

Towards a Group Context-Sensitive Grammar for Genome Analysis (Draft)

Wanhong Huang

huangwanhong.g.official@gmail.com

Abstract

We present the Group-Equivariant Context-Sensitive Grammar (GE-CSG), a formal language framework whose foundational primitive is not an alphabet of concrete symbols but a finite group G —the source of all generativity. Concrete terminal symbols (e.g. $\{A, T, G, C\}$) are derived, not postulated: they arise from a symmetry-breaking mechanism $\mathcal{B}$ applied to the group algebra $\mathbb{C}[G]$. The Wedderburn decomposition $\mathbb{C}[G] \cong \bigoplus_{\rho \in \widehat{G}} M_{d_\rho}(\mathbb{C})$ simultaneously (i) reveals the block-diagonal structure of the grammar, enabling fully parallel parsing across irreducible representation blocks, and (ii) provides the algebraic substrate from which terminal symbols emerge as minimal idempotents. Two qualitatively distinct rewrite arrows are introduced: $\xrightarrow{1}$ for symmetry-preserving generation within $\mathbb{C}[G]$, and $\xrightarrow{2}$ for symmetry-breaking projection onto concrete symbols. We formalise top-down and bottom-up parsing algorithms that exploit orbit compression and block parallelism, and discuss the ontological hierarchy $G \rightarrow \mathbb{C}[G] \rightarrow \Sigma \rightarrow w$ connecting the group-theoretic origin of language to observable sequences.

Contents

1	Motivation and Ontological Stance	3
1.1	The Standard Picture and Its Inversion	3
1.2	The Complexity Wall	3
2	Mathematical Preliminaries	3
2.1	Group Algebras	3
2.2	The Wedderburn Decomposition	4
2.3	Minimal Idempotents and Symmetry Breaking	5
3	The Grammar: Formal Definition	5
3.1	The Five-Tuple	5
3.2	Two Kinds of Rewrite Arrow	6
3.2.1	Generation Arrow $\xrightarrow{1}$: Symmetry-Preserving	6
3.2.2	Breaking Arrow $\xrightarrow{2}$: Symmetry-Breaking	6
3.3	The Two-Phase Derivation	7
4	The Ontological Hierarchy	7
4.1	From <i>The One</i> to Observable Sequence	7
4.2	Why G and $\mathbb{C}[G]$ Are Not Redundant (But Also Not Distinct)	8
5	Parsing Algorithms	8
5.1	Shared Preprocessing: Block Decomposition of Input	8
5.2	Bottom-Up Parsing: Equivariant Earley Algorithm	8
5.2.1	Chart Structure	8
5.2.2	The Three Operations	9
5.2.3	Parse Tree Reconstruction	9
5.2.4	Complexity	9
5.3	Top-Down Parsing: Equivariant Recursive Descent	9
5.3.1	Orbit-Keyed Memoisation	9
5.3.2	Algorithm	10
5.3.3	Generation Mode	10
5.4	Comparison of Parsing Directions	10
6	Residual Symmetry and the Interleaving of Arrows	11
6.1	Breaking Is Not Total	11
6.2	The Vacuum Manifold	11
6.3	Graduated Breaking and Level Interleaving	11
7	Application to Genomic Sequences	11
7.1	Candidate Group for DNA	11
7.2	Symmetry-Breaking Hierarchy for Nucleotides	12
7.3	Biological Interpretation of Grammar Levels	12
8	Integration with Deep Learning (PyTorch)	12
8.1	Equivariant Sequence Encoder	13
8.2	Orbit-Shared Rule Embeddings	13
8.3	Group Fourier Transform Layer	13

9 Summary and Complete Formal Specification	14
9.1 The Grammar in One Place	14
9.2 The Ontological Diagram	14
9.3 Key Properties at a Glance	15
10 Open Questions and Future Directions	15

§ 1 Motivation and Ontological Stance

1.1 The Standard Picture and Its Inversion

Classical formal language theory begins with an *alphabet* Σ of atomic symbols. A grammar is a device for generating strings over Σ . In this picture, the symbols are ontologically primitive: they are given before anything else.

We argue that this picture is *inverted* with respect to the structure actually present in genomic sequences. The nucleotides $\{A, T, G, C\}$ are not featureless atoms; they carry rich symmetry relations—Watson-Crick complementarity ($A \leftrightarrow T, G \leftrightarrow C$), strand reversal, codon positional symmetry, and so on. These symmetries are not incidental decorations on top of an already-given alphabet. They are, in the deepest sense, *prior* to the symbols themselves.

The central ontological commitment of this work is therefore:

Core Thesis

A finite group G —the source of all symmetry—is the true primitive of the grammar. Terminal symbols are *theorems*, not axioms: they are the output of a symmetry-breaking mechanism applied to the group algebra $\mathbb{C}[G]$. The grammar’s generative capacity lives in $\mathbb{C}[G]$; the observable alphabet Σ lives downstream of a breaking event $\mathcal{B}$.

