When Does a Predictable Residual Become a Scientific Correction? A Residual-Transport Audit for AI-Assisted Scientific Discovery

Anonymous authors
Paper under double-blind review

Abstract

Automated scientific-discovery systems can make the residuals of established models predictable, but predictive gain alone does not identify what has been found: a missing scientific contribution, a regime-specific offset, or structured noise. We formulate this downstream problem as *correction acceptance* and introduce a residual-transport audit that evaluates frozen corrections across scientifically meaningful boundaries. The central observation is that held-out-regime performance mixes two forms of evidence: improvement relative to information available before the regime is observed, and explanation of variation within that regime after its own residual level is accounted for. We derive the resulting decomposition and build it into a leakage-safe audit with frozen candidate selection, regime-level replication, uncertainty, and a measurement-noise stopping rule. Across thermodynamics, nuclear physics, mammalian metabolism, cosmology, and a supplementary asset-pricing control, the audit separates corrections that transport from patterns that merely locate regimes. A known equation-of-state correction transports across unseen fluids ($R^2_{\text{pool}} = 0.993$), and a simple correction to Geiger–Nuttall predicts nuclides discovered decades later ($R^2 = 0.635$). In contrast, mammalian temperature is repeatedly selected yet fails within unseen taxonomic orders, while COBE/FIRAS requires no further correction. The contribution is an evaluation layer for AI-assisted science: candidate generation may be flexible, but scientific acceptance should be indexed by the environments in which the corrected model is expected to hold.

1 Introduction

Machine learning has made scientific model proposal inexpensive. Symbolic regression (Schmidt & Lipson, 2009), sparse system identification (Brunton et al., 2016), and physics-informed learning (Karniadakis et al., 2021) search large spaces of equations and interactions with little human effort, and such systems are increasingly deployed as scientific hypothesis generators (Wang et al., 2023). Section 3 discusses this literature in detail. Given the residuals of an established model, such systems can often find something that predicts them. The difficult question comes next: *what kind of scientific evidence does that prediction constitute?*

A predictable residual may indicate a missing physical contribution. It may instead encode differences between fluids, species, instruments, historical periods, or other regimes. A flexible approximation can also exploit support-specific structure that vanishes when the relevant scientific boundary is crossed. These possibilities matter because modifying a scientific representation is a stronger act than improving test error. A proposed correction implicitly makes a transport claim: the term should continue to explain the discrepancy in the environments in which the corrected model is expected to hold.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We call this downstream task *correction acceptance*. It begins after a candidate generator has proposed one or more corrections to a reference-model residual. The generator may be symbolic regression, sparse modelling, a neural system, or a domain-specified candidate family. The acceptance layer asks whether the candidate has survived enough boundary-specific falsification to justify scientific interpretation. This separates our goal from equation discovery, which proposes mathematical structure, and model-discrepancy estimation, which characterizes what an existing model leaves unexplained (Kennedy & O'Hagan, 2001).

The key failure mode is subtle. Random splits allow a model to exploit structure shared between training and test observations (Geirhos et al., 2020), so holding out complete regimes is the stronger protocol when data carry temporal, physical, phylogenetic, or instrumental structure (Roberts et al., 2017). Yet a frozen correction can still score well on an unseen regime simply by predicting that regime's average residual. It may fail to explain any relationship among observations *within* the regime. The classical within–between distinction in fixed-effects modelling makes this algebra explicit (Mundlak, 1978); we use it for a different purpose: interpreting what a frozen machine-learned correction has actually generalized across.

Our contributions are:

- We formulate **correction acceptance** as a distinct evaluation problem for AI-assisted scientific discovery: candidate generation and scientific acceptance are separate stages.
- We derive an **offset decomposition** showing how prospective held-out-regime R^2 can be positive because of between-regime offsets even when the same frozen correction fails within the unseen regime.
- We turn this distinction into a **residual-transport audit**: all data-dependent choices are frozen outside the target regime; evidence is aggregated across scientifically meaningful regimes; uncertainty, stability, and measurement-noise stopping are explicit.
- We demonstrate the resulting decision changes across **real scientific boundaries**: unseen fluids, future nuclear discoveries, unseen elements, unseen taxonomic orders, and measurement noise, with a supplementary historical asset-pricing recovery control.

The intended claim is deliberately narrower than "transport proves mechanism." Passing the audit does not establish causality, theoretical uniqueness, or physical necessity. It establishes that a candidate has survived the boundary implied by the scientific claim well enough to merit domain-level interpretation.

2 Residual-Transport Audit

2.1 Correction acceptance as a post-generation task

Let f_0 be a reference scientific model and

$$r_i = y_i - f_0(x_i) \tag{1}$$

its residual for observation i. Each observation belongs to a prespecified regime $g_i \in \{1, \ldots, G\}$ chosen to match the intended scientific generalization claim—for example, fluid identity, taxonomic order, chemical element, or historical period. A candidate generator receives training-regime data and proposes a correction $h(x)$ to the residual.

For each held-out regime g, the reference model, feature transformations, candidate selection, scaling, and correction coefficients are learned only from the other regimes. The resulting prediction $\hat{r}_i^{(-g)}$ is therefore frozen before any observation from regime g is used for model construction. Figure 1 summarizes the separation between generation and acceptance.

2.2 Prospective and within-regime evidence

For held-out regime g, define

$$E_g = \sum_{i \in g} \left(r_i - \hat{r}_i^{(-g)} \right)^2 . \tag{2}$$

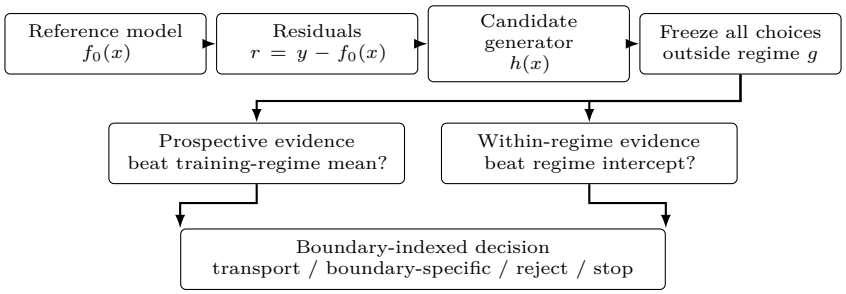


Figure 1: **Correction acceptance is downstream of candidate generation.** A candidate is learned without access to the target regime, then evaluated twice using the same frozen predictions: prospectively relative to information available before the regime was observed, and retrospectively relative to the regime's own residual level. The audit does not certify mechanism; it tests whether the correction survives the declared scientific boundary.

A natural prospective baseline is the mean residual in the training regimes, $\bar{r}_{-g}$:

$$R^2_{g,\text{train}} = 1 - \frac{E_g}{\sum_{i \in g}(r_i - \bar{r}_{-g})^2}. \tag{3}$$

This asks whether the correction improves on information available before regime g is observed. Let

$$W_g = \sum_{i \in g}(r_i - \bar{r}_g)^2, \qquad B_g = n_g(\bar{r}_g - \bar{r}_{-g})^2. \tag{4}$$

Here W_g is variation within the unseen regime and B_g is the offset between its mean residual and the training-regime mean.

Proposition 1 (Offset decomposition)**.** *For every held-out regime,*

$$\sum_{i \in g}(r_i - \bar{r}_{-g})^2 = W_g + B_g, \tag{5}$$

and therefore

$$R^2_{g,\text{train}} = 1 - \frac{E_g}{W_g + B_g}. \tag{6}$$

The same frozen predictions can also be evaluated retrospectively against the best intercept-only description of the observed regime:

$$R^2_{g,\text{within}} = 1 - \frac{E_g}{W_g}. \tag{7}$$

$R^2_{g,\text{within}}$ is not a deployable forecast baseline because $\bar{r}_g$ is unknown before the regime is observed. It is an audit diagnostic: once the regime exists, did the proposed correction explain more than the fact that this regime has its own residual level?

Corollary 1 (Prospective inflation by regime offset)**.** *Whenever $W_g > 0$,*

$$R^2_{g,\text{train}} - R^2_{g,\text{within}} = \frac{E_g B_g}{W_g(W_g + B_g)} \geq 0. \tag{8}$$

The gap grows with the between-regime offset B_g.

Thus a positive regime-holdout result can mean two different things: a common correction explains structure inside the unseen regime, or the correction mainly locates that regime's average residual. For scientific claims that posit a common relationship operating inside new regimes, we require evidence from both views.

2.3 Operational audit

The audit concatenates leakage-safe predictions across outer regime folds and reports

$$R^2_{\text{pool}} = 1 - \frac{\sum_g E_g}{\sum_g (W_g + B_g)}, \tag{9}$$

$$\overline{R^2}_{\text{group}} = \frac{1}{G} \sum_g \left(1 - \frac{E_g}{W_g + B_g}\right), \tag{10}$$

$$\overline{R^2}_{\text{within}} = \frac{1}{G} \sum_g \left(1 - \frac{E_g}{W_g}\right). \tag{11}$$

The primary analyses use complementary-pairs stability selection on training regimes, followed by OLS or ridge refitting of selected reportable terms. The generic proposal mode automatically expands candidate variables into linear, squared, cubic, signed-square-root, log-absolute, and pairwise-interaction terms. This proposal engine is intentionally simple: the contribution is the acceptance layer, not a new discovery algorithm.

For a reportable "transport" decision, the locked protocol requires a stable interpretable term, $R^2_{\text{pool}} \geq 0.01$, $\overline{R^2}_{\text{group}} \geq 0$, $\overline{R^2}_{\text{within}} \geq 0$, a positive lower 95% cluster-bootstrap bound for R^2_{pool}, and positive prospective scores in at least two-thirds of regimes. These thresholds are a prespecified operationalization rather than universal constants; the central sign reversals and boundary comparisons below do not depend on tuning them. If reported measurement uncertainty already explains the discrepancy, the audit stops before proposing another correction. Full selection and uncertainty procedures are in Appendix A.

