What Predicts Whether ML Chemistry Claims Survive External Validation

Elliot Tower
elliot@elliottower.ai

Abstract

We test whether methodological features of ML-for-chemistry papers predict independently observable outcomes. We audited 133 claims across 13 method families and scored each on a five-tier scale from Proposed (benchmark only) to Validated (wet-lab confirmation). Of 55 claims with assignable tiers, 9 reached Validated and 6 were independently disconfirmed. A single feature dominates: whether the study tests against a real experimental endpoint. Claims with a real endpoint survive at 80%; claims without survive at 29% (OR = 10.0, Fisher $p = 0.0003$). In multivariate logistic regression, real endpoint is the only individually significant predictor ($p = 0.012$, pseudo-$R^2 = 0.31$).

To break the circularity of our features predicting our own tier assignments, we correlate raw features with three independently observable citation outcomes: scite.ai supporting/contradicting classifications, Semantic Scholar influential-citation fractions, and keyword-based citation-context analysis. `real_endpoint` predicts more supporting citations (scite: $\rho = +0.578$, $p = 0.008$; S2 influential: $\rho = +0.541$, $p = 0.014$). `pretrained_fm` predicts fewer supporting citations across both scite ($\rho = -0.648$, $p = 0.002$) and S2 keywords ($\rho = -0.532$, $p = 0.034$). The `pretrained_fm` effect is not confounded with institution type: it survives within academic-only papers ($\rho = -0.618$, $p = 0.011$). Total citation count does not predict validity ($\rho = 0.13$, $p = 0.57$); failed papers are cited more often, not less.

Claims published after 2022 survive at 39%, compared to 79% before 2022 ($p = 0.030$). Protein design (Family L) is the exception: 100% of entries test real endpoints, and the family contains 7 of 9 Validated claims.

1 Introduction

RFdiffusion generates protein binders confirmed by cryo-EM and binding assays **?**. Halicin, discovered by a message-passing neural network, cured pan-resistant infections in mice **?**. ProteinMPNN achieves 52.4% sequence recovery versus Rosetta's 32.9%, confirmed by X-ray crystallography **?**. These are real results with real endpoints.

They are also rare. We audited 133 ML-for-chemistry claims across 13 method families—from molecular property prediction and reaction planning to materials discovery and protein design—and found that most have never been tested outside their original benchmark. Of 55 claims with assignable tiers, 73% sit at Underdetermined or below: evaluated only on in-distribution benchmarks, with no independent test of the headline claim.

The standard response is that benchmarks are a starting point and validation will come later. For some claims, it has. For most, it has not. A-Lab claimed autonomous synthesis of 41 novel inorganic compounds at 71% success **?**. Independent reanalysis found nearly all were ordered versions of already-known disordered phases; the actual novelty rate is near zero **?**. DM21, a neural exchange-correlation functional, was claimed to generalize beyond its training chemistry **?**. Independent tests showed it fails on transition metals **?**. MIST reported a 30% improvement in MS/MS annotation; on an independent peer benchmark the improvement was 2% **?**.

Table 1: Validity tier distribution. Of 55 tiered claims, 26 survived (tier 1–2) and 29 did not.

Tier	Label	Criterion	Count
1	Validated	Real-endpoint confirmation	9
2	Causally Suggestive	External or OOD validation	17
3	Underdetermined	Benchmark only	20
4	Contested	Published dispute or vendor-only	3
5	Disconfirmed	Independent replication failed	6
—	Unscored	Insufficient information	78

These are the claims where independent testing happened to exist. The question is: given a new ML chemistry claim with no independent test yet, can we predict whether it would survive one?

We find that one feature dominates: whether the original study tests its claim against a real experimental endpoint—a wet-lab assay, an in-vivo experiment, a clinical measurement—rather than a computational benchmark. Claims with real endpoints survive at 80%. Claims without survive at 29%. The odds ratio is 10 ($p = 0.0003$). In multivariate logistic regression controlling for 9 features simultaneously, real endpoint remains the only significant predictor ($p = 0.012$).

Contributions.

