MULTI-FIDELITY LEARNING UNDER UPDATE BUDGETS: CROSS-FIDELITY SPECTRAL FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Updating a low-fidelity predictor with scarce high-fidelity labels can alter behavior that should be retained. We study multi-fidelity residual learning under an explicit budget on changes to the stored predictor. For fixed kernel models, the weighted norm of the cross-fidelity residual-to-anchor map gives the exact worst-case residual-normalized update, termed the scale-separation index (SSI). We construct a spectral filter that attenuates high-gain residual directions, decreases design-anchor SSI monotonically, and minimizes it among permutations of a fixed shrinkage spectrum. Validation selects the filter and blend step subject to an independently audited update cap. For a finite candidate family, we derive a selection-uniform population bound under independent and identically distributed audit sampling and bounded updates. Across 20 seeds at a normalized cap of 0.20, filtering reduces root mean squared error relative to matched quadratic regularization by 0.590% on QM7b and 0.241% on MF-PCBA, and outperforms equal-spectrum scrambled controls. The filter uses more of the available update budget than matched controls. Stronger external predictors remain more accurate in several comparisons, often with larger changes to the stored predictor. A separate quadratic-control study satisfies all 160 anchor certificates but meets the same cap on only 133 test blocks. Together, these results establish a mechanism for directional update control and delineate the role of anchor coverage in transferring local guarantees.

1 INTRODUCTION

Multi-fidelity learning combines abundant inexpensive observations with a small number of accurate labels. Co-kriging, Δ-learning, and nonlinear autoregressive models all exploit this structure (Kennedy & O'Hagan, 2000; Ramakrishnan et al., 2015; Perdikaris et al., 2017; Meng & Karniadakis, 2020). Their usual objective is high-fidelity prediction error. In some applications, however, the existing predictor also has value: it may have been validated over a broad operating region or encode a response that users want to preserve. A correction trained on a few high-fidelity examples can change that response far from those examples. Predictive error alone does not describe this change.

This leads to a constrained learning problem: improve high-fidelity accuracy while limiting changes to the pre-update function on a specified input distribution. The input distribution and budget express what should be preserved. They are application choices, since retaining a biased low-fidelity response can itself be costly. We measure change relative to the stored predictor, rather than relative to low-fidelity labels, so the measure is zero whenever no update occurs.

For a fixed kernel residual model, the change has a simple structure. The cross-kernel between reference inputs and fine-training inputs maps the fitted residual to the reference outputs. We call the weighted norm of this map the *scale-separation index* (SSI). It gives an exact worst-case coefficient for the normalized update. More importantly, its singular vectors identify which residual directions can produce large changes on the reference inputs. Our method attenuates these directions and selects the remaining update size using validation data. Figure 1 illustrates the constrained selection problem.

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.

The connection to continual learning is functional: both problems ask how a new fit changes a previously learned predictor. Related theory studies neural tangent kernel (NTK) task overlap (Doan et al., 2021; Liu et al., 2026) and forgetting in linear models (Evron et al., 2022; Lin et al., 2023). Orthogonal updates (Farajtabar et al., 2020), structural regularization (Li et al., 2023; 2026), functional penalties (Titsias et al., 2020), and Jacobian sketches (Heckel, 2022) offer ways to preserve prior behavior. Here the object of control is the residual-to-anchor map itself, including its gain and ridge damping. This makes the certificate directly computable for a candidate update. Its kernel interpretation applies when features remain fixed; task-dependent changes in a trained NTK require a new map (Liu et al., 2025).

We make three contributions.

1. **An exact update coefficient.** Weighted SSI characterizes the largest residual-normalized change on a specified anchor measure and is invariant to weight-preserving replication. Independent audit samples give a selection-uniform population bound for a finite family of bounded updates.
2. **A directional control method.** A spectral filter reduces design SSI monotonically and is optimal among assignments of the same shrinkage spectrum to the design operator's singular directions. A validation rule chooses the filter and blend step under the update budget.
3. **A test of the control mechanism.** Twenty-seed QM7b and Multi Fidelity PubChem BioAssay (MF-PCBA) experiments compare spectral alignment with quadratic, trace-matched, and equal-spectrum controls. External predictors, anchor-transfer tests, and neural audits quantify the associated accuracy and preservation trade-offs.

2 Problem: Accuracy, Update, and Concordance

Let f_0 be the predictor stored before the residual-fitting stage, and write the corrected predictor as $f_1 = f_0 + g$. Given high-fidelity training inputs X_f and labels Y_h, the residual target is

$$R = Y_h - f_0(X_f). \tag{1}$$

This target is shared by Δ-learning and sequential kernel regression with the same second-stage kernel and ridge parameter; Appendix A.1 gives the pointwise identity. Any calibration used to construct f_0 is completed before fitting g.

On held-out inputs X, we distinguish three measurements:

$$E_h(f_1; X) = \mathrm{RMSE}(f_1(X), Y_h(X)), \tag{2}$$

$$U(f_1, f_0; X) = \mathrm{RMSE}(f_1(X), f_0(X)), \tag{3}$$

$$C_\ell(f_1; X) = \mathrm{RMSE}(f_1(X), Y_\ell(X)). \tag{4}$$

Here $\mathrm{RMSE}(a, b)$ is the square root of the mean squared coordinate difference on the specified inputs. Equation (2) measures target accuracy, (3) measures the change caused by the update, and (4) measures agreement with low-fidelity labels. An unchanged model has $U = 0$ even when $C_\ell > 0$. Thus agreement with low labels and preservation of the learned function answer different questions.