3 Related Work

Equation discovery and scientific ML. Schmidt & Lipson (2009) recover free-form natural laws from experimental data, and Udrescu & Tegmark (2020) and Cranmer (2023) add physics-inspired priors and scalable search to symbolic regression. Sparse identification instead selects governing terms from a fixed candidate library, for dynamical systems (Brunton et al., 2016) and for partial differential equations (Rudy et al., 2017), while physics-informed learning embeds known structure in the model itself (Karniadakis et al., 2021). Cornelio et al. (2023) additionally require a discovered expression to be derivable from background theory, and La Cava et al. (2021) benchmark such generators against one another. This work is all upstream of our question: these systems *propose* candidates, and we ask what evidence a proposal must survive before it is treated as a scientific correction.

OOD generalization and environment structure. Environment structure is central to work on distribution shift (Quiñonero-Candela et al., 2009): Peters et al. (2016) exploit invariance across environments for causal identification, Arjovsky et al. (2019) turn that invariance into a training objective, and Koh et al. (2021) assemble benchmarks of real-world shift. Bareinboim & Pearl (2016) formalize when a conclusion transports from one population to another. Our distinction is orthogonal to algorithmic invariance: even a genuinely frozen model can achieve positive environment-level performance because its error baseline contains a between-regime offset. We therefore separate prospective usefulness from within-regime explanatory transport.

Within–between decomposition. Mundlak (1978) showed that adding group means to a regression separates within- from between-group effects; Neuhaus & Kalbfleisch (1998) characterize when the two coefficients diverge, and Enders & Tofighi (2007) and Bell & Jones (2015) develop the practice in multilevel modelling. We do not claim this statistical decomposition as new. The contribution is to use it as an acceptance criterion for frozen corrections proposed to modify an existing scientific representation, paired with leakage-safe selection and boundary-indexed evaluation.

Analysis	Boundary	Pros. R^2	Within R^2	Verdict
EOS physical basis	unseen fluid	0.993	0.993	transport
α direct basis	future discovery	0.635	0.602	temporal transport
α direct basis	unseen element	0.379	−0.681	reject universal element transport
BMR temperature	unseen order	0.074	−0.191	reject universal order transport
BMR, AnAge	unseen order	0.176	−1.008	replicated failure
BMR, Capellini–AnAge	unseen order	0.253	−0.015	not accepted
COBE/FIRAS	noise floor	–	–	stop ($\chi^2_\nu = 1.074$)
CAPM control	unseen decade	0.296	0.247	known-factor recovery

Table 1: **Residual transport produces boundary-specific decisions.** Values are generated from the locked analysis outputs. The temporal α test uses a frozen historical split rather than a regime-bootstrap average. CAPM is a supplementary positive control, not a novel scientific claim.

4 Experimental Design

We test the audit on domains chosen to span distinct acceptance outcomes rather than to form a leaderboard. Each boundary corresponds to a different scientific claim.

Equation of state (EOS). The reference is ideal-gas compressibility $Z = 1$. A controlled physical candidate uses reduced density ρ_r and ρ_r/T_r, motivated by low-density van der Waals structure (van der Waals, 1873). Fluid identity is held out. The primary support contains 900 points from ten fluids across six reduced temperatures; broader-support and common-temperature sensitivities are also reported.

Nuclear α decay. The baseline Geiger–Nuttall relation predicts $\log_{10} t_{1/2}$ from $Z/\sqrt{Q_\alpha}$ (Geiger & Nuttall, 1911), evaluated on nuclide properties and atomic masses from NUBASE2020 and AME2020 (Kondev et al., 2021; Wang et al., 2021). Direct correction candidates are Z and neutron number N. The primary temporal test freezes the entire pipeline on 149 nuclides discovered by 1970 and evaluates 94 later discoveries; rolling-origin refreezes test 1950, 1970, and 1990 cutoffs. A separate leave-one-element-out analysis changes the transport boundary while keeping the scientific system otherwise fixed.

Mammalian metabolism. The baseline is a training-fold log–log body-mass model of basal metabolic rate. The proposed correction is inverse body temperature $1/(k_B T)$, the Boltzmann form implied by metabolic theory (Gillooly et al., 2001). It is evaluated across unseen taxonomic orders in three independently assembled compilations: White–Seymour (White & Seymour, 2003), $n = 461$ in nine orders; AnAge (Tacutu et al., 2018; de Magalhães et al., 2024), $n = 318$ in nine orders; and Capellini–AnAge (Capellini et al., 2010), $n = 195$ in six orders.

COBE/FIRAS and CAPM controls. FIRAS tests whether a one-parameter Planck spectrum leaves discrepancy above the reported measurement-noise floor (Mather et al., 1994; Fixsen et al., 1996). A supplementary CAPM analysis asks whether known SMB/HML/volatility-related omitted structure survives held-out historical decades (Sharpe, 1964; Fama & French, 1993). Complete filtering, candidate construction, and data provenance are given in Appendix A.

5 Results

5.1 The same predictive signal can imply different scientific decisions

Table 1 summarizes the locked outcomes. The table is deliberately a decision matrix, not a model ranking: the relevant boundary changes what generalization means.

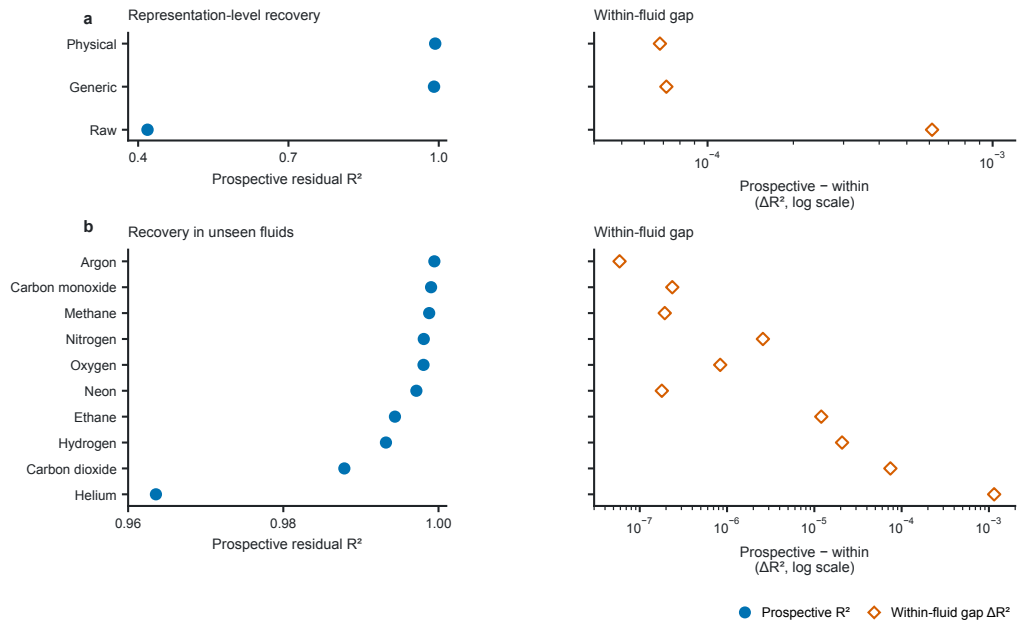


Figure 2: **EOS recovery across unseen fluids.** Prospective and within-fluid scores are nearly identical for the physical basis, so the held-out gain reflects structure inside each fluid rather than primarily fluid-level offsets.

5.2 Known equation-of-state structure transports across unseen fluids

The EOS analysis is a controlled recovery. On 900 state points from ten fluids, the physical basis is selected in every outer fold and obtains $R^2_{\text{pool}} = 0.9929$, $\overline{R^2}_{\text{group}} = 0.9930$, and $\overline{R^2}_{\text{within}} = 0.9928$, with a cluster-bootstrap interval $[0.9856, 0.9977]$. Every held-out fluid is positive (Figure 2). A fixed-intercept comparator yields the expected coefficient signs: positive for ρ_r (0.741, cluster-robust $P = 1.1 \times 10^{-6}$) and negative for ρ_r/T_r (-1.841, $P = 1.7 \times 10^{-9}$).

The result persists on the full available support (1,155 observations from sixteen fluids): $R^2_{\text{pool}} = 0.9917$ and $\overline{R^2}_{\text{within}} = 0.9896$, with all fluids positive. At the single common reduced temperature $T_r = 1.05$, the two physical columns are exactly collinear, so the locked sensitivity fits density alone and still obtains $\overline{R^2}_{\text{within}} = 0.9891$. A generic nonlinear library is also strong ($\overline{R^2}_{\text{within}} = 0.9872$ on full support), preventing an overclaim: the supplied basis transports and is physically interpretable, but transport alone does not establish a unique mathematical representation.

5.3 A simple nuclear correction survives discovery time but not element identity

Freezing the Geiger–Nuttall baseline and correction pipeline on 149 nuclides discovered by 1970 leaves 94 later nuclides untouched. Stability selection retains Z and achieves future $R^2 = 0.6348$ relative to the pre-1970 residual mean and 0.6020 relative to the post-1970 mean, with RMSE 0.839 $\log_{10}$ seconds. About 21% of future feature values lie outside at least one pre-1970 training range. The retained term is consistent with the broad form of Viola–Seaborg systematics, in which nuclear charge enters beyond the simplest Geiger–Nuttall expression (Viola & Seaborg, 1966); we therefore treat it not as a newly discovered nuclear law but as a known-form correction whose temporal transport can be tested. Rolling-origin refreezes at 1950, 1970, and 1990 remain positive: prospective $R^2 = 0.469, 0.678, 0.674$ and within-period $0.427, 0.604, 0.673$ (Figure 3).

