Formation Engine
Canonical Simulation Methodology

§0 — Formal Identity–State Decomposition Framework

A deterministic classical-mechanical framework for the prediction of molecular and crystalline structure from elemental identity and thermodynamic boundary conditions, without recourse to empirical molecular data.

Version 0.1 — First Principles Draft

---

§0 Formal Identity–State Decomposition Framework

0.1 Particle Identity Vector Definition

Define a time-invariant identity vector for particle i:

$$\mathbf{I}_i \in \mathbb{R}^6, \qquad \mathbf{I}_i = \begin{bmatrix} Z_i \\ A_i \\ Q_i \\ \Sigma_i \\ \Lambda_i \\ \Theta_i \end{bmatrix} \tag{1}$$

with components defined as follows.

0.1.1 Nuclear Identity Z_i

$$Z_i \in \mathbb{N}^+$$

Atomic number. Immutable under all non-nuclear transformations. Serves as the identity anchor for all downstream mappings.

0.1.2 Mass Identity A_i

$$A_i \in \mathbb{R}^+$$

Represents either:
- exact mass number (isotopic resolution), or
- effective mass bucket under coarse-graining.

Constraint:

$$A_i^{(\text{fast})} \approx \bar{A}(Z_i), \qquad A_i^{(\text{audit})} = A_i^{\text{exact}}$$

0.1.3 Effective Charge Participation Q_i

$$Q_i \in \mathbb{R}$$

Not an oxidation state. Instead, define Q_i as a participation coefficient modulating interaction strength:

$$Q_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \chi_{ij} \tag{2}$$

where χ_{ij} is an interaction asymmetry measure (electron donation/withdrawal bias) and $\mathcal{N}_i$ is the local neighbourhood.

This allows:

- fractional values,
- environment dependence,
- smooth behaviour under coarse-graining.

0.1.4 Structural Role Signature Σ_i

$$\Sigma_i \in \{0,\, 1,\, 2,\, 3,\, 4\}$$

Discrete structural prior encoding dominant bonding topology:

$$\Sigma_i = \begin{cases} 0 & \text{inert / closed shell} \\ 1 & \text{ionic-dominant} \\ 2 & \text{directional covalent} \\ 3 & \text{metallic / delocalized} \\ 4 & \text{mixed / transitional} \end{cases}$$

This parameter biases search and interaction kernels without explicit bond constraints.

0.1.5 Stability Class Λ_i

$$\Lambda_i \in \{0,\, 1,\, 2,\, 3\}$$

Defines statistical persistence under thermal and configurational perturbation:

$$\Lambda_i = \arg\max_k \; P_i\big(\text{survival} \mid \Delta E,\, T,\, t\big) \tag{3}$$

Interpretation:

Λ	Meaning
0	transient
1	metastable
2	ambient-stable
3	bulk-lattice candidate

0.1.6 Provenance Memory Θ_i

$$\Theta_i \in \mathbb{B}^n, \qquad n \in [8,\, 16]$$

Binary provenance hash encoding transformation lineage:

$$\Theta_i = \mathcal{H}\big(\text{origin},\ \text{generation},\ \text{relaxation depth}\big) \tag{4}$$

Used for:

- audit trails,
- anomaly detection,
- irreversible-history awareness.

0.1.3′ Canonical Particle Container Postulate (CPCP)

0.1.3′.1 Postulate Statement

Postulate 0.1 (Canonical Particle Container)**.** Any time-independent identical atomic or subatomic particle species admits an idealised canonical container representation:

$$\mathcal{C}_i = (\mathbf{I}_i,\ \mathbf{G}_i,\ \mathbf{C}_i^{\text{comp}}) \tag{5}$$

where:
- $\mathbf{I}_i$ = Identity vector (§0.1),
- $\mathbf{G}_i$ = Fundamental internal quantum descriptor set,
- $\mathbf{C}_i^{\text{comp}}$ = Computational annotation payload (non-physical but simulation-critical).