This stance parallels, in spirit, the role of gauge symmetry in physics: the gauge group G is primary; the particle spectrum (“alphabet”) emerges via spontaneous symmetry breaking and vacuum selection.

1.2 The Complexity Wall

A secondary motivation is computational. Context-sensitive grammars (CSGs) are the natural formalism for genomic sequences—long-range dependencies such as splice-site pairing, enhancer-promoter interactions, and tandem repeats all require context sensitivity. However, CSG parsing is PSPACE-complete, and even the most efficient classical algorithms (e.g. Milley’s variant) are limited to sequences of length $\sim 10^3$, far short of the $\sim 10^9$ base-pair scale of eukaryotic genomes.

The group-theoretic approach does not eliminate this wall but *restructures* it. Orbit compression reduces the effective rule set by a factor of $|G|$; the Wedderburn block decomposition makes parsing across blocks fully parallel and independent; and the residual symmetry after breaking can be used to share parse-tree caches across symmetry-equivalent subsequences.

§ 2 Mathematical Preliminaries

2.1 Group Algebras

Let G be a finite group of order $|G|$.

Definition 2.1 (Group Algebra). The *group algebra* $\mathbb{C}[G]$ is the complex vector space with

basis $\{g \mid g \in G\}$, equipped with the bilinear product extending the group multiplication:

$$\left(\sum_{g \in G} \alpha_g g \right) \left(\sum_{h \in G} \beta_h h \right) = \sum_{g, h \in G} \alpha_g \beta_h (gh).$$

Equivalently, viewing elements as functions $f : G \rightarrow \mathbb{C}$, the product is convolution:

$$(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(h^{-1}g).$$

Remark 2.1. G and $\mathbb{C}[G]$ are *not* redundant. They are two descriptions of the same object: G in the language of sets-with-operations, $\mathbb{C}[G]$ in the language of linear algebra. Neither is more “fundamental”; the choice of language is dictated by what one wishes to compute. In this paper we work primarily with $\mathbb{C}[G]$ because the Wedderburn decomposition is most naturally stated there.

2.2 The Wedderburn Decomposition

The key structural theorem is:

Theorem 2.1 (Wedderburn–Artin). Let G be a finite group. Then

$$\mathbb{C}[G] \cong \bigoplus_{\rho \in \hat{G}} M_{d_\rho}(\mathbb{C}),$$

where $\hat{G}$ denotes the set of isomorphism classes of irreducible complex representations (irreps) of G , and d_ρ is the dimension of the irrep ρ . The dimensions satisfy

$$\sum_{\rho \in \hat{G}} d_\rho^2 = |G|.$$

Each block $M_{d_\rho}(\mathbb{C})$ is a full matrix algebra and the blocks are *mutually orthogonal*: elements from different blocks annihilate each other under multiplication.

Example 2.1 ($G = \mathbb{Z}_2$). $G = \{e, g\}$ with $g^2 = e$. The group algebra $\mathbb{C}[\mathbb{Z}_2] = \{\alpha e + \beta g \mid \alpha, \beta \in \mathbb{C}\}$ is two-dimensional. Setting

$$e_+ = \frac{e+g}{2}, \quad e_- = \frac{e-g}{2},$$

one verifies $e_+^2 = e_+$, $e_-^2 = e_-$, $e_+e_- = 0$. Hence

$$\mathbb{C}[\mathbb{Z}_2] \cong \mathbb{C}e_+ \oplus \mathbb{C}e_- \cong \mathbb{C} \oplus \mathbb{C}.$$

Two one-dimensional blocks, corresponding to the trivial representation ρ_+ (even functions on G) and the sign representation ρ_- (odd functions). Applied to DNA: e_+ selects the purine class $\{A, G\}$; e_- selects the pyrimidine class $\{C, T\}$. This classification is not imposed by hand—it is *derived* from the group structure.

Example 2.2 ($G = S_3$ (symmetric group on 3 elements)). $|S_3| = 6$. The irreps are: trivial ($d = 1$), sign ($d = 1$), standard ($d = 2$). Wedderburn gives

$$\mathbb{C}[S_3] \cong \mathbb{C} \oplus \mathbb{C} \oplus M_2(\mathbb{C}), \quad 1^2 + 1^2 + 2^2 = 6. \checkmark$$

2.3 Minimal Idempotents and Symmetry Breaking

Definition 2.2 (Minimal Idempotent). An element $e_{\rho,ij} \in M_{d_\rho}(\mathbb{C}) \subset \mathbb{C}[G]$ is a *minimal idempotent* if it corresponds to the matrix unit E_{ij}^ρ (one in position (i, j) , zero elsewhere). It satisfies

$$e_{\rho,ij} \cdot e_{\sigma,kl} = \delta_{\rho\sigma} \delta_{jk} e_{\rho,il}.$$

Minimal idempotents are the algebraic residue of a basis choice inside each block. Choosing a basis is the algebraic analogue of vacuum selection in spontaneous symmetry breaking: the algebra $M_{d_\rho}(\mathbb{C})$ itself is fully symmetric under conjugation by $GL_{d_\rho}(\mathbb{C})$, but a choice of matrix units breaks this symmetry to a discrete subgroup.