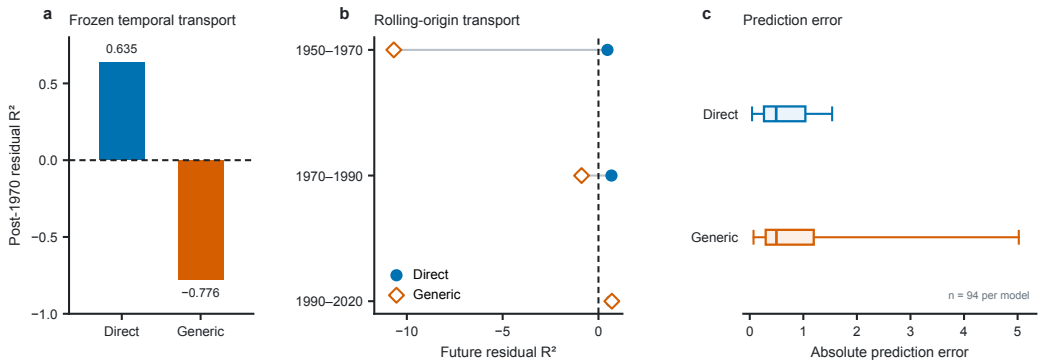


Figure 3: **Temporal transport and functional-form fragility in nuclear α decay.** The direct Z correction remains positive across historical refreezes, while the generic nonlinear proposal engine is unstable under earlier temporal extrapolation.

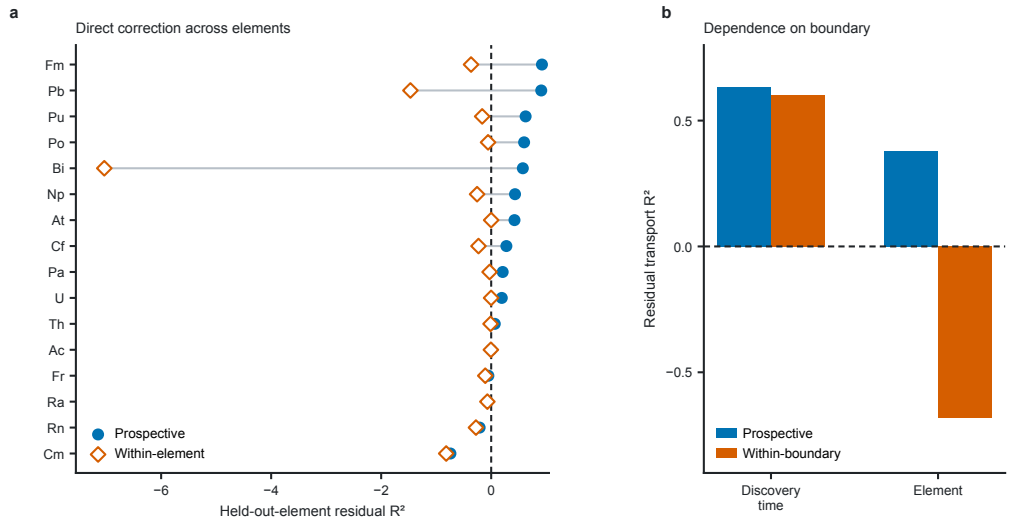


Figure 4: **Scientific generalization is indexed by the boundary.** The same direct nuclear correction transports over discovery time but not as one universal correction across unseen elements.

A richer generic nonlinear library fails the same historical test: $R^2 = -0.776$ after the 1970 freeze, with rolling values -10.681, -0.882, and 0.699. Internal selection stability therefore does not guarantee temporal stability.

Changing only the boundary changes the verdict. In leave-one-element-out evaluation, the direct correction has positive pooled performance ($R^2_{\text{pool}} = 0.379$, 95% interval $[0.159, 0.546]$) and improves eleven of sixteen elements prospectively, yet $\overline{R^2}_{\text{within}} = -0.681$. The generic nonlinear library looks even better prospectively ($R^2_{\text{pool}} = 0.544$, positive in fifteen of sixteen elements) while worsening within-element transport to -2.605 (Figure 4). More flexibility strengthens apparent cross-element performance without establishing a common within-element relationship.

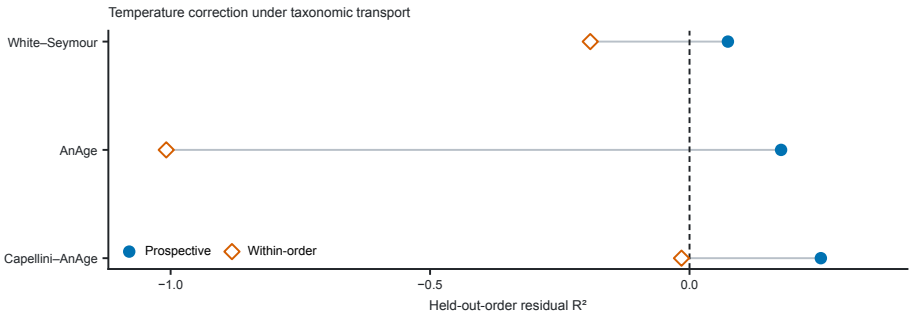


Figure 5: **Pooled-positive temperature corrections fail within-order transport across three mammalian datasets.** The repeated sign reversal distinguishes association and between-order structure from a universal correction that predicts inside unseen orders.

5.4 Mammalian temperature is associated with metabolism but fails universal cross-order transport

Temperature is not absent from the mammalian data. In the 461-species White–Seymour analysis, the temperature term is selected in every held-out-order fold and is nonzero in a fixed-intercept within-order model ($\beta = -0.673$, cluster-robust $P = 0.034$; Webb six-point wild-cluster bootstrap $P = 0.034$). But association does not become predictive transport: $R^2_{\text{pool}} = 0.074$, eight of nine orders are prospectively positive, and $\overline{R^2}_{\text{within}} = -0.191$; the pooled cluster-bootstrap interval includes zero ($[-0.017, 0.391]$).

Two independently assembled analyses repeat the sign pattern. AnAge gives $R^2_{\text{pool}} = 0.176$ with a positive interval $[0.094, 0.352]$ but $\overline{R^2}_{\text{within}} = -1.008$; Capellini–AnAge gives $R^2_{\text{pool}} = 0.253$ and $\overline{R^2}_{\text{within}} = -0.015$ (Figure 5). The evidence therefore supports a narrower interpretation: temperature helps distinguish average residual levels across observed orders, but these data do not establish one frozen temperature correction that predicts residual variation inside a new order.

5.5 A responsible audit must also return "do not correct"

A weighted one-parameter Planck fit to COBE/FIRAS gives $T = 2.725015 \pm 0.000008$ K and reduced $\chi^2 = 1.0737$. Standardized residuals are mostly within ± 2 reported standard deviations. Under the prespecified noise-floor rule, the audit stops rather than sending residuals into an increasingly flexible proposal engine. The FIRAS residual plot and full weighted-fit details are in Appendix E.

The supplementary CAPM control represents the opposite case. Across eleven held-out decades and 29,475 portfolio-month observations, a correction using known SMB/HML and volatility-related structure obtains $R^2_{\text{pool}} = 0.296$, $\overline{R^2}_{\text{within}} = 0.247$, and a 95% interval $[0.223, 0.329]$, with every decade positive. This is not a new asset-pricing discovery; it verifies that the audit can accept known omitted structure under a demanding temporal boundary.

6 Discussion and Limitations

AI changes the economics of scientific discovery: proposing another term or compact expression is increasingly cheap. The bottleneck shifts from generation to adjudication. Our results show why ordinary validation is insufficient for that adjudication. A residual can be predictable in pooled data, improve a frozen model on an unseen regime, and even be

statistically associated after group intercepts are included, yet still fail to explain variation inside a new regime.

The audit deliberately keeps two kinds of evidence separate. $R^2_{g,\text{train}}$ is prospective and deployable: it asks whether the frozen correction improves on information available before the regime exists. $R^2_{g,\text{within}}$ is retrospective: after the regime has been observed, did the correction explain more than an intercept? Neither subsumes the other. A correction can be practically useful for forecasting regime-level shifts without supporting a stronger claim of common within-regime structure.

The nuclear example also makes the boundary dependence concrete. "Generalizes" is not a scalar property of a scientific correction. Time asks whether the relationship persists as the scientific record evolves. Element identity asks whether one common relationship explains isotopic variation inside a nuclear family absent from training. Fluid identity, taxonomic order, and historical decade encode still different claims. Boundaries should therefore be chosen from the intended scientific claim, not selected after observing which split produces acceptance.

There are four main limitations. First, we do not benchmark contemporary equation-discovery engines against one another. The direct and generic proposal modes are intentionally controlled so the acceptance question is not confounded with generator performance. A natural extension is to place PySR, AI-Feynman, SINDy, or language-model-generated candidates upstream of the same audit. Second, the real-domain cluster counts are modest (six to sixteen in several analyses), so cluster-bootstrap uncertainty is interpreted conservatively and BMR uses a small-G wild-cluster check. Third, the nuclear temporal test is pseudo-prospective: discovery dates define a meaningful historical boundary, but the analysis itself is conducted with contemporary knowledge. Fourth, transport is not mechanism. Passing the audit is evidence for boundary-stable predictive structure, not proof of causality, theoretical necessity, or uniqueness.

The acceptance thresholds are likewise not claimed to be universal scientific constants. They define the locked decision protocol used here. The more general contribution is the decision trace itself: declare the boundary, freeze every data-dependent choice outside it, separate prospective from within-regime evidence, quantify regime-level replication and uncertainty, and stop when residual scale is already compatible with measurement noise.

7 Conclusion

A machine-learning system should not earn the right to alter a scientific model merely because it can predict the model's residuals. The relevant question is what that predictability survives. The residual-transport audit makes this question explicit by separating candidate generation from correction acceptance and by distinguishing prospective regime prediction from within-regime explanatory transport. Across five scientific settings, this distinction changes the scientific verdict. As automated discovery systems become more capable, evaluation layers that expose *where* a proposed correction transports may be as important as the systems that generate the correction itself.

8 AI Use Statement

Generative AI tools were used in preparing this manuscript, in three roles. **Writing:** language revision, copy-editing, and restructuring of the manuscript from an earlier long-form version into the present format. **Drafting:** initial drafts of some prose passages, subsequently rewritten and verified by the authors. **Retrieval and literature support:** identifying and checking related work, and verifying bibliographic records against primary sources and citation registries.

