and Electrical Separation, 9(1):39–52, 1998.

[3] David E. Stewart. Rigid-body dynamics with friction and impact. SIAM Review, 42(1):3–39, 2000.

[4] Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Springer, Berlin, Germany, 2nd edition, 2006.

[5] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. Nature Reviews Physics, 3(6):422–440, 2021.

[6] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2024. arXiv:2208.02814.

[7] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. Journal of the ACM, 68(6):1–34, 2021.

[8] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic Learning in a Random World. Springer, New York, NY, USA, 2005.

[9] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5):1189–1232, 2001.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[11] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, pages 2530–2540, 2019.

[12] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2):816–845, 2023.

[13] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A LLVM-based Python JIT compiler. In Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC (LLVM '15), pages 1–6, 2015.