Electronic Supplementary Material for: Ecological Inference Is Structured Empirical Risk Minimization: Generalization Across Space, Nested Units, and Niche Support

R. Craig Stillwell

Independent Researcher

craig.stillwell@gmail.com

Abstract

This electronic supplementary material (ESM) supports *Ecological Inference Is Structured Empirical Risk Minimization* (Stillwell, 2026). It contains full proofs of all formal results in the main text, maps ESM section numbers to main-text labels (Table 1), supplementary tables for the *Stator limbatus* geographic cline vignette, and the semi-synthetic simulation protocol. Analysis code: `https://github.com/rstil2/ecological-inference-structured-erm` (`REPRODUCE.md`).

Contents

1 Shared setup

Data are nested: individual $i \in \{1, \ldots, n_p\}$ in population (site) $p \in \{1, \ldots, P\}$. Write individual response y_{ip}, population mean $\bar{y}_p = n_p^{-1} \sum_i y_{ip}$, site covariates $\mathbf{x}_p \in \mathcal{X} \subset \mathbb{R}^d$, hypothesis class $\mathcal{H}$, and squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$ unless noted otherwise.

Population-level generative model:

$$y_{ip} = \mu(\mathbf{x}_p) + u_p + \varepsilon_{ip}, \qquad \mathbb{E}[u_p \mid \mathbf{x}_p] = 0, \tag{1}$$

with $\mu(\mathbf{x}_p) = \mathbb{E}[\bar{y}_p \mid \mathbf{x}_p]$. Macroecological cline hypotheses concern $\mu(\cdot)$, not within-population noise. Site-level random effect u_p in (1) is distinct from biotic context b_p introduced in Section 6.

Leave-one-population-out (LOPO) risk on population means:

$$R_{\mathrm{LOPO}}(h) = \frac{1}{P} \sum_{p=1}^{P} \left(\bar{y}_p - \hat{h}^{(-p)}(\mathbf{x}_p) \right)^2, \tag{2}$$

where $\hat{h}^{(-p)}$ is fit on all populations except p.

Table 1: Map from ESM sections to main-text labels.

ESM	Main text	Result
§2	Theorem 1	Estimand–CV mismatch
§3	Definition 1	Three estimands
§4.2	Lemma 1	Oracle risk ordering
§4.3	Theorem 2	Proxy rank reversal
§5	Proposition 3	Grinnellian support
§6	Proposition 4	Hidden context
§7	Proposition 5	Plasticity / composite label
§8	Mechanism; two-test	Mechanistic certification
§9–10	—	Tables S1–S4; simulation

2 Estimand–CV mismatch (main-text Theorem 1)

Theorem 1 (Estimand–CV mismatch) *Let $\hat{R}^{\mathrm{kfold}}(h)$ denote K-fold cross-validation with folds over individuals. Let $\hat{R}^{\mathrm{LOPO}}(h)$ denote LOPO on population means. When $\mathrm{Var}(\varepsilon_{ip} \mid p) > 0$ and the scientific estimand is E_2 (new population mean at a new site), $\mathbb{E}[\hat{R}^{\mathrm{kfold}}(h)] \neq \mathbb{E}[\hat{R}^{\mathrm{LOPO}}(h)]$.*

Proof Define population-level risks by estimand.

E_1 **(within-site).** Predict a new individual at a population already represented in training:

$$R_1(h) = \mathbb{E}\Big[\ell\big(h(\mathbf{x}_p), y_{i^\star p}\big) \;\Big|\; p \in \mathcal{P}_{\text{train}}\Big].$$

E_2 **(across-site).** Predict the population mean at a held-out site:

$$R_2(h) = \mathbb{E}\Big[\ell\big(h(\mathbf{x}_{p^\star}), \bar{y}_{p^\star}\big)\Big].$$

Under (1), squared loss at a population mean decomposes as

$$R_2(h) = \mathbb{E}\Big[\big(h(\mathbf{x}_{p^\star}) - \mu(\mathbf{x}_{p^\star})\big)^2\Big] + \mathrm{Var}(\bar{u}_{p^\star}),$$

where $\bar{u}_{p^\star} = n_{p^\star}^{-1} \sum_i u_{i^\star p^\star}$. As $n_{p^\star} \to \infty$, $\mathrm{Var}(\bar{u}_{p^\star}) \to 0$ and the target is $\mu(\cdot)$.