- **The claim catalog (§2).** 133 claims across 13 families, scored on a five-tier scale with 14 binary methodological features per claim.
- **Feature-based prediction (§3).** Real endpoint predicts survival at OR = 10.0. A composite score separates outcomes monotonically ($r = +0.574$, $p < 10^{-4}$).
- **External validation (§4).** Raw features predict citation outcomes across three independent external sources, breaking the circularity of features predicting our own tiers. Institution-type confounding is tested and rejected.
- **The year trend (§5).** Post-2022 claims survive at half the rate of pre-2022 claims ($p = 0.030$).
- **Family profiles (§6).** Protein design has 100% real-endpoint testing and most Validated claims. Reaction prediction has 67% leakage flags. A cross-instrument transportability experiment on MS/MS data demonstrates the mechanistic basis of benchmark collapse.

2 The Claim Catalog

2.1 Scope and scoring

We cataloged 133 ML-for-chemistry claims from the published literature spanning 13 method families (Table 3). Each claim is the unit of analysis; a single paper may contain multiple claims scored independently.

Each claim receives a validity tier. Tier 1 (Validated) requires a physical-world measurement confirming the computational prediction. Tier 2 (Causally Suggestive) requires evaluation on data the model was not trained on, by an independent group, or on a genuinely out-of-distribution chemical space. Tier 3 (Underdetermined) is the default for claims evaluated only on their own benchmarks. Tier 5 (Disconfirmed) requires a published independent study reporting failure to replicate. Of 133 claims, 55 received tiers and 78 had insufficient information for assignment.

Table 2: Methodological features and prevalence across 133 claims.

Feature	Definition	Prevalence
`real_endpoint`	Tests against wet-lab, in-vivo, or clinical measurement	27%
`external_validation`	Evaluated on independent or external dataset	25%
`authored_limits`	Authors explicitly discuss limitations	33%
`strong_baseline`	Compares to competitive non-trivial baseline	18%
`multi_benchmark`	Evaluated on 3+ benchmarks	14%
`pretrained_fm`	Uses a pretrained foundation model	14%
`from_scratch`	Model trained from scratch (not fine-tuned)	14%
`leakage_flag`	Known or suspected data leakage	10%
`scaffold_or_split`	Uses scaffold or temporal split (not random)	8%
`vendor_claim`	Asserted in marketing, not peer review	3%
`code_released`	Source code publicly available	3%
`reports_ci`	Reports confidence intervals	1%

Table 3: The 13 method families.

Family	Label	Claims	Best tier
A	MS/MS & Metabolomics	8	2
B	Clinical Biomarker ML	4	1
C	Molecular Property Prediction	8	2
D	Benchmark Integrity	2	5
E	Reaction Prediction & Generation	6	2
F	Materials & Potentials	12	2
G	Spectral Similarity & NMR	5	2
H	Multi-omics Integration	10	2
I	LLM & NLP for Chemistry	11	1
J	Transportability & Federation	9	2
K	Docking & Structure	4	2
L	Protein Design & pLMs	11	1

2.2 Feature annotation

Each claim is annotated with 14 binary methodological features (Table 2). Two features stand out for their absence: `reports_ci` and `code_released` appear at 1% and 3% prevalence. Almost no ML chemistry papers report confidence intervals on their headline numbers.

2.3 Method families

3 What Predicts Survival

3.1 Univariate predictors

We define survival as reaching tier 1 or 2: the claim has been tested outside its original benchmark and the result held. Of 55 tiered claims, 26 survived and 29 did not. For each feature, we compute survival rates with and without the feature and test with Fisher's exact test.

Table 4: Univariate predictors of survival. Real endpoint produces a 10-to-1 odds ratio.

Feature	n	P(surv|feat)	P(surv|no feat)	OR	p
real_endpoint	**20**	**0.80**	**0.29**	**10.0**	**0.0003**
from_scratch	10	0.80	0.40	6.0	0.035
external_validation	18	0.61	0.41	2.3	0.25
strong_baseline	13	0.62	0.43	2.1	0.34
authored_limits	24	0.54	0.42	1.6	0.42
multi_benchmark	10	0.30	0.51	0.41	0.30
pretrained_fm	10	0.30	0.51	0.41	0.30
leakage_flag	7	0.14	0.52	0.15	0.10

One feature dominates (Table 4).

