

Structural Predictability in the UNNS Substrate: Generative Admissibility, Exclusion Gaps, and Formal Certification

UNNS Research Collective

UNNS Substrate Project

October 2026

Version v011

Abstract

The UNNS (Unbounded Nested Number Sequences) Substrate proposes that a generative structural description can identify an admissible region of a very large nominal state space and support quantitative, falsifiable predictions about the boundary between that region and the physically relevant configurations. This paper tests that proposition of *structural predictability* where every quantity can be computed exactly. Square artificial spin ice is the experimental instrument; the subject of study is predictability. The two-in/two-out ice rule is an established constraint, and it is used because its familiarity makes the experiment reproducible and its outcomes independently assessable, not because the paper concerns square ice.

A generative rule can reduce the number of configurations a system admits by many orders of magnitude. Whether the reduced set still contains what physics regards as relevant is a separate, empirical, and sometimes unfavourable question, and whether its boundary can be predicted in advance is a further one. The test is carried out in a finite, fully specified setting: a square artificial spin ice with free boundary, an all-pair dipolar (dumbbell) Hamiltonian taken from a published simulator, and the local two-in/two-out ice rule as the generative grammar. The grammar is constructed blind to energy and frozen before any physical energy is evaluated; the physical target is defined by energy, as a window above the exact global ground state, not by grammar membership. The Hamiltonian is a frozen model Hamiltonian taken from the same published simulator source, so the two are separate calculations but not of independent provenance.

The organizing result is an exact equivalence: for a finite model whose grammar excludes at least one configuration, the grammar contains every target state at window W if and only if W is smaller than the exclusion gap Δ_{-G} , the energy of the cheapest excluded configuration above the global ground state. Sufficiency of the grammar for every window is therefore decided by one number.

We report four findings, each at its own evidence level. (I) On the 3×3 patch (24 spins) the grammar retains 2,604 of 16,777,216 configurations and constructs them without materializing the rest; an exact combinatorial count extends the contraction to the 4×4 patch, where 193,662 of 2^{40} configurations satisfy the ice rule (99.99998% contraction; a count, not a constructive run and not a Lean result). (II) In a preregistered test, the grammar fails to preserve the physical target: at a window of $1 k_B T$ the target holds 26 states, of which 10 lie in the grammar and 16 are excluded; the failure is exactly predicted by the first grammar-excluded level, which lies at $0.943 k_B T$. (III) A prediction frozen from the 1×1 to 3×3 evidence, that the normalized exclusion gap δ_4 of the 4×4 patch (40 spins, 2^{40} configurations) lies in $[0.790, 1.033]$ and below 1, was

evaluated numerically and then proved. The prediction is prospective, but the planned feasibility campaign was not completed, an earlier exploratory verdict was inconclusive, and the interval is broad. (IV) A Lean 4 proof, assembled from 1,148 independently kernel-checked proof-tree batches, certifies for the exact dyadic encoding of the 780 couplings both global minima, a strictly positive gap $\Delta_{-G} = 12095447728404694790/2^{70}$ in units of the energy scale A , and $\delta_4 < 1$ together with membership in the frozen interval. The certificates fix minimum values; the corner-defect configuration is a verified witness of the excluded minimum, and uniqueness is not claimed. The same certification shows that the lowest grammar-excluded energy originally reported lay $928/2^{70} A \approx 7.9 \times 10^{-19} A$ above the true minimum; the correction replaces the minimizer, moves δ_4 by less than 10^{-15} , and changes no verdict.

The result is two-sided. The predicted finite gap is confirmed, and the same fact, $\delta_4 < 1$, proves that the ice-rule grammar cannot preserve every state within the $1 k_B T$ window of this model. For the program this means that a generative structural description can be assigned an exact domain of physical sufficiency, here every window below Δ_{-G} , and that this domain can be predicted prospectively and then established by numerical evidence combined with formal proof. We distinguish three levels of achievement. Structural predictability, that a frozen generative admissibility defines and predicts a measurable exclusion boundary, is demonstrated for the investigated finite model. Predictive verification, that the predicted boundary can be checked without trusting the numerical optimizer, is achieved by the Lean certification of the encoded model. Cross-domain predictability, that the same methodology predicts analogous relations in unrelated systems, is the ambition of the UNNS methodology and is not established by a single spin-ice case; we identify cross-domain predictive transfer as the next scientific test. We state exactly what the finite-model results establish, what they do not (no universality, no computational advantage, since the comparative timing benchmark was never eligible, and no claim that proof-carrying optimization originates with this work), and how the three evidence layers (audited numerics, machine-checked mathematics, interpretation) are kept apart.

The frozen records and Lean packages are cited at repository commit 4e759d38; the two additive manuscript records are at commit 7610fe5d (Appendix A).

Results at a glance (physics terms). In a finite square artificial spin ice with a full dipolar (dumbbell) Hamiltonian, we ask whether restricting the state space to ice-rule (two-in/two-out) configurations keeps every state within an energy window above the true ground state.

- *The restriction is severe.* It admits 2,604 of 2^{24} configurations on the 24-spin patch and 193,662 of 2^{40} on the 40-spin patch.
- *It is exactly characterized by one energy.* The restriction contains all states within a window W of the ground state if and only if W is below the energy of the cheapest ice-rule-violating state. For the 40-spin model that excitation threshold is $0.9459 k_B T$ at 792 K, so the restriction is sufficient below it and loses at least one state at $1 k_B T$.
- *The threshold was predicted, then certified.* It was predicted before the 40-spin energies were evaluated. Machine-checked proof then fixed the ground-state energy and the lowest rule-violating energy of the encoded 40-spin model exactly, and corrected a previously reported excitation energy by 7.9×10^{-19} of the model's energy unit without changing any conclusion.
- *The cheapest excitations are single-vertex defects.* Reversing one of the eight outer islands that touch a single (corner) vertex costs $0.946 k_B T$ (the eight values agree to 10^{-15} of the gap); reversing an island shared by two vertices costs at least $1.79 k_B T$. This is an exact calculation on one ground state, not a bound over all defects.

All statements concern finite models, one Hamiltonian and one reference temperature.

Contents

1	Introduction	4
2	UNNS and generative admissibility	8
3	Physical system and Hamiltonian	11
4	Experimental design	15
5	Structural contraction and physical sufficiency: the 3×3 study	21
6	The prospective 4×4 investigation	24
7	Formal mathematical certification	28
8	Certified results and the exact correction	34
9	Scientific interpretation	38
10	Discussion and conclusions	46
A	Data provenance and reproducibility	52
B	Formal statements	57
C	Disclosure register	57
	Data and code availability	58

1 Introduction

1.1 The scientific problem

This paper investigates the predictive operation of the UNNS Substrate, not the physics of square ice. The proposition under test is that a generative structural description can identify an admissible region of a large nominal state space and predict, quantitatively and in advance, the boundary between that region and the physically relevant configurations. The spin-ice model is the controlled setting in which the proposition is tested. The remainder of this subsection states the underlying question in its general form.

A generative rule builds configurations by local extension: each step is admitted only if it satisfies a constraint that can be checked using local information. Such rules can be extremely restrictive. In a square lattice of $n = 24$ binary degrees of freedom, a local vertex constraint leaves about 0.016% of the 2^{24} nominal configurations. The restriction is useful for computation if two conditions hold: the surviving set can be *constructed* without first listing the rejected configurations, and the survivors *include what a downstream physical question requires*. The first condition is a statement about generation. The second is a statement about physics, and it does not follow from the first.

The question is easy to state loosely and easy to answer wrongly. If the constraint is itself the physical law that selects the low-energy states, then the answer is true by construction, and nothing has been learned. If the constraint is merely correlated with energy, the answer depends on details: on the energy scale at which one asks for preservation, on how the constraint is violated, and on the cost, in energy, of the lowest configuration it forbids. This paper asks the question in a setting where those quantities can be computed exactly, and it records the outcome whichever way it falls.

Can generative admissibility determine the physically relevant portion of a large configuration space, and can the resulting structural boundaries be established mathematically?

We answer both halves. The answer to the first is partly positive and partly negative, and the negative part is quantitative: it is the size of an energy gap. The answer to the second is positive for one explicitly defined finite model: the boundary, namely the lowest energy that the grammar forbids, can be established by machine-checked proof, and in doing so the proof revises the value that numerical optimization had reported.

1.2 Why the question is worth asking exactly

Restricting a configuration space by a local rule is a common move in computational physics and in combinatorial search: it underlies lattice models with constraints, sampling schemes that stay on a constraint surface, and reduced bases for low-energy sectors. In most uses the restricted set is chosen because it is believed to contain what matters, and the belief is tested, if at all, by the agreement of downstream observables. That test is indirect. It cannot distinguish a grammar that contains the relevant states from one whose omissions happen not to affect the chosen observable, and it says nothing about how far from the ground state the omissions begin.

An exact, finite setting allows a more direct test. The set of states that the grammar omits is explicit, the energy of the cheapest omitted state is a number, and the sufficiency of the grammar for every energy window follows from that number (Section 2). The price is that the setting must be small enough

to certify. The 4×4 patch, with 2^{40} configurations, is at the upper edge of what we could certify with the tools of this program, and it is the setting in which the question is answered here.

1.3 The physical model as a test of UNNS predictability

The purpose of this investigation is not to develop a new theory of square ice or to rediscover its established local constraints. Square artificial spin ice serves as a controlled physical test environment for a broader proposition of the UNNS Substrate: that generative admissibility can identify structurally meaningful regions of a nominal configuration space and support quantitative predictions about the boundaries of their physical sufficiency.

The investigation operationalizes this proposition through a frozen generative grammar, an energy-defined physical target, and a measurable exclusion gap. The grammar is constructed without knowledge of the target energies; its contraction and physical sufficiency are evaluated separately; and a prediction about the exclusion boundary of a larger system is registered before that system is evaluated. The resulting prediction is subsequently subjected to numerical evaluation and formal mathematical certification.

The significance of this procedure lies in the predictive relationship it establishes between structural generation, admissibility, and physical relevance, rather than in the particular constraint chosen to represent admissibility. The two-in/two-out ice rule is an established constraint [1–3], and its mathematical familiarity makes the experiment reproducible and its outcomes independently assessable. We make no claim to have discovered it.

Within UNNS, Structural Complexity Reduction is therefore a test of whether a common generative-admissibility methodology can produce measurable, falsifiable, and certifiable predictions. The present study establishes that capability for the specified finite spin-ice model. Testing whether comparable predictive relations persist across different physical and computational domains is the broader research program to which this investigation contributes.

The system. The physical system is an artificial square spin ice [4, 5]: a lattice of single-domain nanomagnets (islands), each carrying one Ising degree of freedom, arranged on the bonds of a square lattice so that four islands meet at every interior vertex. The generative grammar is the ice rule [1, 2]: at each vertex, two of the four moments point toward the vertex and two point away. We use the finite, free-boundary patch of the Dryad SPMMC v1.00 simulator [6] with all-pair dumbbell interactions, at patch sizes $L \times L$ with $L = 1, \dots, 4$. These have $n = 2L(L + 1)$ spins, namely 4, 12, 24 and 40, and 2^n nominal configurations.

The six-vertex model that counts the ice-rule configurations is classical [3]; the procedure by which the rule is tested as a *generative* admissibility rule is the subject of the study, and is described in Section 4.

1.4 Principal contributions

What the investigation tests. The central proposition is that a generative structural description can identify an admissible region of a large nominal state space and predict a measurable boundary between that region and physically relevant configurations. The study tests it through four connected questions.

Table 1: Three levels of achievement.

Level	Scientific question	Present status
Structural predictability	Can frozen generative admissibility define and predict a measurable exclusion boundary?	Demonstrated for the investigated finite model; the prediction interval is broad (Section 6)
Predictive verification	Can the predicted boundary be verified without trusting the original numerical optimizer?	Achieved by the Lean certification (SCR-IV) of the encoded model
Cross-domain predictability	Can the same methodology predict analogous relationships in unrelated systems?	Research objective; requires cross-domain evidence (Section 10.8)

- Q1.** Can generative admissibility produce substantial state-space contraction without consulting the physical target?
- Q2.** Can the admissibility boundary be translated into a measurable physical quantity?
- Q3.** Can that quantity be predicted from smaller or previously examined systems before the target system is evaluated?
- Q4.** Can the predicted result subsequently be verified independently of the numerical search procedure?

Findings 1 and 2 address Q1 and Q2 (with the gap equivalence of Proposition 2.4 as the translation), Finding 3 addresses Q3, and Finding 4 addresses Q4. The 4×4 result bears most directly on the last two: the normalized exclusion gap was predicted before the 40-spin instance was evaluated, and its value was then established by machine-checked certification.

Three levels of achievement. The questions separate three levels, and the manuscript keeps them apart. Domain independence is the ambition of the UNNS methodology, not something established by a single spin-ice case. The present investigation demonstrates predictive operation in one physical domain; portability across unrelated domains is a distinct, testable proposition.

What “predictive” means here. The exclusion-gap equivalence provides an exact criterion for physical sufficiency, but does not itself constitute an algorithm for predicting an unknown exclusion gap. The predictive achievement of this investigation lies in the UNNS-motivated procedure: a generative grammar was frozen, a quantitative prediction was registered from smaller systems before the larger instance was evaluated, and the resulting prediction was independently tested and formally certified for the specified finite model.

Findings. The study yields four findings. We state them here as claims of the kind each can support; Section 4 gives the rules under which each was evaluated, and Section 9 states what they jointly do and do not establish.

Finding 1 (Generative contraction). The frozen ice-rule grammar reduces the 3×3 configuration space from 16,777,216 nominal configurations to 2,604, and a constructive generator produces exactly this set while materializing no other configuration. An additive exact count extends the contraction to 4×4 (193,662 of 2^{40} configurations; a count only, not a constructive run) (Section 5).

Finding 2 (Physical discrimination). In a preregistered test, the grammar does *not* preserve the physical target. At a window of $k_B T$ above the exact ground energy the 3×3 target contains 26 states; 10 are in the grammar and 16 are excluded. The failure is not an accident of the window: it is equivalent to the statement that the lowest grammar-excluded level lies below the window, and that level lies at $0.943 k_B T$ (Section 5).

Finding 3 (Prospective prediction). Before the 4×4 energies were evaluated, we froze an interval $[0.790, 1.033]$ for the normalized exclusion gap δ_4 and the separate prediction $\delta_4 < 1$. The numerical value 0.9459 satisfies both, and both are proved for the exact encoded model (Sections 6 and 8).

Finding 4 (Mathematical certification). A Lean 4 proof certifies the exact global ground energy and the exact minimum over grammar-excluded configurations of the 4×4 model, the positive gap between them, and both predictions. The proof also corrects the previously reported excluded-state minimum by $928/2^{70} A$ (Sections 7 and 8).

Finding 2 and the second half of Finding 3 are two views of one fact. A prediction of the form $\delta_4 < 1$ is a prediction that the grammar will *fail* to preserve the full $k_B T$ window. The exclusion-gap equivalence (Proposition 2.4) is the mathematical centre of the paper: it turns the sufficiency of the grammar for every energy window into the single number Δ_{-G} . The UNNS achievement is accordingly not that the ice rule excludes almost all spin configurations, which is a property of the constraint. It is that a generative-admissibility framework was used to formulate a prospective, quantitative prediction about the relationship between structural constraints and a physical energy landscape, and that the prediction was subsequently certified; the gap equivalence supplies the mathematical mechanism connecting the two. We therefore regard the confirmed prediction as informative in both directions: it confirms a finite exclusion gap of the expected size, and it establishes that the ice-rule grammar cannot be a sufficient description of the $1 k_B T$ low-energy sector of this model.

1.5 What this paper does not claim

The claims are finite-model claims. Structural contraction does not imply physical completeness (Finding 2 shows the opposite for this model); no computational advantage is claimed, since the comparative speed criterion was never eligible (Section 4); proof-carrying optimization is not claimed to originate here [7]; and the universality of the Unbounded Nested Number Sequences (UNNS) framework, from which the program originates, remains a hypothesis. In particular the domain independence of the predictive methodology is not established by a single spin-ice case. Section 2 places the work within UNNS, and Section 9 gives the full statement of scope and the limits of inference.

1.6 Evidence layers

Three kinds of statement appear in this paper, and we keep them separate throughout.

Layer	What it establishes	Status of statements
SCR-III numerical investigation	3×3 contraction; the primary-window sufficiency failure; the frozen 4×4 numerical results	Audited numerical and experimental evidence; not machine-checked
SCR-IV Lean certification	Exact 4×4 minima of the encoded model, the exclusion gap, the corrected witness, the two prediction inequalities	Machine-checked mathematics in a recorded execution
Additive records (grammar count; single-reversal decomposition; perturbation-bound inputs)	The 4×4 contraction value; the energy cost of single-island reversals; evaluated inputs of the perturbation bound	Exact computation on frozen data, outside the Lean certificate; not optimizations; not preregistered
Scientific interpretation	What the findings suggest about generative admissibility	Interpretation; distinct from theorems

1.7 Organization

Section 2 introduces generative admissibility and its UNNS origin, and sets the constraint within established constrained-model theory. Section 3 specifies the physical system, the Hamiltonian and the energy scales. Section 4 describes the experimental design: the frozen hypotheses, criteria, comparators and null controls. Sections 5 and 6 report the 3×3 study and the 4×4 investigation. Section 7 develops the formal certification, and Section 8 reports its results and the exact correction. Section 9 states the scientific interpretation, and Section 10 compares the approach with related methods and sets out limits and future work. Appendices record data provenance, the formal statements, and the chronology of frozen records.

2 UNNS and generative admissibility

2.1 The substrate and the admissibility chain

The Unbounded Nested Number Sequences (UNNS) framework begins from generated occurrence rather than from isolated value [8]. Its introductory form defines a substrate as a tuple $U = (S, O, D, R, A)$: a space S of nest states, a family O of recursive operations, an admissible domain D , a set R of generated routes, and an admissibility criterion A that assigns structural status to generated occurrences. The original nest operation is

$$M \perp N = MN + \frac{M}{N} + 2M,$$

defined for $N \neq 0$. When $M = kN$ with k and N positive integers, the fractional term collapses and the value lies on an integer ray, $M \perp N = k(N+1)^2$. An occurrence is *admissible* when it survives five explicit checks: the operation is defined on the input; the input belongs to the declared domain; a route of generation is specified; boundary and status are identified; and the occurrence survives the stated test. Admissibility and anchor status are different, the latter strictly stronger.

The program that this paper belongs to takes one idea from that grammar and tests it outside number theory. The idea is a chain,

$$\text{recursive generation} \longrightarrow \text{structural constraints} \longrightarrow \text{admissibility} \longrightarrow \text{physical persistence}.$$

A generative rule produces occurrences; the rule's own structure constrains which occurrences it can produce; the constrained set is the admissible set; and the empirical question is whether the admissible set contains the configurations that persist, or are relevant, in a physical system. The first three arrows are definitions. The fourth is the hypothesis under test, and a negative answer is as informative as a positive one: it locates the boundary at which a structural description stops being sufficient.

2.2 Generative admissibility for finite configuration spaces

We specialize the chain to a finite lattice system with n binary degrees of freedom. Let $S_0 = \{-1, +1\}^n$ be the nominal configuration space.

Definition 2.1 (Local grammar and admissible set). A *local grammar* is a family of constraints $\{c_v\}_{v \in V}$, each $c_v : \{-1, +1\}^{I(v)} \rightarrow \{0, 1\}$ depending only on the spins in a fixed incident set $I(v) \subseteq \{1, \dots, n\}$. The *admissible set* is $S_G = \{s \in S_0 : c_v(s|_{I(v)}) = 1 \text{ for all } v \in V\}$. The *contraction* is $1 - |S_G|/|S_0|$.

Definition 2.2 (Constructive generation). A generator for S_G is *constructive* if it enumerates S_G by ordered local extension, assigning spins in a fixed order and abandoning a partial assignment as soon as some constraint c_v , whose incident set is fully assigned fails, so that configurations outside S_G are never materialized as candidates.

Constructive generation is a statement about the cost of building the admissible set, and it is checkable without reference to energy. Contraction is a statement about size. Neither says anything about physics. The physical quantities enter through a Hamiltonian $H : S_0 \rightarrow \mathbb{R}$ and a window.

Definition 2.3 (Physical target and preservation). Let $E_0 = \min_{s \in S_0} H(s)$ be the global ground energy, taken over the *whole* nominal space, never over S_G . For a window $W \geq 0$ the *physical target* is $S_T(W) = \{s \in S_0 : H(s) \leq E_0 + W\}$. The grammar *preserves* the target if $S_T(W) \subseteq S_G$. A target state outside S_G is a *false exclusion*; their number is FN.

The definition of E_0 matters. If the ground energy were taken over the admissible set, the reference would silently inherit the grammar's omissions, and a grammar that excluded the true ground state could appear to preserve its own target. Every comparison in this paper uses the common, whole-space E_0 .

2.3 The exclusion gap and the tautology guard

Define the *exclusion gap*

$$\Delta_{-G} = \min_{s \notin S_G} H(s) - E_0,$$

with $\Delta_{-G} = +\infty$ if $S_G = S_0$. It is the energy cost of the cheapest configuration that the grammar forbids, measured above the global ground state.

Proposition 2.4 (Gap equivalence). *If $S_G \neq S_0$, then $S_T(W) \subseteq S_G$ if and only if $W < \Delta_{-G}$. If $S_G = S_0$ the inclusion holds for every W .*

Proof. Suppose $W < \Delta_{-G}$ and $s \in S_T(W)$. If $s \notin S_G$, then $H(s) \geq E_0 + \Delta_{-G} > E_0 + W$, contradicting $s \in S_T(W)$; hence $s \in S_G$. Conversely, suppose $W \geq \Delta_{-G}$ and let $s^* \notin S_G$ attain the minimum, so $H(s^*) = E_0 + \Delta_{-G} \leq E_0 + W$. Then $s^* \in S_T(W) \setminus S_G$, and the inclusion fails. In particular the window endpoint $W = \Delta_{-G}$ fails. The finite space guarantees that the minimum is attained. $\square$

Table 2: All 16 configurations of the 1×1 patch, grouped by energy level above the ground energy. Q_v is the vertex charge (inward minus outward arms); the grammar requires $Q_v = 0$. Energies are computed from the frozen model; $k_B T_{\text{ref}} = 0.010831 A$. Spin strings list $s_0 s_1 s_2 s_3$.