For E_1, conditioning on site p,

$$\mathbb{E}\big[(h(\mathbf{x}_p) - y_{i^\star p})^2 \mid p\big] = \big(h(\mathbf{x}_p) - \mu(\mathbf{x}_p)\big)^2 + \mathrm{Var}(u_p \mid \mathbf{x}_p) + \mathrm{Var}(\varepsilon_{ip}).$$

Whenever test individuals share training-site identity,

$$R_1(h) = R_2(h) + \mathbb{E}[\mathrm{Var}(\varepsilon_{ip} \mid p)] + \mathbb{E}[\mathrm{Var}(u_p \mid \mathbf{x}_p)].$$

The within-population noise term $\mathbb{E}[\mathrm{Var}(\varepsilon_{ip} \mid p)]$ enters R_1 but not prediction of $\bar{y}_{p^\star}$ at a new site.

Individual k-fold CV, with multiple individuals per population, places test and training individuals at the same $\mathbf{x}_p$. Its expectation therefore

targets E_1 and averages within-population noise into the reported metric. LOPO removes site p from training when predicting $\bar{y}_p$ and targets E_2.

If $\mathrm{Var}(\varepsilon_{ip} \mid p) > 0$, then $R_1(h) > R_2(h)$ for any h with identical population-level fit. Hence $\mathbb{E}[\hat{R}^{\mathrm{kfold}}(h)] \neq \mathbb{E}[\hat{R}^{\mathrm{LOPO}}(h)]$ when the scientific estimand is E_2. ■

Corollary 2 (Ranking bias) *Model ranking by $\hat{R}^{\mathrm{kfold}}(h)$ need not agree with ranking by $\hat{R}^{\mathrm{LOPO}}(h)$ when models differ in within-site flexibility (e.g. population-specific intercepts).*

3 Three estimands (main-text Definition 1)

Definition 3 (Three estimands)

1. E_1*: new individuals at known populations (individual CV).*
2. E_2*: new population means at new coordinates (LOPO, LOGO).*
3. E_3*: new regions of niche space outside training support (extrapolation / OOD).*

Proof [Classification argument] The estimands differ by where the test point lies relative to training support.

E_1**.** $p \in \mathcal{P}_{\mathrm{train}}$, new individual i: only individual-level noise changes; site identity is known.

E_2**.** New site $p^\star$ with $\mathbf{x}_{p^\star}$ drawn from the geographic process but $p^\star \notin \mathcal{P}_{\mathrm{train}}$: prediction requires transfer of $\mu(\mathbf{x})$; site identity must not leak into training.

E_3. $\mathbf{x}_{p^\star} \notin \mathcal{S}_{\text{train}} = \text{supp}(P_{\mathbf{x}} \mid \text{training})$: support mismatch. No CV scheme on observed sites alone certifies extrapolation without invariance assumptions (Section 5).

Section 2 shows E_1 and E_2 are not interchangeable. ■

4 Gradient confounding and rank reversal

4.1 Setup

Populations lie on a geographic axis g_p. Predictors satisfy $\mathbf{x}_p = \boldsymbol{\phi}(g_p) + \boldsymbol{\epsilon}_p$ with $\boldsymbol{\phi}$ monotone across coordinates (structured collinearity). The generating model is $y_p = \mathbf{x}_p^\top \boldsymbol{\beta}^\star + \eta_p$, $\eta_p \sim (0, \sigma_\eta^2)$.

Let h_j denote the best in-sample univariate predictor on coordinate x_j, and $h^\star$ the population-level OLS predictor on true support $\mathcal{S}^\star = \{k : \beta_k^\star \neq 0\}$.