Claims with a real endpoint survive at 80%. Claims without survive at 29%. The second predictor is `from_scratch` (OR = 6.0, $p = 0.035$): purpose-built architectures outperform fine-tuned foundation models, partly because the same groups that train from scratch tend to run wet-lab validation.

Two anti-predictors: `multi_benchmark` and `pretrained_fm` both have OR = 0.41. Testing on more benchmarks and using a foundation model are uninformative at best.

3.2 Disconfirmation

Of 55 tiered claims, 6 were disconfirmed. No feature reaches significance at this sample size. The strongest signal is `real_endpoint`: zero of 20 claims with real endpoints were disconfirmed, versus 6 of 35 without ($p = 0.076$). Real endpoints predict survival and appear protective against disconfirmation.

3.3 Multivariate logistic regression

We fit logistic regression predicting survival from 9 features with sufficient cell counts. The model achieves pseudo-$R^2 = 0.309$ (AIC = 72.6, $n = 55$). `real_endpoint` is the only individually significant predictor (coefficient = +2.07, $p = 0.012$), corresponding to a 7.9× odds increase consistent with the univariate OR of 10.

3.4 Composite validity score

We define a composite score as validity $= \sum$protective $- \sum$risk, where protective $=$ {`real_endpoint`, `external_validation`, `strong_baseline`, `scaffold_or_split`, `authored_limits`, `from_scratch`} and risk = {`leakage_flag`, `vendor_claim`, `strawman_flag`}.

The composite score predicts survival monotonically (Spearman $r = +0.574$, $p < 10^{-4}$) and separates outcomes at least as cleanly as the multivariate model, requiring no fitting.

4 External Validation

4.1 Breaking the circularity

The analysis in §3 is internally consistent: our features predict our tiers. Both were assigned by the same team. The question is whether the features capture something real about claim quality or reflect our scoring judgments. To test this, we correlate raw features with external citation outcomes that we did not construct.

Table 5: Composite validity score versus outcome. At score 0, one-third are disconfirmed and none survive. At score 3, 82% survive and none are disconfirmed.

Score	n	Survival rate	Disconfirmation rate
-1	3	0%	0%
0	9	0%	33%
1	15	40%	13%
2	16	62%	6%
3	11	82%	0%
4	1	100%	0%

Table 6: Features predicting external citation outcomes. `real_endpoint` replicates across all three sources. `pretrained_fm` anti-predicts across two.

Feature	Outcome source	ρ	p	Direction
`real_endpoint`	Scite: supporting fraction	$+0.578$	0.008	More supporting
	S2: influential citations	$+0.541$	0.014	More influential
	S2 keywords: contradict fraction	-0.496	0.051	Fewer contradictions
	S2 keywords: support ratio	$+0.482$	0.058	Higher support ratio
`pretrained_fm`	Scite: supporting fraction	-0.648	0.002	Fewer supporting
	S2 keywords: support fraction	-0.532	0.034	Fewer supporting
	Scite: contradicting fraction	-0.442	0.051	Both fewer
	S2 keywords: support ratio	-0.469	0.067	Lower ratio

Three data sources were queried programmatically for papers with DOIs ($n = 20$–24 depending on source): (1) scite.ai tallies, which classify each citation as supporting, contradicting, or mentioning; (2) Semantic Scholar influential-citation fractions; (3) a mechanical keyword search over citation contexts, counting confirmation terms ("confirmed," "validated," "reproduced") and contradiction terms ("failed to," "contradicts," "data leakage," "overfitting"). No LLM judgment was used in any source.

`real_endpoint` replicates across all three independent sources: papers that test against physical endpoints accumulate more supporting citations (scite: $\rho = +0.578$, $p = 0.008$), more influential citations (S2: $\rho = +0.541$, $p = 0.014$), and a higher support-to-contradiction ratio in citation contexts ($\rho = +0.482$, $p = 0.058$). `pretrained_fm` shows the inverse: fewer supporting citations across both scite ($\rho = -0.648$, $p = 0.002$) and keyword search ($\rho = -0.532$, $p = 0.034$).