Level above ground	Configurations (spin strings)	States	Q_v	In grammar
0	$-++-$, $+- -+$	2	0	yes (ground doublet)
$0.568 k_B T$ (0.00616 A)	$----$, $+-+-$, $-+-+$, $++$	4	0	yes
$0.871 k_B T$ (0.00944 A)	eight configurations with one or three inward arms	8	± 2	no (first excluded level)
$2.349 k_B T$ (0.02544 A)	$++--$, $--++$	2	∓ 4	no (all in or all out)

The proposition makes the sufficiency question exact. Preservation at a window is decided by one number, the exclusion gap, and the interesting empirical objects are the gap itself and how it behaves as the system grows. It also exposes a *tautology risk* that the experimental design must guard against: if the target were defined as “states satisfying the grammar”, preservation would hold trivially. We therefore keep grammar membership and the energy-window target as separate objects and test the relation between them. The ground state may well lie in S_G ; that is an observation about the model, not an assumption. The target is a set of energies, not a set of rule-satisfying states.

For numerical work the window is expanded by a small tolerance τ (Section 4), giving the operational equivalence $W + \tau < \Delta_{-G}$. The distinction between the mathematical and operational forms is retained wherever certification allows.

2.4 A complete worked example: the 1×1 patch

The smallest instance exhibits every object defined above and can be enumerated by hand. A 1×1 patch has one vertex and four islands: a horizontal island on each side (indices 0 and 2) and a vertical island above and below (indices 3 and 1). There are $2^4 = 16$ nominal configurations. The grammar has one constraint, c_v : exactly two of the four arms point toward the vertex. Six of the sixteen configurations satisfy it, so $|S_G| = 6$ and the contraction is 62.5%. Table 2 lists all sixteen with their energies.

Reading the table, three facts that return in every larger patch are already visible. The ground state is a doublet, related by reversal of all four spins, and it lies in the grammar. The grammar contains a second level, four configurations at $0.568 k_B T$, between the ground state and the first excluded one, so that the admissible set is not only the ground state. And the cheapest configuration the grammar forbids is a vertex with three arms in and one out (or the reverse), at $0.871 k_B T$, which is below the window of $1 k_B T$: with $W = k_B T_{\text{ref}}$, the target is $S_T = \{\text{the 14 configurations up to the first excluded level}\}$, of which only six are in the grammar. In this smallest case the whole of Section 5 is already contained: a strong contraction, a ground state that agrees with the grammar, an exclusion gap that is a single-vertex defect, and a window that opens just past it.

2.5 How the UNNS vocabulary maps onto the experiment

We use the UNNS admissibility checks as an organizing vocabulary for the finite-lattice instance. The mapping below is an operationalization chosen for this study; it is not a theorem of the framework.

UNNS check	Instance in this study
Definedness	Each local extension step and each constraint c_v is a well-defined function of already-assigned spins.
Domain membership	The configuration lies in the declared nominal space $S_0 = \{-1, +1\}^n$ of a fixed patch.
Route specification	Ordered local extension by a fixed assignment order (the constructive generator, Definition 2.2).
Boundary / status	Membership in S_G versus exclusion; the exclusion gap Δ_{-G} is the energetic status of the boundary.
Test survival	Survival of a preregistered physical test: containment of the energy-window target.

Two features of the substrate vocabulary bear directly on the results. First, the framework insists that a boundary is part of the structure and not an artefact to be removed. The finite free-boundary patch is such a case: the lowest grammar-excluded configuration of the 4×4 model is a single-vertex violation at a boundary vertex (Section 8). We note this as an observation consistent with the vocabulary, not as a derivation from it. Second, the framework requires that a claim expose how it can fail. Each criterion in Section 4 was written so that failure is a defined, reportable outcome, and one of them failed.

2.6 The ice rule as an established constraint

The ice rule is a nearest-neighbour constraint on a lattice. In the language of symbolic dynamics, constraints of this kind define subshifts of finite type [9], and the infinite-lattice count of two-in/two-out configurations is the square-ice (six-vertex) entropy [3]. As a physical rule it is the content of Pauling’s description of proton arrangements in ice [1, 2] and of the low-energy manifold of ideal square spin ice [5]. Its role here is that of a *known, simple, local* constraint, chosen because it is frozen, checkable, strongly contracting, and physically motivated, and because it is neither identical to the energy window nor guaranteed to contain it. The validity of each theorem here does not depend on the UNNS vocabulary, and the same is true of any established tool used inside a research program. What the UNNS framework supplies is the proposition under test (that the nominal configuration space and the space that generative structure admits need not coincide, and that the relation between them has a measurable boundary), the architecture that turns this proposition into frozen criteria, and the discipline of freezing, separating, testing and reporting failure. Section 9 states what the investigation establishes for the substrate.

3 Physical system and Hamiltonian

3.1 Artificial square spin ice

Artificial spin ices are lithographically patterned arrays of single-domain ferromagnetic islands whose collective states can be designed and imaged [4, 5]. Each island is an Ising degree of freedom: its moment points along one of two directions of the island’s long axis. In the square geometry the islands sit on the bonds of a square lattice, horizontal islands on horizontal bonds and vertical islands on vertical bonds, so that every interior vertex is shared by four islands. The ice rule, two moments pointing toward a vertex and two away, selects six of the sixteen local arm patterns (Figure 1).

We use finite patches of $L \times L$ vertices with a free boundary. Each vertex has exactly four incident islands, including at the edges: the outward arms at the left and top edges are supplied by $2L$ extra

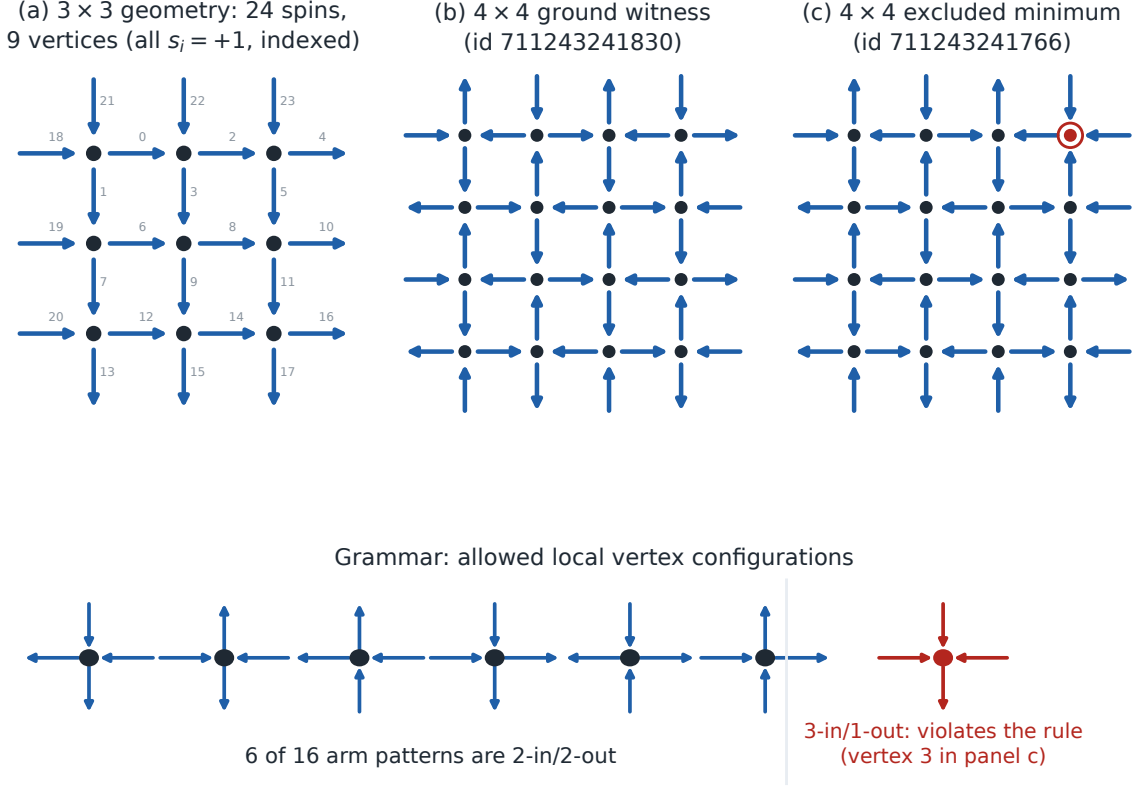


Figure 1: Physical system and generative grammar. (a) The 3×3 patch geometry with island indices (all spins in the reference orientation $s_i = +1$). (b) The 4×4 ground witness (state id 711243241830), drawn from the frozen source geometry; every vertex is two-in/two-out. (c) The corrected lowest grammar-excluded configuration (id 711243241766): the ground witness with island 6 (the right-edge horizontal island of the top row) reversed. Only vertex 3, the top-right corner, is violated (circled; three arrows in, one out). Bottom: the six allowed local vertex configurations and a violating one. All panels were generated from the frozen geometry function and the decoded state ids.

(dangling) islands, and those at the right and bottom edges by the last island of each row and column. The island count is $n = 2L(L+1)$, the vertex count is L^2 , and the number of pair interactions is $\binom{n}{2}$ (Table 3).

3.2 The frozen model Hamiltonian

The interaction model is the all-pair dumbbell model of the Dryad SPMCMC v1.00 simulator [6], without interaction cutoff and with identical islands. Each island is replaced by a pair of opposite magnetic charges $\pm q s_i$ displaced by $\pm(l/3)\mathbf{u}_i$ from its centre $\mathbf{r}_i$, with $q = (\pi/4)M(T)$ and $M(T) = M_0(1 - T/T_C)^\beta$. Here $\mathbf{u}_i$ is the island's unit vector and $s_i = +1$ means the moment points along $\mathbf{u}_i$. The pair energy is the Coulomb energy of the four charge-charge terms,

$$U_{ij} = \frac{\mu_0}{4\pi} q^2 \sum_{a,c=\pm 1} \frac{ac}{r_{ij}^{ac}} s_i s_j, \quad r_{ij}^{ac} = |\mathbf{r}_i + a \frac{l}{3} \mathbf{u}_i - \mathbf{r}_j - c \frac{l}{3} \mathbf{u}_j|,$$

Table 3: Patch sizes studied. All counts are exact. The 4×4 grammar count was not in the frozen records; it was added afterwards by an additive combinatorial record (Appendix A) computed from the frozen vertex table alone. It is a numerical-combinatorial result, not machine-checked in Lean.

Patch	Islands n	Vertices	Pairs	Nominal 2^n	Grammar $ S_G $
1×1	4	1	6	16	6
2×2	12	4	66	4,096	82
3×3	24	9	276	16,777,216	2,604
4×4	40	16	780	1,099,511,627,776	193,662

Table 4: Template parameters of the frozen model (SPMMC v1.00 defaults) and derived energy scales at the reference temperature $T_{\text{ref}} = 792$ K.

Quantity	Symbol	Value
Lattice spacing	a	260 nm
Island length / width / height	l, w, h	160 / 60 / 15 nm
Saturation magnetization, exponent, critical temperature	M_0, β, T_C	7.9×10^5 A/m, 0.5, 820 K
Boltzmann constant (exact SI)	k_B	1.380649×10^{-23} J/K
Energy scale (frozen binary64 value)	$A(792 \text{ K})$	$1.0095767595046513 \times 10^{-18}$ J
Thermal window	$k_B T_{\text{ref}}$	$1.093474008 \times 10^{-20}$ J
Window in units of A	$k_B T_{\text{ref}} / A$	0.0108310...
4×4 coupling sum	$\sum_{i < j} K_{ij} $	0.3159063271129412

and the Hamiltonian is a sum over all pairs,

$$H(s) = \sum_{i < j} U_{ij} = A(T) \sum_{i < j} K_{ij} s_i s_j, \quad K_{ij} = \frac{3}{16} \frac{wh}{l} \sum_{a,c=\pm 1} \frac{ac}{r_{ij}^{ac}}. \quad (1)$$

The energy scale is $A(T) = \kappa \mu_0 M(T)^2 V / 2$ with island volume $V = (\pi/6) l w h$. The coupling matrix K is purely geometric: it depends on the island positions and orientations and on the island dimensions, not on T or on any spin state. The sign in Eq. (1) is that derived from the simulator source; the exploratory versions v001–v004 of the program used an inverted sign, and those outputs are retained only as historical regression evidence. All results in this paper use the locked v005 Hamiltonian. The single-flip energy change is $\Delta E_i = -2A s_i h_i$ with $h_i = \sum_j K_{ij} s_j$, which reproduces the simulator’s reversal update.

3.3 Energy scales and the normalized gap

Three energy scales organize every quantity in the paper. The first is A , the scale of the Hamiltonian: all energies in the exact certification are expressed in units of A . The second is the thermal window $k_B T_{\text{ref}}$ at the reference temperature, which is the width of the physical target; it is $0.01083A$. The third is the numerical tolerance $\tau = 10^{-9} \sum_{i < j} |K_{ij}|$ used to define the operational target. For the 4×4 instance $\tau \approx 3.2 \times 10^{-10} A$.

Because H is $A(T)$ times a temperature-independent form, the ordering of configurations by energy does not depend on T ; temperature enters only through the conversion between A and $k_B T$ and through the choice of the reference $T_{\text{ref}} = 792$ K. The normalized exclusion gap is therefore

$$\delta_L = \frac{\Delta_{-G}(L)}{k_B \cdot 792 \text{ K}} = \frac{\Delta_{-G}^{(A)}(L) A(792 \text{ K})}{k_B \cdot 792 \text{ K}}, \quad (2)$$

Table 5: Pair couplings of the 4×4 patch grouped by island separation d (in units of the lattice spacing a). Computed from the frozen geometry; $\sum |K| = 0.31591$ (units of A).

Separation class	Pairs	Share of $\sum K $	Mean $ K $
$d < 0.75a$ (perpendicular neighbours at a vertex, $d = a/\sqrt{2}$)	64	64.4%	3.18×10^{-3}
$0.75a \leq d < 1.05a$ (parallel and collinear neighbours, $d = a$)	62	22.4%	1.14×10^{-3}
$1.05a \leq d < 1.6a$	144	6.4%	1.40×10^{-4}
$1.6a \leq d < 2.3a$	148	4.3%	9.3×10^{-5}
$d \geq 2.3a$	362	2.5%	2.1×10^{-5}

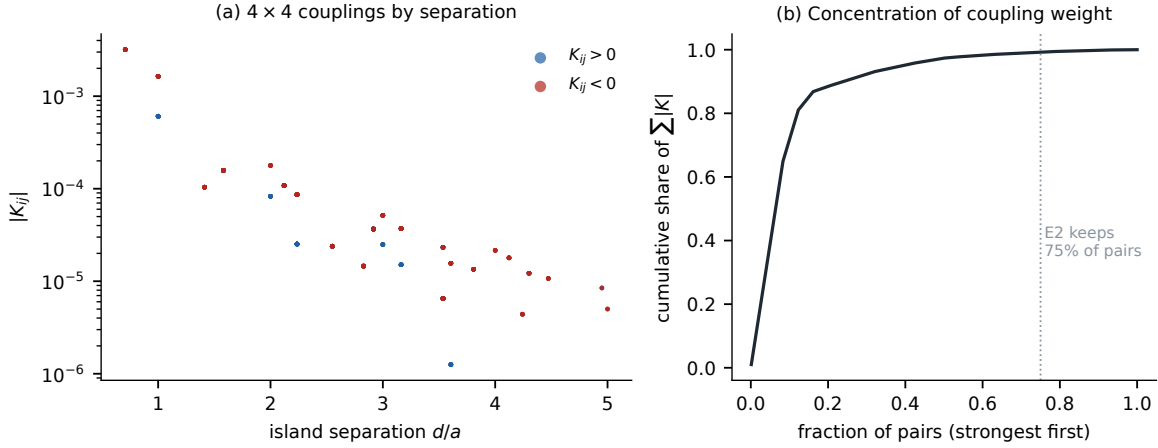


Figure 2: Structure of the 4×4 couplings. (a) Coupling magnitude against island separation, by sign; 59% of the couplings are negative. (b) Cumulative share of $\sum |K|$ carried by pairs ordered from the strongest: the strongest 16% of pairs carry 87%, and the strongest 75% carry more than 99%.

where $\Delta_{-G}^{(A)}$ is the gap in units of A . This is the convention used by the frozen prediction, the numerical solver, and the Lean theorems. As a consistency check, the exact certified gap $\Delta_{-G}^{(A)} = 12095447728404694790/2^{70} \approx 0.0102452427$ converts through Eq. (2) to $\delta_4 = 0.94591722\dots$, matching the frozen numerical value.

3.4 Structure of the couplings

The couplings are not short-ranged, and their structure determines what any exact method must carry. Table 5 groups the 780 pair couplings of the 4×4 patch by island separation, and Figure 2 shows them individually. The geometry alone, with no energy and no spin state, fixes these numbers. As a check, summing $|K_{ij}|$ over the 780 pairs reproduces the frozen instance value 0.3159063271129412.

The nearest 126 pairs (16% of all pairs) carry 86.8% of the total coupling weight; the remaining 654 pairs, separated by more than one lattice spacing, carry 13.2%. The long-range remainder is small per pair and cannot be neglected in aggregate: it exceeds the operational tolerance $\tau = 10^{-9} \sum |K|$ by eight orders of magnitude. Two consequences matter later. First, the model has no natural truncation, and an exact certificate must carry all 780 couplings, which is what the dyadic encoding does (Section 7.1). Second, the comparator E2, which discards the weakest 25% of pairs, discards less than 1% of the coupling weight, so its certified remainder bound is small and its fallback to exact evaluation is rarely needed; it remains an exact method because the bound is rigorous and not because the discarded terms are negligible in principle.

3.5 State encoding

A configuration is a 40-bit (more generally n -bit) integer. Bit i equal to 1 means $s_i = +1$. Islands are indexed as follows: for each row and column in turn, the horizontal island centred at $(a(\text{col} + \frac{1}{2}), -a\text{row})$ with $\mathbf{u} = (1, 0)$, then the vertical island centred at $(a\text{col}, -a(\text{row} + \frac{1}{2}))$ with $\mathbf{u} = (0, -1)$; then the L left-dangling horizontal islands, then the L top-dangling vertical islands. The first $2L^2$ indices are thus the bond islands in row-major order and the final $2L$ indices are the dangling islands. A vertex counts an arm as “in” when the arrow $s_i \mathbf{u}_i$ points toward the vertex. The grammar requires exactly two “in” arms at each of the L^2 vertices. We call the number $Q_v = (\#\text{in}) - (\#\text{out})$ the vertex charge; the ice rule is $Q_v = 0$ at every vertex.

3.6 Provenance and what the model is

The Hamiltonian was derived from the simulator source and cross-checked in two independent implementations: the source’s own functions embedded verbatim in a C++ probe, and a separately written Python evaluator based on explicit charge positions. They agree to relative differences of order 10^{-15} on the 4×4 witnesses, which reflects floating-point rounding. The v005 lock validated the finite-patch model, not the full unmodified simulator and not the published 24×24 production arrays. We therefore call it the *frozen model Hamiltonian*. Grammar membership and energy are computed from the same geometry; they are distinct calculations, not independent provenance. The model is a faithful, exactly specified finite system to which the questions of Section 2 can be put exactly. It is not a claim about any physical array.

3.7 Why this system

The system satisfies the requirements that the program set for a first physical test. The nominal space is finite and exactly solvable at 3×3 , and certifiable at 4×4 by branch-and-bound. The grammar is nontrivial, since it rejects most configurations. It is frozen from local geometry without reference to energy. The physical target, an energy window above the global ground state, is not identical to the grammar. The ice rule is physically motivated but is a strict constraint on a Hamiltonian that does not enforce it: the all-pair dumbbell interaction weights vertex violations by their geometry-dependent energy cost, and whether the cheapest violation is costlier than $k_B T$ is a quantitative question with no obvious a priori answer.

4 Experimental design

4.1 Lineage and chronology

The experiment is the third stage of a program on structural complexity reduction (SCR), whose organizing question is whether structural admissibility can reduce a candidate space before downstream computation while preserving a structural object defined independently of the grammar, in the narrow sense that it is defined by energy and not by grammar membership. The earlier stages fix the vocabulary of success and failure used here.

- **SCR-I-A** tested a benchmark in which the preserved object was aligned with the grammar. Its frozen outcome is that structural compression is confirmed above a representation floor: median

candidate contraction about 92%, with basin recall 1.0 in every cell. Two earlier versions of its protocol were not run, because audits showed that the target oracle was empty (v001) and that candidate identity was the wrong unit of preservation (v002); the preservation unit was changed to basin geometry in v003. At large size, rate-matched random pruning also preserved the broad basin geometry, so the supported claim was preservation plus selective enrichment, not preservation alone.

- **SCR-II** used a deliberately independent arithmetic target and found the opposite boundary: extreme contraction destroyed the target almost completely, while a problem-aligned pruner preserved it. The failure was therefore not a property of pruning in general.
- **SCR-III-A** built a controlled coupling ruler in a synthetic setting: a declared shared generative channel whose realized coverage was measured independently of survival. Its frozen verdict was `COUPLING_MAP_ONLY`: delivered coupling reproducibly improved preservation, but no tested condition reached a valid structural reduction and transfer was not confirmed.

SCR-III-B, the subject of this paper, moved from synthetic coupling to a real physical model. The question it asked was whether grammar and target are aligned in a physical system when the grammar is frozen without access to the target.

Table 6 fixes the order of the records on which the paper rests, distinguishing four phases: exploratory preflight, the primary preregistered study, the prospective 4×4 investigation, and the subsequent formalization. Order is established by the freeze documents and their hash manifests; where a phase falls on a single calendar day, we do not claim a finer ordering than the documents record.