4.2 Oracle risk ordering (main-text Lemma 1)

Lemma 4 (Oracle risk ordering on true support) *Under the linear model above, population-level OLS on $\mathcal{S}^\star = \{k : \beta_k^\star \neq 0\}$ minimizes squared prediction risk among predictors using only $\mathbf{x}_{\mathcal{S}^\star}$ [Hastie et al., 2009]. This optimality does not explain proxy rank reversal; the substantive failure mode is Theorem 5 below.*

Proof For fixed p, when $\beta_j^\star = 0$ and x_j is not sufficient for $\mathbf{x}_{\mathcal{S}^\star}$, any predictor using only x_j incurs Bayes risk at least that of the best linear predictor

on $\mathbf{x}_{\mathcal{S}^\star}$. LOPO averages leave-one-out squared errors; the population-level ordering is preserved up to finite-P estimation variance. ■

4.3 Proxy rank reversal (main-text Theorem 2)

Theorem 5 (Proxy rank reversal) *Under gradient confounding, in-sample $|r|$ or AIC can rank a proxy above the structural model while LOPO ranks them in reverse order.*

Proof Consider $y_p = \beta_1 x_{1,p} + \beta_2 x_{2,p} + \eta_p$ with $x_{2,p} = \rho x_{1,p} + \nu_p$, $\rho \neq 0$, and $\beta_2 \neq 0$, while the proxy coordinate has $\beta_1 = 0$ in the generating model (latitude proxies ecology along the gradient).

In-sample ranking. Pooled OLS on x_1 alone estimates $\hat{\alpha}_1 \approx \mathrm{Cov}(x_1, y)/\mathrm{Var}(x_1)$, which loads on $\beta_2 \rho\, \mathrm{Var}(x_1)$ through collinearity. High $|r|$ and low AIC for the univariate model follow when $|\rho|$ is large and multivariate gain is small on the training sample.

LOPO ranking. At held-out site p, leave-one-out univariate prediction error includes $(\hat{\alpha}_1^{(-p)} x_{1,p} - \beta_2 x_{2,p})^2$. When local deviation ν_p is large, $x_{2,p} \neq \rho x_{1,p}$: the univariate model extrapolates the gradient while the bivariate model retains $\beta_2 x_{2,p}$. Hence $R_{\mathrm{LOPO}}(h_j) > R_{\mathrm{LOPO}}(h^\star)$ at such sites, and model ranking can reverse relative to in-sample $|r|$ or AIC.

Simulation. Semi-synthetic clines (500 replicates; Section 10) yield 84% rank disagreement between strongest $|r|$ and best LOPO model under the stated ecology-driven generating process—a design-specific rate, not an ecological constant. ■

5 Grinnellian support bound (main-text Proposition 3)

Theorem 6 (Grinnellian support bound) *Let training environments* $\mathbf{x}_p \in \mathcal{S}_{\text{train}} \subset \mathcal{X}$. *For Lipschitz* h *with constant* L, *extrapolation error at* $\mathbf{x} \notin \mathcal{S}_{\text{train}}$ *satisfies*

$$|h(\mathbf{x}) - h(\mathbf{x}_0)| \leq L \operatorname{dist}(\mathbf{x}, \mathcal{S}_{\text{train}}),$$

for $\mathbf{x}_0$ *the nearest point in* $\mathcal{S}_{\text{train}}$, *unless invariance assumptions make* h *constant off-support. Range limits occur where* $P(y = 1 \mid \mathbf{x})$ *crosses a decision threshold outside* $\mathcal{S}_{\text{train}}$.

Proof By the triangle inequality, $|h(\mathbf{x}) - h(\mathbf{x}_0)| \leq L\|\mathbf{x} - \mathbf{x}_0\|$. Minimizing $\|\mathbf{x} - \mathbf{x}_0\|$ over $\mathbf{x}_0 \in \mathcal{S}_{\text{train}}$ gives the bound.

For species occupancy, if $h(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$ is Lipschitz in environmental space and training occupied only $\mathcal{S}_{\text{train}}$, predictions outside the convex hull have unbounded epistemic risk as distance grows unless additional invariance (e.g. IRM-style [Arjovsky et al., 2019]) is assumed. A range limit is the geographic boundary where predicted suitability crosses a decision threshold outside observed support—an extrapolation failure, not a within-support optimization problem. ■

