Transactions on Computing Science

https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2

Research Article

N-Body Simulation and Machine Learning-Based Stability Prediction for Multi-Planet Exoplanetary Systems

Zixi Xu[1,*]

[1]Dana Hall School, Wellesley, MA 02482, USA

[*]**Correspondence:** Zixi Xu, lilian.xu.0409@outlook.com

Abstract

We present a reproducible pipeline that combines a fast batched symplectic N-body integrator, a renormalized two-particle MEGNO chaos indicator, and supervised machine learning to predict the 10^4-orbit stability of compact multi-planet systems from only the first 10^2-10^3 orbits. Using 3200 coplanar three-planet systems sampled across mutual-Hill spacings $\Delta \in [3,13]$ and initial eccentricities $e \in [0,0.1]$, we first map how instability fraction rises sharply as Δ decreases. We then show that a histogram-gradient-boosting classifier using twelve physics-inspired features achieves ROC-AUC 0.976 ± 0.002 and balanced accuracy 0.909 from a 10^3-orbit window, outperforming the MEGNO-threshold (AUC 0.671) and a spacing-threshold baseline (AUC 0.904). A label-leakage control that excludes all systems destabilizing inside the feature window still yields AUC 0.952 from 100 orbits, an ablation study finds that 100 orbits already contain most predictive information (AUC 0.971), and transfer to an independent 1600-system four-planet sample retains AUC 0.972 with a well-calibrated Brier score of 0.077. An error analysis shows that the residual failures concentrate within about one mutual Hill radius of the stability transition and are dominated by late, diffusion-driven instabilities. These results suggest that short-run machine-learning surrogates can substantially accelerate stability surveys of exoplanetary systems.

Keywords

Exoplanet dynamics, N-body simulations, Symplectic integrators, MEGNO; Machine learning, Stability prediction

1. Introduction

Multi-planet systems are a defining outcome of planet formation, yet their long-term orbital stability is governed by weakly chaotic, non-linear gravitational interactions that are difficult to predict analytically (Doty et al., 2025). Direct N-body integrations can follow individual systems for gigayears, but the computational cost scales linearly with the integration time and precludes rapid exploration of broad parameter spaces. Recent work has therefore turned to machine learning (ML) as a surrogate model: Tamayo et al. (2020) showed that summary statistics from a 10^4-orbit integration can predict stability over 10^9 orbits, and Gajdoš & Vaňko (2023) found that the SPOCK classifier compares favorably with the MEGNO chaos indicator for real exoplanet systems.

The present paper asks a complementary, more restrictive question: how much of the 10^4-orbit outcome can be learned from the first 10^2-10^3 orbits? A positive answer would let observers and theorists flag candidate unstable architectures almost as soon as they are generated, at roughly a tenfold speed-up over the full integration. To answer it we build a fully reproducible pipeline in Python/Numba, integrate 4800 coplanar systems (3200 three-planet and 1600 four-planet), and compare three ML models against two physics baselines.

The operational setting matters for how such a surrogate would be used. When a single, well-observed system is at stake, one can afford a long, high-accuracy integration with an adaptive integrator. When instead a population of candidate architectures must be screened—as in formation modeling, stability-constrained inference of orbital parameters, or ensemble sensitivity studies—the cost per system multiplies by the ensemble size, and any reduction in the required integration window translates directly into a proportional enlargement of the explorable parameter space. Observed compact multi-planet systems frequently have spacings of only a few mutual Hill radii, so the relevant question is not whether a sharp stability edge exists but how close a given configuration sits to the diffuse, probabilistic transition zone around it. A classifier that outputs calibrated probabilities, rather than a binary verdict, is therefore the natural tool for this regime.

Our contributions are: (i) a vectorized symplectic integrator with a renormalized two-particle MEGNO implementation suitable for thousands of compact systems; (ii) an ML feature set that is independent of planet count and reaches AUC 0.976 from only 10^3 orbits; (iii) a label-leakage control showing that the short-window forecasts remain accurate (AUC 0.952 from 100 orbits) even when systems destabilizing inside the window are excluded; and (iv) evidence that the same model transfers across planetary multiplicity with negligible loss.

The remainder of the paper is organized as follows. Section 2 situates our operating point relative to analytical stability theory, chaos indicators, and existing ML surrogates. Section 3 describes the symplectic integrator, the renormalized MEGNO indicator, and the twelve-feature representation. Section 4 presents the four experimental studies: the stability grid survey, the classification benchmark with an error analysis, the window ablation with cross-architecture transfer, and the leakage-control and convergence checks. Section 5 discusses the physical interpretation, practical implications, and limitations, and Section 6 concludes.

2. Related Work

Analytical studies have clarified the limits of two-planet stability (Gladman, 1993) and the importance of three-body resonances in packed systems (Petit et al., 2020). Because these theories do not extend to eccentric, multi-planet architectures in closed form, numerical integration remains the standard. Symplectic maps for the N-body problem, introduced by Wisdom & Holman (1991), allow efficient long-term integrations when planets are well separated.