What “preregistered” and “prospective” mean here. Both words describe frozen, hash-recorded protocol documents written before the relevant computation, held in the project’s own repository; neither refers to deposition in an external registry. The 3×3 primary protocol was frozen before the 3×3 low-energy spectrum was examined, but not before the public 1×1 and 2×2 gaps (0.871 and 0.952 in units of $k_B T$) were known, and the window of $1 k_B T$ was chosen with that knowledge. The blindness was blindness to the low-energy outcome of the 3×3 patch. For the 4×4 patch, the prediction was frozen before any 4×4 energy was evaluated; Section 6 records the departures from the planned execution protocol.

4.2 The four axes

Four questions are kept separate, reported as independent result fields, and never allowed to compensate for one another.

G, generative contraction. The grammar is generated constructively, matches an independently enumerated grammar exactly, uses no energy or target labels in its construction, and leaves at least 99% of nominal configurations unmaterialized as candidates. Mismatch or energy leakage is a failure; missing evidence is inconclusive.

P, physical preservation. Every independently certified target state lies in S_G ($FN = 0$), with complete nominal-space verification and no unresolved boundary cases. One false exclusion is a failure; incomplete verification is inconclusive, not a pass.

T, independent transfer. Whether the relation holds for an independent family of physical targets.

Table 6: Chronology of frozen records. Dates are those written in the records. Phase labels follow the epistemic role of each record.

Phase	Record(s)	Content and status
Exploratory	Preflights v005–v009 (10 Oct 2026)	v005: Hamiltonian convention lock (the sign correction). v006: feasibility; first public 1×1 and 2×2 gaps. v007: exact 3×3 verifier; grammar count 2,604. v008: solver implementation gate. v009: preregistration closure gate. Local reproduction was confirmed from terminal output supplied by the author; none is a primary result.
Primary preregistration	RATIFY_v001, SPEC_v001_FINAL, PREREG_v001_FINAL (10 Oct)	Design choices ratified and the protocol content frozen with execution explicitly not authorized. Window, tolerances, comparators, null counts and seeds fixed. Execution later authorized in a separate record.
Primary result	PRIMARY_RESULT_v001 and freeze, PRIMARY_AUDIT_v001 (10 Oct)	3×3 run: G PASS, P FAIL, T not tested, C not eligible. Erratum E1 corrects one sentence of wording. An external audit re-checked the excluded states against the frozen v005 module.
Prospective 4×4	GAP-4 protocol; PREDICTION_FREEZE_v001; first run; deviation and witness record (10 Oct)	Prediction interval and the separate directional prediction frozen from 1×1 – 3×3 evidence alone; real 4×4 then evaluated numerically; deviations from the planned protocol disclosed; verdict GAP_INCONCLUSIVE pending certification.
Formalization	SCR-IV, packages v001–v005; addendum; closure (effective 10 Oct)	Lean 4 certification of the exact encoded model; exact-arithmetic correction; both predictions proved. SCR-IV is closed; SCR-III is not (Section 10).

Not tested in this study; the protocol is unfrozen. The gap equivalence (Proposition 2.4) means that preservation of an energy-window target protected by the gap cannot establish transfer.

C, charged computational consequence. Eligible only if P passes. Optimized exhaustive Gray-code search (B0) is the sole denominator. After three warm-up runs, fifteen paired charged timings are taken with one thread, and C passes only if both the median and the minimum of the fifteen ratios $r_i = \text{cost}(B0)_i / \text{cost}(G2)_i$ are at least 1.20. If outputs are equivalent and any ratio is below 1.20 the label is THRESHOLD_NOT_MET, not “no speedup”. No confidence interval is claimed; a nominally 95% bootstrap was found to cover only about 78–85% of the time on synthetic data and was rejected as confirmatory.

4.3 Comparators

All complete-task solvers return the exact set of configuration identifiers in the operational target, using the same E_0 , window and threshold rule.

- **B0:** single-thread exhaustive Gray-code scan of the full nominal space with periodic exact resynchronization.
- **B1:** exact Hamiltonian-aware branch-and-bound with a certified admissible lower bound; a required sensitivity arm.

Algorithm 1 G2: target-blind constructive generation with incremental pair-energy updates (schematic).

Require: assignment order π of the n islands; for each vertex v , the position t_v at which its last incident island is assigned

```

1: extend( $k, s, E$ ):
2:   if  $k = n$  then
3:     emit  $s$  with energy  $E$  return
4:   end if
5:   for  $v \in \{-1, +1\}$  do
6:      $s_{\pi(k)} \leftarrow v$ 
7:      $E' \leftarrow E + v \sum_{i < k} K_{\pi(i)\pi(k)} s_{\pi(i)}$   $\triangleright$  incremental update,  $O(k)$ 
8:     if every vertex  $u$  with  $t_u = k$  satisfies the two-in/two-out rule then
9:       extend( $k + 1, s, E'$ )
10:    end if
11:  end for
12: start: extend( $0, \emptyset, 0$ )

```

- **G1:** scan of the nominal space with grammar filtering, then physical evaluation; returns $S_T \cap S_G$.
- **G2:** the constructive grammar generator with incremental pair-energy updates along its assignment tree; returns $S_T \cap S_G$. It must use competitive update algebra rather than slower direct evaluations.
- **E2:** an energy-aware exact comparator that keeps the strongest 75% of pair couplings by $|K|$ and bounds the omitted remainder rigorously, with exact fallback on ambiguous candidates.
- **C1:** a target-aware oracle, reported as a bound only.

These arms exist so that a computational claim, if one were eligible, would be measured against a strong, honest denominator. Because P failed, the C-axis comparative benchmark was not run (Section 5); this refers only to that benchmark, not to the solver and Lean costs reported elsewhere.

4.4 The comparator algorithms

The comparators are short enough to state. Algorithm 1 is the constructive generator G2; the other two comparators differ from it in what they update and what they bound.

The tree of G2 depends only on the grammar and the assignment order: the energies are carried along but never consulted in a branching decision, which is the sense in which the generator is target-blind. Its output is then filtered by the operational threshold, returning $S_T^{\tau} \cap S_G$. The exhaustive baseline B0 enumerates all 2^n states in Gray-code order, so that consecutive states differ by one spin and the energy changes by $\Delta E_k = -2A s_k h_k$ with an $O(n)$ field update, and it resynchronizes the running energy exactly at a fixed interval to bound drift. The comparator E2 splits $H = H_R + H_{\text{omit}}$, where H_R keeps the strongest 75% of pair couplings by $|K|$ (ties by pair index) and H_{omit} the rest; since $|H_{\text{omit}}(s)| \leq B = \sum_{\text{omitted}} |K_{ij}|$ for every s , a state with $H_R + B \leq E_0 + W + \tau$ is certainly in the target and a state with $H_R - B > E_0 + W + \tau$ is certainly not, and only the ambiguous remainder is evaluated exactly. The branch-and-bound B1 uses the bound of Eq. (3) below, which is the same bound on which the formal certification is built.

4.5 Null controls

Two structural nulls rewire the grammar’s vertex incidence while leaving the physical patch, Hamiltonian, window and target unchanged: N1 preserves vertex degrees, and N2 draws free incidence. Each uses 1,000 valid draws with a cap of 5,000 attempts and a master seed of 20261010, from which sub-seeds are derived deterministically. Null results are descriptive: no confirmatory p -value or null-superiority verdict was preregistered, and none is claimed.

4.6 Tolerances and numerical policy

Four numerical quantities are kept distinct: the physical-energy evaluation error, the incremental-drift bound of solvers that update energies, the reporting-level grouping tolerance, and the operational expansion

$$\tau = 10^{-9} \sum_{i < j} |K_{ij}|,$$

giving the operational target $S_T^{\tau}(W) = \{s : H(s) \leq E_0 + W + \tau\}$ with inclusive equality, and the operational gap condition $W + \tau < \Delta_{-G}$. A state whose membership cannot be certified in either direction is marked unresolved and conservatively counted in the target, and the presence of any unresolved state prevents an unqualified P pass. Numerical uncertainty is never used to remove an eligible target state silently.

4.7 The prospective gap prediction

The 4×4 question was posed in terms of the normalized gap δ_L of Eq. (2). The frozen evidence is $\delta_1 = 0.871411$, $\delta_2 = 0.952382$ and $\delta_3 = 0.942898$. The prediction rule, written before any 4×4 energy was computed, is

$$I_{R2} = [\delta_{\min} - w, \delta_{\max} + w], \quad w = \delta_{\max} - \delta_{\min},$$

which evaluates to $I_{R2} = [0.790439676056526, 1.0333534753990623]$. Two predictions were registered and are adjudicated independently.

- (i) **Range (R2):** $\delta_4 \in I_{R2}$.
- (ii) **Window:** $\delta_4 < 1$, equivalently $W + \tau \geq \Delta_{-G}$ at $W = 1 k_B T$, so that exact grammar preservation would fail on the 4×4 patch.

Because I_{R2} straddles 1, the two can receive different verdicts. The verdict rule is fixed in advance: `GAP_CONSISTENT` if a rigorous certified interval for δ_4 lies entirely within I_{R2} , `GAP_REJECTED` if it lies entirely outside, and `GAP_INCONCLUSIVE` otherwise; solver output without a matching rigorous lower bound cannot establish any verdict. Three data points support no scaling law, and consistency with I_{R2} would not confirm one. The interval is broad (width 0.243), a fact we return to in Section 6.

4.8 Verdict hierarchy and falsification register

The decision hierarchy of the program orders possible verdicts from invalid to strongest: invalid run; preservation fail; no material reduction; reduction without net gain; valid structural reduction; transfer confirmed. Nothing in this study reaches the last two, and the labels are never awarded

Table 7: Falsification register: the pre-declared failure condition of each claim and its outcome.

Claim	Would be falsified by	Outcome
G: contraction (Finding 1)	Generated set differing from independently enumerated membership; energy information used in generation; fewer than 99% of configurations avoided	Not falsified at 3×3 : exact match; 99.984% avoided (4×4 : count only)
P: sufficiency (Finding 2)	Any target state outside the grammar ($\text{FN} \geq 1$)	Falsified: $\text{FN} = 16$ of 26
Range prediction (Finding 3)	δ_4 certified outside I_{R2}	Not falsified: certified inside, margin > 0.087
Directional prediction (Finding 3)	$\delta_4 \geq 1$	Not falsified: $\delta_4 < 1$, margin > 0.054
Certification (Finding 4)	Kernel rejection of any batch or of the assembly; a lower-energy state in either objective	No batch rejected; but the frozen excluded-state value was shown not to be the minimum (Section 8)
C: computation	Failure of the paired-ratio rule at 1.20	Not eligible, since P failed

Table 8: Information state at each freeze. “Known” quantities could have influenced the protocol; “withheld” quantities were not examined before the freeze.

Stage	Known at freeze	Withheld at freeze
3×3 primary protocol	Hamiltonian lock; 1×1 and 2×2 spectra and gaps (0.871, 0.952); grammar count 2,604; 3×3 ground energy for the reference scan	3×3 low-energy spectrum above the ground state; Δ_{-G} ; all $W > 0$ outcomes; null-gap distributions
4×4 prediction	Gaps at sizes 1–3, including the 3×3 primary result; the rule for the interval; the solver design	Every 4×4 energy: ground, excluded minimum, gap
Lean certification	Frozen numerical values for both objectives and both witnesses; both predictions	Nothing relevant to the exact values; the claims being certified were set by the frozen numerics

automatically. Table 7 states, for each finding, the outcome that would have falsified it, written before the corresponding computation, together with what happened.

The register is useful in two ways. It makes visible that the study was able to fail, and that it did, once as a physical claim (P) and once as a numerical record (the excluded-state minimum). And it shows that the last item of the certification row is a falsification of a *previous result by the certification procedure*, which is the reason we describe the correction as part of the findings and not as an erratum.

4.9 What was known when

A prediction is only as informative as the information it was made without. Table 8 records, for each stage, which quantities were known when the protocol was frozen and which were withheld.

The third row is worth reading with care. The certification did not test a prediction made without knowledge of the answer; it tested numerical values that had been reported. That the certificate disagreed with one of them is what makes it informative as a check of the record, and it is the reason

that the corrected value is described as a result of the certification and not as a confirmation.

4.10 Disclosed limitations of the design

We list the departures and weaknesses that a reader should weigh.

- The $1 k_B T$ window was chosen knowing the 1×1 and 2×2 gaps, both of which already implied failure at that window.
- The verifier shares authorship with the grammar code. It is separately implemented and cross-validated against independent models, but it is not externally independent.
- The incremental-drift tolerance in the exact 3×3 verifier was changed post hoc from 10^{-12} to 10^{-10} during preflight v007; this is historical and disclosed.
- The first launch of the 3×3 primary run was killed by the host sandbox during the null phase. Partial artifacts were preserved, and the driver was revised for checkpoint and resume only; a hash comparison shows that no constant, tolerance, window, seed or decision rule changed. The results were recomputed from scratch under the revised driver.
- The held-out discrimination study (axis T) and a synthetic feasibility campaign remain unexecuted.

5 Structural contraction and physical sufficiency: the 3×3 study

5.1 Small patches: complete spectra

Before the preregistered 3×3 study, the complete spectra of the 1×1 and 2×2 patches were available, since 2^4 and 2^{12} configurations can be enumerated at once. These are the frozen exploratory results from which the window and, later, the 4×4 prediction were derived (Table 6); we recomputed both spectra from the frozen energy module for this paper, with agreement to 10^{-15} in δ with the recorded gaps (0.871411 and 0.952382). Figure 3 shows them.

Three features recur in both patches and are what the larger-size tests probe. The ground level is a doublet related by global reversal and lies inside the grammar. Between the ground level and the first excluded level there are admissible levels (a level with four states at 1×1 , one with eight at 2×2): the grammar contains a structured part of the low-energy spectrum, not only the ground state. And the first excluded level sits below $1 k_B T$, at 0.871 and 0.952. Above the gap, admissible and excluded levels interleave at 2×2 , so that no energy cut above δ_2 separates the two sets. The grammar’s boundary is thus energetically sharp only up to the gap.

5.2 The 3×3 instance

This subsection and those that follow report the primary preregistered experiment. The instance is the 3×3 patch ($n = 24$, 276 pair couplings, 16,777,216 nominal configurations). The exact global ground energy, certified before the primary run by an independent exhaustive scan, is $E_0 = -9.271994751803727 \times 10^{-20}$ J. The primary window is $W = k_B T_{\text{ref}} = 1.093474008 \times 10^{-20}$ J, and the numerical tolerance is $\tau = 10^{-9} \sum |K|$. The 3×3 results below are SCR-III numerical results; they are not machine-checked.

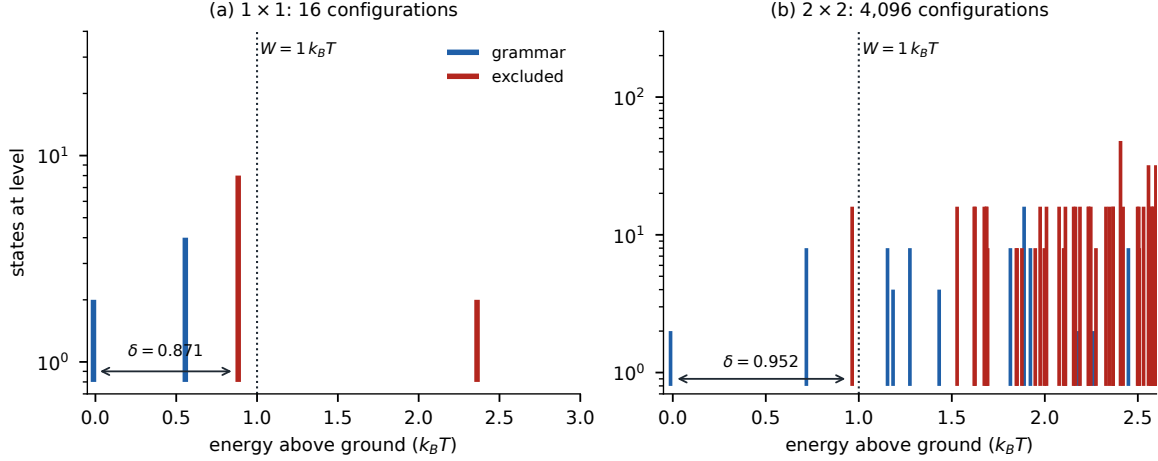


Figure 3: Complete spectra of the small patches. States per energy level above the global ground energy, in units of $k_B T$, separated into grammar-admissible (blue) and grammar-excluded (red) states, on a logarithmic axis. In both patches the ground level (two states, a globally reversed doublet) lies in the grammar and the first excluded level (8 states at 1×1 , 16 at 2×2) lies below the window edge. At 2×2 the spectrum is shown to $2.6 k_B T$.

5.3 Finding I: generative contraction (G passes)

The frozen grammar admits exactly 2,604 configurations. The same set is produced by three routes: the constructive generator G2, the full-scan generator G1, and an independent NUMPY audit. G2 materialized only the 2,604 admissible states, visiting 25,786 tree nodes, so that 99.984% of the 16,777,216 nominal configurations were never constructed as candidates. G2’s traversal is identical at all three target windows: it is blind to the target by construction. Both parts of the G criterion hold: exact match with independently enumerated membership, and contraction well beyond the 99% threshold. The contraction increases quickly with size (Figure 4a): 6 of 16 configurations survive at 1×1 (62.5% contraction), 82 of 4,096 at 2×2 (97.998%), and 2,604 of 16,777,216 at 3×3 (99.9845%). For 4×4 the grammar count was not part of the frozen records. It depends only on the frozen vertex-incidence table and not on the 780 couplings, so it was computed afterwards, without any energy evaluation, by exact counting: 193,662 of $2^{40} = 1,099,511,627,776$ configurations satisfy the ice rule at all 16 vertices, a contraction of 99.99998% (one admitted configuration per 5.68×10^6). Three routes agree: a variable-elimination count and a pruned enumeration over the frozen table of the real instance, and a count from a vertex table rebuilt from the patch geometry; the last also reproduces the counts 6, 82 and 2,604 of the smaller patches, and its 4×4 table equals the frozen one (Appendix A). The counts grow by factors of 13.7, 31.8 and 74.4 from one size to the next while the nominal space grows by 256, 4,096 and 65,536; four sizes are shown and no scaling law is claimed. The 4×4 figure is a count only: no constructive generator was run at 4×4 , so the full G criterion was not re-evaluated there, and the count is not a Lean result.

5.4 Finding II: physical discrimination (P fails)

At $W = k_B T$ the physical target, certified identically by three solvers (B0, B1 and E2 at 75% truncation) and by the audit, contains 26 states. The grammar solvers G1 and G2 return the same 10-state subset. Sixteen target states are false exclusions, so $FN = 16$ and the criterion fails. Table 9 gives the primary

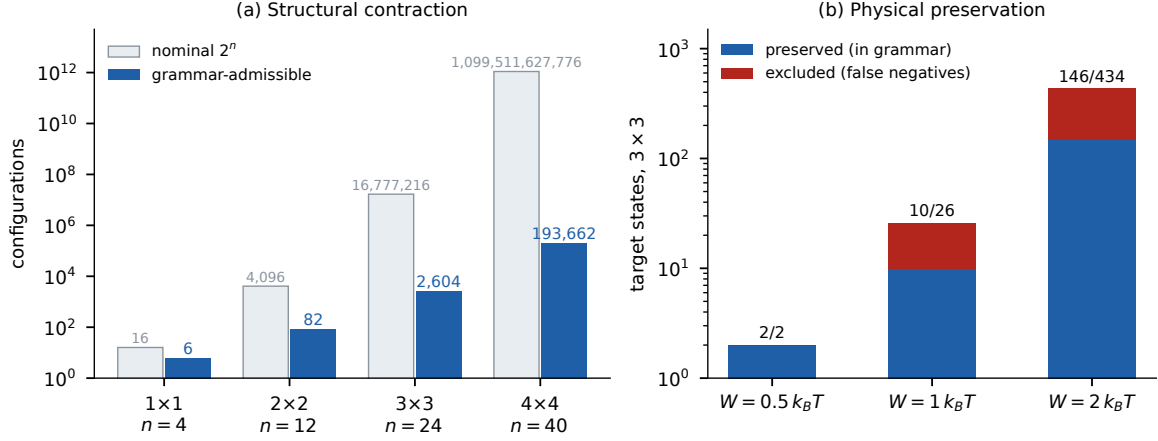


Figure 4: Contraction versus preservation. (a) Nominal and grammar-admissible configuration counts for the four patch sizes (log scale). The 4×4 bar (193,662) is the additive combinatorial count, not a result of the preregistered runs. (b) 3×3 physical target at the three preregistered windows: states preserved by the grammar (blue) and false exclusions (red). The grammar preserves the target only at $0.5 k_B T$, where the target consists of the two globally reversed ground states.

Table 9: 3×3 target and preservation at the three preregistered windows. $\Delta_{-G} = 0.9429 k_B T$.

Window W	Target states	In grammar	FN	W/Δ_{-G}	Preserved
$0.5 k_B T$	2	2	0	0.530	yes
$1 k_B T$ (primary)	26	10	16	1.061	no
$2 k_B T$	434	146	288	2.121	no

window together with the two sensitivity windows, which were registered before the run and are reported regardless of outcome; neither can rescue the primary failure.