6 Realized niche as hidden context (main-text Proposition 4)

Theorem 7 (Realized niche as hidden context) *Suppose biotic context b_p affects the realized response: $y_p = f(\mathbf{x}_p, b_p) + \eta_p$, with b_p unobserved in Grinnellian models that use only $\mathbf{x}_p$. The marginal model $P(y \mid \mathbf{x})$ integrates over b. Transfer to sites where $P(b \mid \mathbf{x})$ shifts induces context-dependent label noise: simultaneous covariate shift in $P(\mathbf{x})$ and label shift in $P(y \mid \mathbf{x})$.*

Proof An observational niche model estimates $\hat{h}(\mathbf{x}) \approx \mathbb{E}[y \mid \mathbf{x}] = \mathbb{E}_{b|\mathbf{x}}[f(\mathbf{x}, b)]$.

At training sites, $P(b \mid \mathbf{x})$ is fixed. At transfer sites, if $P(b \mid \mathbf{x})$ changes—competitors absent, novel mutualists—then $\mathbb{E}[y \mid \mathbf{x}]$ shifts even when $\mathbf{x}$ is unchanged: label shift conditional on $\mathbf{x}$.

Equivalently, the response depends on hidden b; marginalizing b yields a composite target stable only while $P(b \mid \mathbf{x})$ is stable. This formalizes the realized-niche / Eltonian–Grinnellian distinction as hidden-variable risk [Soberón and Nakamura, 2007, Koh et al., 2021]. ■

7 Plasticity as composite-label risk (main-text Proposition 5)

Theorem 8 (Plasticity as composite-label risk) *Field clines follow $y_{ip}^{\text{field}} = g_p^{\text{gen}} + g^{\text{plast}}(\mathbf{x}_p) + \epsilon_{ip}$. Fitting $h(\mathbf{x})$ to y^{field} without decomposition targets a genetic–plastic sum, not the genetic cline g^{gen} alone. Under plasticity,*

$\text{sign}(\partial h/\partial x_j)$ *can disagree with* $\text{sign}(\partial g^{\text{gen}}/\partial x_j)$.

Proof ERM on field phenotypes minimizes $\mathbb{E}[(h(\mathbf{x}) - y^{\text{field}})^2]$. The Bayes-optimal predictor is $h^\star(\mathbf{x}) = g^{\text{gen}}(\mathbf{x}) + g^{\text{plast}}(\mathbf{x})$ —the composite target—not $g^{\text{gen}}(\mathbf{x})$ alone.

If $g^{\text{plast}}(\mathbf{x})$ covaries with $\mathbf{x}$ along the gradient (e.g. temperature-induced developmental plasticity), the fitted slope $\partial h/\partial x_j$ conflates genetic and plastic components.

Remark (Stillwell 2010). Plasticity can generate phenotypic clines without genetic differentiation; ERM on field phenotypes therefore minimizes the wrong risk for an *adaptation* estimand, which requires common-garden or breeding values as the supervision target [Via and Lande, 1985, Stillwell, 2010]. ■

8 Mechanistic adequacy and two-test logic

Theorem 9 (Mechanistic adequacy) *Hypothesis M—that environment e modulates selection on trait z—implies a fitness model with $z \times e$ interaction ($\Delta W \sim z + e + z \times e$). An additive correlational model $W \sim z + e$ can fit a cline in sample yet fail mechanistic tests under cross-environment holdout or experiment.*

Proof If M holds, the relative fitness advantage of phenotype z depends on e: $\partial^2 W/(\partial z\, \partial e) \neq 0$. An additive model sets the interaction coefficient to zero by construction.

Along a narrow environmental range where z and e covary, additive and interactive models can fit similarly in sample. Under manipulation (e.g. selection lines at 22 vs 34 °C) or cross-environment holdout, the interaction term is identifiable.

Stillwell et al. [2008]: large body-size lines had higher fecundity at both temperatures, but the relative Large–Control advantage did *not* increase at low temperature—falsifying the thermal-Bergmann interaction despite size manipulation. ■

Theorem 10 (Two-test logic) *Mechanism certification requires (i) structured holdout (E_2) and (ii) the required interaction under manipulation; neither alone suffices.*

Proof (i) without (ii). A multivariate ecology model can minimize LOPO RMSE (observational generalization) while the env×phenotype interaction required by a thermal-selection story fails under experiment [Stillwell et al., 2007, 2008].