The composite tier does not predict external outcomes as consistently as the raw features (tier vs. scite supporting: $\rho = -0.32$, $p = 0.19$; tier vs. S2 support ratio: $\rho = -0.57$, $p = 0.035$). The raw features are more consistently predictive than the composite judgment.

Total citation count does not predict validity ($\rho = 0.13$, $p = 0.57$). Papers that fail external scrutiny are cited more often, not less (not-survived mean velocity: 1216 citations/year; survived: 175). The features that predict citation *quality* differ from those that predict citation *quantity*.

4.2 Confound test: institution type

An obvious alternative: foundation-model papers originate disproportionately from large technology companies whose publication incentives differ from academic labs. We tested this using author institution data

from OpenAlex ($n = 28$). The confound does not hold. `pretrained_fm` does not correlate with big-tech affiliation ($\rho = -0.105$, $p = 0.66$): six of seven foundation-model papers come from academic labs. Big-tech affiliation shows no relationship with supporting-citation fraction ($\rho = -0.033$, $p = 0.89$). Within the 16 academic-only papers, `pretrained_fm` still predicts fewer supporting citations ($\rho = -0.618$, $p = 0.011$). The effect is about foundation models, not about who built them.

4.3 Case studies

RFdiffusion, ProteinMPNN, Halicin, and Minibinders—all tier 1, all with real endpoints—accumulate 3–17 supporting keyword contexts and 0–2 contradicting. A-Lab (tier 5, no real endpoint) has 2 supporting and 1 contradicting. AlphaFold3 (tier 3, no real endpoint) has 9 supporting but also 3 contradicting, including citations flagging memorization and data leakage.

5 The Year Trend

Newer claims survive less often. Among 47 tiered claims with publication years, Spearman correlation between year and tier is $r = +0.317$ ($p = 0.030$): later publication predicts worse validity tier.

Claims published before 2022 survive at 79% ($n = 14$). Claims published 2022 or later survive at 39% ($n = 33$). Right-censoring is unlikely to explain this: the tiering criterion is whether external evidence exists at all, and pre-2022 claims had more time to accumulate disconfirmations, yet survive at double the rate.

The individual methodological features do not explain the trend. `real_endpoint`, `external_validation`, and `scaffold_or_split` show no significant change over time ($p > 0.5$ for all three). The year effect reflects a broader shift in what gets published, beyond any single measurable feature.

6 Family Profiles

6.1 Protein design: the success story

Family L (Protein Design & pLMs) contains 11 claims, 7 at Validated. Every entry tests a real endpoint (100% prevalence). The family includes RFdiffusion (picomolar binders confirmed by cryo-EM), ProteinMPNN (52.4% sequence recovery, X-ray confirmed), halicin (in-vivo mouse cure), Novokines (functional signaling in human PBMCs), and de novo luciferase design (LuxSit: functional enzyme from scratch). Protein design is hard. The field requires wet-lab confirmation before a claim is considered established, and it works.

6.2 Reaction prediction: the cautionary case

Family E (Reaction Prediction & Generation) contains 6 claims, none above Causally Suggestive. The family has 67% prevalence of leakage flags and 67% strawman flags—the highest of any family. The Molecular Transformer reports >90% accuracy on USPTO, but independent analysis showed test-set contamination reaching 99.97%; removing the leakage dropped accuracy by 4 percentage points Bran & Schwaller (2026). No claim in this family has been tested on a genuinely novel reaction class.

6.3 Spectral ML: benchmark collapse across instruments

Family A (Spectral ML) contains 5 tiered entries (MassFormer, MIST, 3DMolMS, CASMI benchmark, MassGenie). Only one tests a real endpoint (20%). Three use pretrained foundation models (60%). The mean tier is 2.6, and 40% reach tier 1–2. The family is the mirror image of protein design on both key features.