We express the update budget relative to the root mean square (RMS) training residual, $s = \mathrm{RMS}(R) > 0$:

$$\bar{U}(f_1, f_0; X) = U(f_1, f_0; X)/s. \tag{5}$$

The scale s in (5) is fixed using training data, so validation and test updates use the same units. If $R = 0$, no residual correction is needed; the analysis below assumes $s > 0$. The learning objective is to minimize E_h subject to $\bar{U} \le \tau$ on a specified measure, where τ is the allowed change. We report both error and achieved update because two predictors below the same cap may use very different amounts of the budget.

3 Cross-Fidelity Kernel Certificate

Fixed-kernel setting. The formal result assumes that f_0 is fixed and the residual stage is a lazy/kernel ridge update (Jacot et al., 2018; Lee et al., 2019). Let X_c be a declared set of coarse

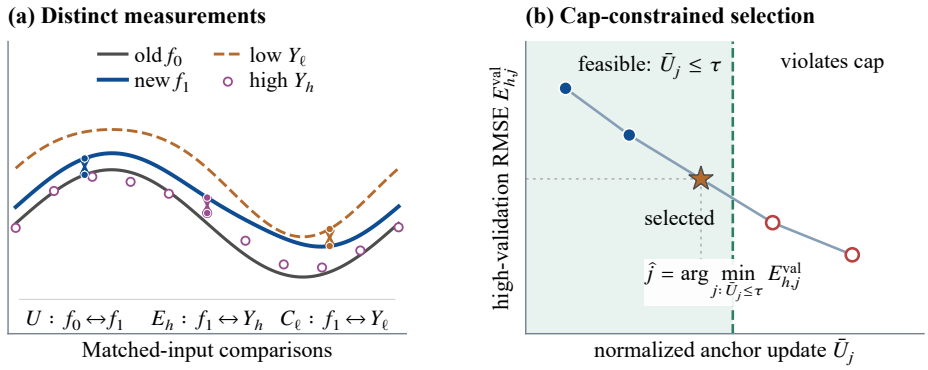

Figure 1: Update measurement and constrained selection. (a) At matched evaluation inputs, the brackets identify update magnitude $U = |f_1 - f_0|$, high-fidelity error $E_h = |f_1 - Y_h|$, and low-label concordance $C_\ell = |f_1 - Y_\ell|$; the three distances are not interchangeable. (b) Candidate updates are ranked by high-fidelity validation RMSE, but only those satisfying the specified normalized anchor-update cap are eligible. Filled blue circles are feasible, open red circles violate the cap, and the star marks the constrained validation minimizer.

anchors and X_f the fine residual training set. Positive diagonal W_c, W_f encode normalized empirical or quadrature weights, with uniform weights recovering RMS norms, and $\|v\|_W = (v^\top W v)^{1/2}$. The matrix 2-norm below is the induced spectral norm. With $\lambda > 0$,

$$\delta f(X_c) = f_1(X_c) - f_0(X_c) = \underbrace{\Theta_{cf}(\Theta_{ff} + \lambda I)^{-1}}_{A} R. \tag{6}$$

Definition 1 (Weighted cross-fidelity interference)**.**

$$\mathrm{SSI}_W = \left\| W_c^{1/2} A W_f^{-1/2} \right\|_2. \tag{7}$$

Theorem 1 (Exact anchor-update coefficient)**.** *For positive diagonal W_c, W_f,*

$$\sup_{R \neq 0} \frac{\|f_1(X_c) - f_0(X_c)\|_{W_c}}{\|R\|_{W_f}} = \mathrm{SSI}_W. \tag{8}$$

Replacing an anchor of weight w by m identical copies of weight w/m leaves SSI_W unchanged.

The maximizing residual is aligned with a leading right singular vector of the weighted map. Weight-preserving replication leaves the empirical measure unchanged for identical rows with conserved total weight, so such duplicates do not increase the certificate. The proof and equality case are in Appendix A.

Proposition 1 (Monotone and equal-spectrum optimal design-SSI filter)**.** *Let A_d map fine residuals to design-anchor updates and let W_d contain positive design-anchor weights. The weighted design operator is $\widetilde{A}_d = W_d^{1/2} A_d W_f^{-1/2} = U \Sigma V^\top$, with nonzero singular values $\sigma_1 \ge \cdots \ge \sigma_r > 0$, with V completed to an orthonormal basis. Set $Q_\gamma = V \,\mathrm{diag}(q_i) V^\top$, where $q_i = \{1 + \gamma(\sigma_i/\sigma_1)^2\}^{-1}$ for $i \le r$ and $q_i = 1$ otherwise. Then, for $\gamma \ge 0$,*

$$\left\| \widetilde{A}_d Q_\gamma \right\|_2 = \max_{i \le r} \frac{\sigma_i}{1 + \gamma(\sigma_i/\sigma_1)^2} \tag{9}$$

is non-increasing and $Q_0 = I$. Take $Q_\gamma = I$ if $\widetilde{A}_d = 0$. More generally, for any fixed shrinkage spectrum $0 \le \bar{q}_1 \le \cdots \le \bar{q}_r$, assigning $\bar{q}_i$ to σ_i minimizes $\max_i \sigma_i \bar{q}_{\pi(i)}$ over all permutations π.

The proof is in Appendix A.2. The permutation result explains the choice to apply the strongest attenuation to the largest-gain directions. It minimizes the design-operator norm among assignments of a fixed shrinkage spectrum; prediction error still depends on the realized residual and validation labels.

From anchors to a population. On the specified anchors, Theorem 1 gives $\|\delta f(X_c)\|_{W_c} \leq \mathrm{SSI}_W \|R\|_{W_f}$ for every residual. SSI therefore includes both the cross-fidelity geometry and the damping from the ridge solve. Whitened overlap, which measures subspace angles, does not retain this gain. Extending the anchor statement to unseen inputs requires sampling assumptions. We use an independent audit set to estimate the squared normalized update for each candidate.