(ii) without (i). Interaction tests at observed sites do not certify transfer to new coordinates; structured holdout remains required for macroecological claims.

The tests are logically independent. Both are needed for mechanistic certification in the observe-then-experiment tradition [Lande and Arnold, 1983, Kingsolver et al., 2001]. ■

9 Supplementary tables: *Stator limbatus* vignette

Source: author reanalysis of Stillwell et al. [2007]. Bootstrap paired LOPO comparisons use $B = 10{,}000$ resamples; protocol matches the main-text Methods section and `scripts/stator_figure2_composite.py`.

Table 2: Table S1. Bootstrap paired LOPO comparisons ($B = 10{,}000$). Negative ΔRMSE favors first model.

Comparison	n	ΔRMSE	95% CI	P(first lower)
Ecology − latitude	93	−0.002	[−0.011, +0.006]	0.68
Ecology − temperature	93	−0.009	[−0.017, 0.000]	0.97
Latitude − temperature	95	−0.007	[−0.018, +0.006]	0.85

Table 3: Table S2. Sample composition: AIC on full survey vs LOPO on matched set. Ecology covariates missing for two populations on full survey.

Sample	Model	n	AIC	AIC rank	LOPO RMSE	LOPO rank
Full survey	Latitude	95	−161.1	1	0.105	2
Full survey	Temperature	95	−147.5	3	0.111	3
Full survey	Ecology	93†	−156.7†	—	0.104†	1
Matched	Latitude	93	−155.7	2	0.106	3
Matched	Ecology	93	−156.7	1	0.104	1

†Ecology fit uses populations with complete seed-size data.

10 Simulation protocol

Each of 500 independent replicates draws $P = 95$ populations. Script: `scripts/simulation_lopo_confounding.py`.

Table 4: Table S3. Semi-synthetic simulation summary (500 replicates × 95 populations). Rates characterize the stated generating process (Section 10), not ecology generally.

Criterion	Best model	Rate
Strongest in-sample $\|r\|$	Latitude (proxy)	100%
Lowest LOPO RMSE	Ecology multivariate	84%
$\|r\|$ rank $\neq$ LOPO rank	—	84%

Table 5: Table S4. Estimand mismatch on *S. limbatus* [Stillwell et al., 2007]: individual 10-fold cross-validation vs. LOPO on population means ($n = 93$). Both schemes rank ecology above temperature, but individual k-fold reports substantially higher RMSE because it answers E_1, not E_2 (see main-text Figure 2B).

CV scheme	Ecology RMSE	Temperature RMSE
10-fold individual	0.150	0.158
LOPO population mean	0.104	0.112

Generating process.

- Ecology predictors (seed size, moisture PC1, seasonality PC1) are correlated with latitude ($r \approx 0.55$).
- Mean annual temperature is weakly correlated with latitude ($r \approx -0.35$) and does not enter the true mean-size function.
- Trait PC1 (population mean body-size axis): $0.35 \cdot \text{seed} + 0.30 \cdot \text{moist} + 0.25 \cdot \text{season} + \mathcal{N}(0, \sigma^2)$.

Candidate models: latitude-only, temperature-only, ecology multivariate, temperature+ecology. **Evaluation:** LOPO via leave-one-out OLS

on standardized population-level predictors; in-sample $|r|$ and AIC computed on the same replicate.

The 84% rank-disagreement rate in Table S3 and Figure S1 is a property of this parameterization (correlation structure, effect sizes, noise level), not a claim about geographic clines in general.