A cross-instrument transportability experiment on MS/MS spectral data provides a mechanistic account of benchmark failure in this family. We computed binary fingerprint embeddings (the standard 2048-dimensional representation used to benchmark MS/MS models) for 2,140 MassBank spectra across 10 instru-

Table 7: Spearman ρ between each distance metric and cross-instrument degradation ($n = 45$ pairs, binary fingerprint embeddings). Five of six metrics reach $p < 0.05$ against match degradation.

	Clf degradation		Match degradation	
Metric	ρ [95% CI]	p	ρ [95% CI]	p
Geodesic (k=10)	−0.149 [−.44, +.18]	.329	−0.040 [−.35, +.25]	.793
Centroid distance	−0.217 [−.54, +.11]	.153	+0.417 [+.10, +.66]	.004
Domain classifier AUC	−0.302 [−.62, +.04]	.044	+0.508 [+.20, +.72]	<.001
MMD (RBF)	−0.217 [−.54, +.11]	.153	+0.417 [+.10, +.66]	.004
Sliced Wasserstein	−0.247 [−.56, +.09]	.101	+0.422 [+.09, +.66]	.004
Proxy $\mathcal{A}$-distance	−0.303 [−.61, +.02]	.043	+0.317 [+.01, +.57]	.034

Table 8: Same analysis excluding 12 pairs with match degradation = 1.0 (EI-B instruments sharing no compounds with partner). $n = 33$. No metric reaches significance.

	Clf degradation		Match degradation	
Metric	ρ	p	ρ	p
Geodesic (k=10)	−0.154	.391	−0.297	.094
Centroid distance	+0.028	.875	−0.162	.367
Domain classifier AUC	−0.075	.680	−0.038	.832
MMD (RBF)	+0.028	.875	−0.162	.367
Sliced Wasserstein	+0.001	.994	−0.142	.432
Proxy $\mathcal{A}$-distance	−0.090	.619	−0.093	.606

ment types spanning three ionization families (ESI: 7 instruments, EI: 2, MALDI: 1), yielding 45 pairwise instrument comparisons. For each pair we measured two degradation targets and six distance metrics.

Degradation targets. Classifier degradation measures the drop in logistic-regression AUC when moving from internal CV to cross-instrument prediction. Spectral matching degradation measures the loss in nearest-neighbor hit@1 accuracy when searching across instruments versus within instrument, restricted to compounds appearing in both.

Distance metrics. We tested six standard metrics from the domain-adaptation literature: geodesic distance on the Grassmannian $\mathrm{Gr}(10, 2048)$, centroid distance, domain classifier AUC, MMD with RBF kernel Gretton et al. (2012), Sliced Wasserstein distance, and proxy $\mathcal{A}$-distance Ben-David et al. (2007). Table 7 reports Spearman correlations with bootstrap 95% CIs.

At first glance, five of six metrics significantly predict match degradation ($\rho = 0.32$–0.51, $p < 0.05$), with domain classifier AUC the strongest ($\rho = +0.508$, $p < 0.001$). Geodesic distance on the Grassmannian predicts neither outcome.

Confound: ionization family. Twelve of the 45 pairs involve EI-B instruments, which share zero compounds with ESI instruments. These pairs have match degradation = 1.0 by construction: no shared compounds means hit@1 drops to zero. Removing these 12 fallback pairs collapses all correlations (Table 8).

Within the ESI family alone (21 pairs, the largest same-ionization group), no metric reaches even $p < 0.25$. The apparent signal in Table 7 is driven by the ionization-family boundary—a coarse binary distinction—rather than gradual distributional differences that a metric could usefully rank. The geodesic distances

between all 45 pairs span 4.22–4.80 out of a theoretical maximum of ∼4.97, a dynamic range of 11.8%: in fingerprint space, every instrument pair looks near-maximally different regardless of actual transferability.

This result is a concrete instance of the paper's central finding. The dominant failure mode in spectral ML (Family A) is benchmark collapse driven by instrument confound (D2). The transportability experiment shows *why* it happens: the standard benchmark embedding does not encode enough spectral structure for any distance metric—geometric, distributional, or discriminative—to track within-family performance loss. The metrics detect whether two instruments use the same ionization method, but this is a property of the chemistry, not a learned representation. The confound in Table 7 mirrors the confound in the broader catalog: apparent predictive power collapses once the structural confounder is controlled.