Theorem 2 (Selection-uniform audit-to-population margin)**.** *Condition on the fine-training data and design anchors. Let P be the retained-input distribution, and let $h_1, \ldots, h_M$ be all update candidates eligible for audit-based selection, including filter–step pairs. They are fixed before audit sampling; let $s > 0$ be the frozen residual scale. Let independent and identically distributed (iid) audit anchors $Z_1, \ldots, Z_n \sim P$ be independent of candidate fitting. Suppose $0 \leq (h_j(x)/s)^2 \leq B$ for every j and every x in the support of P. Define*

$$\widehat{q}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_j(Z_i)}{s}\right)^2, \qquad q_j = \mathbb{E}_{X\sim P}\left(\frac{h_j(X)}{s}\right)^2. \tag{10}$$

Then, with probability at least $1-\delta$, simultaneously for all j,

$$q_j \leq \widehat{q}_j + B\sqrt{\frac{\log(M/\delta)}{2n}}. \tag{11}$$

The simultaneous event also covers a candidate selected using the audit sample. In particular, requiring the empirical squared-update budget to be at most $\tau^2 - B\sqrt{\log(M/\delta)/(2n)}$ implies population normalized RMS update at most τ whenever the right-hand side is nonnegative.

The bound follows from Hoeffding's inequality and a union bound over the candidate family (Appendix A). It accounts for selection among all M candidates, including choices made using the audit sample. Design anchors may shape the filter, but audit anchors must be independent of its fitting. In the bounded-update experiments we enforce the pointwise condition with $\Pi_B(h)(x) = \mathrm{clip}(h(x), -s\sqrt{B}, s\sqrt{B})$ before blending. Pointwise clipping cannot increase the update norm, so the linear certificate remains an upper bound for the clipped predictor.

4 Direct Spectral Control of Residual Updates

Filtering the residual. Let A_X map the fine-training residual to the base kernel update on inputs X, including the ridge term in the kernel solve. With $z_0 = W_f^{1/2} R$ and the design operator $\widetilde{A}_d$, our filter has the following quadratic interpretation when $\sigma_1 > 0$:

$$z_\gamma = \arg\min_z \left\{ \|z - z_0\|_2^2 + \frac{\gamma}{\sigma_1^2}\left\|\widetilde{A}_d z\right\|_2^2 \right\} = Q_\gamma z_0. \tag{12}$$

The first term keeps the residual close to its training value; the second penalizes its effect on the design anchors. Differentiating (12) gives $(I + \gamma \widetilde{A}_d^\top \widetilde{A}_d/\sigma_1^2) z_\gamma = z_0$, whose spectral solution is the filter in Proposition 1. Prediction uses

$$\delta f_\gamma(X) = A_X W_f^{-1/2} Q_\gamma W_f^{1/2} R. \tag{13}$$

At $\gamma = 0$, (13) recovers the base update. As γ grows, high-gain directions shrink most, while design-null-space directions are unchanged. This differs from a scalar step, which shrinks every direction equally. Null-space preservation is local to the design inputs: these directions can still affect other inputs.

Why the observed update and SSI differ. Write $c_i = v_i^\top z_0$ in the design singular basis. Before clipping, the squared update on design anchors is

$$\|\delta f_\gamma(X_d)\|_{W_d}^2 = \sum_{i=1}^{r} (\sigma_i q_i)^2 c_i^2, \qquad \mathrm{SSI}_{d,\gamma} = \max_{i \leq r} \sigma_i q_i. \tag{14}$$

Equation (14) separates the operator gain from the residual's alignment with that operator. A large SSI need not produce a large observed change if the residual has little energy in the leading directions. Conversely, reducing SSI can allow a larger blend step under the same certificate cap. The selected step may then improve accuracy while increasing the realized update. This distinction is central to the experiments.

Certifying the step. For $f_{\gamma,\alpha} = f_0 + \alpha \Pi_B(\delta f_\gamma)$, with $0 \le \alpha \le 1$ and the identity projection when clipping is not used, compute the linear pre-clipping SSI on audit inputs. Denote this upper-bound coefficient by c_γ and the unit-step normalized update after any clipping by u_γ. The feasible step interval is

$$0 \le \alpha \le \alpha_{\max,\gamma} = \min\{1, \tau/u_\gamma, \tau/c_\gamma\}. \tag{15}$$

A zero denominator imposes no restriction. The observed-update variant omits the last term. Recomputing c_γ on audit inputs is essential: monotonic design SSI alone gives no bound for a different anchor set. The optional population margin uses Theorem 2 in place of the nominal empirical budget.

Validation selection. The candidate family and its γ and α grids are fixed before evaluation. Inner high-label validation selects the filter strength among audit-feasible candidates; outer validation selects the blend step in (15). Design inputs determine the filter, audit inputs check its change, and validation labels determine predictive quality. Test labels are used after selection. A zero step returns f_0, so every block has a feasible candidate. For a fixed update direction, the continuous validation optimum is the least-squares step clipped to the feasible interval (Proposition 3); the experiments use a common finite grid across methods.

Mechanism controls. The equal-spectrum scrambled filter permutes the same shrinkage values among residual directions. It preserves the amount of shrinkage and the candidate count, isolating the effect of alignment. The anchor-output quadratic comparator instead penalizes the residual model's squared predictions on design anchors. Its fitted map can rotate as well as shrink as the penalty increases, so its SSI need not be monotone. Appendix A.2 gives its objective and solution. Trace-matched isotropic and training-covariance penalties further separate anchor geometry from overall regularization strength.

