

The Mountain Goats, Every Night

Scraping and modeling 34 years of live setlists from themountaingoats.fandom.com

The dataset

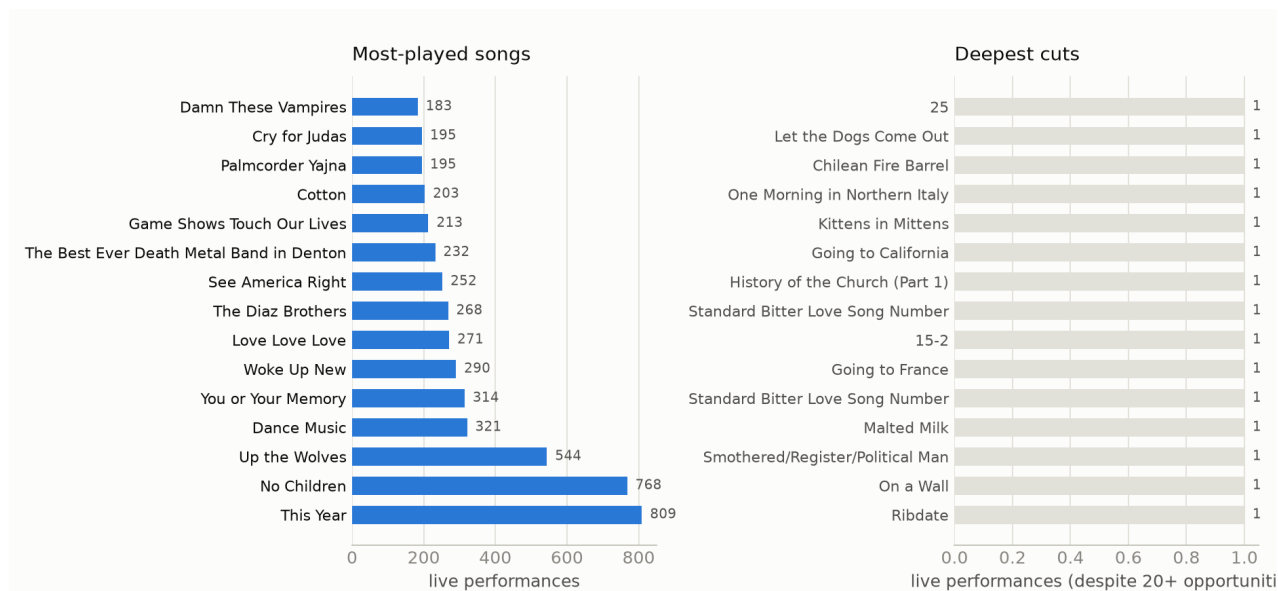
The wiki lists 1,579 live shows from 1992-05-31 through 2026-05-26 — solo John Darnielle sets, full-band tours, radio sessions, festival slots, and Extra Glenns/Extra Lens side-project shows. 1,317 of those have a documented setlist: 22,699 individual song performances across 845 distinct songs, spanning 97 named tours. 1,739 performances have at least one linked YouTube/Vimeo video.

Getting from a wiki category page to those numbers took a few real fixes:

- **Tour isn't in the page text — it's a wiki category** (e.g. *Peter Hughes Farewell Tour 2024*). Two earlier scraper attempts both missed this and scraped a null tour for every show.
- **Encores aren't marked in the setlist table.** They're stated in prose in the Notes section — "The encore was songs 19 through 23" — referencing the table's printed order numbers. A small parser turns ~30 recurring phrasings of that sentence into a per-song encore number.
- **Song identity needed real normalization.** The wiki links each song to its own page, but the same song shows up under underscore-joined slugs, space-joined display text, and case variants of both. Left unmerged, this silently splits a song's play count across rows — a real bug caught mid-project, where "Southwood Plantation Road" initially showed 1 play instead of its real count, 165.
- **Covers needed their own source of truth.** Sniffing setlist notes for the word "cover" only catches about a dozen of them — the wiki actually maintains its own Category:Covers, which tags 156 of the songs played live here. Every song carries an `is_cover` flag from that category, and the prediction model excludes covers from its candidate universe entirely.
- **A disambiguated title quietly duplicated as its own note.** 20 songs have their own wiki page distinct from a same-named album (Tallahassee the song vs. Tallahassee the album); the setlist link's visible text is just the song title, but its title attribute is the disambiguated form. The note-cleanup step was stripping the resolved title from the raw cell text instead of what was actually there, leaving the plain link text stranded as a bogus note on hundreds of performances. Fixed by stripping what's actually in the raw text.

The pipeline is two-stage: **fetch** downloads and locally caches every wiki page via the MediaWiki API (resumable, incremental on later runs), and **build** parses the cached HTML into tidy tables entirely offline.