0.1.3′.2 Fundamental Quantum Descriptor Bundle $\mathbf{G}_i$

For any identical species class:

$$\mathbf{G}_i = \begin{bmatrix} s_i \\ c_i \\ f_i \\ \kappa_i \\ \eta_i \end{bmatrix} \tag{6}$$

Spin s_i.

$$s_i \in \tfrac{1}{2}\,\mathbb{Z}$$

Intrinsic angular momentum. Defines statistics class:

$$s_i \in \mathbb{Z} \ \Rightarrow\ \text{Bosonic regime}, \qquad s_i \in \tfrac{1}{2} + \mathbb{Z} \ \Rightarrow\ \text{Fermionic regime}$$

Colour charge c_i. For quark-level modelling:

$$c_i \in \{r,\ g,\ b,\ \bar{r},\ \bar{g},\ \bar{b}\}$$

or abstracted as $c_i \in \mathbb{Z}_3$. Under coarse projection: $\langle c_i \rangle \to 0$ (enforced colour neutrality constraint).

Flavour index f_i. Discrete particle family label:

$$f_i \in \mathbb{N}_{\text{flavour classes}}$$

For atoms, this collapses to nuclear composition mapping via (Z_i, A_i). For subatomic simulations, it remains explicit.

Internal projection coupling κ_i. Represents coupling between core particle and embedded field carriers:

$$\kappa_i = \begin{bmatrix} \kappa^{(g)} \\ \kappa^{(em)} \\ \kappa^{(w)} \end{bmatrix} \tag{7}$$

mapping to embedded co-particle field contributions (graviton / dimagneton / magneton-style internal carriers).

Projection information density η_i. Minimal information content proxy:

$$\eta_i = \frac{I_{\text{struct}}}{V_{\text{proj}}} \tag{8}$$

Used for:
- collapse behaviour heuristics,
- coarse-graining fidelity loss tracking,
- MIT alignment scoring.

0.1.3′.3 Canonical Container Geometry Approximation

For computational purposes, define an effective interaction envelope:

$$\Omega_i = \left(R_i,\ \mathcal{T}_i(\theta,\phi),\ \rho_i(r)\right) \tag{9}$$

Effective radius.

$$R_i = \alpha_Z\, Z_i^{1/3} + \alpha_Q\, |Q_i| \tag{10}$$

Angular texture.

$$\mathcal{T}_i(\theta,\phi) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} a_{\ell m}\, Y_{\ell m}(\theta,\phi) \tag{11}$$

Radial density envelope.

$$\rho_i(r) = \rho_0\, e^{-r/\lambda_i} \tag{12}$$

This allows identical species to share container topology while differing only in parameterisation.

0.1.4′ Computational Annotation Extension Layer

These attributes are not fundamental physics. They exist because compute is finite and because unstructured metadata degrades into noise.

Define:

$$\mathbf{C}_i^{\text{comp}} = \begin{bmatrix} \Gamma_i \\ \Xi_i \\ \Pi_i \\ \Upsilon_i \end{bmatrix} \tag{13}$$

Complexity class tag Γ_i.

$$\Gamma_i \in \{0,\ 1,\ 2,\ 3,\ 4\}$$

Used for adaptive solver selection:

Value	Meaning
0	inert, low interaction
1	pairwise dominant
2	directional bonding
3	collective electron behaviour
4	multi-scale coupled

Interaction kernel selector Ξ_i. Maps particle to preferred interaction kernel family:

$$\Xi_i \in \{\,\text{LJ},\ \text{Coulomb},\ \text{Screened},\ \text{EAM-like},\ \text{Hybrid}\,\}$$

Allows hot-swapping physics models without identity rewrite.

Solver priority weight Π_i.

$$\Pi_i \in [0,\,1]$$

Used in adaptive timestep / refinement scheduling:

$$\Delta t_i = \frac{\Delta t_{\max}}{1 + \beta\,\Pi_i} \tag{14}$$

Annotation confidence Υ_i. Tracks reliability of assigned metadata:

$$\Upsilon_i = 1 - P(\text{misclassification}) \tag{15}$$

Critical for self-audit subsystems.