Implementation and cost. The direct method requires one factorization of the $n_f \times n_f$ fine Gram matrix and one singular value decomposition (SVD) of the $n_d \times n_f$ weighted design map. Each additional γ reuses these factors and changes only the spectral weights before prediction. For L candidate strengths, this replaces L quadratic-regression solves with one spectral construction and L reweightings. The advantage depends on dimensions: when the residual feature count p is much smaller than n_f, solving the $p \times p$ quadratic systems can be cheaper. Tables 17 and 18 report the measured time and storage grid. Jacobian sketches provide a lower-rank option (Heckel, 2022); their approximation error is measured separately in Appendix C.3.

5 Experiments

The experiments test the local certificate, the benefit of spectral alignment, and the resulting accuracy–update trade-off. We distinguish two protocols: the nested protocol separates filter and step selection, while the pooled QM7b gate protocol examines regularization and transfer using a single validation split. Results from these protocols are reported separately.

5.1 Data and Evaluation Protocols

Benchmarks. QM7b contains 7,211 molecules with quantum-property estimates from several electronic-structure approximations (Montavon et al., 2013). We study Zerner's intermediate neglect of differential overlap (ZINDO) as low fidelity and the Green's-function/screened-interaction approximation (GW) as high fidelity for the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO). High-label budgets are 64, 128, 256, and 512. MF-PCBA assay 1445 pairs a cheap single-dose screen with dose-response labels, using budgets 32, 64, 128, and 256 (Buterez et al., 2023). Both benchmarks use 20 seeds. Appendix B.2 specifies the representations, calibration, and data splits.

Nested mechanism and external comparisons. Design and audit inputs are disjoint. Inner validation selects γ and outer validation selects α. The mechanism comparison uses the specified tight cap $\tau = 0.20$, seven strengths, and the same candidate grids for direct, quadratic, and scrambled

filters. Success requires a reduction in audit SSI and a relative-RMSE degradation of at most 1%, assessed by the paired 95% interval. The full decision rules appear in Appendix B.3.

Sensitivity experiments vary random Fourier feature (RFF) dimension $p \in \{128, 256, 512\}$, ridge $\rho \in \{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$, and design-anchor count $n_d \in \{64, 128, 256\}$ across eleven configurations per benchmark. The external comparison fixes $p = 256$, $\rho = 10^{-4}$, $n_d = 256$, and a 21-point blend grid. It includes exact nonlinear autoregressive Gaussian process (NARGP), first-order autoregressive (AR1) discrepancy regression, and a composite multi-fidelity neural network (MFNN) (Kennedy & O'Hagan, 2000; Perdikaris et al., 2017; Meng & Karniadakis, 2020). Every model is evaluated at its native step and under the common cap. Copper (Cu) potential and trained-network Jacobian experiments examine transfer to other representations.

Pooled gate and frontier comparison. The QM7b gate experiment uses 384 RFFs, a single 80/20 high-label training/validation split, and $\tau = 0.50$. Quadratic regularization is compared with the fixed $\gamma = 0$ residual under the same SSI gate, as well as low-only, high-only, and NARGP predictors under the observed-update cap. A frontier comparison chooses one γ curve by validation hypervolume and evaluates its eleven α points on test against the eleven-point $\gamma = 0$ curve. This keeps curve sizes equal. The nested all-γ envelope is reported separately because it contains the baseline by construction.

Uncertainty. We report high-fidelity RMSE, achieved normalized update, and the fraction of test blocks satisfying the cap. Confidence intervals resample seeds with their task–budget blocks kept together, using 20,000 bootstrap draws for the pooled QM7b study and 10,000 for the robustness and external comparisons. The sensitivity intervals are pointwise, without a multiple-comparison correction. All candidate tuning uses validation data. Detailed grids, tie rules, and estimands are in Appendix B.3.

5.2 Local Certificate and Anchor Weighting

Controlled linear-map tests attain the singular-vector equality to floating-point precision, and every measured anchor update lies below its SSI bound. Across 50 anchor-density trials, weighted SSI has a mean coefficient of variation of 0.0057, compared with 0.478 for the unweighted norm. Normalizing the anchor measure therefore removes most of the sensitivity to replication. In trained-network stress tests, however, weighted SSI has only Spearman correlation 0.163 with measured interference. A fixed-kernel coefficient can describe its own residual update exactly while being a weak proxy for changes made by feature-learning dynamics.

5.3 Spectral Alignment Under a Tight Budget

Table 1 compares direct filtering with quadratic regularization and equal-spectrum scrambled directions at $\tau = 0.20$. The direct filter lowers both prediction error and audit SSI in all four paired comparisons. The error reductions are modest: 0.590% and 0.575% on QM7b, and 0.241% and 0.131% on MF-PCBA. Because the scrambled control has the same shrinkage eigenvalues, these differences measure the benefit of assigning shrinkage to the design operator's high-gain directions.

All four operator/non-inferiority tests pass. Design SSI is monotone in every fitted block, with mean gaps from the exact SVD expression of 7.06×10^{-16} on QM7b and 4.08×10^{-16} on MF-PCBA. Table 2 gives the absolute operating points. At $\tau = 0.20$, the direct filter uses more of the allowed update than these matched controls, consistent with the larger feasible step in (15); all selected models in this comparison satisfy the test cap. The rounded MF-PCBA errors are close, so the paired contrasts in Table 1 are more informative than a rank based on the rounded means.

The effect depends on the allowed change. At larger MF-PCBA caps, some scrambled-control intervals include zero, and the QM7b LUMO-512 subgroup is heterogeneous. Leverage-selected anchors leave test feasibility unchanged but increase relative RMSE by 0.233% [0.100, 0.365] compared with random external anchors. The full threshold and normalization results are in Tables 13 and 12.