The failure is a gap statement. The failure is not an accident of counting; it is predicted exactly by Proposition 2.4. After unblinding, the exclusion gap of the 3×3 patch is $\Delta_{-G} = 1.0310 \times 10^{-20} \text{ J} = 0.9429 k_B T$. The window-to-gap ratio is $W/\Delta_{-G} = 1.0606$, so that $W + \tau > \Delta_{-G}$ and preservation must fail; at $0.5 k_B T$ the ratio is 0.530 and preservation holds. The smallest excluded target state has identifier 2726297, and the sixteen false exclusions lie at $0.9429 k_B T$ above the ground state. An independent follow-up check evaluated the excluded states with the frozen v005 module, using energies from the explicit charge-position model in joules and the ice rule from vertex geometry rather than from the incidence masks used by the solvers: all 26 target states lie within $W + \tau$ of E_0 , exactly 10 satisfy the geometric ice rule and 16 do not, and the 16 are exactly the reported false exclusions. The minimum over the excluded set computed this way, $0.9428984210312306 k_B T$, agrees with the reported gap to about 10^{-14} . Three thousand random geometrically excluded states, drawn as a spot check, were all above the gap.

The part of the target that is preserved. The grammar does preserve the target at $0.5 k_B T$. That target is the two states of the ground-state doublet related by global reversal of all spins. This is the guaranteed-alignment regime: the global ground state lies in S_G , and nothing else is within reach of the window. The test is informative only when the window opens enough to admit the first excited levels, and the first of those contains grammar-violating states.

Table 10: Structural nulls (1,000 valid draws each; descriptive only). Survivor counts are grammar sizes; the real grammar has 2,604. The gap is in units of A (real gap: 0.0102).

Family	Survivors (min / median / max)	Gap median	Gap max	Preserve at $k_B T$
N1 degree-preserving	2552 / 2970 / 7560	0	0.0077	0/1000
N2 free incidence	256 / 2432 / 6336	0	0	0/1000

5.5 Null controls

Table 10 summarizes the two null families. The null results are descriptive. They show that the preservation failure is not peculiar to the real grammar: none of the 2,000 valid nulls preserves the target at $1 k_B T$. They also show what is peculiar to it. In every one of the 1,000 free-incidence nulls the exclusion gap is zero, meaning that the grammar excludes a ground state outright, and in the degree-preserving family the median gap is zero and the maximum is $0.0077A$, below the real gap of $0.0102A$. The real grammar is the only one of the 2,001 grammars (the real one plus 2,000 nulls) with a gap above $0.0077A$. In this specific sense the ice rule is far from arbitrary: among the 2,001 grammars it alone leaves a gap above $0.0077A$, whereas most nulls exclude a ground state outright. That is a descriptive observation, not a significance test.

For N1, 1.6% of the nulls have at most 2,604 survivors. For N2, nine draws were rejected for an empty grammar and 1,009 draws were made in total to obtain 1,000 valid nulls.

5.6 Consequences for the other axes

Because P failed, the computational criterion C was not eligible and the paired B0/G2 benchmark was not run. No speed claim of any kind is therefore made or implied. The criterion T remains untested. The excluded configurations are not claimed to be physically nonexistent: the physical statement is only that the grammar fails to contain every configuration within $1 k_B T$ of the ground state, and, at the 4×4 level, the cheapest excluded state is a single-vertex defect (Section 8).

5.7 What the 3×3 study establishes

It establishes, for this model and window, a pair of facts of different character. A local, energy-blind grammar can contract the space by nearly four orders of magnitude and be constructed efficiently, and the same grammar is not sufficient for the physical sector of interest, because the cheapest excluded state lies just inside the window. The two facts are not in tension: contraction is a property of the grammar's size, sufficiency is a property of the relation between its boundary and the energy landscape, and Proposition 2.4 says exactly where one ends and the other begins.

6 The prospective 4×4 investigation

The 3×3 study ends with a quantity that can be predicted: the normalized gap δ_L , which decides preservation at every window through Proposition 2.4. Three values were in hand, $\delta_1 = 0.8714$, $\delta_2 = 0.9524$ and $\delta_3 = 0.9429$. The 4×4 patch has $2^{40} = 1,099,511,627,776$ configurations, too many for exhaustive scanning on the equipment of this program, and it therefore tests two things at

once: whether the gap scale persists, and whether the gap can be established exactly at a size where enumeration is no longer an option.

6.1 Instance

The 4×4 instance has $n = 40$ spins, 16 constrained vertices and 780 pair couplings, built from the frozen source geometry with orientation and bit conventions inherited from the 3×3 study. The instance file records geometry and couplings only; it contains no energy. The coupling sum is $\sum_{i < j} |K_{ij}| = 0.3159063271129412$ (in units of A), which sets the operational tolerance $\tau \approx 3.2 \times 10^{-10}A$. Two checks confirm that the file reproduces the frozen geometry: the coupling matrix equals the matrix recomputed from the geometry, and the serialized vertex masks equal the masks derived from it.

6.2 Frozen prediction

Before any 4×4 energy operation, the prediction of Section 4 was frozen:

$$\delta_4 \in I_{R2} = [0.790439676056526, 1.0333534753990623],$$

and, separately, $\delta_4 < 1$. The interval was set entirely from unblinded 1×1 – 3×3 evidence. Its width, 0.243, is the full spread of the three observations added to each side. We make two remarks about its strength.

First, the interval is broad, and the value that was eventually found, 0.9459, lies inside the *range of earlier observations* $[\delta_1, \delta_2] = [0.871, 0.952]$ itself, so passing I_{R2} is closer to interpolation than to extrapolation. The interval test is therefore a modest one. Second, the directional prediction $\delta_4 < 1$ is a more consequential statement: it predicts that exact grammar preservation at $1 k_B T$ *fails* on the 4×4 patch. It is nevertheless supported by an observed pattern ($\delta_1, \delta_2, \delta_3$ are all below 1), so what it tests is the persistence of that pattern, not an independent theory. We state both remarks because the value of the exercise lies in the discipline (a rule frozen first, evaluated after) and in the exact certification that followed, not in a claim that the prediction was hard.

6.3 Numerical solver

The 4×4 optimization used a depth-first branch-and-bound solver (`solver_v010`) in double precision. Spins are assigned in index order. At a partial assignment the solver evaluates the bound

$$B(p) = \underbrace{\sum_{\text{assigned pairs}} K_{ij} s_i s_j}_{\text{acc}} - \sum_{j \text{ unassigned}} \left| \sum_{i \text{ assigned}} K_{ij} s_i \right| - \sum_{\substack{j < h \\ \text{unassigned}}} |K_{jh}|,$$

which lower-bounds the energy of every completion in exact arithmetic: any completion adds pair terms among unassigned spins, each at least $-|K|$, and cross terms between assigned and unassigned spins, each at least minus the absolute field. A subtree is pruned when its bound exceeds the incumbent by more than a guard $10^{-11} \max(1, \text{scale})$ (in units of A). Two objectives were solved, each with a budget of 150 seconds: the minimum over all configurations, giving E_0 , and the minimum over configurations that violate the grammar at one or more vertices, giving the excluded-state minimum. For the second objective the solver prunes using a permissive feasibility filter based on the four-spin vertex constraints.

Table 11: Frozen numerical results for the 4×4 instance (solver output, binary64, in units of A) and their independent recomputation. The excluded-state value is the original numerical result; the exact value differs from it by $928/2^{70} A$ (Section 8).

	Ground witness	Excluded-state witness
State id	711243241830	161487427942
Solver energy	-0.16364841082886986	-0.1534031680911189
Recomputed energy (60-digit rational)	-0.16364841082886983	-0.15340316809111887
$ \text{solver} - \text{recomputed} $	3.8×10^{-17}	4.1×10^{-17}
Solver nodes visited	700,405	1,316,073
Violated vertices (grammar)	0	1 (vertex 3)

Before the real run the solver passed the public 1×1 and 2×2 tests against exhaustive scans, with 19 and 31 nodes for the two objectives at 1×1 and 125 and 359 at 2×2 .

The solver does not emit a machine-checkable proof of optimality. A completion flag means that the depth-first search finished under the implemented floating-point pruning rule; the reported equality of lower and upper bounds is an assignment made at completion and is not an independent lower-bound proof. These are exactly the points that the later formal certification addresses.

6.4 Numerical results

Both searches completed. Table 11 gives the witnesses and energies. The numerical gap in units of A is 0.01024524273775096, giving $\delta_4 = 0.9459172223430824$ through Eq. (2). The value lies inside I_{R2} , with a margin of 0.0874 to the nearest interval edge, and below 1, with a margin of 0.0541.

6.5 Independent witness verification

A separate Python checker, which invokes neither the solver binary nor its branch-and-bound code, read each state identifier, evaluated all 780 pair terms with exactly rounded summation, and derived the vertex incidences from the serialized masks. The two witnesses are feasible with the stated energies, and the violating-vertex sets from the masks coincide with those from the geometric ice rule. The charge-position model and the K -matrix model agree to a relative difference of 2.3×10^{-15} . Recomputing the source coupling formula in 60-digit arithmetic from the same binary64 inputs shows how far the stored couplings are from the formula: the largest relative error per pair is 1.9×10^{-13} , the sum of absolute errors relative to $\sum |K|$ is 1.5×10^{-15} , and the resulting bound on the shift of the gap is $9.4 \times 10^{-16} A$, or 8.7×10^{-14} in δ_4 . As consistency checks only, the ground witness is a minimum among its one- and two-spin-flip neighbours, and none of the 817 grammar-violating neighbours of the excluded witness has lower energy. Neither check says anything about global optimality.

6.6 The four sizes side by side

Table 12 collects the quantities of the four patches. The first three rows are the frozen evidence from which the prediction was made; the last is the 4×4 result.

The gap in units of A rises from 0.00944 to 0.01032 between the first two sizes and then stays within 0.5% of 0.0102 at the next two sizes. The 3×3 entry in units of A is derived from the recorded gap in joules, $1.0310349 \times 10^{-20} \text{ J}$, divided by $A = 1.00957676 \times 10^{-18} \text{ J}$.

Table 12: Exclusion gap by patch size. $\delta_L = \Delta_{-G}/k_B T_{\text{ref}}$. Sizes 1–3 are the frozen exploratory and primary results; the 4×4 gap is the certified exact value. “First excluded level” degeneracy is the number of states recorded at that level. FN is the number of target states outside the grammar at $W = k_B T_{\text{ref}}$.

Patch	Spins	δ_L	Gap in units of A	First excluded level	Preserved at $k_B T$
1×1	4	0.871411	0.0094383	8 states	no
2×2	12	0.952382	0.0103153	16 states	no
3×3	24	0.942898	0.0102125	16 states	no (FN = 16)
4×4	40	0.945917	0.0102452	not recorded	no (implied by $\delta_4 < 1$)

6.7 What a numerical certificate could and could not have shown

The audit supplied one rigorous ingredient and one that was missing. The coupling-precision bound is rigorous: recomputing the source formula at high precision from the same binary64 inputs bounds the shift of the gap by $9.4 \times 10^{-16} A$, which is 8.7×10^{-14} in δ_4 . The missing ingredient was a justified global lower bound for the optimization itself. Had the solver’s pruning used directed rounding or interval arithmetic, the resulting certified interval for δ_4 would have been of width of order its guard, 10^{-11} in units of A (about 10^{-9} in δ_4), far narrower than the margins of 0.054 and 0.087. The prediction verdicts were never in numerical doubt; what was missing was the justification, and that is the specific thing that Section 7 supplies. This is worth stating because it clarifies what the formal result adds: not a different answer to the question that the margins had already settled, but a proof that the answer is what the optimizer reported, and an exact value for it.

6.8 Verdict under the frozen rules, and deviations

The frozen decision rule does not permit a solver’s best values to establish a verdict without matching rigorous lower bounds. The recorded verdict at this stage was therefore `GAP_INCONCLUSIVE` for both the range prediction and $\delta_4 < 1$, with both provisionally supported by the computed candidate gap. We did not recast the uncertified floating-point pruning as a physical proof. The deviation record, written after the run, lists the following.

- The prediction freeze stated that no real 4×4 instance could be generated or evaluated without a separate execution authorization. The authorization was subsequently supplied as an explicit request, after the planning documents had been written, and the instance was generated with an explicit authorization flag. The record documents a procedural deviation, not retroactive compliance.
- The planned outcome-blind feasibility campaign (200 synthetic instances, a 90% pass fraction, and a correctness gate at 28 and 32 bits) was not completed. The solver passed the public 1×1 and 2×2 tests and a generator guard check, and was then run on the real instance.
- The solver emitted no optimality proof tree and no interval-arithmetic certificate, and its pruning relied on an ad hoc guard rather than on directed rounding or interval bounds.

The frozen prediction values were not altered, and no replacement experiment was run. In consequence the word “prospective” applies to the prediction, which was frozen before the real evaluation; it does not apply to the execution protocol, whose planned feasibility campaign was not carried out, nor to the certification, which followed the run. We use the word in that restricted sense throughout.

6.9 What the numerical investigation shows

Numerically, the gap scale persists: $\delta_4 = 0.9459$ is within 0.003 of the 3×3 value 0.9429 and inside the range of all three earlier observations. The lowest excluded state found is, as we show in Section 8, a single flipped island at the lattice boundary, so the first excitation outside the grammar is a single-vertex defect at the corner. What the numerical investigation cannot show is that nothing lower exists: that requires a certificate. The next two sections supply it.

7 Formal mathematical certification

The numerical 4×4 investigation leaves a precise gap: two minima over a space of 2^{40} configurations are reported by a floating-point solver that does not justify its pruning. This section develops the certification that closes the gap. The aim is a machine-checked proof of two exact statements about the encoded model: that no configuration has energy below a claimed value U_0 and one configuration attains it, and that no grammar-violating configuration has energy below a claimed value U_n and one attains it. The certification was carried out in Lean 4 [10] using core Lean only, without the mathematical library. The approach is an instance of proof-carrying results [7]: an untrusted program produces a certificate, and a small trusted checker, proved sound once and for all, accepts or rejects it.

7.1 The exact model

The first step is to replace floating-point numbers by exact ones. The 780 couplings K_{ij} of the frozen instance are binary64 numbers. Each is a dyadic rational, and the encoder verified that every one is an integer multiple of 2^{-70} . Define the integer couplings $\hat{K}_{ij} = 2^{70}K_{ij}$ and the scaled Hamiltonian

$$\hat{H}(s) = \sum_{i < j} \hat{K}_{ij} s_i s_j \in \mathbb{Z}, \quad H(s) = A \cdot 2^{-70} \hat{H}(s),$$

where $s_i = \pm 1$ is decoded from the bits of a 40-bit identifier. All energies in the certificates are integers $\hat{H}$; “in units of $2^{-70}A$ ”. The theorems quantify over the identifier $q \in \mathbb{N}$ with $q < 2^{40}$, and nothing in them enumerates identifiers. We call the pair $(\hat{K}, \text{decoding})$ the *encoded model*. Three data enter the encoded model without derivation in Lean, and they define the trusted base of the data: the integer matrix $\hat{K}$, transcribed by a script from the frozen instance file and compared against it; the 16 vertex masks that define the grammar; and the value of $A(792\text{ K})$ as the frozen binary64 number, taken exactly.

7.2 An exact lower bound and sound pruning

Branch-and-bound requires a bound that is valid for every completion of a partial assignment. Fix the assignment order $0, 1, \dots, 39$ and a prefix $p = (p_0, \dots, p_{d-1}) \in \{\pm 1\}^d$. Write $\text{acc}(p) = \sum_{i < j < d} \hat{K}_{ij} p_i p_j$ for the energy among assigned spins, $f_j(p) = \sum_{i < d} \hat{K}_{ij} p_i$ for the field that the prefix exerts on unassigned spin j , and

$$T_d = \sum_{d \leq j < h < 40} |\hat{K}_{jh}|$$

for the total absolute coupling among unassigned spins. The bound is

$$B(p) = \text{acc}(p) - \sum_{j \geq d} |f_j(p)| - T_d. \quad (3)$$

Theorem 7.1 (Lower bound and pruning, `lower_bound_ids`, `prune_sound_ids`). *For every prefix p of length $d \leq 40$ and every completion q of p , $B(p) \leq \hat{H}(q)$. Consequently, if $U \leq B(p)$ then $U \leq \hat{H}(q)$ for every completion q of p .*

Proof idea. Split $\hat{H}(q)$ into the energy among the assigned spins, the cross terms between assigned and unassigned spins, and the terms among unassigned spins. The first equals $\text{acc}(p)$. Each cross term $\sum_{j \geq d} f_j(p)q_j$ is at least $-\sum_j |f_j(p)|$ since $q_j = \pm 1$, and the last sum is at least $-T_d$. The Lean proof carries this out by structural recursion over finite sums of exact integers. $\square$

Proof in more detail. Let q extend p , so $q_i = p_i$ for $i < d$ and $q_j = \pm 1$ for $j \geq d$. Splitting the double sum $\sum_{i < j}$ by whether each index lies below or at least d gives three groups:

$$\begin{aligned} \hat{H}(q) = & \underbrace{\sum_{i < j < d} \hat{K}_{ij} p_i p_j}_{= \text{acc}(p)} + \sum_{j \geq d} q_j \underbrace{\sum_{i < d} \hat{K}_{ij} p_i}_{= f_j(p)} + \sum_{d \leq j < h < 40} \hat{K}_{jh} q_j q_h. \end{aligned}$$

Since $q_j \in \{\pm 1\}$, each term $q_j f_j(p) \geq -|f_j(p)|$, and since $q_j q_h \in \{\pm 1\}$, each term $\hat{K}_{jh} q_j q_h \geq -|\hat{K}_{jh}|$. Summing the three groups gives $\hat{H}(q) \geq \text{acc}(p) - \sum_{j \geq d} |f_j(p)| - T_d = B(p)$. The only properties used are that the spins are ± 1 and the identity of the decomposition, so the bound holds for an arbitrary integer coupling matrix. The Lean development states the decomposition as an equation of finite sums defined by structural recursion, proves it by induction on the split index, and then applies monotonicity of sums for the two inequalities. For pruning, $U \leq B(p) \leq \hat{H}(q)$ gives $U \leq \hat{H}(q)$ for every completion by transitivity of the order on the integers, and the strict version holds by the same argument. $\square$

The theorem is proved for the abstract bound and then specialized in two ways: a bridge theorem identifies the energy of a decoded identifier with the abstract $\hat{H}$, and a loop-faithful transcription of the solver's bound and path accumulator is proved equal to the abstract one, so that the structure of the numerical solver's real-number bound is itself justified. The theorem also covers pruning with ties: a prune keeps every state that ties the incumbent, which is why the certificates establish minimum *values* and not minimizer sets. The solver's binary64 rounding and its outward guard are not formalised; the guard is covered only for the real-number structure, with a guard of any non-negative size.

7.3 Proof-tree certificates

A certificate that every completion has $\hat{H} \geq U$ is a binary proof tree over the fixed assignment order, encoded in pre-order as a natural number. Each node carries a two-bit token:

Token	Name	Condition checked
0	prune-bound	$U \leq B(p)$ at this prefix (at depth 40, $B(p) = \hat{H}(p)$)
1	split	both children, $s_d = -1$ then $s_d = +1$, are certified (depth < 40)
2	grammar-leaf prune	depth 40 and the identifier lies in the grammar (non-grammar objective only)

Algorithm 2 Checker `chk` (schematic). Recomputes every bound exactly; accepts a subtree only if the token is justified.

Require: claimed value U , rows R_d , grammar predicate `leafOK`, certificate bits, state $(\text{acc}, F, T, \text{bits})$ at depth d , position `pos`

```

1:  $t \leftarrow$  token at pos
2: if  $t = 0$  then
3:   return pos + 2 if  $U \leq \text{acc} - \sum |F| - T$ , else fail
4: else if  $t = 2$  and  $d = n$  then
5:   return pos + 2 if leafOK(bits), else fail
6: else if  $t = 1$  and  $d < n$  then
7:    $f \leftarrow$  head of  $F$ 
8:    $\text{pos}_1 \leftarrow$  CHK on child  $s_d = -1$  with  $\text{acc} - f$ , field update  $F' = F_{\text{tail}} - R_d$ ,  $T - \sum |R_d|$ 
9:   return CHK on child  $s_d = +1$  at  $\text{pos}_1$  with  $\text{acc} + f$ , field update  $F' = F_{\text{tail}} + R_d$ , bits |  $2^d$ 
10: else
11:   fail
12: end if

```

After the last token the certificate carries a sentinel bit 1, so the certificate is consumed exactly. The checker (`chk`, `checkAll`) recomputes every bound from the path, with integer arithmetic, carrying the state $(\text{acc}, F, T, \text{bits})$ where $F = [f_d, \dots, f_{39}]$ and `bits` is the identifier of the prefix. A child updates the state in $O(n)$ from one coupling row. Nothing about the search that produced the certificate is trusted.

Theorem 7.2 (Checker soundness, unrestricted form; `checkAll_sound`, from `chk_sound`). *Let P be any property of ± 1 configurations and let `leafOK` be a predicate on depth- n identifiers such that `leafOK`(b) implies that no configuration with property P agrees with b on all n spins. If `checkAll` accepts a certificate for the claim U with this `leafOK`, then $\hat{H}(s) \geq U$ for every ± 1 configuration s with property P . The induction is carried by `chk_sound`, which states the same conclusion for the subtree below any prefix whose threaded state satisfies the invariant.*

The two objectives of this paper are the two instances of Theorem 7.2 that are stated and proved in the package:

- *Ground objective* (`ground_min_of_cert`, `ground_min_of_frontier`): P is trivially true and `leafOK` is constantly false, so the conclusion is $\hat{H} \geq U$ for every configuration. No leaf is ever closed by the grammar rule.
- *Non-grammar objective* (`nonG_min_of_cert`, `nonGT_min_of_frontier`): P is “is not in the grammar”. A depth- n leaf is closed by `leafGrammar`, which is true only if the leaf’s own pattern satisfies the grammar; `leafGrammar_sound` shows that such a leaf cannot be agreed with by a non-grammar configuration. The conclusion is $\hat{H} \geq U$ for every configuration *not* in the grammar, and nothing is asserted about configurations in the grammar.

Remark 7.3 (Why the bound is exact and not approximate). Every quantity in $B(p)$ is an integer computed by additions, subtractions and absolute values from the integer matrix $\hat{K}$. There is no rounding, no tolerance and no floating-point comparison anywhere in the checker. The claim $U \leq B(p)$ is a comparison of two integers. This is what distinguishes the certificate from the numerical solver, whose test is a floating-point comparison with a guard; the guard makes the solver conservative in practice, but it does not make its pruning provably sound.