What gets played, and what doesn't

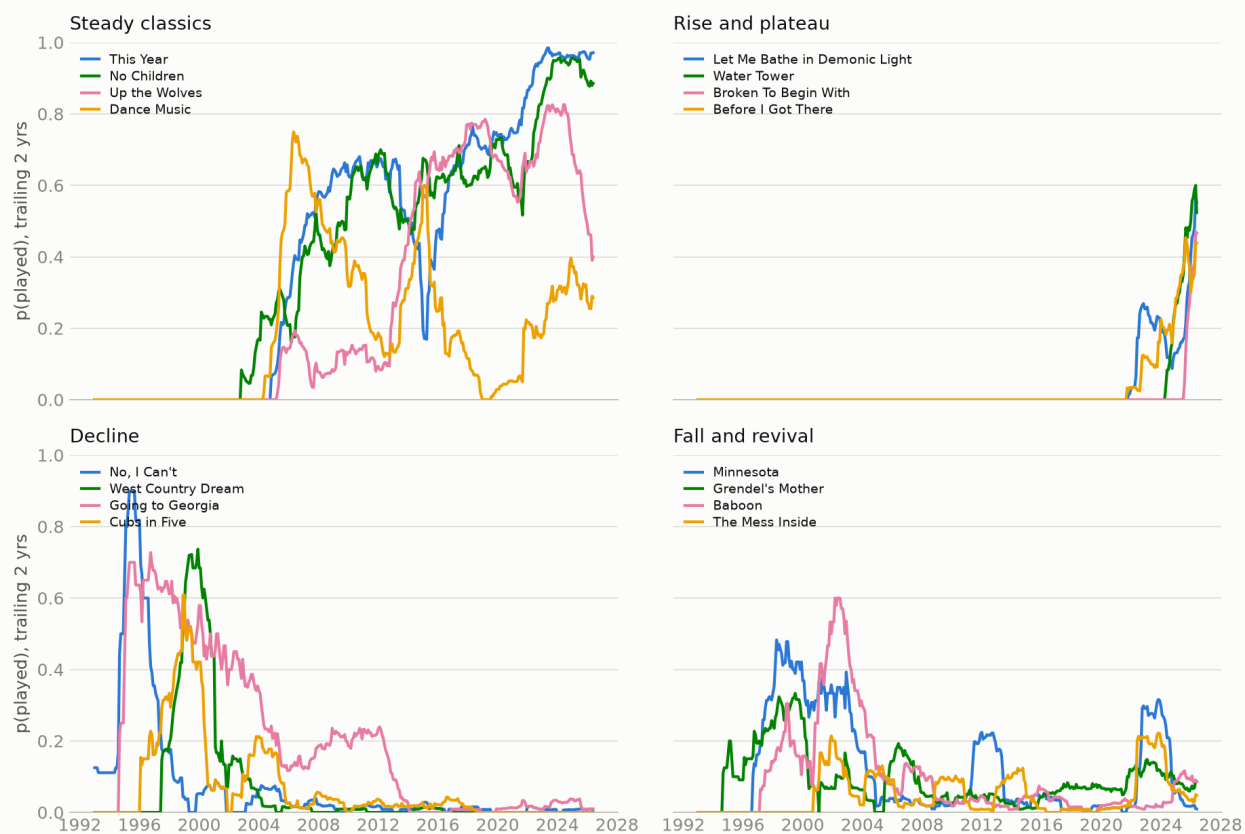


“This Year” and “No Children” anchor the top of the list — both played at a majority of all shows since entering rotation. The deep-cuts panel isn’t just least-played-ever (most of the 845 songs have only 1–2 plays and aren’t interesting on their own); it’s the bottom of an opportunity-adjusted, shrinkage-smoothed play rate among songs with at least 20 chances to be played since their live debut — so a song only counts as a deep cut if it’s been consistently passed over, not just recently written.

Song popularity over time

Treating each song’s live history like a word’s frequency in Google Ngrams — a trailing 2-year play rate sampled monthly — turns 34 years of setlists into trajectories: songs sit at zero before they’re written, rise as they enter rotation, plateau, decline, and sometimes come back.

Song popularity over time



These four groups were picked automatically from the computed trajectories, not hand-selected: the most-played songs overall (steady classics); songs that debuted in the back half of the catalog's history and have climbed straight into rotation since (rise and plateau — a song appearing from nothing and taking hold); the biggest gap between early-career peak and recent play rate (decline); and songs with a real trough between two active periods (fall and revival).

The decline panel caught something the wiki itself corroborates: “Going to Georgia” was a fixture through the mid-2000s, then drops off hard — a 2012 show note has John Darnielle explaining, mid-show, that he considers it a “bullshit song” with an “asshole” narrator and doesn’t want to play it again. The data shows the falloff independent of that anecdote; the anecdote just explains it.