Definition 2.3 (Symmetry-Breaking Mechanism). A *symmetry-breaking mechanism* is a map

$$\mathcal{B} : \mathbb{C}[G] \longrightarrow \mathcal{A},$$

where $\mathcal{A}$ is any target algebraic structure. Three canonical cases:

- (a) **Discrete breaking:** $\mathcal{A} = \Sigma$ is a finite set. $\mathcal{B}$ maps minimal idempotents to concrete symbols.
- (b) **Subgroup breaking:** $\mathcal{A} = \mathbb{C}[H]$ for a subgroup $H \subsetneq G$. The residual symmetry group is H .
- (c) **Continuous breaking:** $\mathcal{A} = C^\infty(G/H)$, the algebra of smooth functions on the coset space.

Remark 2.2 (Graduated Breaking). A single breaking step need not take G all the way to a discrete alphabet. A chain of breakings

$$G = H_0 \supsetneq H_1 \supsetneq H_2 \supsetneq \cdots \supsetneq H_k$$

with $\mathcal{B}_i : \mathbb{C}[H_{i-1}] \rightarrow \mathbb{C}[H_i]$ allows the grammar to operate at multiple levels of symmetry simultaneously. The terminal alphabet Σ appears at the bottom level and carries the *residual symmetry* H_k as a non-trivial group action.

§ 3 The Grammar: Formal Definition

3.1 The Five-Tuple

Definition 3.1 (GE-CSG). A *Group-Equivariant Context-Sensitive Grammar* is a five-tuple

$$\mathcal{G} = (G, V, S, R_1, (\mathcal{B}_i)_{i=1}^k),$$

where:

- G is a finite group (*the One*; all other components are derived from it).
- V is a finite set of *non-terminal* symbols, with $S \in V$ the *start symbol*.
- R_1 is a finite set of *generation rules* (defined below).
- $(\mathcal{B}_i)_{i=1}^k$ is a chain of symmetry-breaking maps $\mathcal{B}_i : \mathbb{C}[H_{i-1}] \rightarrow \mathbb{C}[H_i]$, with $H_0 = G$ and $\Sigma := \text{Im}(\mathcal{B}_k)$.

Note that $\mathbb{C}[G]$, Σ , and the full rule set are all *derived* from the above; they are not independent parameters.

3.2 Two Kinds of Rewrite Arrow

The grammar employs two qualitatively distinct rewrite relations, corresponding to two ontologically different operations.

3.2.1 Generation Arrow $\xrightarrow{1}$: Symmetry-Preserving

Definition 3.2 (Generation Rule). A *generation rule* is a rewrite of the form

$$\alpha A \beta \xrightarrow{1} \alpha \gamma \beta, \quad A \in V, \quad \alpha, \beta, \gamma \in (\mathbb{C}[G] \cup V)^*.$$

The rule operates *entirely within* $\mathbb{C}[G] \cup V$: no concrete terminal symbol appears on either side.

Definition 3.3 (Equivariance Condition). The rule set R_1 is *G-equivariant* if

$$\forall r \in R_1, \forall g \in G : \quad g \cdot r \in R_1,$$

where the group action extends the action on $\mathbb{C}[G]$ to rules component-wise (non-terminals are fixed).

Definition 3.4 (Orbit and Orbit Representative). The *orbit* of a rule r under G is

$$[r]_G = \{g \cdot r \mid g \in G\} \subseteq R_1.$$

The grammar *stores* only one representative per orbit. The full rule set is generated lazily:

$$R_1 = \bigcup_{[r]_G \in R_1/G} [r]_G.$$

By Burnside's lemma, the number of orbits is $|R_1/G| = \frac{1}{|G|} \sum_{g \in G} |\text{Fix}_{R_1}(g)| \leq |R_1|/|G|$.

The Wedderburn decomposition further block-diagonalises R_1 :

$$R_1 = \bigoplus_{\rho \in \hat{G}} R_1^\rho, \quad R_1^\rho \subseteq \text{Rules over } M_{d_\rho}(\mathbb{C}) \cup V.$$

The blocks R_1^ρ are *mutually independent*: no rule in R_1^ρ can produce or consume elements of $R_1^{\rho'}$ for $\rho \neq \rho'$. This enables fully parallel parsing.

3.2.2 Breaking Arrow $\xrightarrow{2}$: Symmetry-Breaking

Definition 3.5 (Breaking Rule). A *breaking rule* is an application of some $\mathcal{B}_i$:

$$x \xrightarrow{2} \mathcal{B}_i(x) \in \mathbb{C}[H_i], \quad x \in \mathbb{C}[H_{i-1}].$$

Breaking rules are the *only* way concrete symbols enter a derivation.