Table 13: Toy proof tree for the claim $\min H \geq -16$ ($n = 5$). B is the bound of Eq. (3) at the prefix; a node is pruned when $B \geq U = -16$. Prefixes are written with $\pm$ for $s_i = \pm 1$.

Prefix	B	Token	Prefix	B	Token
(root)	-24	split(1)	+	-24	split(1)
-	-24	split(1)	+ -	-18	split(1)
--	-16	prune(0)	+ - -	-14	prune(0)
- +	-18	split(1)	+ - +	-12	prune(0)
- + -	-12	prune(0)	+ +	-16	prune(0)
- + +	-14	prune(0)			

The proof is an induction on the remaining depth with an invariant stating that the threaded state equals (acc, F, T) computed from the prefix. The inductive case for a split token combines the two children; the base cases are Theorem 7.1 (token 0) and the definition of the objective (token 2). The theorem is proved once for the generic checker and then instantiated for the real instance.

7.4 A worked toy certificate

A small example shows the objects concerned. Take $n = 5$ spins with integer couplings $K_{01} = 3$, $K_{02} = -2$, $K_{03} = 1$, $K_{04} = -1$, $K_{12} = -4$, $K_{13} = 2$, $K_{14} = 1$, $K_{23} = 5$, $K_{24} = -3$, $K_{34} = 2$, and $H(s) = \sum_{i < j} K_{ij} s_i s_j$. Exhaustive evaluation of the 32 configurations gives the minimum -16 , attained by $(-, -, -, +, -)$ and its global reversal. To certify the claim $U = -16$, the generator explores the tree of prefixes in the fixed order. At the root, $B = -24 < U$, so it splits; at the prefix $(-)$, $B = -24$, so it splits again; at $(--)$, $B = -16 \geq U$, so that subtree is closed by a prune-bound token. Table 13 lists the full tree. It has 11 nodes: five splits and six prunes. In pre-order the token sequence is 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, encoded as a natural number with a sentinel bit.

The checker recomputes each B from the prefix and accepts. It would reject the claim $U = -15$: no certificate exists, since the generator reaches a leaf with $H = -16 < U$ where the bound equals the energy and cannot close the node. The same mechanism rejected the false claims of the real instance (controls below). At the scale of the real instance the trees are larger by five orders of magnitude, but the objects are the same: integer bounds recomputed from prefixes, accepted by a checker whose soundness is proved once.

7.5 Partitioned certificates and kernel batches

A single proof tree for the full instance is far too large for one kernel check: kernel memory grows roughly linearly in the number of checked nodes, at about 0.5–0.7 MB per node, so a monolithic check would exhaust memory. The remedy is a partition. Fix a frontier depth D_0 and certify each of the 2^{D_0} prefixes of length D_0 *independently*, each with its own certificate (`subOK`); a prefix already excluded by a shallower bound is a one-token certificate. Coverage is trivial: every configuration extends exactly one prefix $b < 2^{D_0}$. We use $D_0 = 14$, giving 16,384 prefixes per objective. The soundness of one prefix certificate is `subOK_sound`, and the coverage statement assembles them.

Theorem 7.4 (Frontier assembly). *Suppose the frontier prefixes are listed in batches, and each batch is accepted by `batchOK`. Then every configuration is covered by exactly one accepted prefix certificate, and consequently $\hat{H}(q) \geq U$ for every configuration (respectively every non-grammar configuration) q .*

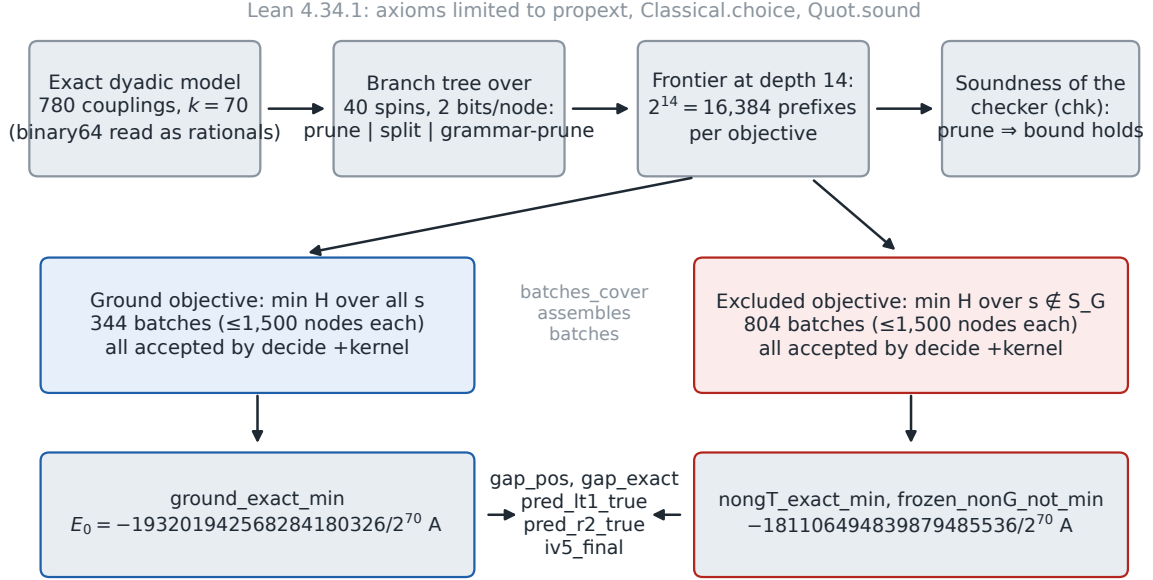


Figure 5: Proof-carrying optimization. The exact dyadic model feeds an untrusted proof-tree generator and a trusted checker whose soundness is proved. The tree is partitioned at depth 14 into $2^{14} = 16,384$ prefixes per objective, grouped into 344 ground and 804 excluded-state batches of at most 1,500 nodes, each accepted by a kernel-checked theorem. Frontier assembly gives the two exact minima; the gap and prediction theorems follow from the minima alone.

The statements are `batches_cover` (coverage) and `ground_min_of_frontier, nongT_min_of_frontier` (the two objectives).

Each batch is a module containing a single theorem of the form `batchOK...=true`, discharged by `decide+kernel`. The batches are capped at 1,500 proof nodes so that peak memory is per-batch and not per-tree. The ground objective requires 344 batch modules and the corrected non-grammar objective 804, in total 1,148. The files `Fin_g` and `Fin_n` import the batch modules and apply Theorem 7.4. Figure 5 shows the architecture.

7.6 The two objectives

For the ground objective the claim is $U_0 = -193201942568284180326$ (in units of $2^{-70}A$). The theorem `ground_exact_min` states that U_0 is a lower bound on $\hat{H}(q)$ for all configurations $q < 2^{40}$ and that the configuration with identifier 711243241830 attains it (the attainment is a closed computation checked by the kernel). For the non-grammar objective the grammar predicate is the frozen convention, two-in/two-out at the 16 listed vertices, and the claim states that U_n lower-bounds $\hat{H}(q)$ over all $q \notin S_G$ and is attained by a configuration that violates at least one vertex. The leaf token 2 handles the leaves in the grammar.

7.7 What is and is not in the trusted base

The theorems are statements about an exactly specified object, and their meaning depends on the fidelity of that specification. We list what Lean checks, what it takes as given, and what is checked outside Lean.

Item	Status
Soundness of bound, pruning, checker, frontier assembly	Proved in Lean. Axioms limited to <code>propext</code> , <code>Classical.choice</code> , <code>Quot.sound</code> ; no <code>sorry</code> ; no <code>native_decide</code> .
Acceptance of the 1,148 batches	Kernel-checked in the recorded execution.
Certificate generator, diagnostic, batch writer, runner	Untrusted; only kernel acceptance matters.
Integer couplings $\hat{K}$	Transcribed from the frozen instance file by script; byte-checked against it. Not derived from the physical formula in Lean.
Grammar convention (16 vertex masks)	Frozen-file convention. Agreement with the geometric ice rule was checked by differential testing, not proved.
$A(792K)$ and the decimal range endpoints	Frozen binary64 value, taken exactly; range endpoints read as exact rationals. The physical derivation of A is not formalised.
Solver's binary64 pruning and guard	Real-number structure formalised; rounding not formalised. The SCR-III numerical audit is not replaced.

Two further limits are intrinsic. The certificates fix minimum *values*; nothing proved here says that the ground state is unique, nor that the excluded-state minimum is attained only by the reported corner defect, because pruning is non-strict. What is verified for the excluded minimum is a *witness*: the configuration of identifier 711243241766 is proved to be valid, outside the grammar, and of exactly the certified energy. And the certification concerns the *encoded model*: it establishes statements about the integer matrix $\hat{K}$ and the frozen grammar, not about any physical array.

7.8 Controls and the recorded execution

A certification pipeline is only credible if it rejects wrong claims. On the real instance, every one of the following controls was rejected by the kernel while its correct twin compiled: the ground claim with $U + 1$; a certificate with a tampered token; a truncated certificate; a certificate with trailing junk; a certificate with wrong coupling rows; a certificate attached to the wrong prefix; and, for the non-grammar objective, a claim of $U + 1$, a claim of the originally reported (frozen) value, and a certificate with the leaf rule removed. A swap of certificates between sibling prefixes is accepted, correctly: the checker recomputes every bound from the prefix, so a valid certificate for the subtree under one prefix can be valid for its sibling; this is not a control and was replaced by a wrong-prefix control.

The recorded execution used Lean 4.34.1 in a cloud container with two cores and 8 GB of memory. The certificate generator produced 503,096 proof nodes for the ground objective (at most 985 for any prefix) and 1,135,204 for the non-grammar objective (at most 2,519). Kernel checking took 1,945 s wall and 3,808 s CPU for the ground objective, with a peak of 3.3 GB per process, and 3,451 s wall and 6,760 s CPU for the non-grammar objective, with a peak of 2.5 GB, running two processes concurrently. The total was about 2.9 CPU-hours. Earlier configurations failed for resource reasons and are recorded: frontier depth 12 with 4,700-node batches, and unbounded build parallelism, were killed for lack of memory; the delivered configuration (depth 14, at most 1,500 nodes per batch, two concurrent processes) completed without a failed batch. An independent re-run in the author's own environment, requiring about 1.5 hours, was deliberately waived. We record this as an archival fact, not as a closure condition: anyone can run the packaged script at any time, and a successful run would add a re-verification note without altering this record.

Table 14: Certified quantities for the exact encoded 4×4 model (Lean 4.34.1, recorded execution). “Numerical” is the frozen SCR-III solver value, converted for comparison.

Theorem	Statement	Exact value (units $2^{-70}A$)
ground_exact_min	Global minimum of $\hat{H}$ over all 2^{40} configurations; attained by id 711243241830	$E_0 = -193201942568284180326$; confirms the frozen ground value
nongT_exact_min	Minimum over grammar-violating configurations; attained by id 711243241766	-181106494839879485536
frozen_nonG_not_min	The frozen value -181106494839879484608 (id 161487427942) is not the minimum	lies 928 above the true minimum
gap_pos, gap_exact	$\Delta_{-G} > 0$ and its exact value	$\Delta = 12095447728404694790$
pred_lt1_true	$\delta_4 < 1$	$1 - \delta_4 > 0.054$
pred_r2_true	$0.790439676056526 \leq \delta_4 \leq 1.0333534753990623$	margins > 0.0874 above, > 0.155 below

8 Certified results and the exact correction

8.1 The certified quantities

Table 14 lists the results of the certification for the exact encoded model of the 4×4 instance, with the Lean theorem that states each. All energies are in units of $2^{-70}A$ unless converted.

The combined statement is the theorem `iv5_final`: for any values U_0 and U_n that are the minima of the two objectives, $U_0 = -193201942568284180326$, $U_n = -181106494839879485536$, $U_n - U_0 = 12095447728404694790 > 0$, and both prediction predicates hold. It is obtained by applying the lemma `iv5_of_minima` to the two accepted minimum theorems; the 1,148 batches are imported, not re-run, and uniqueness of the minimum *value* (`IsMinOn_unique`) lets the exact gap be read off. The headline mathematical result is

$$\Delta_{-G} = \min_{s \notin S_G} H(s) - \min_s H(s) = \frac{12095447728404694790}{2^{70}} A > 0. \quad (4)$$

It is a global statement about the complete finite configuration space of 2^{40} configurations, established from the definitions of the encoded model and not from the output of an optimizer.

8.2 Exact value of the normalized gap and the two predictions

The normalized gap of Eq. (2) is a ratio of exact quantities. The energy scale $A(792\text{ K})$ is the frozen binary64 value, an exact dyadic rational, and $k_B = 1.380649 \times 10^{-23} \text{ J/K}$ is an exact decimal, so $\delta_4 = \text{delta4Num}/\text{delta4Den}$ is a rational number, and the two predictions become inequalities between integers after cross-multiplication. The kernel decides them. The exact value is

$$\delta_4 = 0.94591722234308262\dots,$$

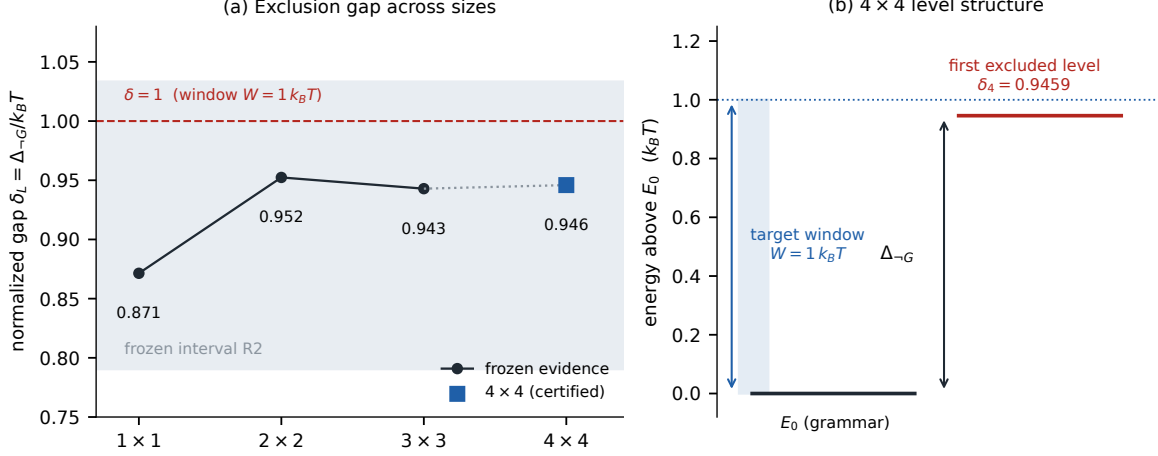


Figure 6: Exclusion-gap geometry. (a) Normalized exclusion gap $\delta_L = \Delta_{-G}/k_B T$ for the four patch sizes. The shaded band is the frozen interval I_{R2} ; the dashed line is the window edge $\delta = 1$. The 4×4 value (square) is the certified exact value. All four values lie below 1, so exact preservation of the $1 k_B T$ window fails at every size studied. (b) Level structure of the 4×4 model in units of $k_B T$ above the ground energy: the target window $W = 1 k_B T$ and the first grammar-excluded level at $0.9459 k_B T$, inside the window.

so that $1 - \delta_4 = 0.05408 \dots > 0.054$, the distance to the lower end of the interval is 0.1555, and the distance to the upper end is 0.0874. The value reported by the numerical solver, 0.9459172223430824, agrees with the exact value to better than 3×10^{-16} . The nearest binary64 number to the exact value ends in $\dots 826$; the reported $\dots 824$ reflects floating-point evaluation in the original calculation and is independent of the correction below.

Under the frozen decision rule of Section 4, a rigorous interval for δ_4 entirely inside I_{R2} yields `GAP_CONSISTENT`, and a rigorous bound below 1 confirms the directional prediction. The certification supplies exactly such bounds, in fact the exact value. The recorded closure verdict is therefore `SCR_IV_COMPLETE/FINITE_GAP_CERTIFIED/PREDICTIONS_PROVEN` for the exact encoded model of the frozen 4×4 instance, in the recorded Lean execution. The earlier verdict `GAP_INCONCLUSIVE` is retained in the record as the correct verdict at the time it was issued.

8.3 Both global minima

The ground minimum confirms the frozen numerical value exactly:

$$E_0 = -\frac{193201942568284180326}{2^{70}}A,$$

attained by the configuration 711243241830, whose vertices are all two-in/two-out. The ground witness therefore lies in the grammar, and no configuration anywhere in the 2^{40} -element space has lower energy. This is the sense in which the generative grammar and the physical ground state agree on this instance.

The excluded-state minimum does not confirm the frozen value. This is the correction.

8.4 The exact correction

The certificate generator for the non-grammar objective was first run with the frozen value as its claim, $-181106494839879484608/2^{70}A$. Generation failed an internal assertion: the diagnostic found a grammar-violating configuration of lower energy. This was a discrepancy in the data, not a defect in the tool. The diagnostic located the true minimum, the configuration 711243241766, and the proof was then run with the corrected claim. The facts, each kernel-checked by a family of `nonGTrue_*` theorems and by `frozen_minus_true`, are as follows.

- The exact minimum over grammar-violating configurations is $-181106494839879485536/2^{70}A$, attained at 711243241766: the ground witness with island 6 reversed. It violates the grammar at vertex 3 alone, which has three inward arms (vertex charge $+2$), while every other vertex is two-in/two-out.
- The frozen value $-181106494839879484608/2^{70}A$ exceeds this by exactly $928/2^{70}A \approx 7.9 \times 10^{-19}A$, a relative amount of about 7.7×10^{-17} of the gap. The frozen witness, 711243241830 with island 39 reversed, is the id 161487427942. It also violates the grammar at vertex 3 alone, but with a single inward arm (vertex charge -2). It is not a minimizer.
- The two excluded witnesses are each a single-spin flip of the ground witness (the flipped islands are the two outer-edge islands that touch only the top-right vertex), and they differ from each other in two bits.
- The exact gap numerator changes from 12095447728404695718 to 12095447728404694790. The correction moves δ_4 by 7.3×10^{-17} , and both prediction verdicts hold for the corrected *and* the frozen value.

Authoritative value and historical record. The corrected value is the authoritative exact result of this paper. The original value remains a documented historical result: it is correct at the resolution of the original solver, whose outward pruning guard is 10^{-11} in units of A , about seven orders of magnitude larger than $928/2^{70}A$ (7.9×10^{-19} against 10^{-11}). The two states are therefore not distinguishable at the solver’s declared resolution. We state this as consistent with the recorded guard; the causal explanation has not been established from the solver’s execution trace, and we do not assert it. No frozen file was modified: the correction is an additive record.

8.5 Numerical and exact values side by side

Table 15 compares the frozen numerical values with the exact certified ones, in units of A (with the exact rational converted to a decimal) and, for the ground energy, in joules.

The numerical and exact ground energies differ by the double-precision rounding of the solver’s accumulation (the recomputation of the witness energy in 60-digit arithmetic differs from the solver’s value by 3.8×10^{-17}), and the two gaps agree to better than the stated bound. The excluded-state entry differs by about 4×10^{-17} , dominated by double-precision rounding of the solver’s accumulated energy; of that, $928/2^{70}A \approx 7.9 \times 10^{-19}$ is the correction, which is therefore invisible in this comparison at the digits shown and is established only by the exact comparison of the two states’ integer energies. The quantity expressed in joules is the charge-position model value of the ground witness from the independent witness check; it is shown to fix the scale ($A \approx 10^{-18}$ J, so that $E_0 \approx -0.164A$).

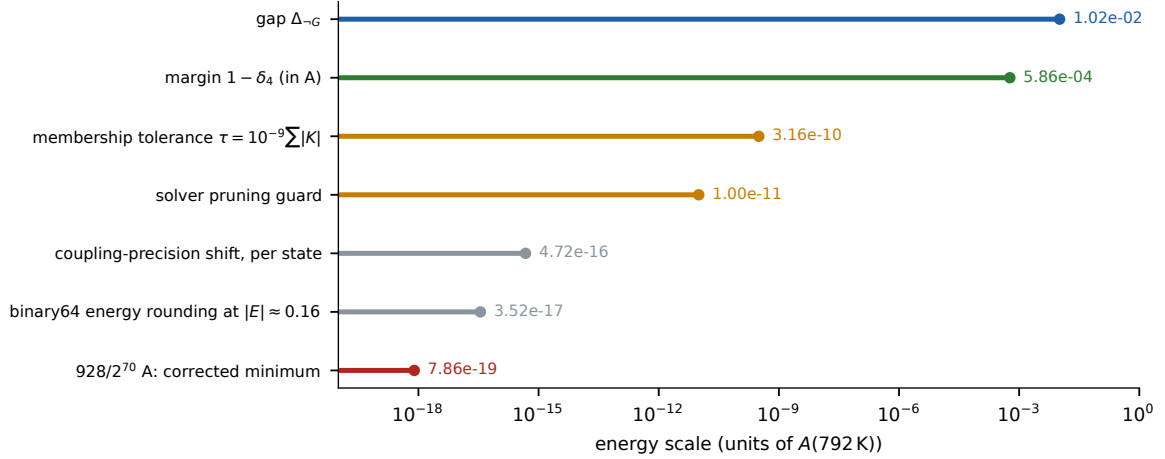


Figure 7: Numerical tolerances and exact certification. Energy scales on a logarithmic axis, in units of A . The correction $928/2^{70} A$ lies far below every numerical tolerance in the pipeline (the solver pruning guard, the membership tolerance τ) and even below the stored-coupling precision and binary64 rounding at the scale of the energies. It is invisible at conventional tolerances, yet exactly real in the encoded model. The decisions of the study (the gap and its margin to the window edge) sit more than fourteen orders of magnitude above the correction.

Table 15: Frozen numerical values and exact certified values for the 4×4 encoded model, in units of A unless stated. The decimal expansions of the exact values are rounded to the digits shown.

