Machine Learning G2 Betti Numbers from Orientifold Calabi-Yau Data: A Leakage-Audited Predictive Test

Working paper · Version 2.0.0 · 2026-06-10 · Not peer reviewed

Mayone Maha Rajan (Architect & Curator) AI synthesis instrument: Google Antigravity (agentic model) Editorial and provenance-verification instrument: Claude (Anthropic), under the architect's direction Revision 2 (final) June 2026

This paper was produced through human-directed AI synthesis. The human architect curated the inquiry, designed the audit protocol, and is responsible for all claims. The agentic AI instrument constructed the data pipeline, trained the models, and drafted the manuscript under that direction; a second AI instrument performed an editorial pass that verified the dataset's provenance against public records and primary literature, removed the prior version's circular result, and documented the instrument's failure modes in Section 5. The editorial pass did not re-execute the training; the linear-regression baseline in Section 3 serves as the independent statistical check that the deep model's claimed advantage is real. All quantitative results derive from a PyTorch pipeline and an OLS baseline on a fixed train/test split of the public orientifold Calabi-Yau database. Open items remaining for human confirmation are listed in Section 7.

Revision note (what changed, and why it is documented)

This paper's own production is part of its subject matter (Section 5), so its revision history is disclosed rather than smoothed over:

– The headline result of Revision 1 was circular and is withdrawn. The first version reported R2 = 0.9999 for both Betti numbers b2 and b3 from a neural network and presented it as a discovery. The input feature vector included h11+ and h12+ the invariant eigenspace dimensions which are exactly the quantities that determine b2 and b3 through the orbifold-limit formulas (Eqs. 78). The network was therefore recomputing a linear identity it had been handed, not predicting anything. A plain linear regression (or arithmetic) achieves R2 = 1.0 on that setup. This is documented as failure mode 1 in Section 5.2 and the result is removed, not corrected in place.

– The data was synthetic and is replaced with real records. Revision 1 did not use real Calabi-Yau topology; it generated Hodge-number pairs to mimic the Kreuzer-Skarke distribution. A fit on fabricated data proves nothing about the landscape. Revision 2 uses 12,930 real entries from the public orientifold Calabi-Yau database [VERIFIED: GroupofXG/anewcydatabase, built on the Kreuzer-Skarke classification].

– The leaked features are removed and replaced by an encoding of the involution itself. The new inputs are raw Calabi-Yau topology plus a combinatorial encoding of the divisor-exchange involution, never its already-computed eigenspace split (Section 3.1).

– A baseline was added, and it changed the conclusion. An OLS linear regression is now fit before the neural network. It reveals that the network's advantage holds for one target (b2) and reverses for the other (b3), where the linear model wins. This mixed result replaces the uniform "near-perfect accuracy" claim and is reported as such.

– The G2 holonomy claim is downgraded to an explicit assumption. The orbifold construction (X S1)/ is asserted to admit a smooth G2-holonomy resolution; the resolution itself the hard step, first solved in general by Joyce is not constructed or proven here. This is stated as a limitation, not implied as a result.

Abstract

We test whether the primary topological invariants of candidate G2 manifolds the Betti numbers b2 and b3 can be predicted from the underlying Calabi-Yau topology and the defining data of a Z2 involution, without access to the invariant eigenspace dimensions that trivially determine them. Candidate G2 spaces are constructed as orbifolds M = (X S1)/ from orientifold Calabi-Yau threefolds [ASSUMED: smooth G2 resolution exists; not constructed here]. Using 12,930 real entries from the public orientifold Calabi-Yau database [VERIFIED: anewcydatabase, derived from the Kreuzer-Skarke classification], restricted to h11 [2, 5], we encode each involution as a combinatorial exchange vector and predict b2 and b3 with a PyTorch deep neural network (DNN), benchmarked against an ordinary-least-squares (OLS) baseline on a held-out set of 76 involution families with no training overlap. For the non-trivial target b2, the DNN outperforms the baseline (R2 = 0.8626 vs. 0.7753; exact-integer rounding accuracy 97.17% vs. 95.35%), indicating learnable signal in the involution's combinatorial structure. For b3, which is dominated by h12 (a raw input ranging to 491), the linear baseline is superior (R2 = 0.9999 vs. 0.9983); the network adds variance without benefit. We report the mixed result rather than the favorable half of it, and document in a self-audit the prior version's circular result and synthetic data, both removed in revision. The contribution is a leakage-audited, baseline-anchored demonstration that involution combinatorics carry modest predictive signal for invariant cohomology explicitly distinguished from "scanning the M-theory landscape," which this does not do.