Generative AI was not used to generate experimental observations or synthetic data, to derive the Proposition or Corollary of Section 2, or to produce any reported numerical result. Every value in this paper and appendix is emitted by the analysis pipeline described in Appendix A under a fixed random seed, and was checked by the authors against the locked

analysis outputs. The authors take full responsibility for all content of this submission, including any AI-assisted text. No AI system is listed as an author.

9 Reproducibility Statement

The complete analysis is governed by a frozen manifest recording the random seed, decision thresholds, scientific boundaries, data filters, and sensitivity analyses. The appendix provides full dataset construction, selection, uncertainty, and comparator procedures. The supplementary package contains analysis-ready datasets, fold traces, regime metrics, term frequencies, input-verification records, software versions, machine-readable result tables, and the analysis notebook. Third-party raw inputs that cannot be redistributed are identified by retrieval location, byte size, and SHA-256 hash so that they can be reconstructed and verified independently.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. https://arxiv.org/abs/1907.02893, 2019. arXiv:1907.02893.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.

Andrew Bell and Kelvyn Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153, 2015. doi: 10.1017/psrm.2014.7.

Ian H. Bell, Jorrit Wronski, Sylvain Quoilin, and Vincent Lemort. Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library CoolProp. *Industrial & Engineering Chemistry Research*, 53(6):2498–2508, 2014. doi: 10.1021/ie4033999.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

Isabella Capellini, Chris Venditti, and Robert A. Barton. Phylogeny and metabolic scaling in mammals. *Ecology*, 91(9):2783–2793, 2010. doi: 10.1890/09-0817.1.

Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler R. Josephson, Joao Goncalves, Kenneth L. Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14:1777, 2023. doi: 10.1038/s41467-023-37236-y.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. https://arxiv.org/abs/2305.01582, 2023. arXiv:2305.01582.

João Pedro de Magalhães, Zoya Abidi, Gabriel Arantes Dos Santos, Roberto A. Avelar, Diogo Barardo, Kasit Chatsirisupachai, Peter Clark, Evandro A. De-Souza, Emily J. Johnson, Inês Lopes, Alex Novoa, Ludovic Senez, Angela Talay, Daniel Thornton, and Paul Ka Po To. Human ageing genomic resources: Updates on key databases in ageing research. *Nucleic Acids Research*, 52(D1):D900–D908, 2024. doi: 10.1093/nar/gkad927.

Craig K. Enders and Davood Tofighi. Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2):121–138, 2007. doi: 10.1037/1082-989X.12.2.121.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. doi: 10.1016/0304-405X(93)90023-5.

Dale J. Fixsen, E. S. Cheng, J. M. Gales, John C. Mather, R. A. Shafer, and E. L. Wright. The cosmic microwave background spectrum from the full COBE FIRAS data set. *The Astrophysical Journal*, 473:576–587, 1996. doi: 10.1086/178173.

Kenneth R. French. Data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, 2026. Accessed 2026-07-17.

Hans Geiger and John M. Nuttall. The ranges of the α particles from various radioactive substances and a relation between range and period of transformation. *Philosophical Magazine*, 22(130):613–621, 1911. doi: 10.1080/14786441008637156.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

James F. Gillooly, James H. Brown, Geoffrey B. West, Van M. Savage, and Eric L. Charnov. Effects of size and temperature on metabolic rate. *Science*, 293(5538):2248–2251, 2001. doi: 10.1126/science.1061967.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

F. G. Kondev, M. Wang, W. J. Huang, S. Naimi, and G. Audi. The NUBASE2020 evaluation of nuclear physics properties. *Chinese Physics C*, 45(3):030001, 2021. doi: 10.1088/1674-1137/abddae.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary symbolic regression methods and their relative performance. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Soumendra N. Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003. doi: 10.1007/978-1-4757-3803-2.

John C. Mather, E. S. Cheng, D. A. Cottingham, R. E. Eplee, D. J. Fixsen, T. Hewagama, R. B. Isaacman, K. A. Jensen, S. S. Meyer, P. D. Noerdlinger, et al. Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. *The Astrophysical Journal*, 420:439–444, 1994. doi: 10.1086/173574.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978. doi: 10.2307/1913646.

National Institute of Standards and Technology. NIST Chemistry WebBook, standard reference database number 69. https://webbook.nist.gov/chemistry/fluid/, 2026. Accessed 2026-07-17.

John M. Neuhaus and John D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645, 1998. doi: 10.2307/3109770.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, Jose J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017. doi: 10.1126/sciadv.1602614.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. doi: 10.1126/science.1165893.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B*, 75(1):55–80, 2013. doi: 10.1111/j.1467-9868.2011.01034.x.

William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964. doi: 10.1111/j.1540-6261.1964.tb02865.x.

Robi Tacutu, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana, Gilad Lehmann, Dmitri Toren, Jingwei Wang, Vadim E. Fraifeld, and João Pedro de Magalhães. Human ageing genomic resources: New and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090, 2018. doi: 10.1093/nar/gkx1042.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Johannes Diderik van der Waals. *Over de Continuiteit van den Gas- en Vloeistoftoestand*. Sijthoff, Leiden, 1873.

Victor E. Viola and Glenn T. Seaborg. Nuclear systematics of the heavy elements—ii: Lifetimes for alpha, beta and spontaneous fission decay. *Journal of Inorganic and Nuclear Chemistry*, 28(3):741–761, 1966. doi: 10.1016/0022-1902(66)80412-8.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. doi: 10.1038/s41586-023-06221-2.

M. Wang, W. J. Huang, F. G. Kondev, G. Audi, and S. Naimi. The AME 2020 atomic mass evaluation (ii). tables, graphs and references. *Chinese Physics C*, 45(3):030003, 2021. doi: 10.1088/1674-1137/abddaf.

Matthew D. Webb. Reworking wild bootstrap based inference for clustered errors. Technical Report 1315, Queen's University, Department of Economics, 2014.

Craig R. White and Roger S. Seymour. Mammalian basal metabolic rate is proportional to body mass$^{2/3}$. *Proceedings of the National Academy of Sciences*, 100(7):4046–4049, 2003. doi: 10.1073/pnas.0436428100.

A Detailed Methods

A.1 Correction-acceptance task

For observations (x_i, y_i) and a reference model f_0, the audit targets

$$r_i = y_i - f_0(x_i).$$

Each observation belongs to a prespecified scientific regime $g_i \in \{1, \ldots, G\}$. A candidate generator receives training-regime data and proposes one or more corrections $h(x)$. The audit returns a verdict: transportable reportable correction, fails the declared transport boundary, structured residual outside the reportable claim, noise-floor stop, or inconclusive.

The task begins after a reference model has been chosen. It does not determine whether that reference is philosophically fundamental or globally optimal. It asks whether a proposed change has enough transport evidence to be advanced for domain-level interpretation.

A.2 Reportable library and stability selection

The direct mode uses the supplied candidate columns without transformation. The generic mode standardizes each candidate within the training data (population standard deviation, standardized values clipped to $[-8, 8]$) and constructs linear, squared, cubic, signed-square-root, and log-of-absolute-value ($\log(1 + |z|)$) transformations together with pairwise interactions. Scaling parameters are learned only from training regimes.

Terms are selected with complementary-pairs stability selection (Shah & Samworth, 2013; Meinshausen & Bühlmann, 2010). Fifty complementary pairs are used in the primary analyses; each pair splits the training rows into two disjoint halves and a cross-validated lasso is fit on each half (Tibshirani, 1996), so the selection probability of a term is the fraction of fitted halves in which its coefficient is nonzero. A term is retained when its selection probability is at least 0.70, with at most ten reportable terms. Selected terms are refit on the complete training fold by ordinary least squares, or by cross-validated ridge regression when the retained design matrix has condition number above 10^8; the refit method used in every fold is recorded in the locked fold tables. If no term clears the threshold, the correction prediction is zero. The CAPM supplement uses month-blocked complementary pairs and ten pairs because observations from the same month share factor values across portfolios.

A.3 Leakage-safe leave-one-regime-out audit

For each held-out regime, the baseline model is fit using only the remaining regimes. Baseline residuals are calculated separately in training and test data. The reportable library, scaling, stability selection, selected terms, and correction coefficients are then learned from the training residuals. The frozen correction predicts the held-out residuals. Concatenating held-out predictions creates one leakage-safe trace covering all observations. For reference, an unconstrained random-forest regressor is fit to the same training residuals in every fold and its held-out predictions are scored identically; this flexible comparator is written to the locked output package as an internal diagnostic and plays no part in any acceptance decision or reported result.

For each regime, we calculate

$$E_g = \sum_{i \in g} (r_i - \hat{r}_i^{(-g)})^2, \quad W_g = \sum_{i \in g} (r_i - \bar{r}_g)^2, \quad B_g = n_g (\bar{r}_g - \bar{r}_{-g})^2.$$

The prospective denominator is $D_g = W_g + B_g$. Summary metrics are

$$R^2_{\text{pool}} = 1 - \frac{\sum_g E_g}{\sum_g D_g},$$

$$\overline{R^2}_{\text{group}} = \frac{1}{G} \sum_g \left(1 - \frac{E_g}{D_g}\right),$$

and

$$\overline{R^2}_{\text{within}} = \frac{1}{G}\sum_{g}\left(1 - \frac{E_g}{W_g}\right).$$

Regimes with zero within-regime variance are undefined for the last quantity and must be reported rather than silently removed.

A.4 Acceptance rule and uncertainty

The locked thresholds are: minimum pooled residual R^2 of 0.01, minimum positive-regime fraction of two-thirds, and minimum stability-selection probability of 0.70. A correction is accepted when it has at least one stable reportable term, $R^2_{\text{pool}} \geq 0.01$, $\overline{R^2}_{\text{group}} \geq 0$, $\overline{R^2}_{\text{within}} \geq 0$, the lower 95% cluster-bootstrap limit for R^2_{pool} exceeds zero, and at least two-thirds of regimes have positive prospective scores. Stability is assessed at two levels. Within a fold, a term must reach the 0.70 complementary-pairs selection probability to enter the correction. Across folds, a correction counts as having a stable reportable term only when some single term is selected in at least two-thirds of the outer held-out-regime folds; the per-fold selection counts are reported in the locked term-frequency tables. Because both criteria concern which terms survive rather than how large the resulting score is, a correction can carry a stable term and still be rejected on transport, which is what occurs in the metabolic and across-element analyses.