The two arrows have sharply contrasting properties:

Property	$\xrightarrow{1}$	$\xrightarrow{2}$
Domain	$(\mathbb{C}[G] \cup V)^*$	$\mathbb{C}[H_{i-1}]$
Codomain	$(\mathbb{C}[G] \cup V)^*$	$\mathbb{C}[H_i]$ or Σ
G -equivariance	$\checkmark$ (by definition)	$\times$ (by definition of breaking)
Invertibility	generally not	<i>strictly</i> not
Information	symmetry preserved	symmetry partially lost
Parallelism	block-parallel	position-independent
Physical analogy	Hamiltonian evolution	measurement / collapse

3.3 The Two-Phase Derivation

A complete derivation of a terminal string w from S proceeds in two strictly ordered phases:

Definition 3.6 (Two-Phase Derivation).

$$S \xrightarrow{\underbrace{1 \cdots 1}_{\text{Phase I: generation}}} x \in \mathbb{C}[G]^* \xrightarrow{\underbrace{2 \cdots 2}_{\text{Phase II: breaking}}} w \in \Sigma^*.$$

Phase I (pure generation): all non-terminals are expanded using R_1 -rules until only group-algebra elements remain. The symmetry group G acts transparently on every intermediate sentential form.

Phase II (breaking): each group-algebra element x_i in the intermediate string is independently mapped to a concrete symbol $w_i = \mathcal{B}(x_i) \in \Sigma$. Breaking at different positions is *independent*, hence naturally parallel:

$$\mathcal{B}(x_1 x_2 \cdots x_n) = \mathcal{B}(x_1) \mathcal{B}(x_2) \cdots \mathcal{B}(x_n).$$

Remark 3.1 (Breaking is not the End of Symmetry). After breaking, the residual symmetry group H_k still acts on Σ . Generation rules R_1 can be re-instantiated at the level of $\mathbb{C}[H_k]$, yielding a *derived grammar* $\mathcal{G}' = (H_k, V', S', R'_1, \mathcal{B}')$. Thus the two arrows can be interleaved across levels:

$$\mathcal{G}(G) \xrightarrow{2} \mathcal{G}(H_1) \xrightarrow{2} \mathcal{G}(H_2) \xrightarrow{2} \cdots$$

§ 4 The Ontological Hierarchy

4.1 From *The One* to Observable Sequence

The complete ontological chain, from abstract symmetry to concrete sequence, is:

$$\underbrace{G}_{\text{The One}} \longrightarrow \underbrace{\mathbb{C}[G]}_{\text{potentiality}} \xrightarrow{\text{Wedderburn}} \underbrace{\bigoplus_{\rho} M_{d_{\rho}}(\mathbb{C})}_{\text{spectral modes}} \xrightarrow{\mathcal{B}} \underbrace{\Sigma}_{\text{symbols}} \xrightarrow{R_1} \underbrace{w \in \Sigma^*}_{\text{sequence}}$$

- G : the group itself, prior to any representation. It is the source of all symmetry but is not itself an element of any alphabet. In the spirit of Plotinus's *The One*, it is the origin that does not appear among the things it originates.

- $\mathbb{C}[G]$: the group algebra, the space of all possible “superpositions” of group elements. The full generative potential lives here.
- $\bigoplus_{\rho} M_{d_{\rho}}(\mathbb{C})$: the Wedderburn blocks are the “frequency modes” of G —analogous to Fourier components. Each block is an independent channel.
- Σ : concrete symbols arise by breaking the continuous block structure to a discrete basis. They carry residual symmetry $H_k \subseteq G$.
- w : a particular derivation—the individual utterance, the phenomenon.

4.2 Why G and $\mathbb{C}[G]$ Are Not Redundant (But Also Not Distinct)

G and $\mathbb{C}[G]$ mutually determine each other: the group can be recovered from the group algebra as its set of group-like elements $\{x \in \mathbb{C}[G] \mid \Delta(x) = x \otimes x\}$. They are two linguistic expressions of the same mathematical object. We keep G in the grammar definition for conceptual clarity (it names the source of symmetry explicitly) and use $\mathbb{C}[G]$ for computation (where the linear structure is essential).

§5 Parsing Algorithms

5.1 Shared Preprocessing: Block Decomposition of Input

Given a terminal string $w = w_1 w_2 \cdots w_n \in \Sigma^*$, both parsers first lift each symbol to $\mathbb{C}[G]$ and project onto blocks:

$$w_i \mapsto \bigoplus_{\rho \in \widehat{G}} \Pi_{\rho}(w_i), \quad \Pi_{\rho}(w_i) \in M_{d_{\rho}}(\mathbb{C}),$$

and compute orbit representatives and witness group elements:

$$[w_i]_G = \min_{g \in G} \{g \cdot w_i\}, \quad g_i \in G \text{ s.t. } g_i \cdot [w_i]_G = w_i.$$

5.2 Bottom-Up Parsing: Equivariant Earley Algorithm

5.2.1 Chart Structure

Definition 5.1 (Equivariant Earley Item). An *equivariant Earley item* is a triple $[A \rightarrow \alpha \bullet \beta, i]$, where $A \rightarrow \alpha\beta \in R_1$, the dot marks progress through the right-hand side, and i is the start position in the input.