Predicting the setlist

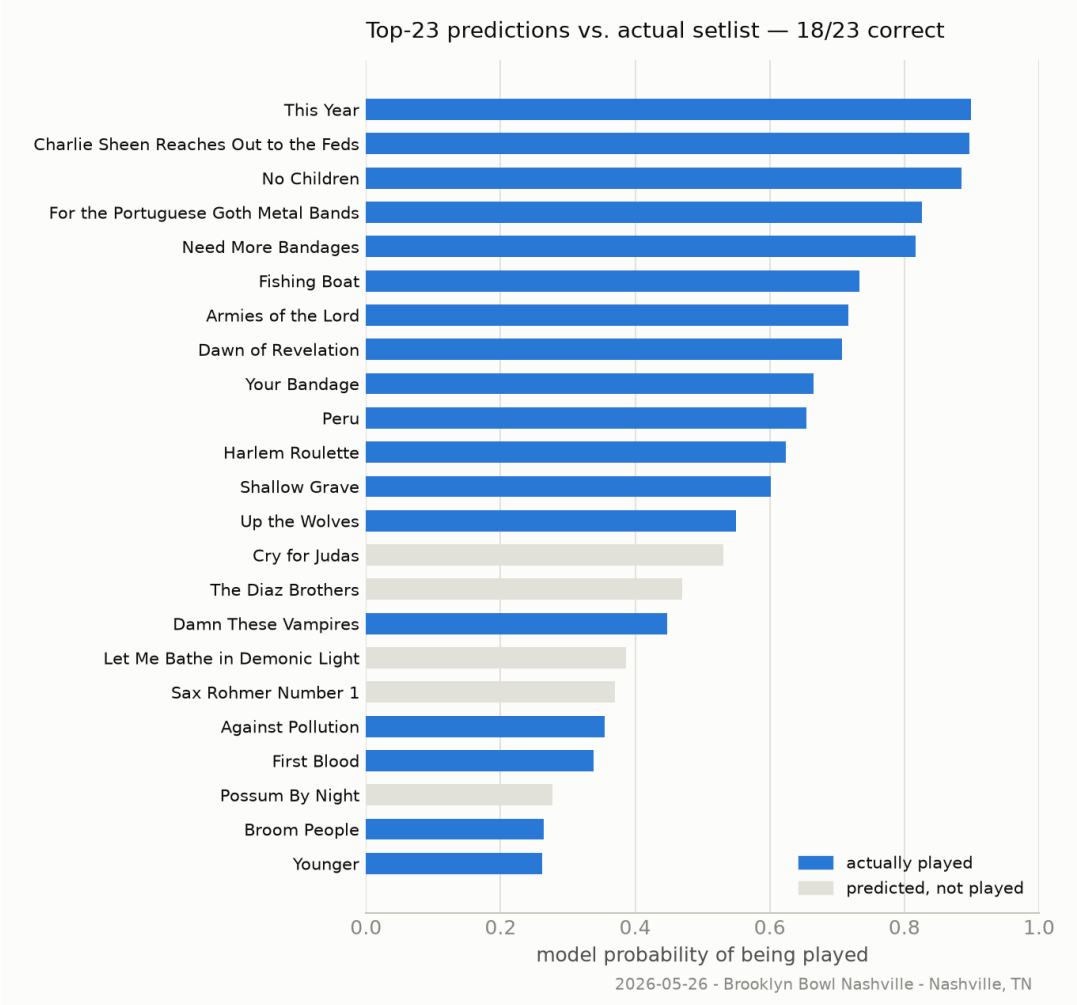
Framed as one binary outcome per (show, song) pair — is this the night song X gets played? — every one of the 1,310 dated, setlisted shows becomes a training example, using only information available before that show: how recently and how often the song's been played, how it's fared on the current tour, its age, and whether it's a shortened-set special appearance (radio/festival/TV).

A logistic regression on those features is compared against a simple but strong baseline — an exponentially-decayed play rate alone — on a strict temporal split (train through 2022, test on the 207 shows from 2023 onward):

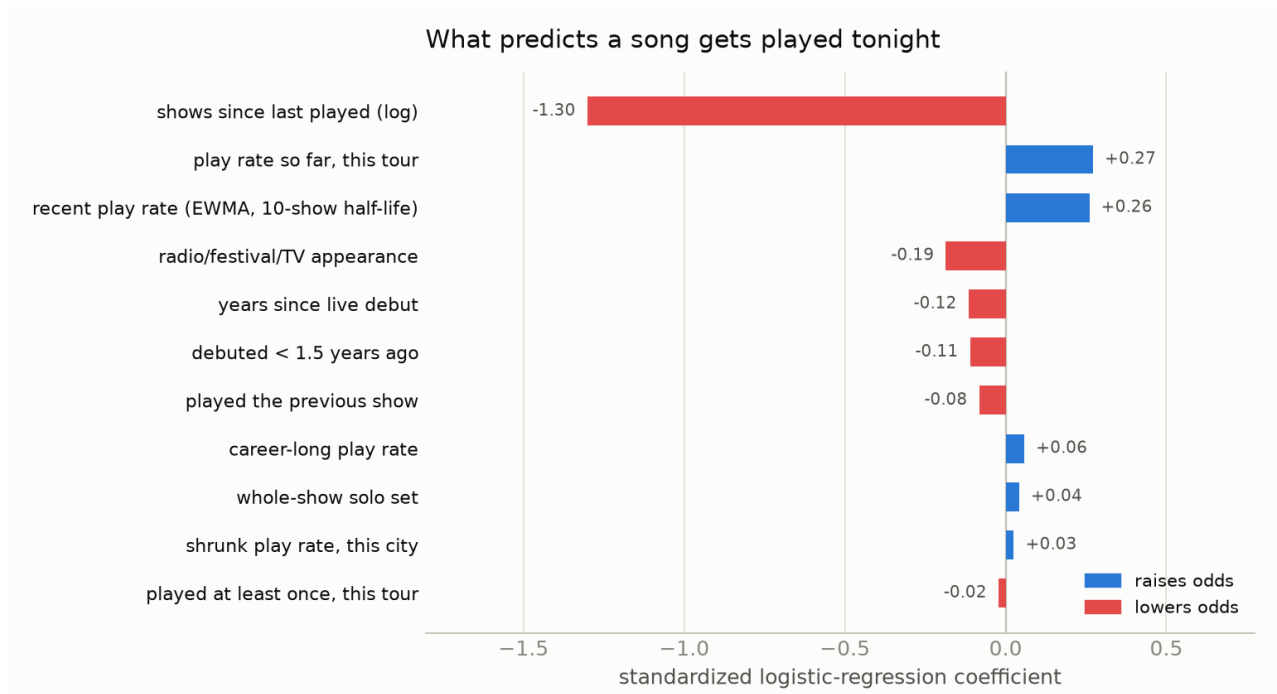
model	log-loss	Brier score	top-n setlist recovery
-------	----------	-------------	------------------------