Cluster-bootstrap intervals resample complete regime metric rows with replacement for 5,000 replicates (Lahiri, 2003). These intervals condition on the leakage-safe selected models and do not rerun selection inside each bootstrap replicate. Results near the boundary are therefore interpreted conservatively.

A.5 Fixed-intercept and Mundlak comparators

To connect predictive transport with established within–between estimation, we fit two comparators to the leakage-safe baseline residual trace. The fixed-intercept model centres each candidate around its regime mean and includes regime indicators. Cluster-robust standard errors use the regime as the cluster. The Mundlak model includes the original predictor and its regime mean (Mundlak, 1978; Bell & Jones, 2015). The within-fluid EOS coefficients and the within-order BMR coefficient are reported in the main text; the complete coefficient tables are provided in Appendix Tables 4 and 11.

With nine BMR orders, asymptotic cluster-robust inference can be unreliable. We therefore additionally use a Webb six-point wild-cluster bootstrap-t with 4,999 replicates (Cameron et al., 2008; Webb, 2014). The primary BMR within-order temperature coefficient has wild-cluster $P = 0.034$.

A.6 Equation-of-state analysis

Thermodynamic state points are derived from NIST reference equations of state (National Institute of Standards and Technology, 2026). They are computed values returned by the WebBook isothermal fluid interface rather than raw laboratory measurements, so this analysis is a controlled recovery against a reference equation of state and not against primary experimental data. Each isotherm was retrieved once, on 17 July 2026, and cached; the archived query log records the exact request URL, the requested reduced temperature and the pressure grid for every isotherm, so the input table can be reconstructed without re-deriving any query. Critical temperatures and densities used to form the reduced variables $T_r = T/T_c$ and $\rho_r = \rho/\rho_c$ are taken from the CoolProp reference implementation (Bell et al., 2014) rather than from the WebBook pages, and are recorded in the archived fluid registry. The outcome is compressibility factor Z and the reference model is the ideal-gas value $Z = 1$. Pressure is not used as a correction predictor because $Z = P/(\rho RT)$ would make the target algebraically available. Candidate variables are density and temperature or their reduced forms.

The locked nominal reduced-temperature grid is $\{1.05, 1.15, 1.30, 1.50, 1.75, 2.00\}$. Floating-point values are canonicalized to eight decimal places only when the adjustment is below 10^{-6}; the unrounded values are preserved. The primary complete dataset contains 900 observations from ten fluids represented at all six nominal temperatures. The full-support sensitivity contains 1,155 observations from sixteen fluids. The common-temperature sensitivity contains 240 observations from sixteen fluids at $T_r = 1.05$.

The confirmatory physical basis is $(\rho_r, \rho_r/T_r)$. At a fixed reduced temperature,

$$\frac{\rho_r}{T_r} = \frac{1}{T_r}\,\rho_r,$$

so the two physical columns are exactly collinear and a two-term recovery is not identifiable. The common-temperature sensitivity therefore fits ρ_r alone and is recorded in the analysis manifest under the label `physical_density_only`. This identifiability decision is fixed before interpretation. Fluid identity is the held-out regime.

A.7 Nuclear α-decay analysis

Nuclear properties come from NUBASE2020 and atomic masses from AME2020 (Kondev et al., 2021; Wang et al., 2021). We retain ground-state nuclides with $Z \geq 82$, positive Q_α, non-estimated half-life and parent/daughter masses, an α-branch symbol of "=" or "∼", and an α branching fraction of at least 90%. The resulting analysis dataset contains 243 nuclides.

The baseline is a linear Geiger–Nuttall model using $Z/\sqrt{Q_\alpha}$ to predict $\log_{10}$ half-life. The direct correction candidates are Z and neutron number N. The generic sensitivity applies the reportable nonlinear library to the same two variables.

For the primary temporal test, training includes discoveries through 1970 ($n = 149$) and testing includes discoveries from 1971 through 2020 ($n = 94$). Every baseline and correction decision is frozen on the earlier data. Rolling-origin windows are 1951–1970 after training through 1950, 1971–1990 after training through 1970 and 1991–2020 after training through 1990. As a post-hoc diagnostic only, we assess calibration by regressing the observed future residuals on the frozen predictions and reporting the resulting intercept and slope (Appendix Table 5); this regression is never used to generate the frozen predictions themselves, which are produced entirely from the pre-cutoff training data without access to any post-cutoff observation. Element identity provides a separate leave-one-element-out boundary. Atomic number is constant inside an element, so a within-element Z coefficient is not identifiable.

A.8 Mammalian metabolic-scaling analyses

The primary dataset is parsed from the White–Seymour mammalian compilation (White & Seymour, 2003). Aggregate taxonomic rows are excluded, species names are normalized and only orders with at least ten observations and non-missing body temperature are retained. The final dataset has 461 species in nine orders. The baseline is a training-fold linear model of log BMR on log body mass. The correction candidate is

$$\frac{1}{k_B T},$$

where $k_B = 8.617333262 \times 10^{-5}$ eV K^{-1}.

The AnAge replication uses mammalian records with positive metabolic rate, body mass and temperature and retains orders with at least ten species, yielding 318 species in nine orders (Tacutu et al., 2018; de Magalhães et al., 2024). The Capellini–AnAge sensitivity joins the Capellini BMR compilation to AnAge temperature by normalized species name, removes records explicitly sourced to White–Seymour and retains orders with at least ten species, yielding 195 species in six orders (Capellini et al., 2010). Taxonomic order is the held-out boundary in all three analyses.

A.9 COBE/FIRAS noise-floor analysis

The Planck spectrum is fit to the FIRAS spectrum using the reported measurement uncertainties as weights (Mather et al., 1994; Fixsen et al., 1996). The only fitted parameter is

temperature. The noise-floor rule stops correction search when reduced $\chi^2 \leq 1.5$. The fitted temperature is 2.725015 K with standard error 8×10^{-6} K, and reduced $\chi^2 = 1.073724$.

A.10 CAPM supplementary control

Monthly Fama–French factors and 25 size–book-to-market portfolio returns are obtained from the Kenneth French Data Library (French, 2026). For each held-out decade and portfolio, the baseline fits excess return on the market excess return using all other decades. Correction candidates are SMB, HML, trailing 12-month market volatility and the portfolio beta multiplied by volatility. Candidate selection uses complete months as resampling blocks. Decades from the 1920s through the 2020s are held out, producing 29,475 held-out portfolio-month observations across eleven regimes.

A.11 Statistics and reproducibility

Sample sizes. No statistical method was used to predetermine sample size; every analysis uses the complete dataset remaining after the prespecified filters described above. The analysis sample sizes are: equation of state, $n = 900$ state points from $G = 10$ fluids (primary), $n = 1{,}155$ from $G = 16$ fluids (full-support sensitivity) and $n = 240$ from $G = 16$ fluids at $T_r = 1.05$ (common-temperature sensitivity); α decay, $n = 243$ nuclides, split into 149 pre-1970 training and 94 post-1970 test nuclides for the temporal boundary and $G = 16$ elements for the element boundary; mammalian metabolism, $n = 461$ species in $G = 9$ orders (White–Seymour), $n = 318$ in $G = 9$ orders (AnAge) and $n = 195$ in $G = 6$ orders (Capellini–AnAge); COBE/FIRAS, the published spectrum fitted with one free parameter; CAPM, $n = 29{,}475$ held-out portfolio-month observations across $G = 11$ decades. No observations were excluded after the filters were applied, and regimes with zero within-regime variance are reported rather than dropped.

Tests, thresholds and P values. All reported P values are two-sided and no multiple-comparison adjustment is applied, because the comparators are reported as descriptive companions to the transport decision rather than as a family of confirmatory hypothesis tests; acceptance is governed solely by the prespecified rule above, which uses no P value. Fixed-intercept and Mundlak comparators are estimated by ordinary least squares with cluster-robust (regime-clustered) standard errors, giving the two-sided P values reported in the text. Because the mammalian analysis has only nine clusters, where asymptotic cluster-robust inference is unreliable, the within-order temperature coefficient is additionally tested with a two-sided Webb six-point wild-cluster bootstrap-t imposing the null, with 4,999 replicates and $\alpha = 0.05$; this yields $P = 0.034$. The wild-bootstrap routine forms its own cluster-robust standard error for the bootstrap-t statistic, using a finite-sample cluster correction that differs slightly from the one applied by the fixed-intercept comparator; both values are reported in Appendix Tables 10 and 11 and agree to within 1%. Interval estimates for R^2_{pool} are 95% percentile cluster-bootstrap intervals over 5,000 replicates that resample complete regime metric rows. Uncertainty on the FIRAS temperature is the standard error of the one-parameter weighted least-squares fit using the reported measurement uncertainties as weights, and the noise-floor decision uses reduced χ^2 rather than a hypothesis test. Descriptive statistics reported in the text are accompanied by the relevant n; quantities following $\pm$ are standard errors.

Replication. The mammalian result is reported across three independently assembled datasets. The nuclear temporal result is reported across three rolling-origin refreezes (training through 1950, 1970 and 1990). The equation-of-state result is reported on the primary support and on two sensitivities. No analysis is reported from a single fit.

Analysis locking. The analysis manifest records random seed 42, all thresholds, the EOS support hierarchy and nominal temperature grid, the BMR boundary and small-cluster test, alpha filtering and temporal windows, the FIRAS stopping threshold, and the CAPM role. All required raw inputs were verified against expected file sizes and SHA-256 hashes before analysis, and execution halts if any file fails verification. The locked output package contains analysis-ready datasets, fold metadata, group metrics, complete prediction traces

where exported, term frequencies, the decision table, input verification and package versions. Software versions are reported in Appendix Table 19.