Definition 5.2 (Equivariant Chart).

$$\text{Chart}[i][j] = \left\{ (A, [w_{i:j}]_G, g_{ij}) \mid A \xrightarrow{*}_1 w_{i:j}, g_{ij} \cdot [w_{i:j}]_G = w_{i:j} \right\}.$$

The key insight: two input substrings in the same G -orbit share a single chart entry.

5.2.2 The Three Operations

For each position $j = 0, \dots, n$, iterate to fixpoint:

Predict.

$$\frac{[A \rightarrow \alpha \bullet B \beta, i] \in \text{Chart}[j] \quad B \rightarrow \gamma \in R_1}{[B \rightarrow \bullet \gamma, j] \in \text{Chart}[j]}$$

Scan.

$$\frac{[A \rightarrow \alpha \bullet a \beta, i] \in \text{Chart}[j] \quad w_{j+1} \in [a]_G}{[A \rightarrow \alpha a \bullet \beta, i] \in \text{Chart}[j+1]} \quad (\text{record } g_{j+1})$$

Complete.

$$\frac{[B \rightarrow \gamma \bullet, k] \in \text{Chart}[j] \quad [A \rightarrow \alpha \bullet B \beta, i] \in \text{Chart}[k]}{[A \rightarrow \alpha B \bullet \beta, i] \in \text{Chart}[j]}$$

5.2.3 Parse Tree Reconstruction

After Chart filling, successful parses are items of the form $[S \rightarrow \gamma \bullet, 0] \in \text{Chart}[n]$. Tree reconstruction back-propagates the witness group elements $\{g_i\}$ to restore the actual input from orbit representatives:

$$T_{\text{orbit}} \xrightarrow{\{g_i\} \text{ back-prop}} T_{\text{actual}}.$$

Each internal node carries: the non-terminal A , the span (i, j) , the orbit representative of the applied rule $[r]_G$, and the specific group element $g \in G$ used.

5.2.4 Complexity

Proposition 5.1 (Equivariant Earley Complexity). For context-free rules, the equivariant Earley algorithm runs in time

$$O\left(n^3 \cdot \frac{|R_1|}{|G|}\right)$$

on a single processor. After Wedderburn block decomposition, the ρ -th block requires

$$O\left(n^3 \cdot \frac{|R_1^\rho|}{|G|}\right),$$

and all blocks run *in parallel*, giving an effective speedup of $|\widehat{G}|$ over sequential processing of the uncompressed grammar.

5.3 Top-Down Parsing: Equivariant Recursive Descent

5.3.1 Orbit-Keyed Memoisation

The top-down parser exploits the symmetry structure through orbit-keyed memoisation:

Definition 5.3 (Equivariant Memo Table).

$$\text{Memo}[A, [w_i]_G, i] = \{(T, j) \mid A \xRightarrow{*} w_{i:j}, T \text{ is a parse tree}\}.$$

Two input positions whose remaining suffixes lie in the same G -orbit share a single memo entry. Since the number of distinct orbits is $|\Sigma^{\leq n}|/|G|$ (approximately), the memo table is compressed by a factor of $|G|$.

5.3.2 Algorithm

Parse(A, w, i) :

1. Compute orbit key: $\text{key} = (A, [w_i]_G, i)$.
2. If $\text{key} \in \text{Memo}$: return cached result.
3. Set $\text{Memo}[\text{key}] \leftarrow \emptyset$ (mark in-progress; prevents left-recursion loops).
4. For each orbit-representative rule $A \rightarrow X_1 \cdots X_k \in R_1/G$:

$$\text{results} \leftarrow \text{MatchRHS}(X_1 \cdots X_k, w, i),$$
5. Update $\text{Memo}[\text{key}] \leftarrow \text{results}$; return.

where MatchRHS is defined recursively:

$$\begin{aligned} \text{MatchRHS}(\varepsilon, w, i) &= \{([], i)\}, \\ \text{MatchRHS}(a \cdot \text{rest}, w, i) &= \begin{cases} \text{MatchRHS}(\text{rest}, w, i+1) & \text{if } w_i \in [a]_G, \\ \emptyset & \text{otherwise,} \end{cases} \\ \text{MatchRHS}(B \cdot \text{rest}, w, i) &= \bigcup_{(T_B, j) \in \text{Parse}(B, w, i)} \text{MatchRHS}(\text{rest}, w, j). \end{aligned}$$

5.3.3 Generation Mode

In *generation* (rather than parsing) mode, the top-down algorithm produces new sequences by sampling:

1. Sample a rule orbit: $[r]_G \sim P([r]_G)$.
2. Sample a group element: $g \sim \text{Uniform}(G)$.
3. Instantiate: $r \leftarrow g \cdot [r]_G$.
4. Recursively expand non-terminals in r .
5. Apply breaking: $w_i \leftarrow \mathcal{B}(x_i)$ for each terminal group-algebra element x_i .