Baseline (decayed play rate)	0.0822	0.0195	51.1%
Logistic regression	0.0673	0.0167	59.4%

Top-n setlist recovery: for each show, take the model's n highest-probability songs, where n is the actual number played that night, and measure what fraction were right. On the most recent show in the data:



What actually predicts a setlist

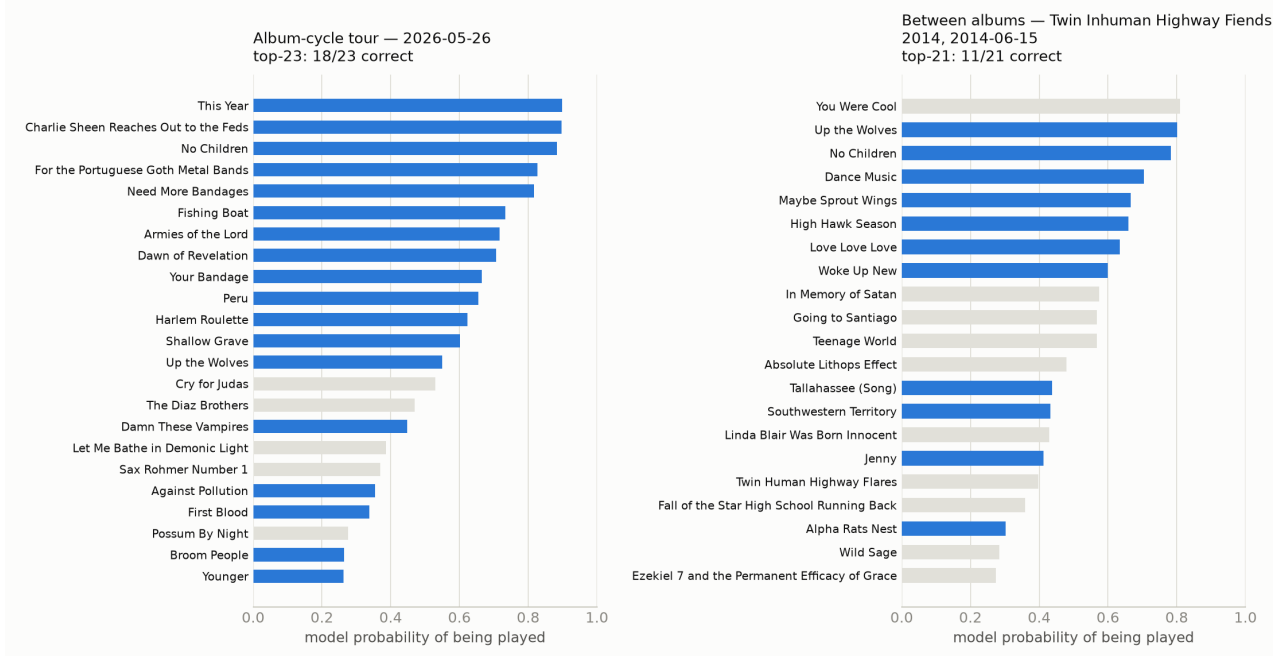


Recency dominates everything else combined — if a song was played a few shows ago, it's overwhelmingly likely to come back soon. That's the visible mechanism behind the tour "rotation": John Darnielle appears to work from a live pool of songs and cycle through it, which is exactly what *tour_rate* (play rate so far, this tour) picking up an independent signal on top of recency confirms. Two smaller, genuinely interesting effects: conditional on recency, brand-new material is actually less sticky than catalog average, and radio/festival/TV appearances reliably favor hits over deep cuts. Two more recent additions, both real but modest: a whole-show solo set slightly raises a song's odds, and a shrunk city-level play rate (see *Does geography matter?* below) picks up a small independent "local favorite" effect.

Is this just tracking album hype?