B Appendix Overview

This appendix reports the complete numerical results behind the main manuscript. The package contains the frozen analysis manifest, analysis-ready data, fold definitions, out-of-fold traces, regime-level metrics, term-selection frequencies, input hashes and environment versions. Values in the manuscript and appendix are taken directly from the locked, version-controlled analysis output package. No result is reconstructed from the prose.

The evidence hierarchy is intentionally asymmetric. Equation-of-state recovery on the complete six-temperature support is the primary controlled recovery. Full-support and common-temperature analyses are sensitivities. Nuclear α-decay is evaluated primarily through frozen future prediction and rolling origins; the element boundary is a separate scientific claim and fails the within-element requirement. Mammalian metabolism is assessed across three independently assembled analyses, with taxonomic order as the transport boundary. FIRAS is a stopping case. CAPM is a supplementary known-omitted-factor positive control.

C Proof of the Offset Decomposition

Section A defines the complete procedures used in this study. This note gives the algebraic derivation supporting the Proposition and Corollary stated in Section 2, which is presented here only to keep the main text readable.

Let $r_i = y_i - f_0(x_i)$ be the residual of a reference model and let g index a held-out regime with n_g observations, regime mean $\bar{r}_g$ and training-regime mean $\bar{r}_{-g}$. Write

$$E_g = \sum_{i \in g} \left(r_i - \hat{r}_i^{(-g)} \right)^2, \tag{12}$$

$$W_g = \sum_{i \in g} (r_i - \bar{r}_g)^2, \tag{13}$$

$$B_g = n_g \, (\bar{r}_g - \bar{r}_{-g})^2 . \tag{14}$$

Proposition (offset decomposition). Adding and subtracting $\bar{r}_g$ inside the prospective denominator and expanding,

$$\sum_{i \in g} (r_i - \bar{r}_{-g})^2 = \sum_{i \in g} \left[(r_i - \bar{r}_g) + (\bar{r}_g - \bar{r}_{-g}) \right]^2 \tag{15}$$

$$= \underbrace{\sum_{i \in g} (r_i - \bar{r}_g)^2}_{W_g} + 2 \, (\bar{r}_g - \bar{r}_{-g}) \underbrace{\sum_{i \in g} (r_i - \bar{r}_g)}_{=0} + \underbrace{n_g \, (\bar{r}_g - \bar{r}_{-g})^2}_{B_g} \tag{16}$$

$$= W_g + B_g. \tag{17}$$

The cross-term vanishes identically because residual deviations sum to zero about their own regime mean. The decomposition is therefore exact and requires no distributional assumption.

Corollary (prospective inflation by regime offset). With $R^2_{g,\text{train}} = 1 - E_g/(W_g + B_g)$ and $R^2_{g,\text{within}} = 1 - E_g/W_g$, and assuming $W_g > 0$,

$$R^2_{g,\text{train}} - R^2_{g,\text{within}} = \frac{E_g}{W_g} - \frac{E_g}{W_g + B_g} = \frac{E_g \left[(W_g + B_g) - W_g \right]}{W_g (W_g + B_g)} = \frac{E_g B_g}{W_g \, (W_g + B_g)} \; \geq \; 0, \tag{18}$$

with equality if and only if $B_g = 0$ or $E_g = 0$. The gap is therefore monotonically increasing in the squared displacement of the held-out regime mean from the training-regime mean. A

held-out regime that sits far from the training regimes inflates the prospective score by an amount that carries no information about whether the correction explains variation inside that regime.

When $W_g = 0$ the within-regime score is undefined; such regimes are reported explicitly rather than silently excluded.

D Complete Numerical Results

D.1 Master decision table

Domain	Analysis	R^2_{pool}	$\overline{R^2}_{\text{group}}$	$\overline{R^2}_{\text{within}}$	Positive	Decision
EOS	Physical basis, complete support	0.993	0.993	0.993	10/10	Accept transportable structure
α decay	Direct basis, future period	0.635	–	0.602	–	Accept temporal transport
α decay	Generic basis, future period	−0.776	–	−0.936	–	Reject functional form
α decay	Direct basis, unseen element	0.379	0.260	−0.681	11/16	Reject across-element correction
Mammalian metabolism	White–Seymour temperature	0.074	0.299	−0.191	8/9	Reject universal correction
COBE/FIRAS	Weighted Planck fit	–	–	–	–	Stop: $\chi^2_\nu = 1.074$
CAPM supplement	Known omitted-factor recovery	0.296	0.259	0.247	11/11	Supplementary positive control

Table 2: Complete decision landscape. The future-period scores use the pre-boundary training mean and the mean of the future observations, respectively. COBE/FIRAS is evaluated against the reported measurement-noise scale rather than by a residual-transport score, so the transport columns are not defined; its decision is the prespecified noise-floor stop at reduced $\chi^2 = 1.074$.

D.2 Equation-of-state recovery and sensitivities

Dataset	Analysis	T_r levels	R^2_{pool}	$\overline{R^2}_{\text{group}}$	$\overline{R^2}_{\text{within}}$	Positive	95% CI
Complete multitemperature	Physical	6	0.9929	0.9930	0.9928	10/10	[0.9856, 0.9977]
Complete multitemperature	Generic	6	0.9906	0.9906	0.9905	10/10	[0.9836, 0.9954]
Complete multitemperature	Raw	6	0.4187	0.4193	0.4181	9/10	[0.0685, 0.6595]
Full available support	Physical	6	0.9917	0.9913	0.9896	16/16	[0.9857, 0.9959]
Full available support	Generic	6	0.9894	0.9890	0.9872	16/16	[0.9834, 0.9934]
Full available support	Raw	6	0.2743	0.2438	0.1187	13/16	[0.0393, 0.4690]
Common temperature	Density only	1	0.9891	0.9891	0.9891	16/16	[0.9839, 0.9938]
Common temperature	Raw	1	0.4885	0.4891	0.4889	15/16	[0.1005, 0.7520]

Table 3: All EOS analyses in the locked package. The complete six-temperature physical analysis is primary. The full-support and fixed-temperature rows are sensitivities. The generic basis also transports strongly, so the result supports recovery of physical structure but not uniqueness of the chosen representation.

Dataset	Status	Reason
Common temperature, $T_r = 1.05$	Two-term physical basis not separately identifiable	ρ_r/T_r is an exact scalar multiple of ρ_r when T_r is fixed. The sensitivity therefore fits ρ_r alone.

Table 4: EOS identifiability statement.

D.3 EOS within–between comparators

Comparator	Term	Coefficient	SE	p
Fixed intercept	$\rho_r - \overline{\rho}_{r,g}$	0.7408	0.0647	1.14×10^{-6}
Fixed intercept	$\rho_r/T_r - \overline{(\rho_r/T_r)}_g$	−1.8414	0.0763	1.73×10^{-9}
Mundlak	ρ_r	0.7408	0.0644	1.11×10^{-6}
Mundlak	ρ_r/T_r	−1.8414	0.0760	1.67×10^{-9}
Mundlak	mean ρ_r	−477.7	39.5	7.25×10^{-7}
Mundlak	mean ρ_r/T_r	631.2	52.2	7.27×10^{-7}

Table 5: EOS fixed-intercept and Mundlak comparators. The within-fluid coefficients are stable and physically signed. The group-mean coefficients are extremely large because the two between-fluid means are highly collinear; they are disclosed for completeness and are not given a physical interpretation.

D.4 Nuclear α-decay temporal transport

Analysis	Train	Test	Future R^2	Within-future R^2	Selected terms
Direct interpretable basis	149	94	0.6348	0.6020	Z
Generic nonlinear sensitivity	149	94	−0.7762	−0.9361	log, root, polynomial and $Z \times N$ terms

Table 6: Primary pre-1970/future-period analysis. The direct correction has RMSE 0.839, calibration intercept −0.374, slope 1.493 and 21.3% of test cases outside the training predictor range. The generic sensitivity has RMSE 1.851 and slope 0.400.

Basis	Train cutoff	Test period	Train n	Test n	Future R^2	Selected terms
Direct	1950	1951–1970	64	85	0.4693	Z
Direct	1970	1971–1990	149	35	0.6780	Z
Direct	1990	1991–2020	184	59	0.6737	Z
Generic	1950	1951–1970	64	85	−10.6809	Cubic Z, N terms
Generic	1970	1971–1990	149	35	−0.8819	Broad nonlinear library
Generic	1990	1991–2020	184	59	0.6991	Linear/log/root terms

Table 7: Rolling-origin temporal validation. The direct Z correction is positive in every historical window. The generic library is internally selectable but fails the first two temporal extrapolations, illustrating functional-form fragility.

D.5 Nuclear boundary dependence

Basis	Role	R^2_{pool}	$\overline{R^2}_{\text{group}}$	$\overline{R^2}_{\text{within}}$	Positive	95% CI
Direct	Primary	0.3789	0.2598	−0.6811	11/16	[0.1585, 0.5458]
Generic	Sensitivity	0.5439	0.4491	−2.6053	15/16	[0.3640, 0.6653]

Table 8: Leave-one-element-out alpha analysis. Both bases appear positive against the training-element mean, but both are decisively negative against each element's own mean. The correction therefore transports over discovery time but is not a universal within-element correction.

Analysis	Variable	Identifiability
Within-element Mundlak decomposition	Atomic number Z	Not identifiable: $Z - \overline{Z}_{\text{element}} = 0$ by construction because every element has one atomic number.

Table 9: Boundary-specific identifiability note for alpha decay.

D.6 Mammalian metabolic replication

Analysis	n	G	R^2_{pool}	$\overline{R^2}_{\text{group}}$	$\overline{R^2}_{\text{within}}$	95% CI
White–Seymour primary	461	9	0.0740	0.2986	−0.1912	[−0.0174, 0.3912]
AnAge replication	318	9	0.1764	0.3123	−1.0084	[0.0937, 0.3518]
Capellini–AnAge	195	6	0.2531	0.3397	−0.0155	[0.1663, 0.4346]

Table 10: Temperature correction under held-out taxonomic orders. Prospective pooled scores are positive in all three analyses, but the correction fails the within-order benchmark in every dataset. This replicated sign pattern supports a taxonomic-offset interpretation rather than a universal law-level addition.