The generation process is thus a two-level stochastic system: the choice of orbit is the “structural” randomness (determined by the grammar’s probability model), while the choice of group element within the orbit is the “symmetry” randomness (determined by the breaking mechanism).

5.4 Comparison of Parsing Directions

Property	Bottom-Up (Earley)	Top-Down (Recursive Descent)
Starting point	$w \in \Sigma^*$	$S \in V$
Output	Parse tree T	Parse tree T or new w
Symmetry use	Scan: orbit match	Memo key: orbit normalisation
Context (CS rules)	Natural via Complete	Explicit context parameter
Ambiguity	Chart retains all parses	All memo branches retained
Best for	Known w , find structure	Known structure, verify/generate w

§ 6 Residual Symmetry and the Interleaving of Arrows

6.1 Breaking Is Not Total

A common misconception is that the breaking arrow $\xrightarrow{2}$ annihilates all symmetry. In general, it does not. The image $\Sigma = \text{Im}(\mathcal{B}_k)$ carries a residual symmetry group $H_k \subseteq G$ defined as the stabiliser of the breaking pattern:

$$H_k = \{h \in G \mid h \cdot \mathcal{B}_k(x) = \mathcal{B}_k(h \cdot x), \forall x \in \mathbb{C}[H_{k-1}]\}.$$

The group H_k acts non-trivially on Σ , and this action is precisely what we call “the symmetry of the alphabet.” For DNA, $H_k \cong \mathbb{Z}_2$ (Watson-Crick complementarity).

6.2 The Vacuum Manifold

The set of all possible breaking choices is parametrised by the *vacuum manifold*:

$$\mathcal{M}_{\text{vac}} = G/H_k.$$

Different points in $\mathcal{M}_{\text{vac}}$ correspond to symmetry-equivalent choices of breaking; they all yield isomorphic grammars related by an outer automorphism of G . The choice of a specific point—a specific breaking $\mathcal{B}_k$ —is the analogue of “vacuum selection” in quantum field theory.

6.3 Graduated Breaking and Level Interleaving

In the fully general setting, the two arrows $\xrightarrow{1}$ and $\xrightarrow{2}$ are *not* strictly separated into two sequential phases. Instead, they interleave across levels:

$$\mathcal{G}(G) : S \xrightarrow{1} \cdots \xrightarrow{1} x_1 \xrightarrow{2} \mathcal{G}(H_1) : S' \xrightarrow{1} \cdots \xrightarrow{1} x_2 \xrightarrow{2} \cdots \xrightarrow{2} \mathcal{G}(H_k) : S^{(k)} \xrightarrow{1} \cdots \xrightarrow{1} w.$$

At each level ℓ , a sub-grammar $\mathcal{G}(H_\ell)$ operates with its own generation rules $R_1^{(H_\ell)}$ and produces intermediate strings over $\mathbb{C}[H_\ell]$, which are then broken down to the next level. This structure is reminiscent of a *renormalisation group flow*: each breaking step integrates out some degrees of freedom (the coset $H_{\ell-1}/H_\ell$) and yields an effective grammar at the lower symmetry level.

The updated grammar definition accommodating this structure is:

Definition 6.1 (Layered GE-CSG).

$$\mathcal{G} = (G = H_0 \supsetneq H_1 \supsetneq \cdots \supsetneq H_k, \{R_1^{(H_\ell)}\}_{\ell=0}^k, \{\mathcal{B}_\ell\}_{\ell=1}^k, S).$$

At each level ℓ , generation rules $R_1^{(H_\ell)}$ are H_ℓ -equivariant; breaking rules $\mathcal{B}_\ell$ are *not* $H_{\ell-1}$ -equivariant (that is their defining property).

§ 7 Application to Genomic Sequences

7.1 Candidate Group for DNA

We propose the following composite group as a natural candidate for the symmetry group of double-stranded DNA at the sequence level:

$$G_{\text{DNA}} = \underbrace{\mathbb{Z}_2}_{\text{complement}} \times \underbrace{\mathbb{Z}_2}_{\text{reversal}} \times \underbrace{\mathbb{Z}_n}_{\text{tandem repeat}} \times \underbrace{S_k}_{\text{exon permutation}}.$$

The four factors capture:

- $\mathbb{Z}_2^{\text{comp}}$: Watson-Crick complementarity $A \leftrightarrow T, G \leftrightarrow C$.
- $\mathbb{Z}_2^{\text{rev}}$: 5'→3' reversal symmetry of double-stranded DNA.
- $\mathbb{Z}_n$: cyclic shift symmetry of tandem repeats of period n .
- S_k : permutation symmetry of k exchangeable exon modules in alternative splicing.

The Wedderburn decomposition of $\mathbb{C}[G_{\text{DNA}}]$ then provides independent parsing channels for each combination of irreps from the four factors.