The 2026 example above is a new-album tour, and *new_material* is a real feature in the model — worth checking whether the strong recovery number is mostly "predict whatever's newest" in disguise. Among all 2014 tours with at least 8 shows, the one whose actually-played songs had the lowest share of new material — picked automatically, not by hand — is the Twin Inhuman Highway Fiends Tour 2014, squarely between Transcendental Youth (2012) and Beat the Champ (2015):

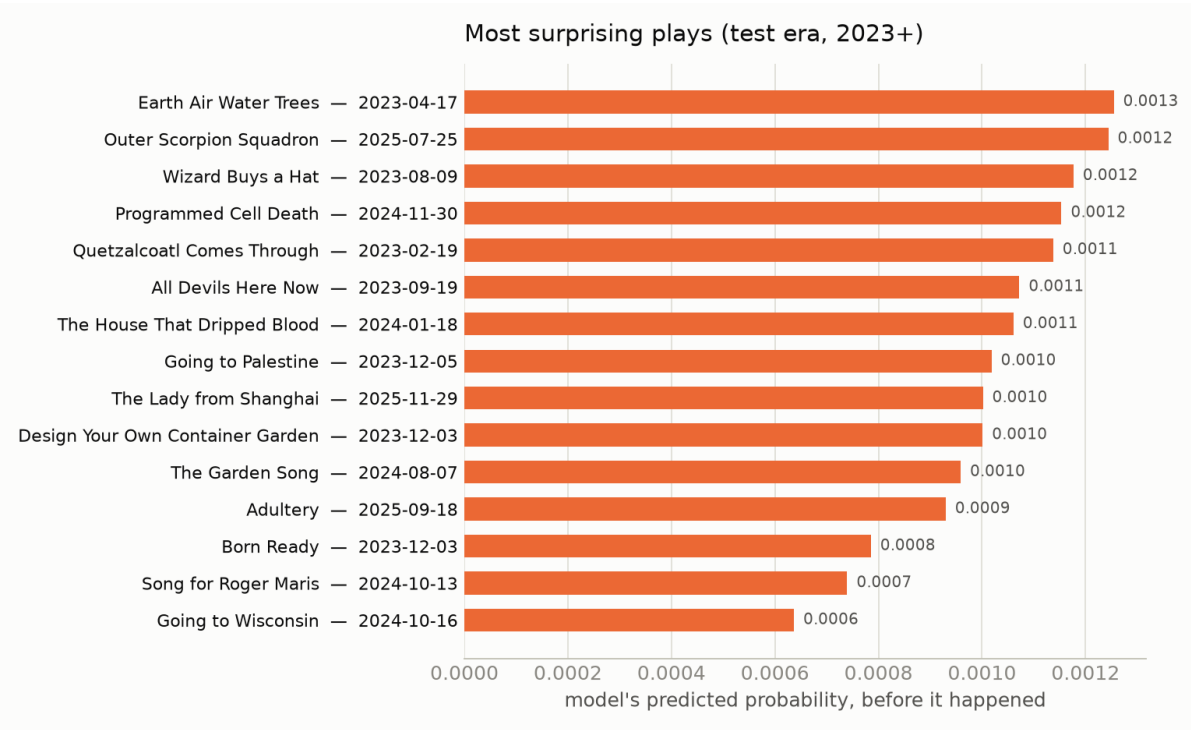
Same model, an album-cycle tour vs. a between-albums tour



The between-albums tour is visibly harder to predict — 55.1% top-n recovery tour-wide, versus 59.4% for the full 2023+ test era — despite 2014 being in-sample (which should make it look artificially easier, not harder). Without a record to promote, the setlist draws more evenly across a wider pool of similarly-loved catalog songs, so there's genuinely more entropy to predict. The model isn't just riding hype; if anything, album cycles are the easy case.

When the model gets surprised

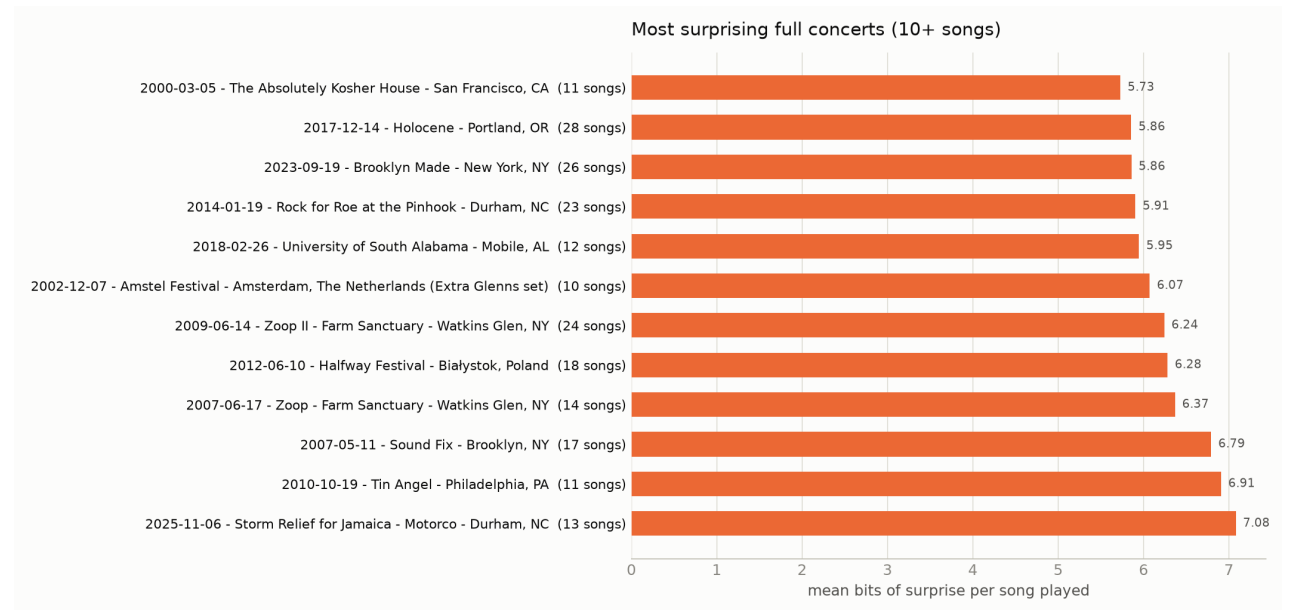
The flip side of a good model is a good list of its misses — nights the model gave a song almost no chance, and it got played anyway:



