

03-Morphology Lookup

Design Document — Morphological Classification & ColorDS Population Script

Constellation Guide Pipeline • Stage 04 • July 2026

1. Purpose

apply_morphology.py is the delivery mechanism for the morphological-research stage of the constellation guide pipeline. Its purpose is to take an observing-list spreadsheet exported from the alias-lookup stage and produce a new spreadsheet in which two columns — "M Class" (morphological classification) and "ColorDS" (B–V color index plus chromaticity) — are populated for every target object. The research itself (catalog lookups, cross-checking, conflict resolution) is performed by the AI assistant per the rules in the companion instruction file Morphological_Research_Task.md; the script deterministically applies those researched values to the spreadsheet so the operation is repeatable, auditable, and reproducible on future constellation datasets.

The script never modifies the input file. Data-integrity problems are never silently dropped: any cell whose value could not be obtained, conflicts between sources, or carries reduced confidence is highlighted yellow and annotated with an explanatory cell comment.

2. What the Script Does

Execution: `python3 apply_morphology.py <input.xlsx> <output.xlsx>`. The script performs six steps:

- **Copy, never modify** — The input workbook is copied to the output path; all writes occur on the copy, preserving the original.
- **Header-name column location** — Row 1 is scanned and the "ID", "ColorDS", and "M Class" columns are found by header name, not position — tolerating column insertions such as the Equipment column added between pipeline revisions. Missing headers abort the run with a diagnostic.
- **Row matching by ID** — Each data row's ID is whitespace-trimmed and matched against the keys of the embedded DATA dictionary (one entry per researched object). Matching by ID, not row number, tolerates row reordering, insertion, and deletion between sheet revisions.
- **Value application** — For matched rows, M Class is always written; ColorDS is written only when the researched entry supplies a value (stellar object types). Non-stellar types leave ColorDS untouched.
- **Flagging** — If a researched entry carries an issue note for either cell, that cell receives a solid yellow fill (FFFF00) and a cell comment containing the full explanation — the source of the conflict, the reason a value is "?", or the confidence caveat.
- **Verification report** — A four-line run report is printed: rows updated, cells flagged, spreadsheet rows with no research match, and research keys that matched no row. The last two lists must be empty (or intentionally explained) before the file is delivered.

3. Expected Inputs

3.1 Input Spreadsheet

An .xlsx workbook whose first worksheet contains a header row (row 1) with at least the following named columns. Additional columns are ignored and preserved.

Column	Requirement
ID	Unique object name per row; the matching key. Multiple-star rows carry the component pair in parentheses, e.g. "Rho Ophiuchi (AB)".
Type	Object type code: GC, OC, DN, PN, CS, DS, MS-X, Star, AS, G, GCl, EN, RN, SRN.
ColorDS	Destination column for B–V / chromaticity values (stellar types only).
M Class	Destination column for the morphological classification.

3.2 Research Data (DATA dictionary)

The DATA dictionary embedded in the script is replaced for each new constellation run. Each entry maps a spreadsheet ID to a 4-tuple:

"Spreadsheet ID": (mclass, colorDS, mclass_issue, colorDS_issue)

- **mclass (str)** — the M Class cell value; "?" when genuinely undetermined.
- **colorDS (str | None)** — the ColorDS cell value, or None for non-stellar types so the cell is left untouched.
- **mclass_issue / colorDS_issue (str | None)** — when present, the corresponding cell is flagged yellow and the text becomes its comment.

4. Values Calculated

4.1 M Class — classification scheme by object type

Type	Scheme	Cell format / example
GC	Shapley–Sawyer Concentration Class	Bare Roman numeral: VIII
OC	Trumpler Classification	Bare class: III2p
DN	Lynds Opacity (1–6)	Opacity=6 (value only, no parentheses)
PN	Balick morphology (Round / Elliptical / Butterfly)	Balick=Butterfly (quadrupolar)
G	De Vaucouleurs class	MC=SA(s)b
CS / Star	Harvard spectral type (+ C-system subtype for carbon stars)	Spec=C5,5 (C-N5)
DS / MS-X	Harvard spectral types of both indicated components	Spec=K2V, K1V
AS	None exists	? (flagged with comment)

4.2 ColorDS — B–V color index and chromaticity

Populated only for stellar types (Star, CS, DS, MS-X); all other types are left blank. Component order matches the spectral types in M Class.