Chaos indicators such as MEGNO quantify the exponential divergence of nearby trajectories and are routinely used to distinguish regular from chaotic motion (Cincotta & Simó, 2000). However, MEGNO requires integrating a shadow or variational trajectory alongside the nominal one, and its short-time value can be ambiguous for weakly chaotic systems. ML stability classifiers, including SPOCK and its Bayesian extension SPOCK II (Cranmer et al., 2021), sidestep this ambiguity by training on summary statistics. They work well for compact systems but typically rely on 10^4-orbit integrations and fixed three-planet architectures. Our work shortens the required integration window and explicitly tests generalization to four-planet systems.

Viewed as points in the (integration window, prediction horizon) plane, the available tools occupy different regimes. Chaos indicators such as MEGNO effectively use the shortest possible window and express their verdict as a divergence rate, but, as our grid survey shows, a thresholded short-run MEGNO value conflates genuine chaos with transient behavior. SPOCK-style classifiers integrate for 10^4 orbits and forecast survival over 10^9 orbits, an extrapolation of five orders of magnitude that is justified by the saturation of instability statistics but is expensive per system. The present work maps the intermediate regime—windows of 10^2-10^3 orbits predicting outcomes over 10^4 orbits—with an explicit label-leakage control and a cross-multiplicity transfer test, two evaluations that we argue should accompany any short-window stability claim. Throughout, we prioritize reproducibility: the entire pipeline is roughly a thousand lines of Python/Numba with fixed random seeds, and every number in this paper can be regenerated from the released scripts.

3. Methods

3.1. Integrator and initial conditions

We adopt units with $G = M_* = a_1 = 1$ so the innermost planet has period $P_1 = 2\pi$. All planets have mass 5 M⊕ and are coplanar. The mutual Hill radius for adjacent planets is

$$R_{H,m}^{(i)} = \frac{a_i + a_{i+1}}{2}(m_i + m_{i+1})/(3M_*)^{1/3} \quad (1)$$

where the semimajor axes are built by fixed-point iteration so that Δ_i = (a_{i+1}-a_i)/R_{H,m}^{(i)}. Eccentricities e and angles ω, M are drawn uniformly. The equations of motion are integrated with a second-order leapfrog (velocity-Verlet) symplectic scheme

$$v_{n+1/2} = v_n + \frac{\Delta t}{2} a(r_n), r_{n+1} = r_n + \Delta t v_{n+1/2}, \tag{2}$$

$$v_{n+1} = v_{n+1/2} + \frac{\Delta t}{2} a(r_{n+1}). \tag{3}$$

The kernel is JIT-compiled with Numba and parallelized over systems with prange. We use a fixed step Δt = P1/150 ≈ 0.0419; the median relative energy drift over the 10^4-orbit integrations is 1.4e-05, with larger excursions confined to systems undergoing close encounters. A system is flagged unstable when any pair of planets approaches within one mutual Hill radius or when a planet exceeds 100 a1 from the star.

The choice of a second-order leapfrog rather than a higher-order or adaptive scheme is deliberate. Leapfrog is symplectic and time-reversible, so its energy error remains bounded and oscillatory rather than drifting secularly, a property that matters when 10^4 orbits must be integrated for thousands of systems at negligible per-system cost. Its second-order accuracy is sufficient here because the prediction target is the occurrence and time of an instability event, not the precise trajectory; our convergence check in Section 4.5 quantifies how label quality responds to a four-fold refinement of the step. The fixed step is also what makes batching and vectorization over systems possible, since all systems can be advanced in lockstep.

The sampling ranges Δ∈[3,13] and e0∈[0,0.1] were chosen to bracket the instability transition rather than to mimic any particular observed population. Classical two-planet Hill-stability arguments require only a few mutual Hill radii for circular orbits, whereas long-term numerical experiments on multi-planet chains show that survival over many orbits typically demands considerably wider separations; our interval spans both regimes. With these ranges the ensemble splits into roughly stable and unstable halves, which keeps the classification task balanced (the realized unstable fraction is 0.421 stable versus 0.579 unstable) and avoids the degenerate case in which a trivial classifier scores well by predicting the majority class.

3.2. Two-particle MEGNO indicator

We evolve a shadow trajectory offset from the nominal orbit by an initially tangent phase-space vector u = (δr, δv) with |u0| = 10^-8, and renormalize when |u| exceeds 10^-4. The MEGNO estimator is

$$Y(t) = \frac{2}{t} \int_0^t s \cdot \frac{u(s) \cdot \dot{u}(s)}{|u(s)|^2} ds, \tag{4}$$

where the dot product is taken over the full phase-space offset. In the discrete form, w = (δr·δv +

$\delta v \cdot \delta a)/(|\delta r|^2+|\delta v|^2)$ and Z accumulates as $Z \mathrel{+}= w \cdot t \cdot \Delta t$, giving $Y = 2Z/t$. For quasi-periodic orbits $Y \to 2$, while for chaotic orbits Y grows linearly with the maximum Lyapunov exponent.

Two implementation details deserve comment. First, the shadow offset is initialized tangent to the flow rather than in an arbitrary direction. A radial offset initially mixes the displacement with the steeply varying gravitational acceleration scale and produces a transient in Y that reflects the choice of offset rather than the intrinsic orbit structure; a tangent offset avoids this artifact and lets Y settle onto its asymptotic behavior within the first few orbits. Second, renormalization rescales the offset back to |u| = 10^-8 whenever it exceeds 10^-4. Rescaling preserves the direction of the offset while resetting only its amplitude, so the divergence information accumulated in Z is unaffected; without renormalization the shadow trajectory would leave the linear regime and contaminate the estimate. Together these choices yield a MEGNO implementation that remains stable over the full 10^4-orbit horizon at no additional cost beyond one extra set of phase-space variables per system.