Term	Coefficient	Robust SE	t	Robust p	Wild p	Clusters
Within-order inverse temperature	−0.6735	0.2616	−2.574	0.0329	0.0340	9

Table 11: White–Seymour small-cluster inference. The Webb six-point wild-cluster bootstrap used 4,999 replicates. The non-zero within-order association does not imply that a model trained without an order predicts that order better than its own mean.

Comparator	Term	Coefficient	SE	p
Fixed intercept	Within-order inverse temperature	−0.6735	0.2642	0.0342
Mundlak	Inverse temperature	−0.6735	0.2622	0.0332
Mundlak	Order-mean inverse temperature	−0.4144	0.3971	0.3272

Table 12: Metabolic fixed-intercept and Mundlak comparators. The within-order association is detectable, while the held-out-order predictive correction is rejected. This is the intended distinction between association and correction transport.

D.7 COBE/FIRAS noise-floor result

Temperature (K)	SE (K)	Total R^2	Reduced χ^2	Decision
2.7250153	7.58×10^{-6}	0.999999725	1.073724	Stop at noise floor

Table 13: Weighted Planck-law fit. The reported R^2 measures how well the one-parameter Planck law reproduces the full measured spectrum, not residual transport across a held-out regime. Because the residual scale is compatible with the reported measurement uncertainty, the audit stops and does not search for a correction.

D.8 CAPM supplementary recovery control

R^2_{pool}	$\overline{R^2}_{\text{group}}$	$\overline{R^2}_{\text{within}}$	Positive	G	95% CI
0.2963	0.2588	0.2467	11/11	11	[0.2231, 0.3290]

Table 14: Overall CAPM known-omitted-factor recovery.

Held-out decade	Observations	Prospective R^2	Within-decade R^2
1920s	775	0.1701	0.1401
1930s	3,000	0.3452	0.3426
1940s	3,000	0.2422	0.2368
1950s	3,000	0.2705	0.2560
1960s	3,000	0.2443	0.2409
1970s	3,000	0.3171	0.3114
1980s	3,000	0.2076	0.2054
1990s	3,000	0.2924	0.2716
2000s	3,000	0.1821	0.1619
2010s	3,000	0.2280	0.2087
2020s	1,700	0.3479	0.3385

Table 15: CAPM recovery by held-out decade. Every decade is positive against both baselines. Complete months, rather than portfolio rows, are the CPSS resampling units.

D.9 Regime-level transport metrics

This subsection reports the per-regime quantities underlying every aggregate score in the main text. The same values are provided in machine-readable form in Supplementary Data workbook (sheets `EOS_Fluid_Metrics`, `Alpha_Element_Metrics` and `BMR_Order_Metrics`).

Table 16: Regime-level transport metrics for the equation-of-state analyses, one row per held-out fluid. For each held-out regime g, E_g is the prospective squared-error sum, W_g the within-regime sum of squares about that regime's own mean, B_g the offset term $n_g(\bar{r}_g - \bar{r}_{-g})^2$, and $D_g = W_g + B_g$ the prospective denominator, so that $R^2_{g,\text{train}} = 1 - E_g/D_g$ and $R^2_{g,\text{within}} = 1 - E_g/W_g$. The offset term is small relative to the within-fluid spread throughout: the median B_g/W_g is 0.0014 in the primary analysis, 0.025 under full support and 7.2×10^{-5} at the common temperature, and B_g never exceeds W_g for any fluid in any of the three analyses. Every fluid has a positive within-fluid score.

Held-out fluid	n_g	E_g	W_g	B_g	D_g	$R^2_{g,\text{train}}$	$R^2_{g,\text{within}}$
Complete six-temperature support (primary)							
Argon	90	5.30×10^{-6}	0.0099	1.09×10^{-6}	0.0099	0.999	0.999
Carbon dioxide	90	1.25×10^{-4}	0.0103	6.30×10^{-5}	0.0103	0.988	0.988

Appendix Table 16 continued.

Held-out fluid	n_g	E_g	W_g	B_g	D_g	$R^2_{g,\text{train}}$	$R^2_{g,\text{within}}$
Carbon monoxide	90	9.61×10^{-6}	0.0100	2.46×10^{-6}	0.0100	0.999	0.999
Ethane	90	5.72×10^{-5}	0.0102	2.18×10^{-5}	0.0102	0.994	0.994
Helium	90	3.67×10^{-4}	0.0098	3.08×10^{-4}	0.0101	0.964	0.962
Hydrogen	90	6.55×10^{-5}	0.0096	2.96×10^{-5}	0.0097	0.993	0.993
Methane	90	1.20×10^{-5}	0.0100	1.61×10^{-6}	0.0100	0.999	0.999
Neon	90	2.83×10^{-5}	0.0099	6.23×10^{-7}	0.0099	0.997	0.997
Nitrogen	90	1.88×10^{-5}	0.0099	1.34×10^{-5}	0.0099	0.998	0.998
Oxygen	90	1.93×10^{-5}	0.0100	4.30×10^{-6}	0.0100	0.998	0.998
Full support (sensitivity)							
Argon	90	9.20×10^{-6}	0.0099	1.70×10^{-4}	0.0101	0.999	0.999
Carbon dioxide	90	7.67×10^{-5}	0.0103	3.90×10^{-4}	0.0107	0.993	0.993
Carbon monoxide	90	1.50×10^{-5}	0.0100	1.83×10^{-4}	0.0102	0.999	0.999
Ethane	90	2.59×10^{-5}	0.0102	2.74×10^{-4}	0.0104	0.998	0.997
Helium	90	4.23×10^{-4}	0.0098	2.63×10^{-5}	0.0098	0.957	0.957
Hydrogen	90	1.18×10^{-4}	0.0096	4.50×10^{-5}	0.0097	0.988	0.988
Isobutane	45	2.92×10^{-5}	0.0055	0.0012	0.0067	0.996	0.995
Isopentane	15	6.58×10^{-5}	0.0025	0.0020	0.0045	0.985	0.974
Methane	90	4.50×10^{-6}	0.0100	1.76×10^{-4}	0.0102	1.000	1.000
Neon	90	5.50×10^{-5}	0.0099	1.63×10^{-4}	0.0101	0.995	0.994
Nitrogen	90	2.30×10^{-5}	0.0099	2.43×10^{-4}	0.0101	0.998	0.998
Oxygen	90	3.24×10^{-5}	0.0100	1.97×10^{-4}	0.0102	0.997	0.997
Propane	75	4.50×10^{-5}	0.0087	4.19×10^{-7}	0.0087	0.995	0.995
n-Butane	45	4.49×10^{-5}	0.0054	9.29×10^{-4}	0.0063	0.993	0.992
n-Hexane	30	1.20×10^{-4}	0.0043	0.0021	0.0064	0.981	0.972
n-Pentane	45	7.16×10^{-5}	0.0055	0.0011	0.0066	0.989	0.987
Common temperature $T_r = 1.05$, density only (sensitivity)							
Argon	15	5.85×10^{-6}	0.0025	1.71×10^{-7}	0.0025	0.998	0.998
Carbon dioxide	15	2.16×10^{-5}	0.0025	5.70×10^{-7}	0.0025	0.991	0.991
Carbon monoxide	15	9.19×10^{-6}	0.0025	1.73×10^{-8}	0.0025	0.996	0.996
Ethane	15	5.06×10^{-6}	0.0025	1.60×10^{-7}	0.0025	0.998	0.998
Helium	15	5.38×10^{-5}	0.0025	4.96×10^{-8}	0.0025	0.979	0.979
Hydrogen	15	7.59×10^{-5}	0.0024	1.47×10^{-6}	0.0024	0.969	0.969
Isobutane	15	1.05×10^{-5}	0.0025	1.08×10^{-8}	0.0025	0.996	0.996
Isopentane	15	4.50×10^{-5}	0.0025	9.49×10^{-7}	0.0025	0.982	0.982
Methane	15	7.89×10^{-7}	0.0025	6.15×10^{-9}	0.0025	1.000	1.000
Neon	15	3.79×10^{-5}	0.0025	1.78×10^{-7}	0.0025	0.985	0.985
Nitrogen	15	8.57×10^{-6}	0.0024	8.51×10^{-7}	0.0024	0.996	0.996
Oxygen	15	1.73×10^{-5}	0.0025	1.04×10^{-7}	0.0025	0.993	0.993
Propane	15	1.33×10^{-5}	0.0025	2.43×10^{-7}	0.0025	0.995	0.995
n-Butane	15	6.35×10^{-6}	0.0024	2.39×10^{-6}	0.0024	0.997	0.997
n-Hexane	15	8.33×10^{-5}	0.0026	4.16×10^{-6}	0.0026	0.968	0.968
n-Pentane	15	4.07×10^{-5}	0.0025	7.35×10^{-8}	0.0025	0.984	0.984

Table 17: Regime-level transport metrics for the leave-one-element-out α-decay analyses, one row per held-out element. For each held-out regime g, E_g is the prospective squared-error sum, W_g the within-regime sum of squares about that regime's own mean, B_g the offset term $n_g(\bar{r}_g - \bar{r}_{-g})^2$, and $D_g = W_g + B_g$ the prospective denominator, so that $R^2_{g,\text{train}} = 1 - E_g/D_g$ and $R^2_{g,\text{within}} = 1 - E_g/W_g$. The offset term is substantial here: the median B_g/W_g is 0.70 and B_g exceeds W_g for 6 of the 16 elements, reaching 26.4 at the extreme. The direct basis yields a negative within-element score for all 16 elements; the generic basis does so for 4 of 16, and its pooled within-element score is negative because the elements with negative scores carry the larger error sums. B_g and W_g depend only on the residuals and the grouping, so they are identical across the two bases; only E_g differs.