Quantity	Frozen numerical		Exact certified	Difference
Ground energy E_0	-0.16364841082886986	-0.16364841082886982...		$\approx 3 \times 10^{-17}$
Excluded-state mini- mum	-0.1534031680911189	-0.15340316809111886...		$\approx 4 \times 10^{-17}$
Gap Δ_{-G}	0.01024524273775096	0.010245242737750961...		$< 2 \times 10^{-17}$
Normalized gap δ_4	0.9459172223430824	0.94591722234308262...		$< 3 \times 10^{-16}$
Ground energy E_0 (J)	—	$-1.6521563 \times 10^{-19}$		—

8.6 Why the correction matters, and why it does not change the science

Figure 7 puts the correction on the scale of the tolerances. At conventional numerical tolerances the original answer is a legitimate result; the discrepancy is an example of what a floating-point optimizer with an outward guard cannot resolve and an exact check can. It is of no consequence for any verdict: the gap changes by a relative 7.7×10^{-17} , δ_4 by 7.3×10^{-17} , and the margin of 0.054 to the window edge is more than 10^{14} times larger. It is nonetheless a correction to an exact value and to the identity of a minimizer, and it came from the certification, not from a second numerical run. We regard this as the methodological content of the investigation: a scientific optimization claim can carry a formally checked certificate, and exact verification can reveal discrepancies that are invisible at the tolerances in use. The record is better for it, because it now says exactly which state is the minimizer and by how much the earlier number was off.

8.7 Verification of controls on the final theorem

The final theorem was also tested against false claims. Five false variants were rejected by the kernel while their true twins compiled: the frozen gap numerator attached to the corrected pair of minima, a

claim that δ_4 lies below its true value, a false lower bound for the interval, the claim $\delta_4 < 0.9$, and the wrong reference temperature. The exact-arithmetic lemmas of IV-5 use no axioms; the core theorem uses `propext` and `Quot.sound`, and the final theorem uses only the three permitted axioms.

9 Scientific interpretation

This section separates what the theorems and numerical audits establish from what we read into them. Statements in the first category carry the status of their evidence layer (Section 1); statements in the second are interpretation.

9.1 What the finite-model results establish

Taken together, the four findings establish the following, each for the explicitly defined models of Section 3.

- E1.** Generative admissibility can be defined operationally for a finite lattice system and evaluated exactly: grammar membership, constructive generation and contraction are all well-defined and checkable without reference to energy (Definitions 2.1 and 2.2; 3×3 study).
- E2.** A generative boundary can be compared with a physical energy landscape without circularity. By keeping grammar membership and the energy-window target distinct, and by taking the ground energy over the whole space, the comparison reduces to the single number Δ_{-G} (Proposition 2.4).
- E3.** An exclusion gap can be certified for a complete finite model of 40 binary degrees of freedom (Eq. (4)), together with a verified witness of the excluded minimum.
- E4.** A structural hypothesis can produce both a falsified sufficiency claim (3×3 : 16 of 26 target states excluded) and a successful quantitative prediction (4×4 : δ_4 inside the frozen interval and below 1), within one preregistered framework, and the two are consistent with each other.
- E5.** Exact certification can improve a numerical record that is correct at its declared resolution (the correction of $928/2^{70}$ A).
- E6.** A frozen generative description can support a quantitative prediction, registered before the target instance is evaluated, of the boundary of its own physical sufficiency, and the predicted boundary can be verified independently of the numerical search (structural predictability and predictive verification, Table 1), for this finite model.

9.2 Register of claims

Table 16 lists the six principal claims of the study with their evidence and layer. The first two are SCR-III numerical findings and the last four are SCR-IV certifications.

These are claims about explicitly defined mathematical and physical models. None is a claim about the universe, about real arrays, or about a general principle.

	G	P	T	C	gap prediction	exact certificate
SCR-III-B primary 3 × 3 (24 spins)	PASS	FAIL	not tested	not eligible	(gap 0.943)	n/a
SCR-III GAP-4 4 × 4 (40 spins)	count only	implied by $\delta_4 < 1$ (numerical)	not tested	not run	pass (numerical)	none: solver only
SCR-IV Lean 4 × 4 encoded model	count only (not in Lean)	fails: $\delta_4 < 1$ proved	not tested	not run	PASS proved	PASS 1,148 batches

Figure 8: G/P/T/C outcomes. Rows are the three evidence stages; columns are the four axes, the gap prediction and the exact certificate. Green marks a pass, red a failure, amber a numerical result without certificate, grey an axis not tested, not eligible, not evaluated or not applicable. For the 4 × 4 patch the axis-P failure is implied by $\delta_4 < 1$ (numerically at the SCR-III stage, proved at the SCR-IV stage); for 4 × 4 axis G is shown amber because only the additive combinatorial grammar count (193,662) exists; no constructive generator was run and the count is not in Lean.

9.3 Does generative admissibility determine the relevant portion of the space?

The question posed in the Introduction has a precise answer in this model, and it has two parts that Proposition 2.4 makes exact.

Below the gap, yes. For every window $W < \Delta_{-G}$ the physical target lies entirely inside the grammar. In thermal units, the ice-rule grammar contains the whole low-energy sector of the model up to $0.9459 k_B T_{\text{ref}}$ above the ground energy at 4 × 4, and up to $0.9429 k_B T_{\text{ref}}$ at 3 × 3. Within that window the generative description is not merely a plausible approximation to the physical sector; it contains it, and it does so with a contraction of nearly four orders of magnitude at 3 × 3 (the 4 × 4 grammar count of 193,662 was established subsequently through the separately archived combinatorial analysis, not as part of the original frozen experiment).

At and beyond the gap, no. At $W = k_B T_{\text{ref}}$ the target contains states that the grammar forbids, because the first forbidden level sits just inside it. The grammar is therefore not a sufficient description of the thermally accessible sector at the reference temperature, and the shortfall is not marginal in thermodynamic terms: the Boltzmann weight of a single state at the first excluded level, relative to a single ground state, is $e^{-\delta_4} \approx 0.39$ at the reference temperature, and the first excluded level is degenerate at the sizes where this was recorded (8, 16 and 16 states at 1 × 1, 2 × 2 and 3 × 3). The statement concerns weights of individual states and is not a computation of any thermodynamic observable.

The two parts together are what the exclusion gap means. The gap is the width, in energy, of the region above the ground state in which the structural description and the physics agree.

Table 16: Register of principal claims, evidence and status.

#	Claim	Evidence	Status
C1	A local generative grammar produces a very large contraction of the nominal space	3×3 : 2,604 of 16,777,216 retained	Established (numerical)
C2	Structural contraction alone does not guarantee physical sufficiency	3×3 preregistered window: 16 of 26 target states excluded	Established negative finding (numerical)
C3	The 4×4 ground energy is established exactly for the encoded Hamiltonian	344 batches accepted; <code>ground_exact_min</code>	Certified
C4	The lowest excluded energy is established exactly, with a correction of $928/2^{70}A$	804 batches accepted; <code>nongT_exact_min</code> , <code>frozen_nonG_not_min</code>	Certified
C5	A strictly positive finite gap separates the ground energy from the first grammar-excluded level	<code>gap_pos</code> , <code>gap_exact</code>	Certified
C6	The frozen 4×4 gap predictions are fulfilled	<code>pred_lt1_true</code> , <code>pred_r2_true</code>	Certified

Table 17: Sufficiency map of the ice-rule grammar: whether $S_T(W) \subseteq S_G$, from Proposition 2.4 and the gaps of Table 12. Entries for the primary windows (0.5, 1 and $2k_B T$) of the 3×3 study agree with the measured outcomes of Table 9.

Window W ($k_B T_{\text{ref}}$)	1×1 ($\delta = 0.871$)	2×2 (0.952)	3×3 (0.943)	4×4 (0.946)
0.5	preserved	preserved	preserved	preserved
0.8	preserved	preserved	preserved	preserved
0.9	fails	preserved	preserved	preserved
0.95	fails	preserved	fails	fails
1 (primary)	fails	fails	fails	fails
2	fails	fails	fails	fails

9.4 A sufficiency map

Proposition 2.4 converts the four gaps into a complete map of where the ice-rule grammar is sufficient. For a window W (in units of $k_B T_{\text{ref}}$) and a patch with normalized gap δ_L , the grammar preserves the target if and only if $W < \delta_L$ (up to the operational tolerance, which is of order 10^{-9} in A and does not affect the entries below). Table 17 lists the outcome for a range of windows.

Two observations follow. First, the patches disagree only inside narrow bands. At $W = 0.9$ the 1×1 patch fails while the other three pass; at $W = 0.95$, which lies above δ_3 and δ_4 and below $\delta_2 = 0.952$, the 2×2 patch passes while the 3×3 and 4×4 patches fail. Outside the interval from 0.871 to 0.952 all sizes agree. Second, what separates success from failure is not the window alone but its position relative to a number that varies by only about 0.08 over the four sizes. The grammar is sufficient at all four sizes for every window below $0.871 k_B T_{\text{ref}}$, and at the three larger sizes for every window below $0.942 k_B T_{\text{ref}}$; the reference window of $1 k_B T_{\text{ref}}$ is above all four gaps.

9.5 The two-sided reading of the prediction

The confirmed prediction $\delta_4 < 1$ is read in two ways, and both belong in the record. It confirms the predicted finite exclusion-gap behaviour: the gap is positive, of the size the earlier sizes indicated, and now proved. It also shows that, for this model, the ice-rule grammar cannot preserve every state inside the registered $1 k_B T$ window: exact physical target preservation fails. A reader who sees only the first reading will overestimate the grammar; a reader who sees only the second will underestimate what the prediction accomplished. Neither reading is the more fundamental one; they are the same inequality read from two sides.

9.6 What the lowest excluded state is

The lowest grammar-excluded configuration of the 4×4 model is the ground configuration with a single outer-edge island reversed, so that exactly one vertex, the top-right corner, has three inward arms. In the language of spin ice, it is a single vertex defect carrying a net charge. We note three observations and one interpretation.

1. The lowest excluded configuration, original and corrected, is a single-island reversal; its grammar violation is confined to one vertex. What the Lean certificates establish is the minimum *value*; the single-island reversal is a verified witness attaining that value (valid, outside the grammar, exactly the certified energy, one vertex with three inward arms); it is one of eight nearly degenerate single-vertex reversals (Section 9.7). Uniqueness of the minimizer is not established, and whether this defect is the only minimizer is not established.
2. The reversed island is an *outer-edge* island touching a single constrained vertex. Reversing a bulk island would change the charge of two vertices; reversing an edge island changes the charge of only one. This is a consequence of the geometry.
3. The normalized gap differs by only 0.003 between 3×3 and 4×4 (0.9429 and 0.9459), and the four values span 0.081 (0.871 to 0.952).

Interpretation (hypothesis, not a result). The size-insensitivity of δ_L is what one would expect if the first excluded level is a *local* defect whose creation energy is set by the interactions in the neighbourhood of one vertex and is weakly affected by the size of the patch. This reading is consistent with four data points and with the single-vertex nature of the lowest state, but four sizes do not establish a scaling law, no larger patch has been computed, and the free boundary may favour an edge defect whose energy would differ from a bulk defect's in an infinite lattice. We therefore treat the persistence of the gap scale as an empirical regularity of this model on these sizes. The single-reversal decomposition of Section 9.7 adds exact, but limited, support for the local reading.

9.7 Single-island reversals of the ground state

Evidence status. Supplementary analysis, recorded additively as SCR_MS_SUPPLEMENT_v001: exact rational arithmetic on the frozen couplings and the Lean-certified ground witness. It is not a Lean theorem, not an optimization, and it does not alter the certified gap.

Reversing island i of the ground witness changes the energy by $\Delta E_i = -2s_i \sum_{j \neq i} K_{ij} s_j$ (units of A), a sum of one pair contribution per partner j . Every one of the 40 reversals leaves the grammar, because

Table 18: Energy above the ground witness for the 40 single-island reversals (792 K). The ratio is to the certified gap $\delta_4 = 0.9459 k_B T$.

Class	Islands	Energy range ($k_B T$)	Ratio to gap
One vertex, corner vertex	8	0.94592 (spread 7×10^{-16} of the gap)	1.000
One vertex, other	8	1.04889	1.109
Two vertices	24	1.78944 to 1.90986	1.892 to 2.019

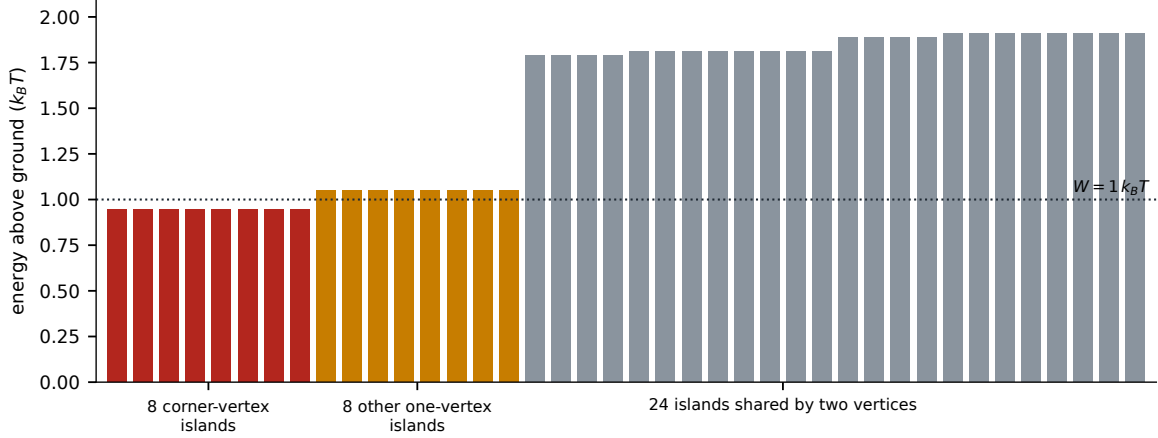


Figure 9: Single-island reversals of the certified ground witness. Energy above the ground state of each of the 40 single reversals, sorted, coloured by class (red: islands at corner vertices; amber: other one-vertex islands; grey: islands shared by two vertices). Computed in exact arithmetic from the frozen couplings (additive record SCR_MS_SUPPLEMENT_v001); no optimization.

an island touching m constrained vertices changes the inward count at those m vertices; each ΔE_i is therefore at least Δ_{-G} . Sixteen islands (the outward arms at the patch edge) touch one constrained vertex, and 24 touch two. Table 18 and Figure 9 give the 40 values.

Three observations follow from the exact values.

1. The eight lowest reversals are exactly the eight islands that touch a single constrained vertex that is a corner vertex, and their energies agree to within $8562/2^{70}A$, or 7×10^{-16} of the gap. The certified minimizer (island 6) is one of them.
2. The numerically reported minimizer (island 39) is another, lying $928/2^{70}A$ above island 6, which is exactly the certified correction of Section 8. The correction therefore reflects the ordering within a group of nearly degenerate states and not a different kind of defect. The near-equality at the 10^{-16} level suggests that these states are related by a symmetry of the idealized geometry and split only by the binary64 representation of the couplings; this reading was not verified here.
3. Every reversal of an island shared by two vertices costs at least $1.789 k_B T$, which is 1.89 times the gap.

Table 19 splits two of these costs by interaction distance. For island 6, the nearest shell (the two diagonal neighbours at $a/\sqrt{2}$) contributes $1.175 k_B T$, which is 124% of the gap; the shell at distance a partly cancels it; all pairs beyond $2a$ contribute a net $+0.013 k_B T$ (1.4% of the gap). The lowest two-vertex island has four nearest diagonal neighbours instead of two, and its nearest-shell contribution is exactly twice as large.

Table 19: Pair decomposition of the reversal cost (units of $k_B T$; a is the lattice constant). Island 6 is the certified minimizer; island 2 is the lowest of the 24 islands shared by two vertices.

Distance	Island 6		Island 2	
	partners	contribution	partners	contribution
$a/\sqrt{2}$	2	+1.1746	4	+2.3492
a	2	−0.1911	3	−0.4942
$\sqrt{2}a$	1	+0.0192	2	+0.0383
$\sqrt{5/2}a$	3	−0.0873	6	−0.1745
$2a$	2	+0.0176	3	+0.0505
beyond $2a$	29	+0.0129	21	+0.0202
Total	39	0.9459	39	1.7894

What this shows and does not show. It shows that, for the certified ground state, the cheapest reversals are those that create a single charged vertex at the patch corner, that the cost is dominated by the nearest-neighbour interaction at the defect, and that creating two charged vertices at once costs roughly twice as much. It is consistent with the local-defect reading of the near-constant gap, but it does not establish it. Single reversals of one ground state are not all excluded states, and the certificate gives the minimum over all of them, not the minimum over excluded states whose violations avoid the boundary; that restricted minimum would need its own certificate and is left open. Whether a bulk defect has an energy comparable to the boundary one in an infinite lattice is likewise not addressed.

9.8 Sensitivity of the gap to the Hamiltonian

Evidence status. The proposition below is an elementary statement proved in the text; it is not formalized in Lean. Its inputs are evaluated from the frozen geometry and couplings in the additive record `SCR_MS_SUPPLEMENT_v001`. No perturbed model has been solved, so no statement is made about the actual gap of any perturbed model beyond what the bound implies.

Proposition 9.1 (Perturbation bound). *Let H and H' be energy functions on the same finite configuration space, with the same grammar $S_G \neq S_0$, and let Δ_{-G} and Δ'_{-G} be their exclusion gaps. If $\varepsilon = \max_s |H'(s) - H(s)|$, then $|\Delta'_{-G} - \Delta_{-G}| \leq 2\varepsilon$. More generally, if $H' = \lambda H + R$ with $\lambda > 0$ and $\rho = \max_s |R(s)|$, then $|\Delta'_{-G} - \lambda \Delta_{-G}| \leq 2\rho$.*

Proof. Write $\Delta_{-G} = m_{-G} - m_0$ with $m_{-G} = \min_{s \notin S_G} H$ and $m_0 = \min_s H$. A pointwise bound $|H' - H| \leq \varepsilon$ moves each minimum by at most ε , so $|m'_{-G} - m_{-G}| \leq \varepsilon$ and $|m'_0 - m_0| \leq \varepsilon$, and the difference moves by at most 2ε . For $H' = \lambda H + R$, the minima of λH are λm_{-G} and λm_0 and each moves by at most ρ under R . $\square$

Corollary 9.2. *If $2\varepsilon < \Delta_{-G}$ then $\Delta'_{-G} > 0$. By Proposition 2.4, $S'_T(W) \subseteq S_G$ for every $W < \Delta_{-G} - 2\varepsilon$, and the inclusion fails for every $W \geq \Delta_{-G} + 2\varepsilon$.*

For pair couplings, $\varepsilon \leq \sum_{i < j} |A'K'_{ij} - AK_{ij}|$. The grammar must be unchanged; a different vertex rule is outside the statement.

Evaluation for the frozen model. The total coupling weight is $\sum_{i < j} |K_{ij}| = 0.3159A$, and the gap is $\Delta_{-G} = 0.01025A$, which is 3.2% of it. A perturbation of the couplings that is bounded by

Table 20: Worst-case bound for three perturbations of the frozen 4×4 model at 792 K. “Ratio” is $2\varepsilon/\Delta_{-G}$ for truncation and $2\rho/(\lambda\Delta_{-G})$ for the island-dimension rows, with λ the best rescaling. A ratio of 1 or more means the bound cannot exclude a change of the size of the gap.

Perturbation	λ	Ratio	Consequence
Keep only the nearest 126 pairs (remove 13.2% of the weight)	1	8.12	None: the worst-case bound is uninformative
Island width w or thickness h scaled by $1+p$	$(1+p)^2$	≈ 0	Exact rescaling of every pair energy (here $K \propto wh$ and $A \propto lwh$), so Δ_{-G} scales by $(1+p)^2$
Island length l scaled by 1.01	1.0225	0.055	Δ'_{-G} within 5.5% of $1.0225\Delta_{-G}$
Island length l scaled by 1.05	1.1163	0.279	Δ'_{-G} within 28% of $1.1163\Delta_{-G}$

$\varepsilon \leq 2.93 \times 10^{-4}A$, which is 0.093% of the total weight, therefore leaves the failure at $1k_B T$ in place. Table 20 evaluates the bound for the perturbations a reader is most likely to ask about.

Two consequences are worth stating. First, within this model the failure at $1k_B T$ and fixed temperature persists under increases of w or h up to 2.8%, since $\delta_4(1+p)^2 < 1$ holds until $(1+p)^2 = 1/0.9459$. Second, the nearest-neighbour truncation is the case the bound cannot decide: the long-range pairs carry 13.2% of the coupling weight, which is large compared with the gap, so a truncated model must be solved to learn its gap. The sign-inverted exploratory versions of the Hamiltonian are a different model, with a different ground state, and the bound says nothing about them.

9.9 What Structural Complexity Reduction establishes for the UNNS Substrate

The investigation is a direct mathematical and computational test of structural predictability within the UNNS (Unbounded Nested Number Sequences) Substrate. The proposition under test is that the nominally possible configurations of a system and the configurations its generative structure admits need not coincide, and that the boundary between them can be predicted quantitatively. The study gives the proposition an operational form. Recursive generation becomes constructive configuration building; structural admissibility becomes membership in a frozen local grammar; the admissibility boundary becomes the distinction between generated and excluded configurations; and physical persistence is tested against the energy-defined target of a specified Hamiltonian. The four steps of the chain of Section 2 are each realized and tested in a model that can be computed exactly.

The central result. The study establishes a quantitative, predictable relation between generative admissibility and physical sufficiency. For a finite configuration space with at least one grammar-excluded state, the exclusion gap determines the complete domain of energy-window preservation,

$$S_T(W) \subseteq S_G \iff W < \Delta_{-G}.$$

Because the whole domain is fixed by one number, the boundary can be the object of a prediction. This makes a central UNNS question decidable: not only how much of the nominal space a grammar excludes, but exactly how far that exclusion can be reconciled with the physical energy structure of the system. The grammar contracts the 4×4 space to 193,662 of 2^{40} configurations, is sufficient for every window below $\Delta_{-G} = 0.9459k_B T$, and fails above it, so that its failure at the registered window of $1k_B T$ locates the boundary of its sufficiency and does not negate its significance.