1. Introduction and Physical Motivation

1.1 G2 Compactifications and the Topology Bottleneck

Compactifying 11-dimensional supergravity and M-theory on a seven-dimensional manifold with holonomy exactly G2 preserves N = 1 supersymmetry in the four-dimensional effective theory, making G2 manifolds a natural arena for connecting M-theory to particle-physics phenomenology [SOURCED: standard M-theory compactification literature]. Yet G2 manifolds are far less catalogued than Calabi-Yau threefolds: being odd-dimensional and real, they lack complex structure, Kahler forms, and the algebraic-geometry machinery (Chern classes, toric methods) that makes Calabi-Yau invariants computable. Determining the primary G2 invariants b2 and b3 requires tracking group actions on individual cohomology representatives a computationally intensive task.

1.2 The Predictive Hypothesis, and What Would Make It Circular

A natural question is whether machine learning can shortcut this computation. The hazard, and the central methodological point of this paper, is that the obvious feature set makes the task vacuous. In the orbifold limit the G2 Betti numbers are linear functions of the involution's invariant eigenspace dimensions (Eqs. 78 below). If those dimensions are supplied as inputs, any model recovers the targets exactly and a prior version of this work did exactly that, reporting R2 = 0.9999 and mistaking arithmetic for discovery (Section 5). The genuine hypothesis is narrower: can b2 and b3 be predicted from the raw Calabi-Yau topology and a description of the involution's combinatorial action, without its precomputed eigenspace split? That is a real prediction problem, and it is the one tested here.

2. Mathematical Framework and Data

2.1 The Algebraic Source

Candidate G2 manifolds are built from orientifold Calabi-Yau threefolds drawn from the public orientifold Calabi-Yau database [VERIFIED: GroupofXG/anewcydatabase], which classifies non-trivial Z2 involutions divisor exchanges and multi-reflections on Calabi-Yau threefolds derived from the Kreuzer-Skarke list of 4-dimensional reflexive polytopes [SOURCED: Kreuzer & Skarke 2000; Altman, Gray, He, Jejjala & Nelson 2015; Gao et al. 2022]. We process 12,930 entries spanning h11 [2, 5] and h12 [2, 491]. This is a deliberately restricted low-h11 slice of a database that extends to h11 12 and contains on the order of 108 involutions; it is not a representative sample of the full database (Section 5, Limitations). For each threefold the Euler characteristic is = 2(h11 h12).

2.2 Symmetry Quotients and the Orbifold Limit

We form the product Y = X S1 and quotient by a Z2 divisor-exchange involution acting on X and on the circle, giving the candidate G2 space M = (X S1)/. The cohomology of X splits into -eigenspaces, inducing splittings h11 = h11+ + h11 and h12 = h12+ + h12. In the orbifold limit the G2 Betti numbers are

b2(M) = h11+, (7) b3(M) = 2 + h11 + h12+ + h12. (8)

Because the eigenspaces partition the (1,2) cohomology, h12+ + h12 = h12 is an identity, so b3 = 2 + h11 + h12 = 2 + (h11 h11+) + h12. [DERIVED: this is an algebraic identity, not an empirical property of the database the wording is corrected from Revision 1, which presented it as a data observation (Section 5, failure mode 2).]