Held-out element	n_g	E_g	W_g	B_g	D_g	$R^2_{g,\text{train}}$	$R^2_{g,\text{within}}$
Direct Z correction							
Pb	6	1.152	0.4667	12.320	12.786	0.910	−1.468
Bi	5	12.811	1.595	28.557	30.152	0.575	−7.034
Po	22	14.635	13.847	22.593	36.440	0.598	−0.057
At	16	9.394	9.388	6.907	16.295	0.424	−0.001
Rn	19	11.449	8.966	0.4939	9.459	−0.210	−0.277
Fr	19	8.056	7.272	0.3902	7.662	−0.051	−0.108
Ra	21	9.356	8.748	0.0016	8.749	−0.069	−0.070
Ac	20	8.481	8.433	0.0245	8.457	−0.003	−0.006
Th	23	15.621	15.440	1.268	16.708	0.065	−0.012
Pa	16	10.303	10.011	3.010	13.021	0.209	−0.029
U	18	18.164	18.137	4.358	22.496	0.193	−0.001
Np	9	10.411	8.287	10.081	18.368	0.433	−0.256
Pu	8	1.699	1.460	3.082	4.541	0.626	−0.164
Cm	8	16.402	9.023	0.3976	9.421	−0.741	−0.818
Cf	7	11.266	9.153	6.413	15.566	0.276	−0.231
Fm	7	2.312	1.692	28.222	29.915	0.923	−0.366
Generic library							
Pb	6	18.213	0.4667	12.320	12.786	−0.424	−38.029
Bi	5	11.076	1.595	28.557	30.152	0.633	−5.945
Po	22	9.570	13.847	22.593	36.440	0.737	0.309
At	16	6.070	9.388	6.907	16.295	0.628	0.353
Rn	19	5.042	8.966	0.4939	9.459	0.467	0.438
Fr	19	3.368	7.272	0.3902	7.662	0.561	0.537
Ra	21	6.233	8.748	0.0016	8.749	0.288	0.288
Ac	20	7.048	8.433	0.0245	8.457	0.167	0.164
Th	23	10.374	15.440	1.268	16.708	0.379	0.328
Pa	16	6.858	10.011	3.010	13.021	0.473	0.315
U	18	9.766	18.137	4.358	22.496	0.566	0.462
Np	9	5.214	8.287	10.081	18.368	0.716	0.371
Pu	8	3.285	1.460	3.082	4.541	0.277	−1.250
Cm	8	7.131	9.023	0.3976	9.421	0.243	0.210
Cf	7	6.834	9.153	6.413	15.566	0.561	0.253
Fm	7	2.518	1.692	28.222	29.915	0.916	−0.488

Table 18: Regime-level transport metrics for the mammalian temperature correction, one row per held-out taxonomic order, in all three datasets. For each held-out regime g, E_g is the prospective squared-error sum, W_g the within-regime sum of squares about that regime's own mean, B_g the offset term $n_g(\bar{r}_g - \bar{r}_{-g})^2$, and $D_g = W_g + B_g$ the prospective denominator, so that $R^2_{g,\text{train}} = 1 - E_g/D_g$ and $R^2_{g,\text{within}} = 1 - E_g/W_g$. Within-order scores are negative for 4 of 9 orders in White–Seymour, 4 of 9 in AnAge and 2 of 6 in Capellini–AnAge, and the pooled within-order score is negative in all three. B_g exceeds W_g for 3 of 9, 5 of 9 and 1 of 6 orders respectively, which is the offset contribution that makes the prospective pooled score positive while the within-order score is not.

Held-out order	n_g	E_g	W_g	B_g	D_g	$R^2_{g,\text{train}}$	$R^2_{g,\text{within}}$
White–Seymour							
Carnivora	38	5.469	4.093	5.114	9.206	0.406	−0.336
Chiroptera	67	73.897	70.005	0.2055	70.211	−0.053	−0.056
Insectivora	38	9.267	12.337	0.5940	12.931	0.283	0.249
Primates	14	1.552	2.124	0.1419	2.266	0.315	0.269
Rodentia	233	21.049	21.460	0.3737	21.834	0.036	0.019
Xenarthra	15	1.470	0.5499	2.927	3.477	0.577	−1.674
Dasyuromorpha	21	2.777	1.509	2.726	4.235	0.344	−0.841
Didelphimorphia	11	0.2541	0.4292	0.1025	0.5318	0.522	0.408
Diprotodontia	24	1.091	1.437	0.0312	1.468	0.257	0.241
AnAge							
Artiodactyla	19	4.167	0.9223	5.677	6.599	0.369	−3.518
Carnivora	46	6.019	6.139	0.1422	6.282	0.042	0.020
Chiroptera	31	1.819	2.454	0.2173	2.672	0.319	0.259
Dasyuromorphia	20	1.986	0.6594	3.226	3.886	0.489	−2.011
Didelphimorphia	10	0.0961	0.2265	0.5201	0.7466	0.871	0.576
Diprotodontia	20	1.477	1.082	1.337	2.419	0.389	−0.366
Primates	22	1.972	2.085	0.3669	2.452	0.196	0.054
Rodentia	134	14.362	15.754	0.0024	15.756	0.088	0.088
Soricomorpha	16	12.610	2.436	10.795	13.231	0.047	−4.177
Capellini–AnAge							
Carnivora	29	3.520	4.274	0.8789	5.153	0.317	0.176
Chiroptera	25	1.904	1.705	0.0555	1.761	−0.081	−0.117
Rodentia	102	11.407	12.106	2.087	14.193	0.196	0.058
Dasyuromorpha	16	1.306	0.7446	1.933	2.677	0.512	−0.754
Didelphimorphia	10	0.1257	0.2490	0.2436	0.4926	0.745	0.495
Diprodontia	13	0.8809	0.9252	0.4284	1.354	0.349	0.048

D.10 Frozen inputs and computational environment

All nine required raw input files matched their expected byte sizes and SHA-256 hashes. The verified inputs were the AME2020 mass table, AnAge archive, White–Seymour HTML source, Capellini BMR table, CAPM archive, FIRAS spectrum, common-support EOS CSV, full EOS archive and NUBASE2020 evaluation file. Exact byte sizes and hashes are listed in the `Input_Verification` sheet of Supplementary Data workbook.

Package	Version	Package	Version
Python	3.12.13	NumPy	2.0.2
Pandas	2.2.2	SciPy	1.16.3
scikit-learn	1.6.1	statsmodels	0.14.6
Matplotlib	3.10.0	Additional packages	See locked CSV

Table 19: Core environment versions. The complete package inventory is in `table_environment_versions.csv`.

E FIRAS Residual Diagnostic

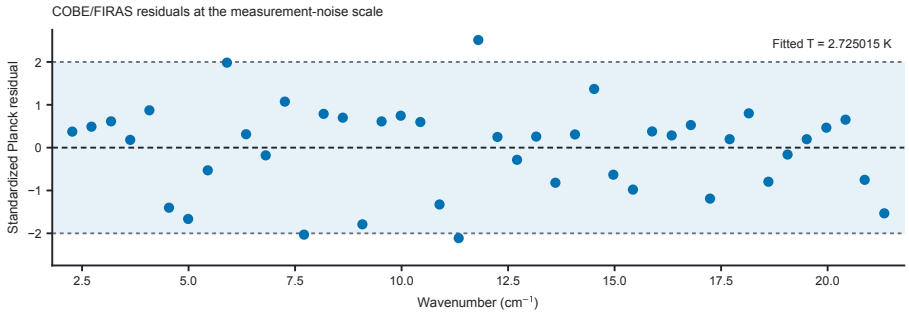


Figure 6: **COBE/FIRAS residuals remain at the reported measurement-noise scale.** Standardized residuals of the weighted Planck fit are shown against wavenumber; the residuals lie almost entirely within the ±2 reported-standard-deviation band, triggering the prespecified noise-floor stop.

F Map from manuscript items to locked output files

Manuscript item	Locked source files
Figure 2, EOS	`table_eos_summary.csv`; physical, generic and raw `group_metrics.csv`; `table_eos_identifiability_note.csv`
Figure 3, alpha temporal	`table_alpha_direct_vs_generic_future.csv`; `table_alpha_direct_vs_generic_rolling.csv`; alpha analysis dataset
Figure 4, alpha boundary	`table_alpha_element_boundary.csv`; direct and generic element group metrics; `table_alpha_identifiability_note.csv`
Figure 5, metabolism	`table_bmr_replication.csv`; three BMR group-metric files; `table_bmr_wild_cluster_bootstrap.csv`
Appendix Figure 6, FIRAS	`table_firas_summary.csv`; original weighted spectrum input verified by hash

Manuscript item	Locked source files
CAPM supplement	`table_capm_supplement_summary.csv`; `table_capm_supplement_decades.csv`; `table_capm_supplement_terms.csv`
Comparator claims	`table_valid_fixed_mundlak_comparators.csv`
Master decisions	`table_analysis_decisions.csv`
Input provenance	`table_input_verification.csv`; `analysis_manifest.json`; `table_environment_versions.csv`

G Supplementary Data

Supplementary Data workbook. `supplementary_data.xlsx`. Source values for every quantity reported in the main manuscript and in this appendix, exported directly from the locked analysis output package. The workbook contains twenty sheets:

- `README` — purpose, locking statement and interpretation notes.
- `Main_Decisions` — the decision landscape reproduced as main-text Table 1.
- `EOS_Summary`, `Alpha_Future`, `Alpha_Rolling`, `Alpha_Elements`, `BMR_Summary`, `BMR_Wild_Cluster`, `FIRAS`, `CAPM_Summary`, `CAPM_Decades`, `CAPM_Terms` — per-domain aggregate results.
- `EOS_Fluid_Metrics`, `Alpha_Element_Metrics`, `BMR_Order_Metrics` — regime-level metrics, tabulated as Appendix Tables 16–18.
- `Comparators` — fixed-intercept and Mundlak coefficients, standard errors and P values.
- `Identifiability` — the recorded identifiability decisions.
- `Input_Verification` — expected and observed byte sizes and SHA-256 hashes for all nine raw inputs.
- `Environment` — the complete package inventory.
- `Chart_Data` — the plotted values underlying Figures 2–5 and Appendix Figure 6.