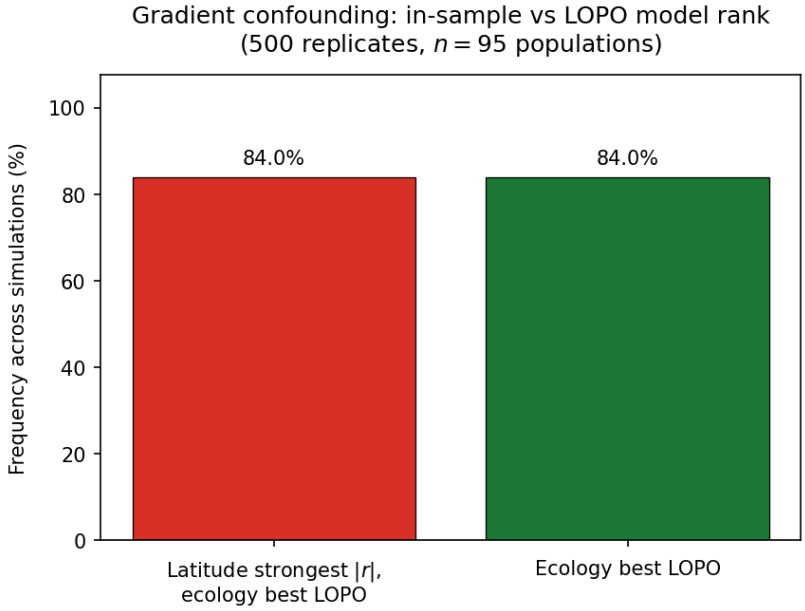

Figure 1: Semi-synthetic validation (500 replicates; Figure S1): under the stated generating process, the strongest in-sample univariate correlate (latitude) mis-ranks models relative to LOPO in 84% of replicates (main-text Figure 1).

References

Martin Arjovsky, Léon Bottou, Ishan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

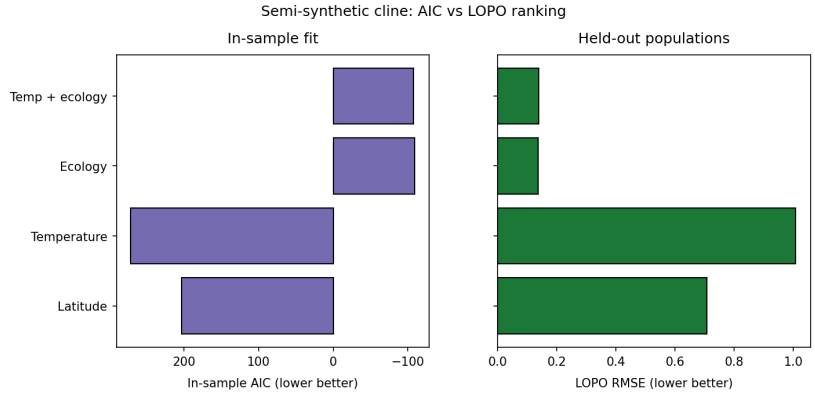

Figure 2: Single-replicate example (Figure S2): in-sample AIC and $|r|$ rankings versus LOPO RMSE rankings under gradient confounding.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

Joel G. Kingsolver, David W. Pfennig, and Maria R. Servedio. Migration, local adaptation and the evolution of plasticity. *Trends in Ecology & Evolution*, 16(9):472–473, 2001. doi: 10.1016/S0169-5347(01)02215-2.

Pang Wei Koh, Shiori Sagawa, Henrik Xiao, Marvin Zhang, Aaron Bordes, Florian Clavera, Ian Mendes, Quoc V. Le, Ruslan Salakhutdinov, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pages 5637–5664, 2021.

Russell Lande and Stevan J. Arnold. The measurement of selection on correlated characters. *Evolution*, 37(6):1210–1226, 1983. doi: 10.2307/2408782.

Jorge Soberón and Yukinori Nakamura. Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 104(suppl. 1):12644–12650, 2007. doi: 10.1073/pnas.0700527104.

R. Craig Stillwell. Are latitudinal clines in body size adaptive? *Oikos*, 119 (1):184–190, 2010. doi: 10.1111/j.1600-0706.2009.17725.x.

R. Craig Stillwell, Charles W. Fox, and Timothy A. Mousseau. Geographic variation in body size and sexual size dimorphism of a seed-feeding beetle. *The American Naturalist*, 170(3):358–369, 2007. doi: 10.1086/519396.

R. Craig Stillwell, Charles W. Fox, and Timothy A. Mousseau. Selection does not favor large body size at lower temperature in a seed-feeding beetle. *Evolution*, 62(10):2534–2544, 2008. doi: 10.1111/j.1558-5646.2008.00474.x.

Sara Via and Russell Lande. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*, 39(3):505–522, 1985. doi: 10.2307/2408649.