Table 1: Direct-spectral comparisons at $\tau = 0.20$. Entries are direct filter minus matched control, shown as seed-cluster means with paired 95% intervals over 20 seeds. Negative relative RMSE and Δ SSI favor the direct filter; positive $\Delta\bar{U}$ indicates a larger realized update. Bold cells have intervals entirely below zero.

Benchmark	Matched control	Relative ΔE_h (%)	Δ SSI	$\Delta\bar{U}$
QM7b	quadratic surrogate	$\mathbf{-0.590}$ $[\mathbf{-0.763}, \mathbf{-0.436}]$	$\mathbf{-0.340}$ $[\mathbf{-0.412}, \mathbf{-0.269}]$	0.011 $[0.009, 0.014]$
QM7b	scrambled directions	$\mathbf{-0.575}$ $[\mathbf{-0.690}, \mathbf{-0.459}]$	$\mathbf{-0.433}$ $[\mathbf{-0.529}, \mathbf{-0.341}]$	0.016 $[0.013, 0.019]$
MF-PCBA	quadratic surrogate	$\mathbf{-0.241}$ $[\mathbf{-0.394}, \mathbf{-0.099}]$	$\mathbf{-0.405}$ $[\mathbf{-0.530}, \mathbf{-0.251}]$	0.004 $[-0.001, 0.010]$
MF-PCBA	scrambled directions	$\mathbf{-0.131}$ $[\mathbf{-0.252}, \mathbf{-0.020}]$	$\mathbf{-0.219}$ $[\mathbf{-0.375}, \mathbf{-0.051}]$	0.009 $[0.003, 0.015]$

Table 2: Absolute performance at $\tau = 0.20$. Cells give seed-cluster means and 95% intervals. Every selected model satisfies the test cap. Audit SSI is the unit-step coefficient; the reported update includes the selected blend.

Method	High RMSE	Audit SSI	Normalized update
QM7b			
fixed $\gamma = 0$	0.955 $[0.948, 0.962]$	2.166 $[2.137, 2.195]$	0.050 $[0.048, 0.052]$
quadratic surrogate	0.953 $[0.946, 0.960]$	1.728 $[1.688, 1.766]$	0.063 $[0.061, 0.065]$
direct spectral	0.948 $[0.940, 0.955]$	1.388 $[1.307, 1.472]$	0.074 $[0.072, 0.077]$
scrambled directions	0.953 $[0.946, 0.960]$	1.821 $[1.757, 1.890]$	0.058 $[0.056, 0.060]$
MF-PCBA			
fixed $\gamma = 0$	0.254 $[0.248, 0.261]$	1.898 $[1.863, 1.939]$	0.021 $[0.017, 0.026]$
quadratic surrogate	0.254 $[0.248, 0.261]$	1.393 $[1.348, 1.440]$	0.037 $[0.032, 0.042]$
direct spectral	0.253 $[0.248, 0.260]$	0.988 $[0.858, 1.139]$	0.042 $[0.034, 0.049]$
scrambled directions	0.253 $[0.248, 0.260]$	1.207 $[1.103, 1.297]$	0.033 $[0.027, 0.039]$

5.4 Robustness and Strong Predictive Baselines

Across the eleven configurations on each benchmark, direct filtering has lower mean RMSE than quadratic regularization in 22/22 comparisons and lower mean RMSE than scrambled directions in 21/22. All pointwise quadratic-contrast intervals and 16/22 pointwise scrambled-contrast intervals lie below zero. The reversed point estimate occurs on MF-PCBA at $p = 512$, $\rho = 3\times10^{-4}$. Table 14 reports the configuration-level estimates.

Table 3 compares the direct filter with external predictors under the same cap. On QM7b, exact NARGP, AR1, and MFNN are more accurate, while the direct filter changes the stored predictor less. On MF-PCBA, AR1 is more accurate; the NARGP and MFNN error contrasts include zero. Thus the same admissible budget can produce different operating points. Both error and realized update are needed to interpret these comparisons.

The native and capped predictions in Table 16 show how much each family relies on blending. This is particularly important for MFNN, whose native predictions are unstable in the small-data implementation used here.

On Cu potentials, direct filtering has 3.681% higher RMSE than the quadratic control but 5.875% lower RMSE than scrambled directions, accompanied by opposite update trade-offs. The trained-

Table 3: External comparisons at $\tau = 0.20$: direct filter minus control, with paired seed-cluster 95% intervals (20 seeds). Negative values mean lower error or update, respectively. Native and capped predictions appear in Table 16.

Benchmark	External control	Relative ΔE_h (%)	$\Delta \bar{U}$
QM7b	exact NARGP	11.590 [11.083, 12.095]	−0.102 [−0.107, −0.098]
QM7b	AR1 discrepancy	11.807 [11.477, 12.178]	−0.106 [−0.111, −0.102]
QM7b	composite MFNN	6.301 [5.560, 7.016]	−0.065 [−0.076, −0.052]
MF-PCBA	exact NARGP	0.105 [−0.321, 0.562]	−0.037 [−0.053, −0.021]
MF-PCBA	AR1 discrepancy	1.887 [1.377, 2.389]	−0.087 [−0.103, −0.069]
MF-PCBA	composite MFNN	−0.159 [−0.459, 0.163]	0.011 [−0.005, 0.026]

network Jacobian experiment is nearly tied with both controls: the mean error differences are +0.088% and +0.091%. The numerical intervals and implementation details are in Table 15 and Appendix C.2.