These are basically a machine-generated “rarities and one-offs” list: songs that came back from years of dormancy for a single night, often at a show with some specific occasion.

Most surprising concerts

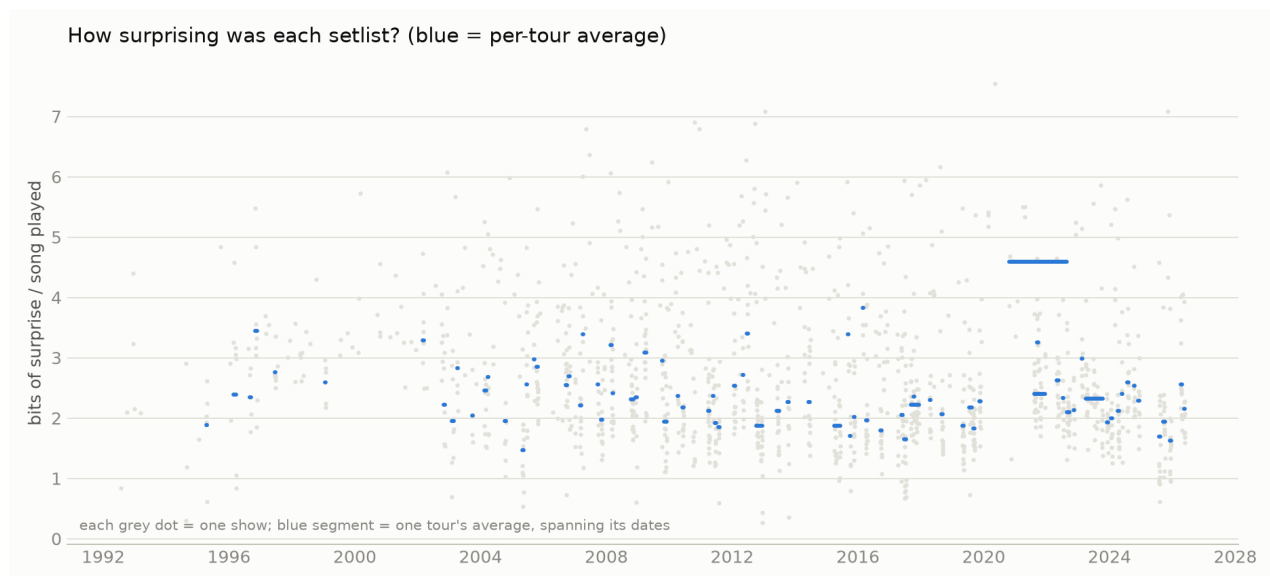
Individual surprising plays are one thing; whole surprising concerts are a different, more interesting question. Averaging each song's pre-show surprisal across a whole setlist answers this directly, filtered to real setlists (10+ songs) so a 1-song guest cameo can't trivially top the list:



These cluster hard around benefit shows, festivals, and untoured one-offs — a Hurricane Jamaica relief show, a “Rock for Roe” benefit, Merge Records' 35th-anniversary show, farm-sanctuary benefit gigs. Special-occasion shows pull deep cuts the regular tour rotation wouldn't produce, which is exactly what this metric should catch.

How surprising was each setlist as a whole?

Zooming out further: the same pre-show probabilities give a surprise score for an entire night — the average bits of information across the songs actually played. Aggregated per tour (touring is bursty, so a fixed show-count window saws at tour boundaries):