The predictive achievement. The UNNS achievement is not that the ice rule excludes almost all spin configurations. It is that a generative-admissibility framework was used to formulate a prospective, quantitative prediction about the relationship between structural constraints and a physical energy landscape, and that the prediction was subsequently certified. The gap was predicted from smaller systems before the 4×4 instance was evaluated, located numerically, and then fixed exactly by machine-checked proof that did not rely on the numerical optimizer, with the qualifications on the strength of the prediction stated in Section 6 (see Section 1.4 for what “predictive” does and does not mean here): the planned feasibility campaign was not completed, an earlier exploratory verdict was inconclusive, and the frozen interval is broad.

Three levels. Table 1 separates what has been shown from what has not. Structural predictability is demonstrated for the investigated finite model. Predictive verification is achieved for the encoded model, since the certificate does not depend on trusting the optimizer. Cross-domain predictability is not established; it is the broader objective of the methodology, it requires evidence from systems unrelated to spin ice, and it is the next scientific test (Section 10.8). The distinction does not weaken the present result; it identifies where the study sits within the larger program.

What this adds to the substrate. Structural selection is not the same as physical selection, and the relation between them has a measurable and predictable boundary. A generative grammar can be physically sufficient within a precisely determined energy domain and insufficient beyond it. A boundary of the admissible set is therefore not a statement that everything outside it is physically absent: the excluded configurations here are physically accessible states, and the grammar remains an exact description of a narrower energy sector. This replaces an unrestricted expectation of structural reduction with a testable, quantitative relation, and it refines what admissibility means within the substrate.

Roles of the established methods. The ice rule belongs to established statistical mechanics, and the certification uses established ideas from exact optimization and formal verification; their validity does not depend on UNNS terminology. Their integration here serves a UNNS research objective, namely to investigate whether generation, admissibility, boundaries and physically relevant structure are related in a way that can be predicted. The substrate supplies the organizing proposition and the research architecture, and the established methods supply the means of testing it.

Scope. The result concerns one grammar, one Hamiltonian and finite patches of at most 40 spins. It does not test the broader hypotheses of the program: that the substrate is a universal description of physical law, that recursion generates physical constants, or that the predictive methodology is domain independent. These lie beyond what a finite-model study of one grammar can reach, and they are neither supported nor refuted here. No computational advantage is claimed, since the comparative benchmark was never run.

9.10 Limits of inference

- **Finite models.** All statements concern patches of at most 40 spins with a specified Hamiltonian. Extrapolation to larger arrays or to real samples is not supported by this study.

- **One grammar, one Hamiltonian, one reference temperature.** The result at $W = k_B T_{\text{ref}}$ depends on the position of the window relative to the gap; a different reference temperature would move the window, and for a window below the gap the grammar preserves the target.
- **Independence.** The grammar is constructed blind to energy and was frozen before any physical energy was evaluated (target-blind construction). The Hamiltonian is a frozen model Hamiltonian from a published simulator source, and both are computed from the same geometry. They are distinct calculations, not independent provenance; the verifier shares authorship with the grammar code. “Independently defined target” means only that the target is defined by energy rather than by grammar membership.
- **Prediction strength.** The frozen interval is broad and contains values already observed (Section 6). Success is a successful finite-model prediction; it is not an independent prediction about nature.
- **Transfer.** Whether an admissibility grammar frozen blind to a target family tracks an independently measured or computed family (axis T) is untested.

10 Discussion and conclusions

10.1 Relation to established approaches

The aim of this comparison is to place the contribution, not to rank approaches; each of the following addresses a different question, and several can be used together.

Symbolic dynamics. In symbolic dynamics a local constraint on admissible symbol arrangements defines a subshift of finite type, and the entropy of the constrained space is a central invariant [9]. The ice rule on a square lattice is a two-dimensional instance, and its entropy density is the subject of Lieb’s exact solution of square ice [3]. Asymptotically, the fraction of nominal configurations that satisfy the rule decays exponentially with the number of vertices at a rate set by that entropy, with finite-patch counts modified by the free boundary. The contraction we report is thus an unsurprising property of the constraint, and we do not offer it as a discovery. What symbolic dynamics does not ask is how the *energy* of the unconstrained system relates to the constrained set. Our exclusion gap is a quantity of that kind: it couples the combinatorial boundary of a subshift-like set to a Hamiltonian that does not enforce it.

Constraint optimization. In constrained optimization the constraints are part of the problem, and one solves for the best feasible point. Branch-and-bound [11], constraint programming and their descendants all exploit constraints to prune the search. Our setting differs in the status of the constraint: it is a *hypothesis* about where the answer lies, not a requirement of the problem, and the object of study is whether the unconstrained optimum, and the neighbourhood of it that a physical question needs, lies inside the constrained set. In optimization language, the ground energy over all configurations is the unconstrained optimum, the ground state lies in the feasible set (the constraint is inactive at the optimum), and Δ_{-G} is the penalty paid by the best *infeasible* point relative to the optimum. It is a second-best-infeasible gap. Computing it exactly is itself a constrained optimization problem over

the complement of the grammar, and our certification of that problem is an application of branch-and-bound with an exact bound. Ground-state problems for broad classes of Ising spin-glass models are computationally hard in the worst case [12]; we claim no polynomial-time structure for the all-pair dumbbell instance, and the 4×4 branch-and-bound visits 7×10^5 and 1.3×10^6 nodes in the numerical run for this instance, and no scaling is claimed.

Renormalization-based model reduction. Renormalization-group methods reduce a many-body problem by integrating out degrees of freedom and flowing to effective couplings for the retained ones [13]. A generative grammar reduces the problem differently: it restricts the *set of configurations* and keeps the Hamiltonian unchanged. The two are not competitors; a reduction that keeps all configurations within a window of the ground state and a reduction that preserves long-distance observables answer different questions. The comparison is nevertheless instructive. A renormalization scheme comes with a control: its errors are organized by the flow, and the omitted degrees of freedom are quantified in terms of irrelevant operators. The analogue for a grammar is the exclusion gap: it quantifies what the restriction omits, as the energy of the cheapest omitted configuration, and Proposition 2.4 converts it into a sufficiency statement for every window.

Artificial-spin-ice sampling and ground-state studies. Several established approaches work inside the ice manifold. Cluster and loop-type algorithms were developed for ice (vertex) models, in which the ice rule is part of the model, so that every update preserves it [14], and the directed-loop algorithm has been adapted to vertex models [15]. Restricting the state space to the six-vertex manifold in this way is the same restriction that our grammar represents. Such methods answer questions about the constrained system; by construction they do not ask what the omitted configurations cost. In artificial spin ice itself, ice-rule-violating vertices are reported as the elementary excitations of square ice [16], and in lithographic arrays isolated defects can fail to anneal [17]. In that vocabulary the exclusion gap of this paper is the energy of the cheapest defect state of a finite model, and the contribution is the test of an energy-blind restriction against an energy-defined window together with a proof of where it stops preserving that window. The present paper does not review the literature exhaustively, and it makes no priority claim: a systematic search for earlier evaluations of the omitted excitation energies of ice-rule restrictions has not been carried out.

Proof-producing solvers and formal verification. Proof-carrying code established the pattern of an untrusted producer and a small trusted checker [7]. Formal verification of large computations has precedent in the formal proof of the Kepler conjecture, which verified computer-assisted components in proof assistants [18]. Lean 4 itself is a proof assistant and programming language [10]. Our use of the pattern is an application: a certificate format for branch-and-bound over ± 1 configurations, a soundness proof for its checker, and a partitioning scheme that brings the kernel’s resource use within a modest machine. We do not claim that proof-carrying optimization originates here, nor that it is specific to UNNS, nor that it confers a computational advantage: the certified computation costs about 2.9 CPU-hours against the 150 seconds per objective of the original numerical run. Its value lies in what it adds to the record, namely a statement that does not depend on trusting the optimizer, and, in this instance, a correction.

10.2 Implications for UNNS

Section 9.9 states what the study establishes for the substrate: structural predictability in one domain, and its verification. Four consequences for the program follow from it, each checkable. First, and principally, a generative structural description can be assigned an *exact domain of physical sufficiency*. For the ice-rule grammar on the 4×4 model this domain is every energy window below $\Delta_{-G} = 0.9459 k_B T$ (Proposition 2.4); numerical evidence located and predicted the boundary, and formal proof fixed it exactly and corrected the record. Second, the vocabulary of admissibility has been operationalized end to end (Section 2), and a boundary of the admissible set has acquired a measured energetic status. Third, the program has reported an unfavourable outcome and a favourable one within the same framework, as the falsifiability requirement of the substrate vocabulary demands. Fourth, a hypothesis has been narrowed. In the chain of recursive generation, structural constraints, admissibility and physical persistence, the last arrow holds for the ice-rule grammar and this Hamiltonian below a quantified energy and fails above it. Whether some other generative rule determines a larger part of the physically relevant sector is a question for a different grammar with its own frozen test; the present result supplies the instrument for asking it, namely the exclusion gap, together with a method for certifying it.

10.3 Implications for computational physics

Two methodological points may be useful beyond the present model. The first is the discipline of separating the generative constraint from the energy-defined target and using the whole-space ground energy, which prevents a restricted solver from certifying its own omissions. The gap equivalence turns the sufficiency of any local constraint on a finite model into the computation of one number. The second is the demonstration that exact certification can sit alongside a numerical campaign without replacing it. The numerical optimizer found the right ground energy, a minimizer of the excluded objective that was correct to the resolution declared, and a verdict-bearing gap; the certificate confirmed the first, corrected the identity of the second at a level seven orders of magnitude below the solver's guard, and upgraded the status of the third from inconclusive to proved. A group using floating-point branch-and-bound with a guard should expect that its reported minimizer may be correct only to the guard, and an exact check is the means of knowing.

10.4 Methodological lessons

Five practices of the study seem to us transferable, whatever one thinks of the particular grammar.

1. **Take the reference from the whole space.** A restricted solver cannot certify its own omissions. Every comparison used the global ground energy, never a grammar-restricted one.
2. **Separate the structural object from the physical target.** Keeping grammar membership and the energy window distinct, and reducing their relation to one number, makes the question exact and removes the tautology.
3. **Freeze, then evaluate, then record what was known.** Table 8 is a template: for each protocol, say which quantities were available when it was frozen.
4. **Report the axes separately.** A favourable contraction cannot compensate for a failed sufficiency test, and a failed sufficiency test does not erase the contraction. The study could fail on one axis

and succeed on another, and it did.

5. **Let the certificate correct the record.** When exact verification disagreed with a frozen value, the response was to add a record, state exactly what changed and what did not, and keep the earlier value as history.

10.5 Limitations

The inferential limits (finite models, one grammar, Hamiltonian and reference temperature, provenance of the Hamiltonian, prediction strength, transfer) are stated in Section 9 and are not repeated. The following limits concern the record and are collected here.

1. **Prediction provenance.** The prediction was frozen before the 4×4 evaluation, but the planned feasibility campaign was not completed, the certification audit followed the run, and an earlier exploratory verdict was inconclusive. The window was chosen knowing the 1×1 and 2×2 gaps.
2. **Axes T and C.** Axis T is untested; because the gap equivalence protects an energy-window target, the present preservation test cannot establish transfer. Axis C was not eligible, so the comparative B0/G2 benchmark was never run.
3. **Certification scope.** The Lean results are exact-arithmetic statements about the encoded dyadic model. The original solver’s binary64 pruning is not formalised, and the SCR-III numerical audit is not replaced. The grammar convention was matched to the geometric ice rule by differential testing, not proof. Only minimum values are certified, not minimizer sets; the corner defect is a verified witness.
4. **Re-verification.** The Lean runs were executed in a cloud container; an independent re-run in the author’s environment was waived and is an archival fact.
5. **Unrecorded and unchecked items.** The 4×4 grammar count (193,662) was added after the frozen records by an additive combinatorial computation from the frozen vertex table; it is not machine-checked and no constructive 4×4 generator run exists. The 3×3 results (2,604 of 16,777,216 retained; 16 of 26 target states excluded) are SCR-III numerical results and were not re-checked in Lean.

10.6 Open items and future work

SCR-III is not closed by the closure of SCR-IV. It remains open on three items: the held-out discrimination study (axis T), whose protocol exists in draft form and whose default settings are awaiting confirmation (mapping of target families, the standard-error rule for a single block, handling of patches, prefix sub-instances, threading, and execution host); the synthetic feasibility campaign for 4×4 -scale certification; and the evaluation of the grammar against independent target families. None has been executed. The following directions follow from this study and are listed as proposals, not as results.

- **Cross-domain predictive transfer.** The principal next test, set out in Section 10.8: the same protocol applied to a system unrelated to spin ice. It has not been run.
- **Larger patches.** A 5×5 patch has 60 spins, and a 6×6 patch 84. Whether the gap scale persists, and whether the certificate machinery scales, are open; no such computation has been run (see the outlook below).

- **Decomposing the gap.** Splitting the first excluded level into boundary and bulk defects would test the local-defect interpretation of Section 9. This needs a certificate for a restricted objective and has not been done.
- **Other grammars.** A grammar that tolerates a bounded number of vertex violations would be a different generative rule, with a different gap. The protocol of this study forbids modifying the grammar to recover excluded states after the fact; such a grammar would need its own frozen test.
- **Strengthening the prediction.** A future prediction should be frozen with an interval tighter than the spread of the observations it is derived from, and for a size at which the answer is not already inside the observed range.
- **Re-verification.** Running the packaged certification in an independent environment would add a note to the record.

10.7 Scalability outlook

The certification of Section 7 has three cost drivers: the number of nodes in the proof tree, the kernel memory per checked node (which forced the partition into batches), and the need to check every bound independently of the untrusted solver. The 4×4 run gives one measurement of each, and the paper does not extrapolate a runtime or a size limit from a single instance. Qualitatively, the asymmetry between the two directions of the problem is the useful fact. An upper bound on the gap needs only a witness configuration, and its energy is checked exactly in a single evaluation; a lower bound needs the whole search, and it is the lower bound that the exhaustive certificate pays for. Four directions might extend the reach, none of which has been tried: stronger bounds that prune earlier, a restricted objective that targets only the states near the excluded boundary, symmetry reduction, and a more compact certificate representation. A probabilistic or approximate certificate would have to supply a verified lower bound with a stated guarantee, for example from a relaxation evaluated in exact arithmetic; sampling alone can produce witnesses but cannot prove that nothing lies lower, so a statement obtained that way would have a different evidentiary status from the theorems here and should be labelled accordingly.

10.8 The next scientific test: cross-domain predictive transfer

The present study demonstrates structural predictability in one physical domain. Whether the predictability belongs to the UNNS methodology or to the particular properties of square spin ice cannot be decided from one case. The decisive test is therefore transfer: to apply the same procedure to a system that is unrelated to spin ice and to ask whether the predictive relation reappears. We state what such a test would require, so that it can be registered before it is run. This subsection is a proposal and reports no result.

A transfer test would need five components, each the counterpart of an element of the present study.

1. **A generative grammar fixed blind to the target.** A rule that constructs admissible configurations of the new system by local extension, frozen before any quantity of the target is evaluated.
2. **An independently defined target.** A target set defined by a quantity of the new system (an energy, a cost, a score) and not by grammar membership, so that the tautology guard of Section 2 applies.

3. **A measurable exclusion quantity.** The analogue of Δ_{-G} : the cost of the cheapest configuration the grammar excludes, measured from the reference taken over the whole space.
4. **A prediction frozen from smaller instances.** An interval for the normalized exclusion gap of a larger instance, registered before that instance is evaluated, with an interval narrower than the spread of the observations it is derived from.
5. **Verification independent of the search.** A certificate for the larger instance that does not rely on trusting the search procedure.

The outcomes are informative whichever way they fall. A frozen prediction that is fulfilled in an unrelated domain would be the first evidence that the predictive relation is not specific to spin ice. A prediction that fails would locate the limit of the methodology, and would be recorded as the unfavourable result of the present study was. Neither outcome is anticipated here. Until such a test has been run, the cross-domain predictability of the UNNS methodology remains a hypothesis, and the claims of this paper concern the finite spin-ice model only.

10.9 Conclusions

This study asked whether the UNNS Substrate can generate quantitative, falsifiable predictions about structural admissibility and its relationship to physical behaviour. Square artificial spin ice served as the controlled test environment, with a frozen all-pair Hamiltonian and the ice rule as grammar; the ice rule is the instrument, and UNNS structural predictability is the subject.

In this finite model, a local grammar frozen without reference to energy contracts the space by nearly four orders of magnitude at 3×3 and is constructed without materializing what it excludes; an exact count shows that at 4×4 it admits 193,662 of 2^{40} configurations, a contraction of 99.99998% (the count is combinatorial, not a constructive run). It contains the exact global ground state at every size examined and the whole physical target for every window below a finite exclusion gap. It does not contain the whole target at the thermal window of $1 k_B T$, because the cheapest excluded state lies at $0.9459 k_B T$; the certified minimum is attained by a verified single-vertex boundary defect (a witness, not a proof of uniqueness). Equivalently, by the gap equivalence, the grammar is sufficient for exactly those windows $W < \Delta_{-G}$. The gap was predicted from smaller sizes before the largest instance was evaluated, and the exact value, $\Delta_{-G} = 12095447728404694790/2^{70} A$, was then established by a machine-checked proof that also corrected the numerical minimizer of the excluded objective by $928/2^{70} A$ without changing any verdict.

The UNNS achievement is not that the ice rule excludes almost all spin configurations. It is that a generative-admissibility framework was used to formulate a prospective, quantitative prediction about the relationship between structural constraints and a physical energy landscape, and that the prediction was subsequently certified, with the exclusion-gap equivalence supplying the mathematical mechanism. In the levels of Table 1, structural predictability is demonstrated for the investigated finite model, and predictive verification is achieved by the Lean certification of the encoded model. Both statements concern one grammar, one Hamiltonian and finite patches; the prediction interval was broad, and the planned feasibility campaign was not completed.

These are exact results about an explicitly defined finite model, a successful prospective prediction within that model, a meaningful falsification of sufficiency at a registered window, and an auditable integration of a generative-admissibility hypothesis with physical optimization. They are not a proof

that structural contraction implies physical completeness, evidence of a computational advantage, or support for a universal organizing principle.

Domain independence is the ambition of the methodology and is not established by a single spin-ice case. The next scientific test is therefore cross-domain predictive transfer: a frozen generative grammar, an independently defined target, a prediction registered before evaluation, and verification independent of the search, applied to a system unrelated to spin ice (Section 10.8). Whether the predictive relation reappears there, or fails and so marks the limit of the methodology, is the question to which the present investigation leads.

A Data provenance and reproducibility

A.1 Canonical source reference

The canonical public source for every record cited in this paper is the repository `ukbbi/UNNS` at the immutable commit

4e759d381875c7341e2f26cb72319b44d87aed15 (4e759d38)

under the project directory `projects/structural_complexity_reduction`, available at the address below [19].

https://github.com/ukbbi/UNNS/tree/4e759d381875c7341e2f26cb72319b44d87aed15/projects/structural_complexity_reduction

Paths below are relative to that directory. The repository publishes the *extracted* contents of the packages, not the original ZIP archives; the file names and contents cited here were checked against this commit.

Record	Path at commit 4e759d38
Final Lean package (v005), with the 4×4 batch modules and certificates	<code>scr_iii_b/SCR_IV_LEAN_v005/</code> (status record <code>CERT_STATUS.md</code> ; batch modules in <code>Bat/</code> ; assembly <code>Fin_g.lean</code> , <code>Fin_n.lean</code> ; final theorems <code>SCR_IV_V5.lean</code> , <code>SCR_IV_V5Final.lean</code>)
Exact-gap addendum	<code>scr_iii_b/GAP4_EXACT_ADDENDUM_v001.md</code> (and a copy in <code>SCR_IV_LEAN_v005/</code>)
Preserved earlier package (v004) and v001–v003	<code>scr_iii_b/SCR_IV_LEAN_v004/</code> , <code>SCR_IV_LEAN_v001/</code> to <code>v003/</code>
Closure records	<code>scr_iii_b/SCR_IV_CLOSURE_v001.md</code> , <code>SCR_IV_CLOSURE_FREEZE_v001.md</code>
4×4 prediction freeze, instance, first run, witness records	<code>scr_iii_b/GAP4_v010/</code> (including audit/ and output/)
Witness and certification audit; result freeze	<code>scr_iii_b/GAP4_AUDIT_v001/</code> , <code>scr_iii_b/GAP4_RESULT_FREEZE_v001.md</code>
Primary-result freeze and audit; ratification	<code>PRIMARY_*_v001.md</code> , <code>RATIFY_v001.md</code> , <code>SCR_III_B_SPEC_v001_FINAL.md</code> , <code>SCR_III_B_PREREG_v001_FINAL.md</code>
Preflight freezes	<code>FREEZE_v005.md</code> to <code>FREEZE_v009.md</code>

A.2 Frozen artifacts and what the hashes identify

The hashes below are those recorded in the closure record of SCR-IV. The two ZIP hashes identify the *frozen original deliveries*, that is, the archives as they were handed over and closed. They are not hashes of the repository directories: the repository holds the extracted files, so the ZIP digests cannot be recomputed from a checkout. For the two single-file records, which are not archives, the file in the repository has the recorded digest (checked at commit 4e759d38: the addendum, both copies, and the result freeze).

