VLMS Companion Analysis System

Testing binary-origin pathways for planetary-mass companions around very low-mass stars (VLMS)

1) Scientific motivation

Close Saturn/Jupiter–mass companions around ultra–low-mass M dwarfs pose an apparent tension with disk-based planet formation when framed solely as "planets from a circumstellar disk." This repository implements a quantitative test of an alternative: **mass-asymmetric turbulent cloud fragmentation** ("failed binary") followed by **post-birth migration** (disk torques and/or high-eccentricity cycles plus tides). The analysis is deliberately modest in scope but statistically explicit and fully reproducible.

Key questions addressed

1. **Demographics:** Do companions to VLMS hosts (0.06–0.20 M⊙) exhibit bimodality in $(log q, log a)$ consistent with a binary-like cohort (fragmentation) distinct from a planet-like cohort?
2. **Orbital architecture:** Are eccentricity distributions $e(a)$ systematically different between low- and high-mass-ratio companions?
3. **Migration plausibility:** Are there credible regions of external perturber parameter space where **Kozai–Lidov (KL) cycles + tides** can shrink orbits to $a \sim 0.05$ AU within ~Gyr, and/or can early disk torques do so within a protoplanetary-disk lifetime?
4. **Classification:** Can a transparent, minimal **origin classifier** that assigns a probability of "binary-like" origin to individual systems (including TOI-6894b) be created?

2) Data provenance (observational, not simulated)

- NASA Exoplanet Archive TAP (PSCompPars; official TAP endpoint): https://exoplanetarchive.ipac.caltech.edu/TAP Column reference: https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html
- Brown Dwarf Companion Catalogue (dataset landing): https://ordo.open.ac.uk/articles/dataset/Brown_Dwarf_Companion_Catalogue/24156393 Code/mirror: https://github.com/adam-stevenson/brown-dwarf-desert

Primary variables used: host mass $M_\star$ (M⊙), companion mass M_c (M_J; true or $m \sin i$, flagged), semi-major axis a (AU), eccentricity e, discovery method, [Fe/H] where available. We form $q = M_c/M_\star$ (with $1 M_\odot = 1047.56 M_J$) and restrict to **VLMS hosts** ($0.06 \leq M_\star/M_\odot \leq 0.20$).

Selection / cleaning summary

- Drop rows lacking any of {$M_\star$, M_c, a, e}.

- Retain both true-mass and $m \sin i$ (flagged); sensitivity checks exclude $m \sin i$.
- Clip $e \in [0, 1)$; handle upper limits in robustness tests (see §8).

Candidate requirements

The candidates that are counted and processed by the new interactive and percentage modes must satisfy:

- **VLMS host criteria:** Stellar mass between 0.06–0.20 M☉
- **Data completeness:** Must have stellar mass, companion mass, and semimajor axis values
- **Physical plausibility:** Stellar temperature 2000–4000K, reasonable metallicity (-2.5 to +0.7 dex)
- **Orbital validity:** Semimajor axis > 0, eccentricity in range [0,1)

3) Installation & environment (CPU-optimized)

Use a BLAS-backed scientific Python stack. Example with conda:

```
conda create -n toi6894 python=3.11 numpy scipy pandas scikit-learn statsmodels numba matplotlib reque:
conda activate toi6894
```

Threading (avoid oversubscription):

```
export OMP_NUM_THREADS=1
export OPENBLAS_NUM_THREADS=1
export MKL_NUM_THREADS=1
export NUMEXPR_NUM_THREADS=1
export NUMBA_NUM_THREADS=<n_cores>   # e.g., 24 on Threadripper 2970WX
```

On multi-die NUMA CPUs (e.g., AMD 2970WX), interleave memory:

```
numactl --interleave=all python panoptic_vlms_project.py --fetch --outdir results
```

4) End-to-end usage

4.1) Interactive candidate counting and percentage selection

Report the number of candidates that meet requirements and interactively specify what percentage to process:

```
python panoptic_vlms_project.py --count-candidates --outdir results
```

This mode will:

1. Count candidates from both NASA Exoplanet Archive and Brown Dwarf Catalogue
2. Display the total number that meet the candidate requirements
3. Wait for user input to specify what percentage (0-100) to process
4. Allow the user to type 'exit' to quit without processing