0.1.4′.5 Container Invariance Condition

For identical species:

$$\mathcal{C}_i \equiv \mathcal{C}_j \quad \text{iff} \quad \mathbf{I}_i = \mathbf{I}_j \ \wedge\ \mathbf{G}_i = \mathbf{G}_j \tag{16}$$

But computational layers may differ:

$$\mathbf{C}_i^{\text{comp}} \neq \mathbf{C}_j^{\text{comp}}$$

This is deliberate and necessary for performance scaling.

0.1.4′.6 Practical Consequence

Physics layer. Identical particle = identical fundamental container.

Compute layer. Identical particle can be simulated differently depending on context.

That is exactly how you triple effective data density without multiplying particle count or neighbour graph size.

0.2 Minimal Cell Container Notation (Periodic / Molecular Cells)

0.2.1 Cell Container Postulate

A simulation "cell" (molecular box, periodic crystal cell, or derived supercell) is represented as a minimal container:

$$\mathcal{K} = (\mathbf{K}^{\text{id}},\ \mathbf{K}^{\text{geom}},\ \mathbf{K}^{\text{comp}}) \tag{17}$$

where:

- $\mathbf{K}^{\text{id}}$ is immutable lineage / provenance,
- $\mathbf{K}^{\text{geom}}$ is the geometric definition of the cell,
- $\mathbf{K}^{\text{comp}}$ is the minimal computational annotation payload (caches are explicitly *not* part of the container).

0.2.2 Immutable Cell Identity (Lineage) $\mathbf{K}^{\text{id}}$

$$\mathbf{K}^{\text{id}} = \begin{bmatrix} \Theta_{\mathcal{K}} \\ \Pi_{\mathcal{K}} \\ F_{\mathcal{K}} \end{bmatrix} \tag{18}$$

Provenance hash.

$$\Theta_{\mathcal{K}} = \mathcal{H}(\text{sources},\ \text{recipe},\ \text{version})$$

Pipeline / recipe identifier.

$$\Pi_{\mathcal{K}} \in \mathbb{B}^n, \qquad n \in [64,\ 256]$$

(hash of ordered construction steps).

Source layer flag.

$$F_{\mathcal{K}} \in \{Z,\ A,\ C\}$$

matching the `XYZFormat` enumeration.

This is the cell-level analogue of the particle provenance memory Θ_i, except it fingerprints *structure lineage*, not a single site.

0.2.3 Cell Geometry $\mathbf{K}^{\text{geom}}$

The cell geometry is defined by a lattice matrix and a boundary mode.

Lattice matrix. Let the lattice vectors be $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c} \in \mathbb{R}^3$. Then:

$$\mathbf{L} = \begin{bmatrix} \mathbf{a}^\top \\ \mathbf{b}^\top \\ \mathbf{c}^\top \end{bmatrix} \in \mathbb{R}^{3\times 3} \tag{19}$$

Cell volume: $V = \det(\mathbf{L})$.

Boundary condition mode.

$$\mathcal{B} \in \{\text{vac},\ \text{pbc},\ \text{slab}\}$$

The geometry bundle is:

$$\mathbf{K}^{\text{geom}} = \left[\, \mathbf{L},\ \ \mathcal{B},\ \ \mathbf{u} \,\right] \tag{20}$$

where $\mathbf{u}$ denotes units metadata (e.g. Å).

Replication (supercell operator). For constructed cells:

$$\mathbf{n} = (n_x,\ n_y,\ n_z) \in \mathbb{N}^3, \qquad \mathbf{L}' = \operatorname{diag}(\mathbf{n})\,\mathbf{L} \tag{21}$$

0.2.4 Minimal Computational Annotation $\mathbf{K}^{\text{comp}}$

Intentionally small and non-physical:

$$\mathbf{K}^{\text{comp}} = \begin{bmatrix} \Gamma_{\mathcal{K}} \\ \Xi_{\mathcal{K}} \\ \Upsilon_{\mathcal{K}} \end{bmatrix} \tag{22}$$

Complexity class. $\Gamma_{\mathcal{K}} \in \{0,\ 1,\ 2,\ 3\}$ (0: molecule, 1: small periodic, 2: bulk candidate, 3: coupled / heterogeneous).