3.3. Features and machine-learning models

From each short integration window we extract twelve features that do not depend on the number of planets, so that a model trained on three-planet systems can be applied unchanged to four-planet systems. They fall into four groups: initial geometry (Δmin, Δmean, e0 mean/max), chaos indicators from the MEGNO trajectory (Y at window end, Ymax), orbital diffusion statistics (maximum and mean eccentricity excursion, maximum relative semimajor-axis drift, AMD variation, eccentricity at window end), and the closest pairwise approach recorded during the window (dmin/RH). Table 1 lists the definitions. Systems that destabilize inside the window contribute their last recorded pre-event values, mirroring the operational early-warning setting in which a running forecast must be issued from whatever data exist at the moment. The label is whether the system becomes unstable within 10^4 orbits.

Table 1

The twelve planet-count-agnostic features extracted from each integration window

Feature	Definition	Group
Δmin	minimum pairwise spacing in mutual Hill radii	initial geometry
Δmean	mean pairwise spacing in mutual Hill radii	initial geometry
e0_mean	mean initial eccentricity	initial geometry
e0_max	maximum initial eccentricity	initial geometry
Y_end	MEGNO value at the window end	chaos indicator
Y_max	maximum MEGNO value within the window	chaos indicator
de_max	maximum per-planet eccentricity excursion in the window	orbital diffusion

WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2

de_mean	mean per-planet eccentricity excursion in the window	orbital diffusion
log10 AMD var	log10 of the AMD range within the window	orbital diffusion
dmin/RH	closest pairwise approach in the window, in mutual Hill radii	close approaches
da_rel_max	maximum relative semimajor-axis drift in the window	orbital diffusion
e_end_max	maximum eccentricity at the window end	orbital diffusion

We compare three supervised classifiers: logistic regression with standardized features, random forest, and histogram gradient boosting (HGB). Hyperparameters are kept close to library defaults (random forest: 400 trees, min_samples_leaf = 5; HGB: 300 iterations, learning rate 0.1; logistic regression: C = 1). Two physics baselines are included for reference: a MEGNO threshold $Y > 2$ at the window end, and a Δmin threshold tuned on the training split. All reported scores are means and standard deviations over five 70/30 stratified splits.

3.4. Reproducibility

The complete pipeline is roughly a thousand lines of Python. The integrator and MEGNO indicator are JIT-compiled with Numba; the classifiers, baselines, and metrics come from scikit-learn; figures are produced with Matplotlib. The experiments reported here were run with Python 3.13, NumPy 2.5, Numba 0.67, and scikit-learn 1.9 on a single desktop machine, with no GPU or cluster resources required. Each experiment uses a fixed random seed for both the sampling of initial conditions and the train/test splits (seeds 20260902, 20260903, 20260904, 20260904, and 20260905 for Experiments 1-5, respectively), so every table and figure in this paper can be regenerated bit-for-bit from the released scripts. The raw trajectory samples and extracted feature matrices are a few hundred megabytes and are retained alongside the code.

4. Experiments and Results

4.1. Stability grid survey

We first integrate 3200 three-planet systems with $\Delta\text{min} \in [3,13]$ and $e0 \in [0,0.1]$. The overall unstable fraction is 0.579. Figure 1 shows the instability-fraction map in the (Δmin, e0) plane: instability rises sharply toward small spacings and large eccentricities. A logistic fit gives $\Delta50 \approx 6.74$ and 10.21 mutual Hill radii for e0 = 0.05 and 0.10, respectively.

Figure 1

Instability fraction within 10^4 orbits as a function of minimum mutual-Hill spacing Δmin and mean initial eccentricity e0. Contours mark 0.25, 0.50, and 0.75