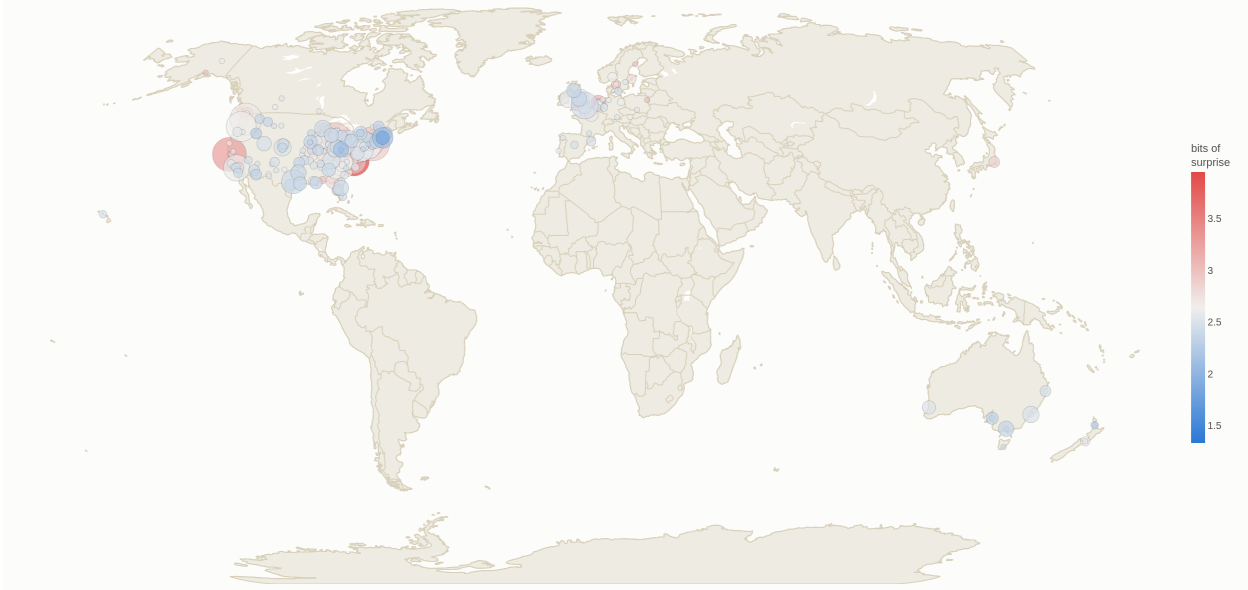
The tours with the highest average surprise are almost all solo, stripped-down tours — Winter Solo Tour 2024, the Ghost Cave Incubation Chamber solo non-tour, All Roads Lead to Lincoln Solo Mini-Tour — which is what motivated adding *is_solo* to the model above. Its coefficient came back small, though, meaning a whole show's unpredictability isn't fully explained by “solo” alone.

Does geography matter?

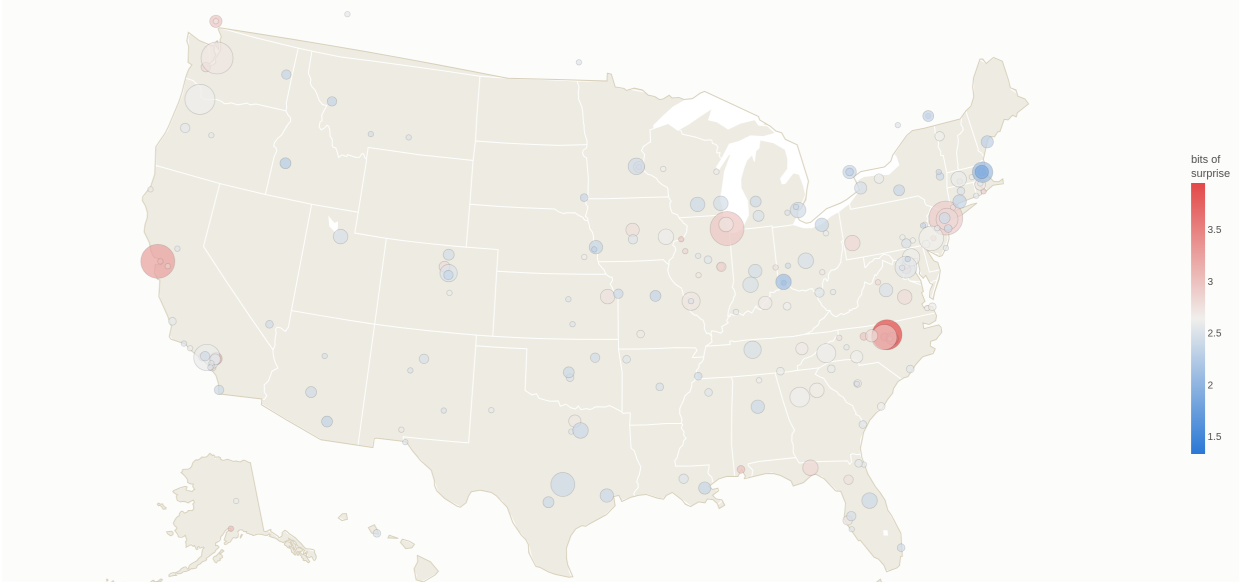
Two different questions hide inside “does geography matter”: where JD plays what (*city_song_rate*, already a feature inside the prediction model — genuinely predictable local patterns like a song that runs hot in one city) versus how surprising a place is even after the model's best effort (the residual left over once prediction has already used everything it knows, including geography). That residual is computed once per show and never revisited — what can legitimately be adjusted is how individual shows get aggregated into a per-city summary, since a two-show average is mostly noise. San Francisco, in particular, prompted the check: is its reputation real, or availability bias?

City-level data is thin (312 distinct cities, median 2 shows each), so each city's mean is empirical-Bayes shrunk toward a prior, weighted by its own sample size. The prior matters: administrative boundaries are a poor proxy for “shared local scene” — pooling every North Carolina city together would let a random small town 150 miles from Durham borrow Durham's specialness just because of a shared state line. The right unit is actual geographic proximity: cities within 25km of each other (real coordinates, DBSCAN + haversine distance) pool as one scene; anywhere with no real neighbor in the tour history stays fully on its own, shrinking straight to the global mean and weighted only by its own sample size — deliberately conservative, favoring under-grouping over erasing a genuinely distinct local signal.