5.5 Gate Strength and the Accuracy–Update Frontier

The pooled QM7b experiment isolates the effect of quadratic regularization under a common SSI gate (Figure 2). Relative to the fixed residual gate, regularization reduces high-RMSE by 0.009 [0.005, 0.013] and increases normalized update by 0.122 [0.107, 0.137]. Mean operator gain falls from 1.937 to 1.110, while the selected blend step rises from 0.223 to 0.518. A nonzero penalty is selected in 132/160 blocks. The regularized gate therefore permits more of the learned correction to pass, improving accuracy at the cost of a larger realized change.

Against the unchanged low-only predictor, the gate reduces RMSE by 0.101 [0.095, 0.107], winning 156/160 blocks. Compared with selection using only the observed-update cap, the SSI gate reduces normalized update by 0.147 [0.132, 0.162] and raises RMSE by 0.060 [0.055, 0.066]. These two comparisons quantify the benefit of updating and the cost of a worst-case constraint.

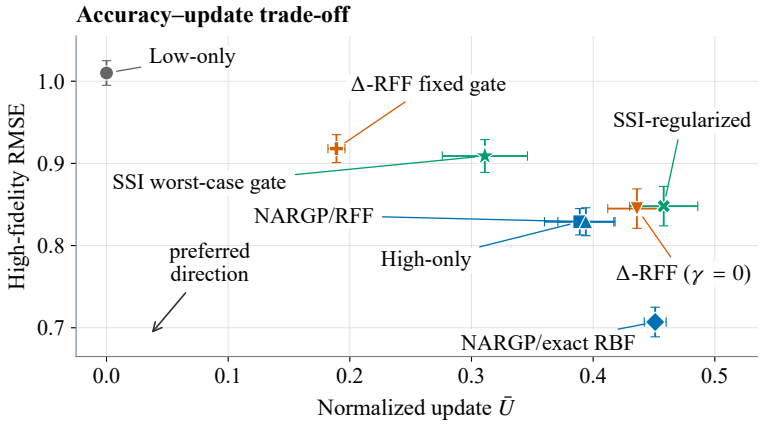

Figure 2: Accuracy–update trade-off in the pooled QM7b gate comparison. Points are cluster means and horizontal/vertical bars are cluster standard deviations over 20 seeds, two target pairs, and four high-label budgets. Lower values on both axes are preferred. Every nontrivial method uses the shared anchor cap $\bar{U} \leq 0.50$.

The equal-sized frontier comparison gives a different result: test hypervolume changes by −2.61% [-3.97, -1.48], with 1 positive, 47 negative, and 112 tied blocks. The corresponding RMSE and normalized-update changes are 0.003 [0.001, 0.006] and 0.022 [0.015, 0.029]. Pooling all γ curves gives a 1.81% [1.73, 1.88] larger envelope, but that envelope includes the baseline curve. The validation-selected equal-sized comparison is the relevant test of whether regularization improves the frontier.

In the nested protocol at $\tau = 0.50$, quadratic-minus-fixed RMSE is 0.0006 [-0.0022, 0.0041] on QM7b and 0.0008 [0.0002, 0.0014] on MF-PCBA. Neither joint accuracy/operator criterion is met, and trace-matched isotropic regularization is more accurate on QM7b. Operator suppression alone therefore does not determine which penalty predicts best.

5.6 Predictive Alternatives in the Pooled Comparison

Validation selects exact NARGP as the strongest feasible baseline in 131 blocks, NARGP/RFF in 15, high-only in 13, and Δ-RFF in one. The quadratic SSI model has 0.137 [0.127, 0.146] higher RMSE than this selection, with normalized-update difference 0.003 [-0.010, 0.014]. It wins 0/160 error comparisons and 64/160 update comparisons. The complete absolute and task–budget results are in Tables 4, 6, and 7.

5.7 Transfer of the Update Budget

All 160 pooled QM7b blocks satisfy the nominal anchor certificate. On test molecules, the same numerical cap holds in 126/160 blocks for observed-update SSI selection and 133/160 for the worst-case gate. Gate feasibility decreases from 40/40 at $n_h = 64$ to 39/40, 34/40, and 20/40 at budgets 128, 256, and 512. The larger-budget failures coincide with larger selected steps. They show that a cap enforced on the measured anchors need not hold on held-out inputs, especially when validation selects a larger step.

In the disjoint-audit nested experiments, the nominal quadratic SSI gate is already test-feasible in every block on both benchmarks. The finite-family margin leaves feasibility unchanged and adds 0.0151 [0.0134, 0.0167] to QM7b RMSE. Both specified margin-success criteria fail. Under its iid and bounded-update conditions the bound remains valid, but its conservative allowance brings no measured benefit at these operating points; its practical value depends on the audit sample size, update bound, and candidate count.

6 Discussion

The filter is useful when operator gain constrains an otherwise validation-preferred step. It then permits a larger correction without exceeding the certificate budget. If the unfiltered step is already feasible, attenuating its residual can instead remove useful predictive information. The value of the filter therefore depends on both the budget and the residual direction, consistent with the threshold-dependent results.

Using the method requires a meaningful retained input measure and update budget. A biased stored model calls for a larger allowed change or a different retained region. If preservation is unnecessary, unconstrained high-fidelity error is the relevant objective.

Finite anchors can miss regions where the update is large. The population bound addresses sampling uncertainty under iid audits and bounded updates, but not distribution shift. Evidence is limited to two QM7b targets, one assay, and narrow Cu and neural tests. The weak SSI–interference correlation in trained networks and unstable native MFNN predictions warrant broader feature-learning experiments. Appendix C.3 gives practical diagnostic checks.

An immediate direction is to apply the filter to a stronger residual model and compare it with scalar blending at matched observed update. For NARGP or AR1, this requires accounting for the fitted correlation model in the full residual-to-update map.