https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2

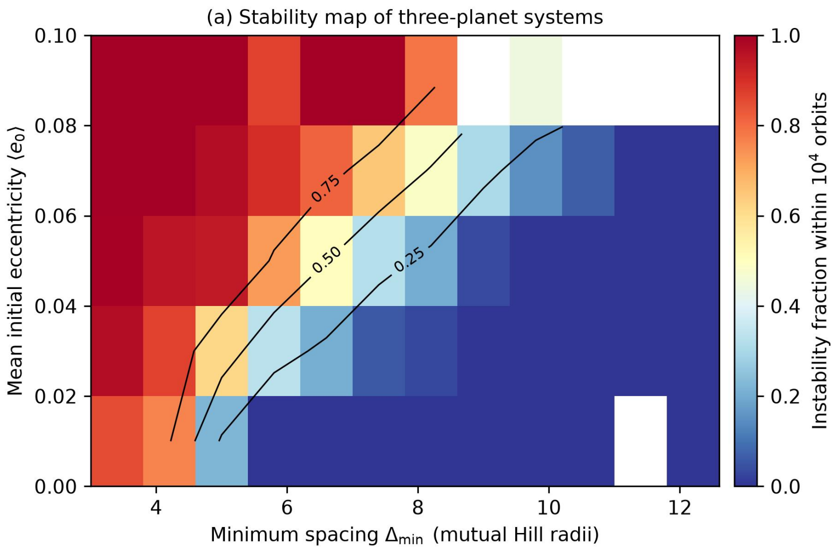
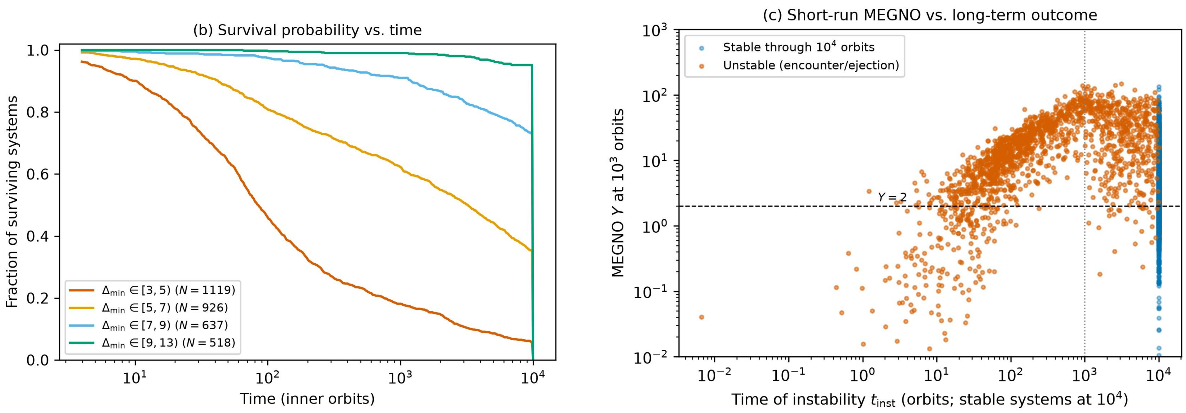


Figure 2a gives the survival probability as a function of time for four Δmin bins. Tightly packed systems (Δmin < 5) lose more than 90% of their members within a few hundred orbits; systems with Δmin∈[9,13] are almost entirely stable over the same horizon. Figure 2b shows the short-run MEGNO value at 10^3 orbits versus the final instability time. Early-unstable systems occupy the upper right, but many stable systems also have Y > 2, indicating that a simple Y-threshold misses slowly diffusing instabilities.

Figure 2

Early dynamical signatures of instability: (a) survival probability versus time for four Δmin bins; (b) short-run MEGNO value at 10^3 orbits against the final instability time



Two features of Figure 2 shape the rest of the study. First, the survival curves are far from single exponential decays: the tightest bin loses most of its members within the first few hundred orbits and then flattens, indicating that a subpopulation is destabilized almost immediately by first-encounter-scale dynamics while the remainder survives on a much slower diffusive timescale. Any predictor built from a short window must therefore capture two distinct physical mechanisms. Second, the MEGNO scatter shows why a thresholded chaos indicator is a weak baseline in this regime: the populations with

Y(10^3) above and below 2 overlap strongly in instability time, so no single cut separates them. The information needed for accurate classification is spread across several weak indicators rather than concentrated in one, which is precisely the setting in which supervised multivariate models are expected to help.

The measured transition spacings are also instructive when set against the two-planet Hill criterion, which guarantees stability of a circular pair only above $2\sqrt{3} \approx 3.46$ mutual Hill radii. Our three-planet Δ50 lies 3.3 radii above that floor at e0 = 0.05 and 6.8 radii above it at e0 = 0.10, confirming that a chain of planets needs markedly more room than any adjacent pair in isolation. The strong dependence of Δ50 on eccentricity reflects the geometric fact that free eccentricity brings apocenters and pericenters closer together, effectively eroding the clearance between neighboring orbits; at e0 = 0.10 the transition has moved so far that most of our sampled range becomes unstable. Both trends—multiplicity inflating the required spacing, and eccentricity inflating it further—are the qualitative signatures expected from secular resonance overlap arguments, here measured directly on a 3200-system grid.

4.2. Machine-learning prediction from 10^3 orbits

Table 2 reports the classification performance on the 3200-system dataset. HGB and random forest both achieve AUC ≈ 0.97–0.98, while MEGNO alone only reaches AUC 0.671. The Δmin-threshold baseline is competitive (AUC 0.904) because our parameter space is dominated by geometry, but it cannot capture the additional information in eccentricity growth and close-approach history.