7.2 Symmetry-Breaking Hierarchy for Nucleotides

The nucleotide alphabet arises through a two-step breaking chain:

$$G_{\text{DNA}} \xrightarrow{\mathcal{B}_1} H_1 = \mathbb{Z}_2 \times \mathbb{Z}_2 \xrightarrow{\mathcal{B}_2} H_2 = \mathbb{Z}_2 \xrightarrow{\mathcal{B}_3} \{A, T, G, C\}$$

- $\mathcal{B}_1$ breaks to purine/pyrimidine classes: $e_+ \mapsto \{A, G\}, e_- \mapsto \{C, T\}$.
- $\mathcal{B}_2$ further distinguishes by hydrogen-bond count: $\{A, G\} \mapsto \{A(2), G(3)\}, \{C, T\} \mapsto \{C(3), T(2)\}$.
- $\mathcal{B}_3$ (residual $\mathbb{Z}_2$) maps to the four letters, retaining the complementarity action $A \leftrightarrow T, G \leftrightarrow C$ as the final residual symmetry.

7.3 Biological Interpretation of Grammar Levels

Grammar Level	Symmetry Group	Biological Meaning
$\mathcal{G}(G_{\text{DNA}})$	Full composite	Abstract gene structure
$\mathcal{G}(H_1)$	$\mathbb{Z}_2 \times \mathbb{Z}_2$	Purine/pyrimidine patterns
$\mathcal{G}(H_2)$	$\mathbb{Z}_2$	Complementary strand grammar
$\mathcal{G}(H_3 = \{e\})$	Trivial	Concrete $\{A, T, G, C\}$ sequences

§ 8 Integration with Deep Learning (PyTorch)

The block structure of $\mathbb{C}[G]$ maps naturally onto the architecture of modern neural networks.

8.1 Equivariant Sequence Encoder

Each block $M_{d_\rho}(\mathbb{C})$ defines an independent representation channel. A sequence w is encoded as a tuple of tensors, one per block:

$$\text{Encode}(w) = \{\Pi_\rho(w) \in \mathbb{R}^{n \times d_\rho^2}\}_{\rho \in \widehat{G}}.$$

Group actions act as *permutation matrices* on one-hot encodings:

$$\text{Encode}(g \cdot w) = P_g \cdot \text{Encode}(w), \quad P_g \in \{0, 1\}^{|\Sigma| \times |\Sigma|},$$

a relation that can be *unit-tested* during training to verify equivariance.

8.2 Orbit-Shared Rule Embeddings

Rules in the same G -orbit share a single embedding vector:

$$\text{emb}(g \cdot r) = \text{emb}(r), \quad \forall g \in G.$$

This reduces the embedding table from $|R_1|$ entries to $|R_1/G|$ entries—the same compression factor as the grammar itself.

8.3 Group Fourier Transform Layer

The Wedderburn decomposition is implemented as a learnable (or exact) group Fourier transform layer:

$$\hat{f}(\rho) = \sum_{g \in G} f(g) \rho(g) \in M_{d_\rho}(\mathbb{C}), \quad \forall \rho \in \widehat{G}.$$

Forward passes in this basis are block-diagonal matrix multiplications, which can be dispatched as independent GPU kernels—one per irrep.

§ 9 Summary and Complete Formal Specification

9.1 The Grammar in One Place

Complete GE-CSG Specification

Primitive: A finite group G (“The One”).

Derived objects:

- Group algebra: $\mathbb{C}[G] \cong \bigoplus_{\rho \in \widehat{G}} M_{d_\rho}(\mathbb{C})$
- Terminal alphabet: $\Sigma = \mathcal{B}_k(\mathbb{C}[G])$
- Residual symmetry: $H_k \subseteq G$ acting on Σ

Grammar tuple: $\mathcal{G} = (G, V, S, R_1, (\mathcal{B}_i)_{i=1}^k)$

Two rewrite arrows:

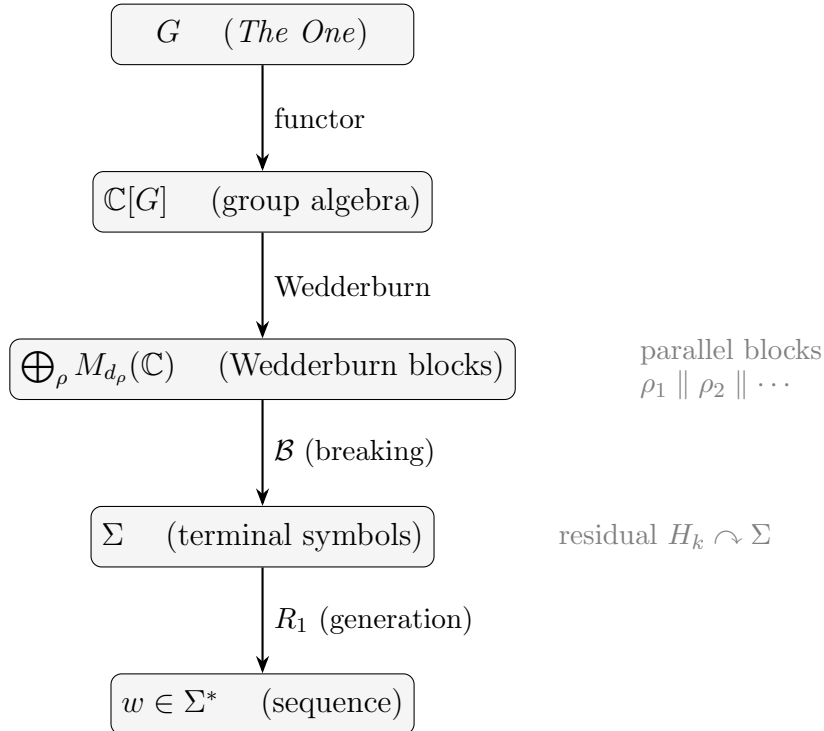
$$\alpha A \beta \xrightarrow{1} \alpha \gamma \beta, \quad \alpha, \beta, \gamma \in (\mathbb{C}[G] \cup V)^*, \text{ } G\text{-equivariant}$$