7 Conclusion

Cross-fidelity kernel interference turns a functional update budget into a computable operator constraint. Its singular directions yield a spectral filter with monotone design SSI and fixed-spectrum permutation optimality. Experiments show small but consistent tight-budget improvements over matched quadratic and scrambled controls, together with clear trade-offs against strong external predictors. The resulting approach connects a local guarantee to a practical validation rule and makes the choice between accuracy and preserving the stored function explicit.

Artificial Intelligence (AI) Use Statement

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

Ethics Statement

The experiments use public molecular and simulation data without personal information. Applications must choose the retained input region and update budget using domain knowledge, since preserving an inaccurate reference predictor can perpetuate its errors.

Reproducibility Statement

Appendices A and B specify proofs, splits, candidate grids, selection rules, and estimands. Appendix B.3 reports computational resources. The recorded outcomes include every scheduled seed and a feasible $\alpha = 0$ option; update measurements use the stored predictor for each method.

References

David Buterez, Jon Paul Janet, Steven J. Kiddle, and Pietro Liò. MF-PCBA: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *Journal of Chemical Information and Modeling*, 63(9):2667–2678, 2023. doi: 10.1021/acs.jcim.2c01569.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 2021.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 2022. URL https://proceedings.mlr.press/v178/evron22a.html.

Mehrdad Farajtabar, Navid Azizan, Alexander Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Reinhard Heckel. Provable continual learning via sketched jacobian approximations. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10448–10470. PMLR, 2022. URL https://proceedings.mlr.press/v151/heckel22a.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Marc C. Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. doi: 10.1093/biomet/87.1.1.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 513–533. PMLR, 2023. URL https://proceedings.mlr.press/v232/li23b.html.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Memory–statistics tradeoff in continual learning with structural regularization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qfEqXJnlB4.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100. PMLR, 2023. URL https://proceedings.mlr.press/v202/lin23f.html.

Jingren Liu, Zhong Ji, YunLong Yu, Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. Parameter-efficient fine-tuning for continual learning: A neural tangent kernel perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8297–8314, 2026. doi: 10.1109/TPAMI.2026.3670952.

Yuzhi Liu, Zixuan Chen, Zirui Zhang, Yufei Liu, and Giulia Lanzillotta. Reactivation: Empirical NTK dynamics under task shifts. *arXiv preprint arXiv:2507.16039*, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *Journal of Computational Physics*, 401:109020, 2020. doi: 10.1016/j.jcp.2019.109020.

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15: 095003, 2013. doi: 10.1088/1367-2630/15/9/095003.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017. doi: 10.1098/rspa.2016.0751.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxCzeHFDB.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

A Theoretical Foundations

The proofs use the notation of Sections 2 and 4.

A.1 Residual Updates and Anchor Certificates

Residual identity.

Proposition 2 (Residual identity)**.** *For a fixed first-stage predictor, Δ-learning and two-task sequential kernel regression use the same residual target. If their second stages use the same kernel and ridge parameter, their predictions coincide pointwise.*

Let a fixed first-stage predictor be f_0 and let the second-stage kernel be K. Delta learning fits the target $R = Y_h - f_0(X_f)$ and returns

$$f_\Delta(x) = f_0(x) + K(x, X_f)(K_{ff} + \lambda I)^{-1} R. \tag{16}$$

Sequential kernel regression presents the second task after the first task has produced f_0. Its residual target at X_f is therefore the same R. Using the same kernel and ridge parameter gives the same expression pointwise.

Exact weighted anchor-update coefficient. Write $A = \Theta_{cf}(\Theta_{ff} + \lambda I)^{-1}$ and $z = W_f^{1/2} R$. Then

$$\begin{aligned} \frac{\|AR\|_{W_c}}{\|R\|_{W_f}} &= \frac{\left\| W_c^{1/2} A W_f^{-1/2} z \right\|_2}{\|z\|_2} \\ &\le \left\| W_c^{1/2} A W_f^{-1/2} \right\|_2 . \end{aligned} \tag{17}$$

Equality is attained when z is a top right singular vector of the weighted operator. Hence the supremum is exactly SSI_W.

For replication invariance, suppose row $a_i^\top$ of A has coarse weight w_i and is replaced by m identical rows with weights w_i/m. Their total contribution to the right Gram matrix is

$$\sum_{k=1}^{m} (w_i/m) a_i a_i^\top = w_i a_i a_i^\top . \tag{18}$$

The singular values are unchanged. This result concerns exact duplication. A new support location can change the empirical measure and SSI.

A.2 Spectral Filtering, Transfer, and Quadratic Control

Exact monotone design-SSI filter. Let $\widetilde{A}_d = U \Sigma V^\top$ and write $Q_\gamma = V \operatorname{diag}(q_i) V^\top$ as in Proposition 1. Orthogonal invariance gives

$$\left\| \widetilde{A}_d Q_\gamma \right\|_2 = \|\Sigma \operatorname{diag}(q_i)\|_2 = \max_{i \le r} \sigma_i q_i. \tag{19}$$

Each $\sigma_i q_i = \sigma_i / \big(1 + \gamma(\sigma_i/\sigma_1)^2\big)$ is non-increasing for $\gamma \ge 0$, proving the claim. At $\gamma = 0$, every $q_i = 1$ and the matched residual update is recovered exactly. Null-space coordinates do not affect the design output and are left unchanged. This proof concerns the pre-projection linear operator. The shared pointwise clipping is non-expansive, but it does not make the post-projection update a linear map.