Kernel family selector. $\Xi_{\mathcal{K}} \in \{\text{pairwise},\ \text{screened},\ \text{EAM-like},\ \text{hybrid}\}$

Annotation confidence. $\Upsilon_{\mathcal{K}} = 1 - P(\text{misclassification})$

Explicit exclusion: Runtime caches (bonds, neighbour lists, acceleration structures) are not part of $\mathcal{K}$; they live in a separate mutable cache set:

$$\mathcal{M}_{\mathcal{K}}(t) = \{\,\text{bonds, neighbour lists, BVH, RDF bins, } \ldots\,\} \tag{23}$$

0.2.5 Cell Content Map

A cell contains particles (or clusters) each with its own container:

$$\mathcal{K} \;\Rightarrow\; \{\mathcal{C}_i\}_{i=1}^{N}$$

Optionally with a connectivity relation $\mathcal{E}$ (bonds / adjacency) treated as derived:

$$\mathcal{E} = \text{Infer}(\{\mathbf{S}_i\},\ \{\mathbf{I}_i\},\ \mathbf{L},\ \mathcal{B}) \tag{24}$$

0.2.6 Comparison to Typical Implementation

Remark 0.2 (Alignment with `Crystal.hpp`). The existing implementation already aligns with the container split:

Code field	$\longleftrightarrow$	Formal layer
`xyz_path, xyzA_path, xyzC_path`		$\mathbf{K}^{\text{id}}$ (immutable references)
`lattice, replication`		$\mathbf{K}^{\text{geom}}$ (geometry)
`ConstructionRecipe{steps, hash}`		$\Pi_{\mathcal{K}},\ \Theta_{\mathcal{K}}$ (recipe lineage)
`mutable inferred_bonds, bonds_computed`		$\mathcal{M}_{\mathcal{K}}(t)$ (caches, explicitly mutable)

The notation imposes one discipline: caches never become "truth." Humans enjoy turning caches into truth, then acting surprised when it rots.

If you keep your cell container this minimal, you get: deterministic rebuild decisions, clean coarse-graining transitions, and provenance you can actually defend in writing.

0.3 Minimal World Container Notation (Multi-Cell / Run / Dataset Layer)

0.3.1 World Container Definition

Define a world (a simulation "instance" that may contain many cells, phases, or time segments) as:

$$\mathcal{W} = (W^{\text{id}},\ W^{\text{top}},\ W^{\text{ctrl}},\ W^{\text{comp}}) \tag{25}$$

- W^{id}: immutable provenance + dataset identity,
- W^{top}: topology — cells, links, domains, boundaries,
- W^{ctrl}: run controls — integrators, schedules, objectives,
- W^{comp}: minimal compute annotation (not caches).

Runtime caches are again excluded and live in $\mathcal{M}_{\mathcal{W}}(t)$.

0.3.2 Immutable World Identity W^{id}

$$W^{\text{id}} = [\,\Theta_{\mathcal{W}},\ \Pi_{\mathcal{W}},\ D_{\mathcal{W}},\ \tau_{\mathcal{W}}\,] \tag{26}$$

World provenance hash.

$$\Theta_{\mathcal{W}} = \mathcal{H}(\{\Theta_{\mathcal{K}_m}\},\ \text{controls},\ \text{code version}) \tag{27}$$

Pipeline identifier.

$$\Pi_{\mathcal{W}} \in \mathbb{B}^n, \qquad n \in [128,\ 512]$$

Dataset manifest.

$$D_{\mathcal{W}} = \{(\text{path}_j,\ \text{fmt}_j,\ \text{hash}_j)\}_{j=1}^{J} \tag{28}$$

Timebase / epoch definition. Not "`created_utc`"—a real timebase:

$$\tau_{\mathcal{W}} = (t_0,\ \Delta t,\ \mathcal{T}) \tag{29}$$

where $\mathcal{T}$ is time domain type (continuous, bucketed, event-driven).