$$x \xrightarrow{2} \mathcal{B}_i(x) \in \mathbb{C}[H_i], \quad \text{non-equivariant, irreversible}$$

Derivation: $S \xrightarrow{1} \dots \xrightarrow{1} x \in \mathbb{C}[G]^* \xrightarrow{2} \dots \xrightarrow{2} w \in \Sigma^*$

Parsing: Both bottom-up (equivariant Earley) and top-down (orbit-keyed recursive descent) exploit G -orbit compression and Wedderburn block parallelism.

9.2 The Ontological Diagram



9.3 Key Properties at a Glance

Property	Statement
Rule compression	$ R_1/G \leq R_1 / G $ (Burnside)
Block independence	$R_1 = \bigoplus_{\rho} R_1^{\rho}$, blocks non-interacting
Parallel parsing	$ \widehat{G} $ independent parse channels
Orbit cache sharing	Memo key = $(A, [w_i]_G, i)$
Breaking irreversibility	$\xrightarrow{2}$ is non-invertible by definition
Residual symmetry	$H_k \subseteq G$ acts on Σ after breaking
Universal grammar	Same R_1 , different $\mathcal{B} \Rightarrow$ different organisms
Torch-compatible	Block $\leftrightarrow$ GPU kernel; orbit $\leftrightarrow$ shared embedding

§ 10 Open Questions and Future Directions

- Q1. Grammar Induction.** Given a corpus of annotated genomic sequences, how does one learn the orbit-representative rules R_1/G and the breaking parameters $(\mathcal{B}_i)$? The equivariant Inside-Outside algorithm (an EM procedure respecting orbit structure) is the natural candidate, but convergence guarantees under the orbit projection step require analysis.
- Q2. Group Discovery.** How does one infer G from data? Spectral clustering on sequence similarity graphs, algebraic topology (Vietoris-Rips complex automorphisms), and approximate symmetry detection (allowing evolutionary noise) are candidate approaches.
- Q3. Treewidth of the Dependency Graph.** The genomic dependency graph is sparse. If its treewidth is bounded, CSG parsing complexity degrades to $O(n \cdot f(\text{tw}))$, potentially making chromosome-scale parsing feasible.
- Q4. Cross-Species Universality.** Do different organisms share a common G with different $\mathcal{B}$? The framework predicts yes—the “universal grammar” of life is the group G_{DNA} ; species differ in their symmetry-breaking patterns.
- Q5. Non-Abelian Groups.** When G is non-abelian (e.g. S_k for exon permutations), the Wedderburn blocks have $d_{\rho} > 1$. The matrix structure of $M_{d_{\rho}}(\mathbb{C})$ encodes non-commutative context dependencies not expressible in standard CSGs. Their biological interpretation is an open problem.
- Q6. The Origin of G .** Why is the symmetry group of life G_{DNA} and not some other group? Is this an empirical fact (to be discovered by experiment) or a structural necessity (derivable from physical chemistry)? This is the deepest open question of the framework.

References

- [1] Wedderburn, J.H.M. (1908). On hypercomplex numbers. *Proceedings of the London Mathematical Society*, 6(2), 77–118.
- [2] Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2), 94–102.
- [3] Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124.
- [4] Serre, J.-P. (1977). *Linear Representations of Finite Groups*. Springer.
- [5] Cohen, T.S., & Welling, M. (2016). Group equivariant convolutional networks. *ICML 2016*.
- [6] Milley, R. (1999). Parsing with context-sensitive rules. *Proceedings of ACL*.
- [7] Sakakibara, Y. (2003). Grammatical inference in bioinformatics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4), 520–525.
- [8] Lari, K., & Young, S.J. (1990). The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, 4(1), 35–56.
- [9] Dolan, L., & Nappi, C.R. (2001). Noncommutative field theory and the Ward conjecture. *Nuclear Physics B*, 600, 291–305.
- [10] Plotinus. *Enneads*, VI.9 (*On the Good or the One*). Trans. A.H. Armstrong. Harvard University Press, 1988.