Foundation models such as LSM-MS2 Asher et al. (2025) and DreaMS Bushuiev et al. (2025) train transformer encoders via masked spectral-peak reconstruction on millions of MS/MS spectra. Whether their learned embeddings produce distances that predict within-family degradation—the specific failure that fingerprint embeddings cannot resolve—remains untested.

6.4 Materials and potentials: fragmentation under scrutiny

Family F (Materials & Potentials) has the most entries (12) and the highest evidence diversity. The family is fragmented: the DM21 neural functional (claimed, debated, disconfirmed in three sequential papers) coexists with CG force-field transferability claims (accurate in-distribution, unstable over long trajectories) and the disciplined alternative DeePKS (honest about scope limitations). More evidence has not produced convergence.

7 Discussion

7.1 The real-endpoint gap

ML chemistry claims that test against physical reality survive. Claims that stay on benchmarks mostly do not. The 10-to-1 odds ratio is the largest effect in the dataset, survives multivariate control, and predicts independently observable citation outcomes across three external sources. The protein design community has closed this gap. The reaction prediction community has not. The difference is disciplinary norms about what counts as evidence.

7.2 What does not predict survival

Two anti-predictors: testing on more benchmarks (OR = 0.41) and using a pretrained foundation model (OR = 0.41). The foundation-model effect is confirmed externally (fewer supporting citations) and is not confounded with institution type (survives within academic-only papers, $\rho = -0.618$, $p = 0.011$). Computational sophistication and the number of evaluations are uninformative at best.

7.3 Citations measure influence, not quality

Total citation count does not predict validity ($\rho = 0.13$, $p = 0.57$). Papers that fail external scrutiny are cited more often than papers that survive (1216 vs. 175 citations/year). High citation counts reflect influence, controversy, and benchmark ubiquity—not correctness.

7.4 The year trend as a field diagnostic

The halving of survival rates after 2022 tracks the scaling of foundation models and generative AI in chemistry. Publication volume has roughly doubled while experimental validation per claim has dropped. The methodological features do not explain the trend: the year effect is not reducible to a decline in any single measurable practice.

8 Limitations

The catalog has uneven coverage across families. Tier assignments involve judgment, though we applied strict criteria and documented edge cases. The composite score weights all protective features equally. The year trend may partly reflect ascertainment bias.

The external validation covers 20–24 papers with DOIs, a subset of the catalog. The keyword search is mechanical: terms such as "overfitting" can appear unrelated to the cited paper's validity. Scite.ai's contradicting classification is conservative: A-Lab and DM21 both show zero contradicting citations despite published rebuttals. These limitations bias toward the null, making the significant correlations more notable.

The principal sample-size limitation is not the 133-entry catalog but the 55 tiered claims. We base central claims on rank-based tests (Fisher's exact, Spearman) that remain valid at small n, not on logistic regression point estimates alone.

9 Conclusion

One number summarizes the state of ML for chemistry: 10. That is the odds ratio for survival when a claim tests against a real experimental endpoint versus when it does not. The same feature predicts independently observable citation outcomes across three external data sources, breaking the circularity of features predicting our own tiers. Foundation-model papers accumulate fewer supporting citations. Highly cited papers are not higher-quality papers. The field's best-validated results come from protein design, where wet-lab testing is the norm. Its least-validated come from reaction prediction, where benchmark leakage is the norm. The difference is cultural. Newer claims survive at half the rate of older ones.

The odds ratio says the rest: test claims against reality.

References

Gabriel Asher, Devesh Shah, Amy A. Caudy, et al. Lsm-ms2: A foundation model bridging spectral identification and biological interpretation. *arXiv*, 2025. arXiv:2510.26715.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007.

Andres M. Bran and Philippe Schwaller. Syncat: Synthesis-aware contamination testing for reaction prediction, 2026. Preprint.

Roman Bushuiev, Anton Bushuiev, et al. Learning from unstructured mass spectrometry data with self-supervised transformers. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02663-3.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.