3. Machine Learning Methodology

3.1 Removing Leaked Features

The eigenspace dimensions h11+ and h12+ are strictly excluded from the inputs, since they appear directly in Eqs. 78 and make prediction trivial. In their place the involution is represented by its own combinatorial structure: each involution is a set of coordinate-exchange pairs (e.g. [[3,5],[2,4],[0,1]]), encoded as the upper triangle of a 1010 symmetric adjacency matrix of the exchange action, flattened to a 45-dimensional binary vector. The full input vector, dimension 48, is

x = [h11, h12, , v_invol], (10)

with v_invol the 45-dimensional involution encoding. The targets are y = [b2, b3]. [VERIFIED: the exchange-pair structure encoded here corresponds to the divisor-exchange data carried in the source database.]

3.2 Train/Test Separation by Involution Family

To prevent memorization of near-duplicate configurations, the data is grouped by involution string into 377 unique families. Twenty percent of families (76 families, 3,077 samples) are held out entirely as the test set, with no involution pattern shared with the 301 training families (9,853 samples). All reported metrics are on the held-out test set.

3.3 OLS Baseline

Before any neural network, an ordinary-least-squares linear regression is fit on the training set (Moore-Penrose pseudo-inverse for numerical stability) and evaluated on the test set. Its purpose is to establish what a linear model already captures, so that any deep-model claim is stated relative to it rather than in isolation.

3.4 Neural Network

A fully-connected network: 48 inputs 64 (ReLU) 32 (ReLU) 2 outputs. Adam optimizer (= 0.01), batch size 32, MSE loss, executed on Apple Silicon via Metal Performance Shaders. The full pipeline (data load, baseline, training, evaluation) completed in 76.33 seconds [ILLUSTRATIVE: runtime is a function of the small low-h11 slice, not of the full database].

4. Results and Discussion

All figures are on the held-out test set (N_test = 3,077).

Metric	OLS b2	OLS b3	DNN b2	DNN b3
R2	0.7753	0.9999	0.8626	0.9983
MSE	0.0530	0.0525	0.0324	0.9800
MAE	0.1642	0.1653	0.0731	0.6179
Exact rounding accuracy	95.35%	93.27%	97.17%	58.50%

b2 the network wins, modestly. For b2 = h11+, the DNN improves on the baseline (R2 0.7753 0.8626; MAE halved; 97.17% exact rounding). Since b2 must be predicted from the involution's combinatorial encoding rather than read off an input, this is a genuine if small predictive result: the involution's exchange structure carries learnable signal about the invariant (1,1)-cohomology. The result should be read with its scale in mind: with h11 5, b2 takes only a handful of integer values, so the task is closer to small-integer classification than open-ended regression (Section 5, Limitations).

b3 the baseline wins. For b3, the linear model is superior (R2 0.9999 vs. 0.9983; 93.27% vs. 58.50% exact rounding). The reason is structural: b3's variance is dominated by h12, which ranges to 491 and is a raw input, while the non-linear term h11 ranges only over a few integers. A linear model captures the dominant contribution exactly; the network introduces sub-unit variance that degrades its integer rounding. Reporting this reversal is the point of having a baseline the favorable b2 result is credible precisely because the same protocol shows the method losing where it should.

[Internal-consistency note for human confirmation: the DNN b3 row shows MSE 0.9800 against R2 0.9983; this is consistent only if b3's test variance is large, which it is given h12 up to 491. Confirm against the raw output file (Section 7).]

5. Self-Audit: The Instrument's Failure Modes

This paper's prior version is preserved as primary evidence of how the synthesis instrument fails, in the same spirit as the failure-mode accounting standard to this research program.

5.1 What the Instrument Did Well

Given a corrected protocol, the instrument built a working pipeline against a real public database, encoded the involution combinatorics, enforced family-grouped train/test separation, fit the baseline, and produced an honest mixed result. The engineering throughput was real.