0.3.3 World Topology W^{top}

World contains a set of cells:

$$\{\mathcal{K}_m\}_{m=1}^{M}$$

and a connectivity graph describing interactions / adjacency / exchange between cells:

$$G_{\mathcal{W}} = (V,\ E), \qquad V = \{1, \ldots, M\} \tag{30}$$

Each edge $e = (u, v) \in E$ carries a boundary operator:

$$B_{uv} = (X_{uv},\ \ R_{uv},\ \ S_{uv}) \tag{31}$$

- X_{uv}: exchange type (matter, charge, heat, constraints),
- R_{uv}: mapping operator (index map, ghost region, interpolation),
- S_{uv}: schedule (when exchange occurs).

Minimal world topology bundle:

$$W^{\text{top}} = [\,\{\mathcal{K}_m\},\ \ G_{\mathcal{W}}\,] \tag{32}$$

This is how you model multi-phase systems, slabs + vacuum cells, reservoirs, defect domains, coarse-grained / atomistic coupling — without duct-taping everything into one mega-cell.

0.3.4 World Control Bundle W^{ctrl}

Controls are not "global variables"; they are a formal program.

$$W^{\text{ctrl}} = [\,\mathcal{I},\ \ \mathcal{O},\ \ \mathcal{S},\ \ \mathcal{A}\,] \tag{33}$$

Integrator family. $\mathcal{I} \in \{\text{MD},\ \text{MC},\ \text{hybrid},\ \text{relax}\}$

Objectives (scoring / constraints).

$$\mathcal{O} = \{O_k\}_{k=1}^{K}, \qquad O_k : \mathcal{W} \to \mathbb{R} \tag{34}$$

Schedule (what runs when).

$$\mathcal{S} = \{(t_j,\ \text{op}_j)\}_{j=1}^{J} \tag{35}$$

Audit policy (self-checks).

$$\mathcal{A} = \{A_\ell\}_{\ell=1}^{L}, \qquad A_\ell : \mathcal{W} \to \{0,\, 1\} \tag{36}$$

This makes the always-on self-auditing system first-class, not an afterthought.

0.3.5 Minimal Compute Annotation W^{comp}

$$W^{\mathrm{comp}} = \begin{bmatrix} \Gamma_{\mathcal{W}} \\ \Xi_{\mathcal{W}} \\ \Upsilon_{\mathcal{W}} \end{bmatrix} \tag{37}$$

Complexity class, kernel family selector, annotation confidence. Same pattern as particle and cell. Consistency is boring, which is why it works.

0.3.6 Explicit Runtime Cache Set (Excluded from Truth)

$$\mathcal{M}_{\mathcal{W}}(t) = \{\, \text{neighbour lists, domain decompositions, GPU buffers, BVHs, RDF histograms, FFT plans, } \ldots \tag{38}$$

These are derivable artifacts. If you serialise them, it is only as optional acceleration, never as truth.

0.3.7 Nano–Meso Boundary Extensions

0.3.7.1 Surface–Interior Decomposition.

For any cell or domain $\mathcal{K}$ containing N sites, define a partition into surface and interior subsets:

$$\mathcal{K} = \mathcal{K}^{\partial} \cup \mathcal{K}^{\circ}, \qquad \mathcal{K}^{\partial} \cap \mathcal{K}^{\circ} = \varnothing \tag{39}$$

A minimal deterministic definition uses a local coordination deficit:

$$\delta_i = z_{\mathrm{bulk}}(Z_i,\, \Sigma_i) - z_i \tag{40}$$

where z_i is measured coordination, and z_{bulk} is the expected interior coordination for species class $(Z_i,\, \Sigma_i)$. Then:

$$i \in \mathcal{K}^{\partial} \iff \delta_i \geq \delta_{\min}, \qquad \text{else } i \in \mathcal{K}^{\circ}$$

This split matters because the surface carries:

- dominant chemical reactivity,
- defect nucleation and healing,
- adsorption and morphology selection,

while the interior carries:

- bulk stiffness,
- long-time stability regimes,
- lattice ordering and transport baselines.

Coarse graining that erases this split produces beads with "average" behaviour that matches nothing.

0.3.7.2 Formation-Condition Determinism as a First-Class Map.

At the nano–meso boundary, emergent structure is highly sensitive to formation conditions. The design goal is not to "fit" outcomes, but to model a deterministic projection from formation controls to mesoscopic identity.