Where the surprising shows happen (marker size = shows played)



Same data, continental US



San Francisco checks out: clustered with Oakland into a Bay Area scene, SF's own 57-show, 3.28-raw average barely moves (3.27 shrunk) — its own sample dominates — ranking 5th of 125 places with a real sample. But it's not the strongest signal: Durham, NC is #1 (clustered with Chapel Hill/Carrboro/Cary/Raleigh, 3.95 shrunk bits) — John Darnielle's home turf, tracking with hometown shows disproportionately including benefit gigs and extended sets, the same occasion that dominates the most-surprising-concerts list above. Notably, Pittsboro NC ranks #4 on its own (too far from Durham to share a cluster) and Watkins Glen NY is #2 entirely alone — all 4 of its shows are the "Zoop" Farm Sanctuary benefit concerts, a genuinely different kind of show that pooling with distant NY cities would have erased. This motivated adding city_song_rate to the prediction model itself, not just this descriptive analysis — its coefficient is real but small. Cities were geocoded against an offline database, not fabricated or hand-typed — an earlier fallback silently produced wrong coordinates for 74 towns (Pittsboro → Charlotte; Montreal → Toronto) by guessing "the biggest city in the state"; fixed to exact-match-only, honest coverage 91.6% of setlisted shows.

Where in the set does a song land?

Two related but distinct questions, both descriptive — neither feeds the prediction model, since encore/position status isn't known before a show happens. Encore rate is restricted to the 429 (of 1,317) shows whose Notes

prose actually spells out the encore breakout — most don't say, so an unrestricted average would be diluted by shows where the split is simply unknown, not shows without an encore. Within that subset, each song's rate is empirical-Bayes shrunk toward the global rate (21.6% of tracked plays are encore songs), restricted to songs with at least 10 tracked plays. Spent Gladiator 2 tops the list (89% — main-set exactly once in 60 tracked plays), followed by No Children (81%, the classic closer) and This Year (63%). At the other extreme, Moon Over Goldsboro, Clemency for the Wizard King, and Andrew Eldritch Is Moving Back to Leeds have never once been recorded as an encore in 50+ tracked plays each — genuinely 0%, and that's what the webapp shows; the shrunk estimate (around 2%, not quite 0) exists only to order the ranking sensibly, not to relabel a real, clean 0% as something else.

Set position is broader: defined for any performance where the show's full length is known, not just the encore-tracked subset, as $(seq-1)/(n_songs-1)$ — 0 = opener, 1 = closer, with encores naturally landing near 1. Same shrinkage, toward a 0.50 global mean. It lines up with encore rate where you'd expect (No Children averages 0.89; This Year, 0.86) but isn't redundant with it — a song can lean very late in the set without technically ever being the encore (The Best Ever Death Metal Band in Denton and Houseguest both land in the top 10 latest-leaning songs without cracking the encore-rate list at all). At the other end, Younger — 3% encore rate, 0.21 average position — is about as reliable an opener as this catalog has. The webapp's Song Explorer also shows a KDE density curve (boundary-corrected by reflecting the data across 0 and 1 before fitting, since a plain KDE badly underestimates density right at the edges) above the position marker, for any song with at least 20 position-eligible plays — useful for catching a song that's genuinely bimodal (sometimes an opener, sometimes a late surprise) rather than reliably mid-set.

Caveats

- The wiki is fan-maintained and lists most, not all, shows — coverage is noticeably better from the mid-2000s on.
- The model has no access to anything outside setlist history — no lyrics, no album themes, no explicit “songs currently being toured.” tour_rate and song_age are proxies for that.
- Song identity is deduplicated via wiki link slugs where available; a handful of never-linked rarities may still have unmerged spelling variants.
- is_solo is deliberately conservative, not exhaustive: it only fires on an explicit “solo show” note or a tour branded “Solo,” so it likely under-flags true solo shows the wiki didn't call out as exceptional — especially pre-2002, before a full-time backing band was the norm.
- City-level analyses only cover geocoded shows (91.6%) and are inherently thin for most cities (median 2 shows) — that's exactly what the shrinkage is for.
- The 25km metro-clustering radius is a judgment call, not a principled optimum — chosen deliberately conservative rather than tuned against ground truth.
- Encore rate only “sees” the 429 shows whose Notes explicitly spell out the breakout — a song that's always encored, but only at shows where nobody noted that, would be invisible to this measure entirely, not scored as 0%.

Reproducing this

```
PY=/Users/bdoughty/opt/miniconda3/envs/tmg-scrape/bin/python
$PY scrape.py fetch && $PY scrape.py build # refresh the raw data
$PY analyze.py # deep cuts, tour summaries
$PY predict.py # the setlist model + surprisal
$PY geocode_cities.py # offline city geocoding
```

```
$PY geography.py          # city/region surprisal, shrunk
$PY timeseries.py         # popularity-over-time series
$PY plot_report.py        # this report's figures
$PY plot_map.py           # the surprise maps
```

Full data dictionary in README.md; wiki-scraping gotchas and repo layout in CLAUDE.md.