Object class	Cell format / example
Single star / carbon star	BV=+1.71 Color=red
Double / multiple pair	BV= +0.86, +1.15 Color= orange, orange-red

Chromaticity vocabulary: blue, blue-white, white, yellow-white, yellow, yellow-orange, orange, orange-red, red, deep red. B–V for carbon stars is treated as critical and researched exhaustively; because these are pulsating variables, the recorded value is a representative catalog value, phase-dependent by nature. When only a combined-system B–V exists for a pair, it may stand in for the dominant primary with "?" for the secondary — always flagged with a comment.

5. Website Resources Utilized

Resource	Used for
SIMBAD (simbad.cds.unistra.fr)	Spectral types, B and V fluxes (B–V), cross-identifications for all stellar objects.
Corwin GGCl list (haroldcorwin.net/globulars/GGCl.txt)	Authoritative Shapley–Sawyer concentration classes ("CC" column, Arabic → Roman conversion).
Lynds' Catalogue of Dark Nebulae (via Wikipedia table)	LDN ↔ Barnard cross-match and opacity values for dark nebulae.
Stelle Doppie (stelledoppie.it)	Priority source for double/multiple-star component spectral types ("Spectra Class" field).
Washington Double Star catalog (WDS) data	Component magnitudes, separations, and spectral types for pairs.
Wikipedia / SEDS (spider.seds.org)	Cross-checks for all types; Sky Catalogue 2000.0 Trumpler classes; object infoboxes.
WEBDA / DAML02 (Dias) catalogs	Primary Trumpler classifications for open clusters.
Balick (1987, AJ 94, 671) + morphology literature	Planetary nebula Round/Elliptical/Butterfly categories.
GCVS / AAVSO VSX	Variable and carbon star spectral types and designations.
Barentine carbon star list (johncbarentine.com)	SIMBAD-sourced B–V values for carbon stars.
SAC Red Stars, Yale BSC, NED, VizieR, theskylive.com , in-the-sky.org	Supplementary verification and gap-filling.

6. Data Integrity Validation

Integrity is enforced at three levels — during research, in the spreadsheet, and at run time:

- **Research level** — Every value is cross-checked against at least one independent source where possible. Pre-existing sheet values are re-verified, not trusted; demonstrably wrong entries (e.g., a carbon star recorded as M4V) are overwritten with the corrected value.
- **Fill-with-flags, never blank-or-guess** — No value is ever fabricated or silently guessed. Unobtainable values are recorded as "?" and the cell is highlighted yellow with a comment stating exactly why (not published, too faint, scheme does not exist) and what follow-up would resolve it.
- **Run-time verification** — The four-line run report proves completeness: unmatched spreadsheet rows expose unresearched objects; unmatched research keys expose typos. Both must be empty before delivery. A post-run check confirms every yellow cell carries a comment.

Yellow flags fall into four defined categories, each named in its comment: (1) value does not exist in the literature; (2) source conflict, with both sources cited; (3) lower confidence, e.g. secondary references only or mass-based estimates; (4) combined-system B–V standing in for a component value.

7. Source Selection Logic

When sources disagree or coverage is partial, data is selected by the following rules:

- **Type-specific priority** — Each object type has a designated authority (Section 5): Corwin GGCI for globulars, Lynds for dark nebulae, Stelle Doppie for pairs, GCVS for carbon-star types. Its value wins unless demonstrably in error.
- **Two-source rule for conflicts** — When the designated authority is silent or ambiguous, the value supported by two or more independent sources is adopted; the losing value and its source are preserved in the cell comment (e.g., Barnard 59 opacity: Pipe-Nebula sources 6 vs. raw Lynds 5 → 6 adopted, conflict noted).
- **Primary over secondary** — Original catalogs and peer-reviewed literature outrank aggregators; aggregators outrank observing guides. A value resting only on secondary references is adopted but flagged medium-confidence (e.g., Collinder 350/359 Trumpler classes).
- **Alternatives preserved** — Where dated classifications disagree (Trumpler 26: Ruprecht 1966 II2p vs. Lyngå 1982 III1m), the better-corroborated value is adopted and the alternative is recorded in the comment rather than discarded.
- **Honest unknowns** — A cited "?" beats an uncited number: if no source of acceptable quality yields a value, the cell is flagged rather than estimated — with the sole documented exception of combined-system B–V for a dominant primary, which is always disclosed in a comment.