4.2) Non-interactive percentage processing

Process a specific percentage of candidates without interaction:

```
python panoptic_vlms_project.py --fetch --percent 50 --outdir results
```

Or with local files:

```
python panoptic_vlms_project.py --ps pscomppars_lowM.csv --bd BD_catalogue.csv --percent 25 --outdir re
```

4.3) Standard usage modes

Fetch fresh catalogs and run full analysis:

```
python panoptic_vlms_project.py --fetch --outdir results
```

Run on local CSVs you already have:

```
python panoptic_vlms_project.py --ps pscomppars_lowM.csv --bd BD_catalogue.csv --outdir results
```

Customize the plotted marker for TOI-6894b (host mass, companion mass, and "final" a for figure annotations):

```
python panoptic_vlms_project.py --fetch --toi_mstar 0.08 --toi_mc_mj 0.30 --toi_a_AU 0.05 --outdir resu
```

Provide the system age (in Gyr) to activate age–orbit comparisons against the rest of the catalog:

```
python panoptic_vlms_project.py --fetch --toi_age_gyr 5.0 --outdir results
```

When TOI-6894's age is supplied, the pipeline emits `results/age_comparison.csv` summarizing Δage, semimajor axis, and eccentricity for every system with a measured host age.

The script prints a summary and writes all artifacts to `results/` (filenames listed in §7).

5) Data model (column schema after preprocessing)

The stacked VLMS dataset (`vlms_companions_stacked.csv`) contains at minimum:

- `host_mass_msun` (M⊙), `companion_mass_mjup` (M_J), `mass_ratio` q,
- `semimajor_axis_au` (AU), `eccentricity` (unitless),
- `discovery_method` (string), `metallicity` (dex, may be NaN),
- `host_age_gyr` (Gyr, when available),
- **Derived quantities:** `log_mass_ratio` , `log_semimajor_axis` , `log_host_mass` , `above_deuterium_limit` , `high_mass_ratio` ,
- **Age analysis features:** `age_group` ∈ {Young, Intermediate, Old, Unknown}, `log_host_age_gyr` , `tidal_timescale_proxy` , `migration_efficiency` , `potential_migrator` ,
- **TOI comparison:** `age_delta_vs_toi_gyr` , `is_younger_than_toi` (when TOI age provided),
- `data_source ∈ {NASA, BD_Catalogue, TOI}` .

We also write object-level probabilities `P_binary_like` after classification (§6.4).

6) Analysis methods (statistical spine)

6.1 Mixture in $(log q, log a)$

We fit 1-component and 2-component **Gaussian Mixture Models (EM)** and compare by **BIC**:

\mathbf{z}_i=(\log q_i,\log a_i),\qquad p(\mathbf{z}_i)=\sum_{k=1}^{K}\pi_k\,\mathcal{N}(\mathbf{z}_i\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),\ K\in\{1,2\}.

Deliverable: `gmm_summary.json` (BICs, winner), plus labels/responsibilities used in downstream plotting.

6.2 Eccentricity architecture

We model e in **two subsets** (split at $q = 0.01$ by default):

$$e \mid z = k \sim \mathrm{Beta}(\alpha_k, \beta_k), \quad k \in \{\text{low-}q,\ \text{high-}q\},$$

with MLE via log-parametrization; uncertainty from nonparametric bootstrap (optional extension). A **KS two-sample test** compares the empirical CDFs. Deliverables: `beta_e_params.csv` (parameters), `ks_test_e.txt` (KS statistic, p-value).

Every run now also performs a **bootstrap bagging** pass (default 500 resamples, 80% sampling fraction) on the eccentricity split. This reports the stability of the fitted Beta parameters and the KS/Mann–Whitney statistics: `beta_e_bootstrap_summary.json` captures aggregate moments and detection rates, while `beta_e_bootstrap_distributions.csv` stores the individual bootstrap draws for custom diagnostics.