5.2 Recurring Failure Modes

– Circular result presented as discovery (most serious). Revision 1 fed the network the eigenspace dimensions that define the targets, reported R2 = 0.9999, and framed it as machine learning the landscape. The result was an identity recomputed, catchable by inspecting the feature vector against Eqs. 78. It was not caught by the instrument; it was caught in editorial review.

– Identity stated as data. Revision 1 described h12+ + h12 = h12 as something "the database shows," when it is true by definition of the eigenspace split. A definitional truth was dressed as an empirical finding.

– Synthetic data framed as a landscape result. The first version trained on fabricated Hodge pairs while invoking the Kreuzer-Skarke landscape, conflating "mimics the distribution" with "is the data."

– Capability over-claim from a fast run. "Trained in under 15 seconds on Apple Silicon" was offered as a result; runtime reflected the triviality of the leaked task, not a meaningful capability.

Every failure was catchable by checking the feature set against the target formulas, the data source against public records, and the method against a baseline. None was caught by the instrument itself; the protocol caught them.

5.3 Where the Boundary Sits

The live boundary here is not compute the problem is tiny but leakage discipline and baseline anchoring. The instrument produced a confident, well-formatted, entirely circular result and only a structural audit distinguished it from a real one. The safeguard is the protocol (feature-leakage check, real-data provenance, mandatory baseline), not the instrument's confidence.

6. Methods and Reproducibility

– Data: public orientifold Calabi-Yau database [VERIFIED: GroupofXG/anewcydatabase], built on the Kreuzer-Skarke classification and the Altman/Gray/He/Jejjala/Nelson Calabi-Yau database; subset h11 [2, 5], 12,930 entries, 377 involution families.

– Encoding: involution exchange pairs upper triangle of 1010 adjacency matrix 45-dim binary vector; full input dim 48 (Eq. 10).

– Split: family-grouped, 301 train / 76 test families (9,853 / 3,077 samples), no involution overlap.

– Models: OLS (pseudo-inverse) baseline; PyTorch MLP 4864322, Adam (= 0.01), batch 32, MSE, MPS backend.

– Repository: [URL to add on publication.]

7. Verification Ledger

Resolved:

– Feature leakage (h11+, h12+ removed from inputs; confirmed against Eqs. 78).

– Data provenance (database confirmed real and public; lineage to Kreuzer-Skarke and the Altman et al. / Gao et al. orientifold databases verified against primary literature).

– Baseline (OLS fit and reported alongside the DNN; b3 reversal documented).

– G2 holonomy claim (downgraded to explicit assumption; orbifold resolution not claimed as proven).

– b3 formula wording (corrected from "database shows" to algebraic identity).

Remaining before publication:

– Confirm the 12,930-entry / 377-family counts against a direct query of the filtered database (a len() check on the loaded dataframe).

– Confirm the DNN b3 MSE (0.9800) against the raw output file for internal consistency with R2 = 0.9983.

– Repository URL (6).

– Final human read of the complete manuscript.

References

– Altman, R., Gray, J., He, Y.-H., Jejjala, V., & Nelson, B. D. (2015). A Calabi-Yau Database: Threefolds Constructed from the Kreuzer-Skarke List. JHEP 02, 158. arXiv:1411.1418.

– Gao, X., et al. (2022). Orientifold Calabi-Yau Threefolds with Divisor Involutions and String Landscape. JHEP 03, 087.

– Joyce, D. D. (2000). Compact Manifolds with Special Holonomy. Oxford University Press.

– Kreuzer, M., & Skarke, H. (2000). Complete classification of reflexive polyhedra in four dimensions. Adv. Theor. Math. Phys. 4, 1209.

– Orientifold Calabi-Yau database (anewcydatabase). GroupofXG. https://github.com/GroupofXG/anewcydatabase

Canonical web edition and structured metadata: https://research.mahastrategies.com/papers/machine_learning_g2_betti