Table 2

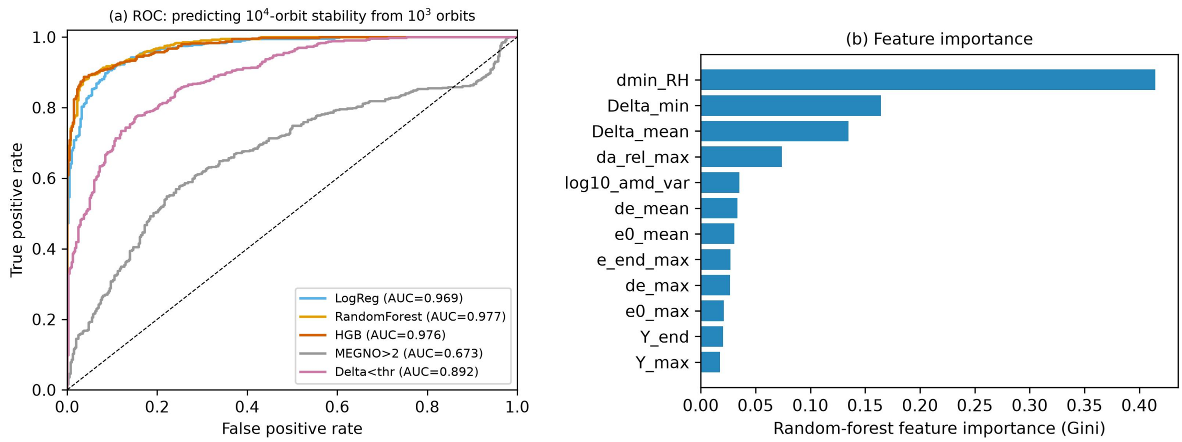
10^4-orbit stability prediction from a 10^3-orbit window (mean ± std over 5 splits)

Method	ROC-AUC	Bal. Acc.	F1	Accuracy
LogReg	0.973±0.003	0.912±0.007	0.926±0.007	0.914±0.007
RandomForest	0.979±0.002	0.916±0.006	0.927±0.004	0.916±0.005
HGB	0.976±0.002	0.909±0.004	0.922±0.003	0.910±0.004
MEGNO>2	0.671±0.006	0.586±0.002	0.708±0.004	0.620±0.003
Delta<thr	0.904±0.008	0.814±0.011	0.849±0.013	0.821±0.013

Figure 3a shows the ROC curves. The ML curves hug the upper-left corner, while the MEGNO baseline is far closer to the diagonal. Figure 3b gives the random-forest feature importances: the closest relative Hill approach dmin/RH dominates, followed by initial spacing and semimajor-axis diffusion. Notably, the MEGNO features Yend and Ymax rank lowest, confirming that a single global chaos indicator is less informative than the full early orbital history.

Figure 3

Classifier performance on the 3200-system three-planet dataset: (a) ROC curves for the three ML models and two baselines; (b) random-forest feature importances



Three observations follow from Table 2 and Figure 3. First, logistic regression trails HGB by less than 0.01 AUC despite its simplicity, indicating that the stability boundary in this twelve-dimensional feature space is close to linear in the neighborhood of the transition; the tree ensembles earn their small advantage on the curved flanks of the boundary, where interactions between geometry and diffusion matter. Second, the ROC curves of all three ML models dominate both baselines over the entire operating range, so the improvement does not depend on threshold tuning. Third, the balanced accuracies (≈0.91) are markedly lower than an AUC of 0.98 might suggest: with a 0.5 probability threshold, every error type identified in Section 4.3—transition-zone architectures and near-miss survivors—contributes, and the gap between the two metrics is itself a measure of how diffuse the boundary region is.

4.3. Where the classifier fails

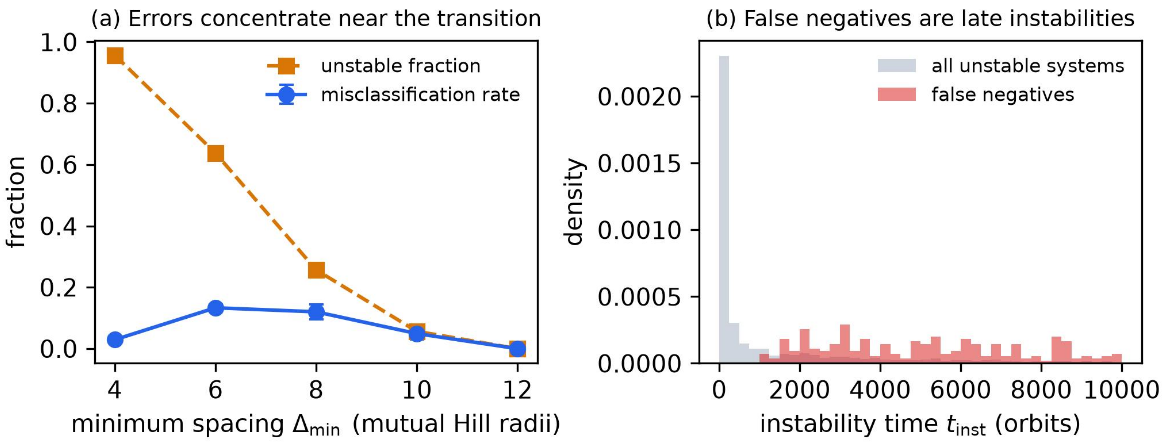
Aggregate scores say little about failure modes, so we examined the errors of the 10^3-orbit HGB classifier directly (Figure 4). Misclassification is strongly localized in the parameter space: pooled over five splits, the error rate is 13.3% in the Δmin bin [5,7) and 12.0% in [7,9), but only 2.9% for Δmin < 5 and zero for Δmin > 11. In other words, systems deep inside either regime are classified almost perfectly, and the residual errors straddle the stability transition. Consistently, misclassified systems lie a mean 0.98 mutual Hill radii from the logistic Δ50(e0) boundary fit of Section 4.1, versus 2.35 for correctly classified systems: the errors are architectures whose fate is genuinely marginal rather than model pathologies spread uniformly over the space.