6.3 Migration feasibility (KL + tides; plus a disk-torque sanity band)

- **Kozai–Lidov timescale** (quadrupole, order-of-magnitude):

$$t_{\rm KL} \sim \frac{M_\star + M_c}{M_{\rm out}} \frac{P_{\rm out}^2}{P_{\rm in}} \left(1 - e_{\rm out}^2\right)^{3/2}.$$

We explore a grid over $(M_{\rm out}, a_{\rm out})$ and randomize $e_{\rm out}$ (and a proxy for inclination) to estimate the **fraction of draws** that (i) satisfy $t_{\rm KL} \leq T$ and (ii) achieve periapsis r_p below a critical threshold.

- **Tidal shrink (intuition):**

$$t_a \approx \frac{2Q'_\star}{9} \frac{M_\star}{M_c} \left(\frac{a}{R_\star}\right)^5 \frac{1}{n}, \quad n = \sqrt{\frac{GM_\star}{a^3}}.$$

At $a \approx 0.05$ AU and $Q'_\star \sim 10^{6-7}$, stellar tides alone are **too slow** unless high-e phases produce very small periastron; hence the dual emphasis on **KL-assisted** or **early disk** migration.

Deliverable: `fig3_feasibility.png` (heat-map of feasibility fraction) + `feasibility_map.npz` . The script uses a conservative periastron criterion (default $r_{\rm crit} \sim 5R_\star$) and a 1 Gyr horizon, both user-tunable in code.

Disk torques: We also report order-of-magnitude Type-I–like timescale bands in the paper text using:

$$t_{\rm mig} \sim C \frac{M_\star}{M_c} \frac{M_\star}{\Sigma a^2} \left(\frac{H}{a}\right)^2 \Omega^{-1}, \qquad \Omega = \sqrt{\frac{GM_\star}{a^3}},$$

for M-dwarf-appropriate $\Sigma(a)$, H/a, and C. (This is documented in the manuscript; the current script emphasizes the KL+tide feasibility map for reproducibility.)

6.4 Minimal, testable origin classifier

We publish a **regularized logistic** model giving $P(\text{binary-like})$ using features

$$x = \left(\log q,\ \log a,\ e,\ \log M_\star,\ [\text{Fe/H}],\ \text{method dummies}\right).$$

Training is performed on heuristic anchors (high-q vs low-q) as a **fallback**; with labeled anchors available, swap in that label vector. We report **5-fold AUROC** and write per-object probabilities to `objects_with_probs.csv` . This is intended as a practical, transparent tool—coefficients can be exported for community use.

6.5 Age–orbit correlation study

- Ingest `st_age` (PSCompPars) or catalogue ages mapped onto `host_age_gyr` when available; derive Δage ≡ age − age_TOI.
- Flag systems younger than TOI-6894b and assess how Δage co-varies with semimajor axis and eccentricity (Pearson correlations, median Δage, younger fraction).
- Deliverables: `age_comparison.csv` (rows with age, Δage, a, e, source) and an "Age comparison" block inside `SUMMARY.txt` with the summary statistics.

6.6 Age-migration regression analysis

Introductory statistical approach preceding the physics-based migration modeling:

- **Simple correlations:** Pearson and Spearman correlations between stellar age and orbital parameters (semimajor axis, eccentricity).
- **Linear regression models:**
  - $\log a \sim \log(\mathrm{age})$: Power-law relationship between orbital distance and stellar age
  - $e \sim \log(\mathrm{age})$: Eccentricity evolution with stellar age
  - **Multiple regression:** $\log a \sim \log(\mathrm{age}) + e + \log M_\star$: Combined age and stellar property effects
- **Deliverables:** `age_regression_summary.json` (coefficients, R², p-values), `age_regression_report.txt` (detailed analysis report)

6.7 Age-dependent migration physics

Physics-based approach incorporating stellar evolution effects:

- **Age-dependent stellar properties:**
  - Stellar radius: $R_\star(t) = R_{\mathrm{MS}} \times [1 + 0.1\log_{10}(t/1\,\mathrm{Gyr})]$ (young stars larger, contract with age)
  - Tidal Q-factor: $Q_\star(t)$ increases from $\sim 10^5$ (young) to $\sim 10^7$ (old) as magnetic activity declines
- **Age-dependent migration timescales:**
  - **Kozai-Lidov cycles:** Timescale independent of age, but available migration time = min(KL timescale, stellar age)
  - **Tidal evolution:** $t_{\mathrm{tidal}} \propto Q_\star(t) \times (a/R_\star(t))^5$ — young systems migrate faster due to larger radii and lower Q-factors
- **Enhanced feasibility analysis:** 3D parameter space (perturber mass, separation, **stellar age**) to identify optimal migration scenarios
- **Migration efficiency indicators:** Systems classified by ratio of tidal timescale to stellar age — efficient migrators have ratios $\lesssim 10$