Artifact	Status	SHA-256
SCR_IV_LEAN_v005.zip (original delivery)	Frozen	b8dcd220ff7beb1e8e209627a8882010 7510aef0dd45d62fcfa30320db0740fd
GAP4_EXACT_ADDENDUM_ v001.md	Frozen	b6471c06804448be4eafecb78d424a5f 7649bbdc38093eda8ddf3d1e45c0440d
SCR_IV_LEAN_v004.zip (original delivery)	Preserved	fc2197ff5ef947736d75f23ae2829167 5cf3ffb940a1094bb8e4b2f26e4fa262
GAP4_RESULT_FREEZE_ v001.md	Preserved	9771ab659c38c098afcbab9a54351c66 69564b0377a8be5893db0f3f080b77c3

The packages SCR_IV_LEAN_v001 to v003 are preserved unchanged; their hashes are in the packaged manifests. The original numerical records (GAP4_v010, the witness audit, the primary-result freeze, the preflight freezes v005–v009, the ratified specification and preregistration) are read-only and were not modified for this paper.

A.3 Additive record: the 4×4 grammar count

The count $|S_G(4)| = 193,662$ is documented in a separate additive record, SCR_MS_GRAMMAR_COUNT_4X4_v001, consisting of a record file, a self-contained script and the output of the recorded run. The record reads only the frozen vertex table of GAP4_v010/output/real_4x4.txt (sha-256 3d47ed1256da7620292a5fe4c64e0bc226b9836768ac87df0ff4fdbf91c5f7a2, equal to the frozen instance hash) and the frozen geometry module, and touches no coupling or energy. It is *not* part of commit 4e759d38, which predates it; it was added at commit 7610fe5d under scr_iii_b/manuscript_evidence/SCR_MS_GRAMMAR_COUNT_4X4_v001/. Its files have the following SHA-256 digests.

File	SHA-256
grammar_count_4x4_v001.py	f974be076f3af44f63d3d921905c1ce2 bcd522229f337ed871b185fb7f81e0ae
grammar_count_4x4_v001_output.json	810ec552829fd1bb8900cee05eaded00 dd014f30853c3a21410c0c8c205cb718
RECORD_v001.md	f093fa3f5f8b0aae8cbba09f4657fae7 75749d8310102ca5eaf70626e0697eea

A.4 Additive record: single-reversal analysis and perturbation-bound inputs

The analyses of Sections 9.7 and 9.8 are documented in the additive record SCR_MS_SUPPLEMENT_v001: a record file, a self-contained script and the output of the recorded run. The script reads only

GAP4_v010/output/real_4x4.txt and the frozen geometry module, uses the couplings as exact dyadic rationals, minimizes nothing and enumerates nothing over 2^{40} . Like the grammar-count record, it is not part of commit 4e759d38; it was added at commit 7610fe5d under `scr_iii_b/manuscript_evidence/SCR_MS_SUPPLEMENT_v001/`. Its status is exact computation on frozen data outside the Lean certificate.

File	SHA-256
supplement_analysis_v001.py	d742f679d428efb4801c216cd58a4111b7b720d858d73a7c947b44253625aa74
supplement_analysis_v001_output.json	1d691c8f027a6cfbc5eb78ae84938b098b203ed3f2486db54a5c4d4a3aeaff7a
RECORD_v001.md	49c2d396abfb167a503328673b260287058521eb1e51ffb7ba0afc51123c831c

A.5 What the reproducibility package contains

The repository directory `scr_iii_b/SCR_IV_LEAN_v005/` holds the full set of 1,148 batch modules (344 for the ground objective, 804 for the excluded objective), the proof-tree certificates from which they are built (in `certs/`), the untrusted generation and verification scripts, the frozen instance copy and provenance copies of the energy module and solver source, and logs of the recorded run. They are not reproduced in this manuscript. Running `FULL=1 bash RUN_TEST.sh` in the package re-checks all batches in about 1.5 hours on two cores with at most 3.3 GB of memory per process. A default run regenerates the certificates, checks that they equal the shipped ones, audits the axioms of the theorems of the earlier stages and runs the negative controls, without re-checking the 1,148 batches.

A.6 Statement of the data used in each section

Section	Source records
3	Frozen energy module (<code>physical_energy_v005</code>); the v005 Hamiltonian lock
5	Primary result and its freeze; primary audit; preflights v006–v007
6	GAP-4 protocol, prediction freeze, first-run output, witness report and deviation record, witness audit
7, 8	<code>scr_iii_b/SCR_IV_LEAN_v005/</code> package; status record; addendum; closure record (paths above)

A.7 Figure provenance

Figures 1–7 and Figure 8 were generated from the frozen geometry function, the decoded state identifiers, and the certified quantities of Table 14; no figure uses a value that is not recorded in a frozen source, with two exceptions: the 4×4 grammar count of Figure 4 and the values of Figure 9 come from the additive records described above. The counts of Figure 4 are those of Table 3 and Table 9.

The figure scripts are not part of the frozen records. Figures 1–7 and Figure 8 draw only the quantities named above. The toy certificate of Section 7.4 and the small-patch spectra of Figure 3 were computed afresh from the frozen geometry and the frozen energy module for this paper, with no change to any frozen file.

A.8 Revision note (v002)

This version supersedes the first draft v001, which is kept unchanged. The changes are editorial and technical, not new results: the checker-soundness theorem is stated in its unrestricted and non-grammar forms as in the Lean sources; the commit above is the canonical source reference; the gap equivalence and the notion of an exact domain of physical sufficiency are foregrounded; the scope of the statements about timing, independence of the target and the identity of the excluded minimizer is made precise; and repeated limitation passages are consolidated. No frozen numerical value was changed.

A.9 Revision note (v003)

Adds the exact 4×4 grammar count and the corresponding contraction (Table 3, Figure 4, Figure 8, Section 5, the limitations and the disclosure register), computed from the frozen vertex table without any energy evaluation and recorded additively. No verdict, gap, prediction or certificate changed. v001 and v002 are kept unchanged.

A.10 Revision note (v004)

Adds a short physics-terms summary after the abstract; a comparison with methods used for artificial spin ice, with verified references and no priority claim; the single-island reversal analysis of the 4×4 ground state (Section 9.7, Figure 9); the perturbation bound with its evaluated inputs (Section 9.8); a scalability outlook; a Data and Code Availability section that separates the fixed research commit from the additive records; and a shorter statement of what the paper does not claim. The new analyses carry their own evidentiary status (Section 1, Appendix C). No frozen result, verdict, certificate or numerical value changed. v001, v002 and v003 are kept unchanged.

A.11 Revision note (v005)

Rewrites Section 9.9 and Section 10's implications for UNNS so that the importance of the study for the UNNS Substrate is stated through what it establishes (the exact relation between generative admissibility and physical sufficiency, and its certified finite boundary), with the roles of the established methods and the scope of the claims stated without treating the independence of those methods as distance from UNNS. The corresponding closing sentence of Section 2 is aligned, and the statement that the certificate fixes “the identity of a minimizer” is corrected to a verified witness. No result, verdict, certificate or value changed. v001 to v004 are kept unchanged.

A.12 Revision note (v006)

Corrects the SHA-256 table of Appendix A, in which the record file of the grammar-count package had been given the digest of the record file of the supplement package. All digests in both tables were recomputed from the delivered files and compared with the manuscript text by a script. The disclosure register numbering is corrected (the final entry is now 14). No result, verdict, certificate or value changed. v001 to v005 are kept unchanged.

A.13 Revision note (v007)

Reframes the manuscript so that it leads with UNNS structural predictability, uses artificial spin ice as the rigorous demonstration, and ends by identifying cross-domain predictive transfer as the next scientific test. Changes: the Abstract; a new Section 1 subsection “The physical model as a test of UNNS predictability” replacing the paragraph that disclaimed the originality of the ice rule; the four questions Q1–Q4 and the three levels of achievement (Table 1) in the principal contributions; Section 9.9; a new Section 10.8 specifying what a cross-domain transfer test would require (a proposal, no result); and the Conclusions. No result, verdict, certificate, record or numerical value changed. v001 to v006 are kept unchanged.

A.14 Revision note (v008)

Changes the title to “Structural Predictability in the UNNS Substrate: Generative Admissibility, Exclusion Gaps, and Formal Certification”, which reflects the predictive purpose stated in Section 1. The earlier title, “Structural Complexity Reduction: Generative Admissibility, Physical Exclusion Gaps, and Machine-Certified Optimization”, is retained as the title of v001 to v007. The program and record names (SCR-III, SCR-IV) are unchanged. No other text, result, verdict, certificate or value changed.

A.15 Revision note (v009)

Adds one clarification, in Section 1.4, distinguishing the exact criterion given by the exclusion-gap equivalence from the prospective prediction of an unknown gap, with a cross-reference from Section 9.9. No result, verdict, certificate, record or numerical value changed. v001 to v008 are kept unchanged.

A.16 Revision note (v010)

Records the public commit of the two additive records, 7610fe5d (full identifier in Section C), in the Abstract, the Data and Code Availability section and this appendix, and states that the original research evidence remains anchored to commit 4e759d38. The SHA-256 digests in the tables above were re-checked against the files at the new commit and all eight match; the recorded outputs of both records were also reproduced byte for byte from the frozen inputs. No text, result, verdict, certificate or numerical value changed otherwise. v001 to v009 are kept unchanged.

A.17 Revision note (v011)

Two minor corrections. The sentence in Section 9 stating that the 4×4 grammar count “was not recorded” is replaced by a statement that the count of 193,662 was established subsequently through the separately archived combinatorial analysis, not as part of the original frozen experiment. The reproducibility commands in Section C are set in a verbatim block so that their spaces are preserved. No result, verdict, certificate, record, commit reference or numerical value changed. v001 to v010 are kept unchanged.

B Formal statements

This appendix records the principal Lean theorems by name with their informal content. The names are those of the package at commit 4e759d38 (Appendix A). Their proofs are in the package; the full statements are Lean source and are not reproduced.

Name	Content
<code>lower_bound_ids</code>	$B(p) \leq \hat{H}(q)$ for every prefix p and completion q (identifier level)
<code>prune_sound_ids</code> , <code>prune_strict_ids</code>	Pruning with ties; strict pruning
<code>solver_prune_sound_ids</code>	The numerical solver’s bound, transcribed loop for loop, prunes soundly for any guard $g \geq 0$ (real-number structure)
<code>chk_sound</code>	Soundness of the generic checker for the subtree below a prefix whose threaded state satisfies the invariant (induction on remaining depth)
<code>checkAll_sound</code>	Whole-tree form: if <code>checkAll</code> accepts, every ± 1 configuration with property P has $\hat{H} \geq U$, provided <code>leafOK</code> closes only leaves that no P -configuration can reach
<code>leafGrammar_sound</code>	A leaf accepted by <code>leafGrammar</code> cannot be agreed with by a non-grammar configuration
<code>ground_min_of_cert</code> , <code>nonG_min_of_cert</code>	Instances of <code>checkAll_sound</code> for the ground objective (P trivial, no leaf rule) and the non-grammar objective ($P = \text{not in the grammar}$)
<code>subOK_sound</code>	Soundness of a single prefix certificate
<code>batches_cover</code>	Every configuration extends exactly one frontier prefix
<code>ground_min_of_</code> <code>frontier</code> , <code>nonGT_</code> <code>min_of_frontier</code>	Assembly of accepted batches into a global minimum claim
<code>ground_exact_min</code>	Exact global minimum E_0 of the encoded Hamiltonian
<code>nongT_exact_min</code>	Exact minimum over grammar-violating configurations
<code>frozen_nonG_not_min</code>	The frozen non-grammar value is not the minimum
<code>frozen_minus_true</code>	Frozen minus true minimum equals 928
<code>gap_pos</code> , <code>gap_exact</code>	The gap is positive and equals $12095447728404694790/2^{70} A$
<code>pred_lt1_true</code> , <code>pred_</code> <code>r2_true</code>	$\delta_4 < 1$; $\delta_4 \in [0.790439676056526, 1.0333534753990623]$
<code>iv5_final</code>	Conjunction of the above for any pair of values that are the two minima

Axioms. The theorems depend on at most `propext`, `Classical.choice` and `Quot.sound`. The exact-arithmetic lemmas of the gap and prediction stage need no axioms. There is no `sorry` and no `native_decide`. The recorded axiom audit is part of the package. Stage counts, in the package’s own accounting: 25 theorems for the model and witnesses, 44 for pruning, 24 for the certificate design and checker.

C Disclosure register

The register lists, in one place, each departure from a plan, each correction, and each limitation that affects how a result should be read. Each item is recorded in the source records named.

#	Item	Disclosure
1	Sign convention	Exploratory versions v001–v004 used an inverted Hamiltonian sign; the v005 lock derived the sign from the simulator source. Earlier outputs are retained as historical regression evidence and are not used.
2	Window selection	The $1k_B T$ window was selected with the 1×1 and 2×2 gaps known. Blindness was to the 3×3 low-energy outcome only.
3	Post hoc tolerance	The incremental-drift tolerance of the exact 3×3 verifier was changed from 10^{-12} to 10^{-10} during preflight v007.
4	Aborted first run	The first 3×3 primary launch was killed by the host sandbox in the null phase. The driver was revised for checkpoint and resume only (hash comparison: no constant, tolerance, window, seed or rule changed). Results were recomputed.
5	Wording erratum	“All five solvers returned identical target sets” was corrected (Erratum E1): B0, B1 and E2 return the 26-state target and G1, G2 its 10-state subset. Verdicts unchanged.
6	Verifier authorship	The verifier shares authorship with the grammar code. It is cross-validated, not externally independent.
7	4×4 execution	The prediction freeze stated that the real instance was not authorized; the execution was subsequently requested. The planned feasibility campaign was not completed. Recorded in the deviation record.
8	Original solver	No optimality proof tree, no interval-arithmetic certificate; ad hoc pruning guard 10^{-11} .
9	Excluded-state minimum	The original numerical value lies $928/2^{70} A$ above the true minimum. The correction is additive and replaces the minimizer; no verdict changes.
10	Re-verification	Independent re-run of the Lean batches in the author’s environment was waived. Archival fact, not a closure condition.
11	Encoded model	Exact-arithmetic statements about the encoded dyadic model. Binary64 pruning, the physical derivation of $A(792\text{ K})$, and the geometric ice rule versus the frozen grammar convention are not formalised.
12	Additive 4×4 count	The 4×4 grammar-only count (193,662) was computed after the frozen records, from the frozen vertex table, by an additive record outside commit 4e759d38 (added at commit 7610fe5d). Not machine-checked; no constructive 4×4 generator run.
13	Supplementary analyses	The single-reversal decomposition (Section 9.7) and the inputs of the perturbation bound (Section 9.8) are exact computations on frozen data in the additive record SCR_MS_SUPPLEMENT_v001, outside the Lean certificate and not preregistered. The proposition is proved in the text, not in Lean. No perturbed model was solved.
14	Open	Held-out discrimination study (T); synthetic feasibility campaign; six defaults awaiting confirmation.

Data and code availability

Two archival layers are distinguished, because they have different status.

Layer 1: fixed research commit. The frozen SCR-III and SCR-IV evidence is public at the immutable commit 4e759d381875c7341e2f26cb72319b44d87aed15 of the repository `ukbbi/UNNS`, project directory `projects/structural_complexity_reduction` [19]. Paths are relative to that directory.

Item	Path at commit 4e759d38
Geometry and energy module (frozen, source-derived)	scr_iii_b/GAP4_v010/physical_energy_v005.py
Couplings and vertex table of the 4×4 instance (780 couplings, 16 vertices)	scr_iii_b/GAP4_v010/output/real_4x4.txt, with its hash in INSTANCE_SHA256.txt; exact dyadic encoding in scr_iii_b/SCR_IV_LEAN_v005/SCR_IV_Data.lean
Prediction freeze, first run, witness and deviation records	scr_iii_b/GAP4_v010/, scr_iii_b/GAP4_AUDIT_v001/, scr_iii_b/GAP4_RESULT_FREEZE_v001.md
Lean packages, 1,148 batch modules and certificates	scr_iii_b/SCR_IV_LEAN_v001/ to v005/
Primary-result records and manifests	PRIMARY_RESULT_v001.md, PRIMARY_RESULT_FREEZE_v001.md, PRIMARY_RESULT_MANIFEST_v001.json, PRIMARY_MANIFEST_v001.json
Closure records	scr_iii_b/SCR_IV_CLOSURE_v001.md, SCR_IV_CLOSURE_FREEZE_v001.md

The SHA-256 digests of the original ZIP deliveries recorded in the closure record (Appendix A) identify those archives as delivered. The repository holds the extracted directories, so those two digests are not hashes of the repository content; single-file records were checked against their recorded digests directly.

Layer 2: additive manuscript records. Two records postdate the commit of Layer 1 and carry their own evidentiary status (Appendix A): SCR_MS_GRAMMAR_COUNT_4X4_v001 (the 4×4 grammar count) and SCR_MS_SUPPLEMENT_v001 (the single-reversal decomposition and the inputs of the perturbation bound). They are public at the immutable commit 7610fe5d98a49d9781a35beb65f98cfb0d8c1a3f of ukbbi/UNNS, under projects/structural_complexity_reduction/scr_iii_b/manuscript_evidence/. Each consists of a record file, a self-contained script, the output of the recorded run and a checksum file; their SHA-256 digests are listed in Appendix A and match the committed files. Between commit 4e759d38 and commit 7610fe5d the only changes are an expanded project README and the addition of the manuscript_evidence/ directory; no frozen SCR-III or SCR-IV file was modified. The second commit documents additional manuscript analyses and does not alter the research freeze, whose evidence remains anchored to commit 4e759d38. At the new depth the scripts take the directory scr_iii_b as an argument; from inside each package directory:

```
python3 grammar_count_4x4_v001.py ../..
python3 supplement_analysis_v001.py ../..
```

They read the frozen inputs without modifying them, and re-running them reproduces the recorded outputs byte for byte (verified by the authors for both records).

Source of this paper. The manuscript is distributed as a single $\text{\LaTeX}$ file with its figures and as a multi-file source package that includes the figure scripts. Figures 1–7, 8 and 9 are generated from the frozen quantities and records named above. The Hamiltonian is derived from the source code of the Dryad SPMCMC v1.00 simulator [6]; the derived frozen module is part of Layer 1.

References

- [1] J. D. Bernal and R. H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *The Journal of Chemical Physics*, 1(8):515, 1933. doi: 10.1063/1.1749327.
- [2] Linus Pauling. The structure and entropy of ice and of other crystals with some randomness of atomic arrangement. *Journal of the American Chemical Society*, 57(12):2680–2684, 1935. doi: 10.1021/ja01315a102.
- [3] Elliott H. Lieb. Residual entropy of square ice. *Physical Review*, 162(1):162–172, 1967. doi: 10.1103/PhysRev.162.162.
- [4] R. F. Wang, C. Nisoli, R. S. Freitas, J. Li, W. McConville, B. J. Cooley, M. S. Lund, N. Samarth, C. Leighton, V. H. Crespi, and P. Schiffer. Artificial spin ice in a geometrically frustrated lattice of nanoscale ferromagnetic islands. *Nature*, 439(7074):303–306, 2006. doi: 10.1038/nature04447.
- [5] Cristiano Nisoli, Roderich Moessner, and Peter Schiffer. Colloquium: Artificial spin ice: Designing and imaging magnetic frustration. *Reviews of Modern Physics*, 85(4):1473–1490, 2013. doi: 10.1103/RevModPhys.85.1473.
- [6] Anthony Hurben, Ioan-Augustin Chioar, Cristiano Nisoli, and Peter Schiffer. A tale of three temperatures: superparamagnetism in an interacting nanomagnet array, 2026. *Physical Review Letters* (2026); data and code (Dryad repository) accompanying the publication; DOI 10.1103/f92d-w6fn.
- [7] George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, Paris, France, 1997.
- [8] UNNS Research Division. Introduction to the UNNS substrate: Nested generation, route, boundary, and admissibility. Manuscript, UNNS Substrate Research Program, <https://unns.tech>, 2026.
- [9] Douglas Lind and Brian Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995.
- [10] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pages 625–635. Springer, Cham, 2021. doi: 10.1007/978-3-030-79876-5_37.
- [11] A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960.
- [12] Francisco Barahona. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241–3253, 1982. doi: 10.1088/0305-4470/15/10/028.
- [13] Kenneth G. Wilson. The renormalization group: Critical phenomena and the Kondo problem. *Reviews of Modern Physics*, 47:773, 1975. doi: 10.1103/RevModPhys.47.773.
- [14] G. T. Barkema and M. E. J. Newman. Monte carlo simulation of ice models. *Physical Review E*, 57:1155, 1998. doi: 10.1103/PhysRevE.57.1155.

- [15] Olav F. Sylljuåsen and M. B. Zvonarev. Directed-loop monte carlo simulations of vertex models, 2004. arXiv:cond-mat/0401491.
- [16] Jason P. Morgan, Aaron Stein, Sean Langridge, and Christopher H. Marrows. Thermal ground-state ordering and elementary excitations in artificial magnetic square ice. *Nature Physics*, 7: 75–79, 2011. doi: 10.1038/nphys1853.
- [17] G. Möller and R. Moessner. Artificial square ice and related dipolar nanoarrays. *Physical Review Letters*, 96:237202, 2006. doi: 10.1103/PhysRevLett.96.237202.
- [18] Thomas Hales, Mark Adams, Gertrud Bauer, Tat Dat Dang, John Harrison, Le Truong Hoang, Cezary Kaliszyk, Victor Magron, Sean McLaughlin, Tat Thang Nguyen, Quang Truong Nguyen, Tobias Nipkow, Steven Obua, Joseph Pleso, Jason Rute, Alexey Solovyev, Thi Hoai An Ta, Nam Trung Tran, Thi Diep Trieu, Josef Urban, Ky Vu, and Roland Zumkeller. A formal proof of the Kepler conjecture. *Forum of Mathematics, Pi*, 5, 2017. doi: 10.1017/fmp.2017.1.
- [19] UNNS Research Collective. UNNS repository, project structural_complexity_reduction, commit 4e759d38. https://github.com/ukbbi/UNNS/tree/4e759d381875c7341e2f26cb72319b44d87aed15/projects/structural_complexity_reduction, 2026. Immutable commit 4e759d381875c7341e2f26cb72319b44d87aed15; accessed October 2026.