Define a formation condition vector:

$$\mathcal{E} = \big[\, T(t),\;\; c(t),\;\; p(t),\;\; P,\;\; R \,\big] \tag{41}$$

where:

- $T(t)$ is the thermal history (dominant driver),
- $c(t)$ is composition / concentration history,
- $p(t)$ is pressure history (or constraint proxy),
- P denotes pathway class (reaction / assembly route),
- R denotes environmental regime (solvent, vacuum, ionic melt, etc.).

The core postulate for the boundary scale is:

$$\boxed{M(t^*) = \mathcal{F}(\mathcal{E})} \tag{42}$$

where $M(t^*)$ is a mesoscopic descriptor at the end of formation (or at a chosen observation epoch t^*).

This is framed positively because it is exactly what makes the nano–meso boundary scientifically valuable: formation is not noise, it is structure selection.

0.3.7.3 Lower Meso Descriptor (Precise Boundary Definition).

Define "lower meso" (nano–meso boundary) as the regime where:

- individual atoms are no longer the preferred degrees of freedom,
- yet the structure is not fully bulk-continuum.

Operationally, represent the system as domains D_k built from atoms:

$$D_k \subset \{1, \ldots, N\}, \qquad k = 1, \ldots, K$$

Each domain carries a minimal descriptor vector:

$$M_k = \big[\, \phi_k,\;\; \psi_k,\;\; \chi_k,\;\; \omega_k,\;\; E_k \,\big] \tag{43}$$

where:

Surface fraction.

$$\phi_k = \frac{|D_k^{\partial}|}{|D_k|} \tag{44}$$

Ordering / crystallinity proxy. Any stable scalar order parameter:

$$\psi_k = \langle \Psi(\mathbf{r}_i) \rangle_{i \in D_k} \tag{45}$$

Defect density.

$$\chi_k = \frac{1}{|D_k|} \sum_{i \in D_k} \mathbf{1}[\delta_i \geq \delta_{\min}] \tag{46}$$

Composition / identity mean.

$$\omega_k = \sum_{i \in D_k} w_i \, \mathbf{I}_i \tag{47}$$

Identity entropy (heterogeneity retained).

$$E_k = \mathrm{Var}(\mathbf{I}_i)_{i \in D_k} \tag{48}$$

This is the boundary-scale sweet spot: you can now distinguish
- uniform crystal domains vs. mixed disordered aggregates,
- smooth surfaces vs. high-defect catalytic skins,
- pathway-selected polymorphs vs. temperature-quenched glassy states.

0.3.7.4 Temperature as the Dominant Deterministic Driver.

Treat temperature not as a scalar input but as a history operator:

$$T(t) \;\Rightarrow\; \beta(t) = \frac{1}{k_B \, T(t)} \tag{49}$$

Formation and relaxation become biased flows on the state manifold:

$$\frac{d\mathbf{S}}{dt} = -\nabla U(\mathbf{S};\, \mathbf{I}) + \eta(t) \tag{50}$$

but in deterministic coarse graining, the randomness term $\eta(t)$ is replaced by a reproducible stencil driven by the provenance hash:

$$\eta(t) \equiv \mathcal{G}\big(\mathcal{H}(\mathbf{I},\, \mathbf{S},\, t)\big) \tag{51}$$

So the mapping $\mathcal{F}(\mathcal{E})$ remains stable across reruns.

This is precisely why temperature is "the main one" at this boundary: thermal history controls which basins are accessible, which defects persist, and whether the surface freezes into metastable motifs.

0.3.7.5 Emergence Is a Feature, Not a Problem.

The hard part is not that formation conditions influence outcome. The hard part is doing it without losing determinism and without lying about coarse-grained meaning.

This framework treats emergence as a structured output:

$$\{M_k\}_{k=1}^{K} \text{ is the emergence layer}$$

and preserves the full trace:

$$\mathcal{E} \;\longrightarrow\; M \;\longrightarrow\; \text{CG beads / fields} \tag{52}$$

So rather than coarse graining being a destructive step, it becomes the formal bridge between formation physics and mesoscopic realism.

Formation Engine Methodology v0.1 — End of §0 (Pre-Section Theory Layer)