The two error types have distinct physical signatures. The 219 false negatives (unstable but predicted stable) are dominated by late instabilities: their median instability time is 4939 orbits, against 151 for unstable systems overall, and 49% of them destabilize only after 5×10^3 orbits—

systems whose early evolution looks benign because the destabilizing diffusion has not yet had time to act. The 161 false positives, conversely, are near-miss architectures: their median closest approach within the window is 2.98 mutual Hill radii, compared with 5.28 for stable systems overall, meaning they experience deep encounters yet survive the full horizon. Both groups are intrinsically hard cases situated on opposite sides of the same diffuse boundary, and their existence explains most of the gap between the observed AUC ≈ 0.98 and perfection.

Figure 4

Error analysis of the 10^3-orbit HGB classifier, pooled over five splits: (a) misclassification rate and unstable fraction versus Δmin bin, showing that errors concentrate in the transition zone; (b) instability-time distributions of all unstable systems and of the false negatives, showing that missed systems are late instabilities



The ROC curve also supports a concrete screening operating point. In a survey setting, missed instabilities are costlier than unnecessary caution, so the decision threshold should be chosen to bound the missed-instability rate. Tuning the 10^3-orbit classifier so that at most 5% of unstable systems evade detection lets 38% of the ensemble be declared stable—and thus released from further integration—while 16% of stable systems are conservatively carried forward. Relaxing the missed-instability bound to 10% releases 45% of the ensemble at a false-alarm rate of only 7%. Together with the error localization above, this means the classifier can reliably retire the clear cases and concentrate expensive integration on the genuinely ambiguous minority near the transition.

4.4. Window ablation and generalization to four-planet systems

To test how much integration time is actually needed, we retrained HGB with feature windows of 100, 300, 1000, and 3000 orbits. Table 3 and Figure 5a show that even 100 orbits yield AUC 0.971, only 0.014 below the 3000-orbit score. This implies that the dynamical signature of future instability is already present in the earliest approach events and secular drifts.

Table 3

Transactions on Computing Science
https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2

HGB performance versus length of the feature-integration window

T_win (orbits)	ROC-AUC	Balanced Acc.
100	0.971±0.002	0.902±0.005
300	0.972±0.003	0.902±0.005
1000	0.976±0.003	0.911±0.007
3000	0.985±0.002	0.934±0.005

Finally, we generated an independent sample of 1600 four-planet systems (unstable fraction 0.724) and evaluated transfer from a three-planet-trained HGB. The four-planet sample draws spacings, eccentricities, and masses from the same distributions as the three-planet sample, so its higher unstable fraction reflects the larger number of interacting pairs rather than a distribution shift. Table 4 and Figure 5b show that the cross-architecture model achieves AUC 0.972, essentially identical to a model trained and tested on four-planet systems. Adding a small amount of four-planet data during training gives a marginal improvement to 0.975. The transferred model is also well calibrated, with a Brier score of 0.077 on the four-planet test set, close to the 0.070±0.006 of the within-architecture model.

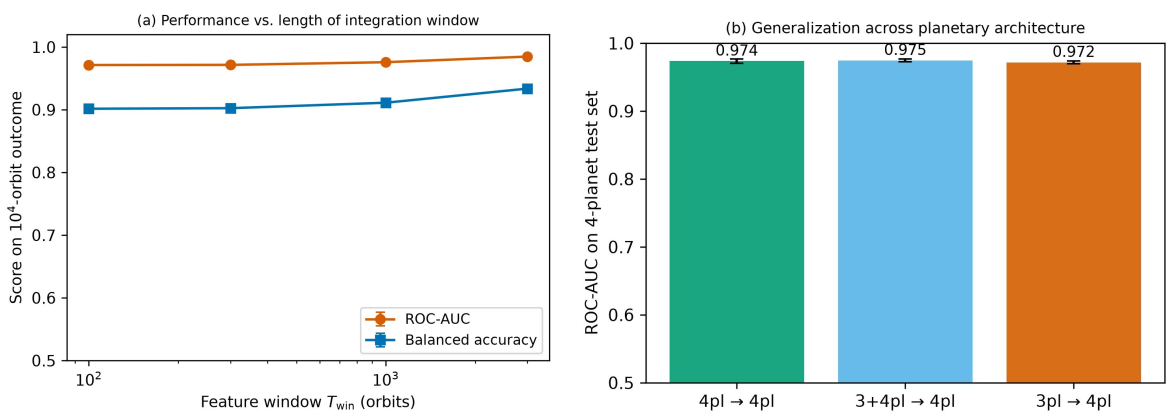
Table 4

Generalization to four-planet systems (AUC on held-out 4-planet test set)

Training set	ROC-AUC
3-planet ⟶ 4-planet	0.972±0.002
3+4-planet ⟶ 4-planet	0.975±0.002
4-planet ⟶ 4-planet	0.974±0.003

Figure 5

(a) HGB ROC-AUC versus length of the feature-integration window; (b) ROC-AUC on the held-out four-planet test set for models trained on three-planet, mixed, or four-planet data



https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2

The transfer result is consistent with the design of the feature set. Every feature is either an initial-condition statistic (Δ, e0) or an orbit-agnostic time-series statistic (excursions, MEGNO, closest approach), so a four-planet system presents exactly the same twelve coordinates to the model as a three-planet system; the only difference is that more interacting pairs contribute to each pooled statistic. A four-planet chain has six pairwise interactions rather than three, which raises the number of opportunities for a deep encounter and hence the unstable fraction, but the local physics sampled by each feature is unchanged. The classifier therefore does not need to learn new physics to score four-planet systems—it merely extrapolates the same geometry-versus-diffusion boundaries to a denser pair network. This is an encouraging indication that such models may generalize to still higher multiplicities, although we have not tested that here.