7) Outputs (reproducibility artifacts)

- **Figures** `fig1_massmass.png` — $M_\star$ vs M_c (log–log), with 13 M_J and 0.075 M☉ lines; TOI-6894b marked. `fig2_ae.png` — e vs a (log a), styled by mass ratio and discovery method. `fig3_feasibility.png` — KL + tides feasibility fraction across $(M_{\mathrm{out}}, a_{\mathrm{out}})$.
- **Data tables** `vlms_companions_stacked.csv` — Combined cleaned catalog for VLMS hosts with enhanced age analysis features. `objects_with_probs.csv` — Each object with q, $P_{\mathrm{binary\_like}}$, and metadata. `age_comparison.csv` — Systems with measured ages, Δage vs TOI-6894b, a, e.
- **Model summaries** `gmm_summary.json` — BIC(1-comp) vs BIC(2-comp); chosen model. `beta_e_params.csv` — $(\hat{\alpha}, \hat{\beta})$ by subset. `ks_test_e.txt` — KS statistic and p-value on e distributions. `age_regression_summary.json` — Age-migration regression coefficients, R², and statistical tests. `age_regression_report.txt` — Detailed age-migration regression analysis report. `feasibility_map.npz` — Arrays used to render Fig. 3. `SUMMARY.txt` — One-page recap including

source URLs (see §2), age-correlation metrics, age-regression results, and the three headline numbers you'll quote in the paper.

8) Robustness and selection-effect controls

- **Detection method stratification:** Repeat mixture and e analyses excluding each method (RV / transit / imaging / astrometry) to show stability.
- **Inclination censoring:** Repeat with true-mass subset only (drop $m \sin i$); qualitative conclusions unchanged in tests to date.
- **Upper limits on e:** Provide two passes—(a) exclude limits; (b) EM-style treatment with truncated likelihood. Expect the high-q skew to persist.
- **Heterogeneous uncertainties:** Main results are unweighted; a heteroscedastic extension (optional) yields consistent partitions.
- **Sensitivity of KL map:** Re-run for $T \in \{1, 3, 5\}$ Gyr and $r_{\mathrm{crit}}/R_\star \in \{3, 5, 7\}$; report coverage fractions.

9) Performance guidance

Typical end-to-end run (few hundred systems) is CPU-bound and fast:

- GMM / Beta / logistic + CV: seconds to minutes.
- KL map (100×100 grid, ~200 draws per cell): minutes; vectorized NumPy suffices. Use `NUMBA_NUM_THREADS` and `numactl --interleave=all` on Threadripper-class CPUs.

10) Troubleshooting

- **KeyError on column names:** Ensure your local CSVs expose `st_mass, pl_bmassj, pl_orbsmax, pl_orbeccen` ; the Brown Dwarf CSV loader maps catalogue-specific names onto these. If a mass column in Earth masses is required downstream, we derive it from M_J via $1 M_J = 317.828 M_\oplus$.
- **Too few VLMS rows:** Confirm the ADQL host-mass filter ($0.06 \le M_\star/M_\odot \le 0.20$) and that `pl_bmassj` is not NULL in your export.
- **Runtime/memory spikes:** Check you haven't set conflicting thread env vars; keep BLAS threads at 1 and let joblib/NumPy parallelize hot loops.

11) How to extend

- Replace the heuristic training labels with a curated anchor set (wide imaged BDs vs disk-formed sub-Neptunes).
- Add Gaia DR3 NSS outer-perturber cross-matches for systems with astrometric companions: https://www.cosmos.esa.int/web/gaia/dr3-non-single-stars

- Promote the KL+tide toy criterion to a proper secular code with tidal evolution (e.g., add a lightweight integration for a subset and compare feasibility fractions).

12) Citation and code availability

Please cite the analysis note and repository if you use any part of this pipeline:

```
Johnson, R.S. (2025). Binary-Origin Substellar Companions Around M Dwarfs: Evidence from Demographics,
```