4.5. Label-leakage control and numerical convergence

A short-window classifier could in principle look trivially accurate: any system that already destabilizes inside the feature window is “seen” rather than predicted. Among the 3200 three-planet systems, 801 (25.0%) destabilize within 100 orbits and 1327 (41.5%) within 10^3 orbits. To isolate genuine forecasting, we repeated the training after excluding all systems whose instability time falls inside the window, so that the task becomes predicting instability strictly between T_win and 10^4 orbits. Table 5 and Figure 6 show that the conditional AUC of HGB is 0.952 at T_win = 100 and 0.921 at T_win = 1000, lower than the unconditional values but still far above both physics baselines, which converge toward 0.81–0.84 in the same setting. The bulk of the predictive power therefore reflects genuine early warning rather than window-internal events.

Table 5

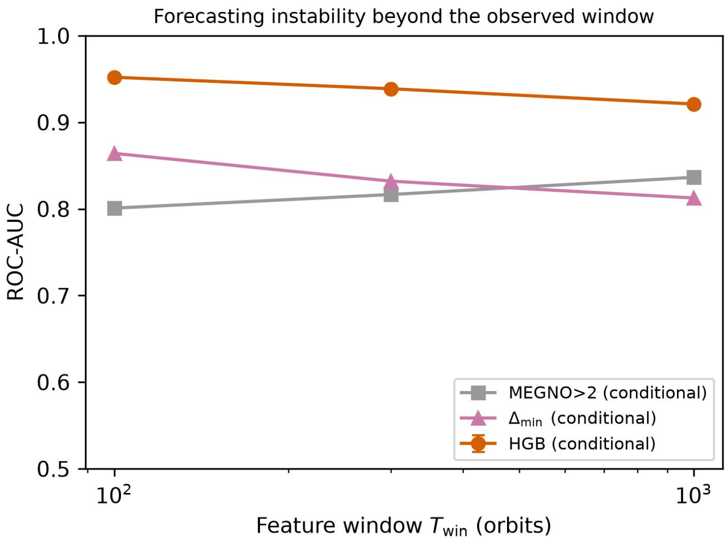
Conditional prediction: systems destabilizing before the window end are excluded; AUC for predicting instability in (T_win, 10^4]

T_win (orbits)	Excluded (%)	HGB AUC	MEGNO AUC	Δmin AUC
100	25.0	0.952±0.005	0.801±0.009	0.864±0.012
300	34.8	0.939±0.004	0.817±0.007	0.832±0.006
1000	41.5	0.921±0.007	0.837±0.016	0.813±0.009

Figure 6

Conditional ROC-AUC versus feature window after excluding systems that destabilize inside the window, compared with the two physics baselines evaluated on the same subsets

Transactions on Computing Science
WISDOM ACADEMIC https://doi.org/10.63808/tcs.v2i2.520 Volume 2 , Issue 2



Because the instability labels themselves come from our fixed-step integrator, we verified numerical convergence by re-integrating a stratified subsample of 300 systems with a four-times smaller step (Δt = P1/600). The instability labels agree for 90.7% of the subsample; among systems found unstable at both resolutions the instability times correlate with Spearman rank coefficient 0.82, and the maximum relative energy error drops to 6e-06. The disagreements are concentrated near the chaos boundary, where instability times are inherently sensitive, and do not alter the qualitative conclusions.

The convergence check also clarifies the role of the energy error in this pipeline. At the refined step the worst-case relative energy error over the subsample is about 6×10^-6, smaller than the median error at the production step (about 1×10^-5), confirming that the step size, rather than close-encounter chaos per se, sets the energy-error scale. The roughly 9% of labels that change under refinement belong to architectures near the chaos boundary, where neighboring trajectories diverge and instability times are inherently sensitive; the labels of systems far from the transition are insensitive to the step. This is the expected behavior for a symplectic scheme with a bounded, non-accumulating energy error, and it means that the classifier is learning the stability structure of the equations of motion rather than artifacts of a particular discretization.

5. Discussion

The strong performance of dmin/RH and Δmin is physically expected: compact multi-planet systems destabilize when planets first enter each other's Hill spheres or when secular drift brings their orbits into crossing configurations (Smith & Lissauer, 2009). The ML models learn these boundaries from data without requiring an analytical closed form. That 100 orbits suffice for AUC 0.971 (0.952 after the leakage control of Section 4.5) suggests that the most unstable systems reveal themselves almost immediately through close approaches or rapid eccentricity growth, while slowly diffusing systems require longer windows. Accordingly, the conditional AUC of the ML model declines as T_win grows, because the surviving prediction set becomes increasingly dominated by slowly diffusing—and therefore harder—cases; meanwhile the MEGNO baseline creeps upward and the Δmin baseline drifts downward, both converging toward 0.81–0.84 once the easiest window-internal events have been removed from their evaluation set.

The error analysis of Section 4.3 adds a practical dimension to these scores. Because failures concentrate within roughly one mutual Hill radius of the Δ50 boundary, the classifier's output is best interpreted as a proximity-to-transition score rather than a verdict, and downstream users can treat the predicted probability as a continuous risk estimate—the Brier score of the transferred model confirms that these probabilities are calibrated. This suggests a natural two-stage screening strategy for large ensembles: a first pass with 100-orbit windows flags clear cases in either regime at negligible cost, a second pass with 10^3-orbit windows resolves most of the transition zone, and only the few percent of genuinely marginal architectures—those within about one mutual Hill radius of the boundary—require the full 10^4-orbit integration. Since the long-run integration is the dominant cost, replacing it for most of the ensemble converts directly into survey throughput.

Relative to existing ML surrogates, our setting answers a shorter-horizon question from a much cheaper window, and the two are complementary rather than competing. A SPOCK-style classifier that forecasts 10^9-orbit stability from a 10^4-orbit integration could itself use our short-window forecasts as a pre-filter, skipping the full integration for architectures that are already unstable at 10^4 orbits. Conversely, our conditional-AUC analysis shows what information is genuinely absent from short windows: late, diffusion-driven instabilities are the dominant failure mode of the 100-orbit model, and they are precisely the systems for which longer integration—or a model trained on longer-horizon labels—is irreplaceable. A leakage control of the kind used in Section 4.5 should, in our view, accompany any claim that short-run features predict long-run outcomes.

Our study has several limitations. The ensemble is coplanar, equal-mass, and restricted to super-Earths; real systems have inclinations, mass ratios, and tidal or general-relativistic effects that may alter the early signature of instability. The fixed time step becomes costly if one resolves arbitrarily close encounters, and our convergence check found that about 9% of labels change under a four-fold step refinement, with the disagreements concentrated near the chaos boundary; future work should combine our integrator with a Bulirsch-Stoer or hybrid method (Rein & Tamayo, 2015). Finally, the 10^4-orbit horizon is short compared with the Gyr lifetimes of observed systems; extending the label horizon will test whether the same features remain predictive.

6. Conclusion

We have built and validated a fast, reproducible pipeline for predicting the short-term (10^4-orbit) stability of compact multi-planet systems. A symplectic N-body integrator with a renormalized MEGNO indicator generates the training data, and a gradient-boosted classifier turns the first 10^3 orbits into an accurate stability forecast. The model outperforms both a MEGNO threshold and a pure Hill-spacing rule, retains most of its accuracy with windows as short as 100 orbits even after excluding window-internal instabilities, and transfers from three-planet to four-planet systems with no meaningful degradation. The residual errors are not random: they concentrate within about one mutual Hill radius of the stability transition, and the missed systems are overwhelmingly late, diffusion-driven instabilities—precisely the cases for which longer integration is irreplaceable. These properties define a concrete division of labor between cheap short-window surrogates and full N-body integration in future stability surveys of exoplanetary systems.

Acknowledgments

Not applicable.

Author contributions

Zixi Xu individually undertook all roles including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, original draft writing, review & editing, visualization, supervision and project administration for this study, and has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability

The data are not publicly available due to privacy.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional information

Received: 4 September 2026

Accepted: 8 September 2026

Published online: 15 September 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Cincotta, P. M., & Simó, C. (2000). Simple tools to study global dynamics in non-axisymmetric galactic potentials. A&AS, 147(2), 205-228.

[2] Cranmer, M. D., Tamayo, D., Rein, H., Battaglia, P., & Hadden, S. (2021). A Bayesian neural network predicts instability in compact planetary systems. *Proc. Natl. Acad. Sci. U.S.A., 118*(6), e2016951118.

[3] Doty, M. J., Weiss, L. M., He, M. Y., & Petit, A. C. (2025). From stability to instability: characterizing the eccentricities of multi-planet systems in the California Kepler Survey as a means of studying stability. ApJ, 2025, in press (arXiv:2501.06358).

[4] Gajdoš, P., & Vaňko, M. (2023). Chaos in multiplanetary extrasolar systems. *MNRAS, 518*(2), 2068-2075.

[5] Gladman, B. (1993). Dynamics of systems of two planets. *Icarus, 106*(1), 247-263.

[6] Petit, A. C., Pichierri, G., Davies, M. B., & Johansen, A. (2020). The path to instability in compact multi-planetary systems. *A&A, 641*, A176.

[7] Rein, H., & Tamayo, D. (2015). WHFast: a fast and unbiased implementation of the Wisdom-Holman integrator. *MNRAS, 452*(1), 376-388.

[8] Smith, A. W., & Lissauer, J. J. (2009). Orbital stability of systems of closely-spaced planets. *Icarus, 201*(1), 381-394.

[9] Tamayo, D., Silburt, A., Valencia, D., et al. (2020). Predicting the long-term stability of compact multi-planet systems. *Proc. Natl. Acad. Sci. U.S.A., 117*(31), 18194-18199.

[10] Wisdom, J., & Holman, M. (1991). Symplectic maps for the N-body problem. *AJ, 102*, 1528-1538.