For the equal-spectrum statement, consider two singular directions with $\sigma_i \ge \sigma_j$ and two shrinkages $a \le b$. The oppositely ordered assignment satisfies

$$\max\{\sigma_i a, \sigma_j b\} \le \max\{\sigma_i b, \sigma_j a\} = \sigma_i b. \tag{20}$$

Starting from any permutation with an inversion, swapping that inverted pair therefore cannot increase the largest product. Repeating the exchange removes all inversions and yields $\min_\pi \max_i \sigma_i \bar{q}_{\pi(i)} = \max_i \sigma_i \bar{q}_i$ for $\sigma_1 \ge \cdots \ge \sigma_r$ and $\bar{q}_1 \le \cdots \le \bar{q}_r$. Because the proposed q_i have exactly this opposite order, the direct filter is bottleneck-optimal among all assignments with the same shrinkage eigenvalues.

Selection-uniform audit-to-population margin. Condition on the fine-training data and design anchors. This conditioning fixes the M candidate functions before the independent audit anchors are drawn. For candidate j, write

$$\ell_j(x) = \left(\frac{h_j(x)}{s}\right)^2, \qquad \widehat{q}_j = \frac{1}{n}\sum_{i=1}^{n} \ell_j(Z_i), \qquad q_j = \mathbb{E}_P \ell_j(X). \tag{21}$$

Under $0 \le \ell_j \le B$, the one-sided Hoeffding inequality gives

$$\Pr(q_j - \widehat{q}_j \ge \epsilon) \le \exp\left(-\frac{2n\epsilon^2}{B^2}\right). \tag{22}$$

Taking a union bound over $j = 1, \dots, M$ and setting $\epsilon = B\sqrt{\log(M/\delta)/(2n)}$ makes the failure probability at most δ. The event holds simultaneously, so it also holds for any index selected using the audit sample. If $\widehat{q}_{\widehat{j}} \le \tau^2 - \epsilon$, then $q_{\widehat{j}} \le \tau^2$ and hence $\sqrt{q_{\widehat{j}}} \le \tau$.

Clipping every candidate at $s\sqrt{B}$ enforces the pointwise bound used above. The finite family must include every candidate eligible for audit-based selection. Conditioning on a single candidate chosen after seeing audit inputs would not justify the same union-bound argument.

SSI-regularized linear feature map. For residual features Φ_f on n_f fine-training inputs and Φ_c on n_c design anchors, the quadratic comparator in Section 4 fits

$$w_\gamma = \arg\min_w \frac{1}{n_f}\|\Phi_f w - R\|_2^2 + \frac{\gamma}{n_c}\|\Phi_c w\|_2^2 + \rho\|w\|_2^2. \tag{23}$$

Here $\gamma \ge 0$ penalizes anchor update and $\rho > 0$ is the ridge coefficient. The stationary condition gives

$$w_\gamma = \left(\frac{\Phi_f^\top \Phi_f}{n_f} + \frac{\gamma \Phi_c^\top \Phi_c}{n_c} + \rho I\right)^{-1} \frac{\Phi_f^\top R}{n_f}. \tag{24}$$

Multiplication by Φ_c yields the residual-to-anchor map A_γ. Consequently the same induced-norm proof applies for every γ.

B Experimental Protocol and Evaluation

B.1 Constrained Selection Protocol

Closed-form shared-cap paths.

Proposition 3 (Closed-form shared-cap path selection)**.** *Fix a candidate direction $d = \widetilde{f} - f_0$, validation residual $e = Y_h - f_0$, and positive validation weights W_v. Let u be the residual-normalized audit update of d at unit step and let c be any valid unit-step certificate coefficient. Omit the certificate constraint when none is available. Put $\alpha_{\max} = \min\{1, \tau/u, \tau/c\}$, where a zero denominator imposes no restriction. If $\|d\|_{W_v} > 0$, the minimizer of $\|e - \alpha d\|_{W_v}^2$ over $[0, \alpha_{\max}]$ is*

$$\alpha^* = \mathrm{clip}\left(\frac{\langle e, d\rangle_{W_v}}{\|d\|_{W_v}^2}, 0, \alpha_{\max}\right). \tag{25}$$

If $\|d\|_{W_v} = 0$, every feasible step has the same validation loss and the tie rule selects $\alpha^* = 0$.

For a fixed direction d, both the observed audit update and any homogeneous operator certificate scale linearly with nonnegative α. Their constraints therefore reduce to $0 \le \alpha \le \alpha_{\max}$. Expanding the validation objective gives

$$\|e - \alpha d\|_{W_v}^2 = \|e\|_{W_v}^2 - 2\alpha\langle e, d\rangle_{W_v} + \alpha^2 \|d\|_{W_v}^2. \tag{26}$$

For $\|d\|_{W_v} > 0$, the unconstrained quadratic minimizer is $\langle e, d\rangle_{W_v} / \|d\|_{W_v}^2$. Projection onto the closed interval $[0, \alpha_{\max}]$ gives the stated clipped solution. When $\|d\|_{W_v} = 0$, every feasible α has the same validation loss and the tie rule selects zero. The argument applies unchanged to Gaussian process, neural, and SSI candidates.

Grid selection and frontier calculation. For the pooled QM7b gate protocol, the blend grid is $\{0, 0.1, \ldots, 1\}$ and the penalty grid is $\{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Among candidates satisfying the applicable constraints, ties in validation RMSE are broken by smaller anchor update and then smaller α. The equal-sized frontier calculation selects one penalty curve by validation hypervolume and evaluates that curve on test with the same reference point as the fixed-residual curve. The union of all penalty curves is recorded as a descriptive envelope.

Nested selection and penalty controls. For each seed, the low-stage predictor excluded audit and test examples, while the high-label budget was split into residual-training, inner-validation, and outer-validation subsets. MF-PCBA used an additional calibration subset to map the cheap single-dose (SD) assay to dose-response (DR) units and define the stored pre-update function. Design anchors could enter the residual penalty, whereas a disjoint audit-anchor set was excluded from fitting.

We fit the Cartesian family of three penalty geometries and seven γ values using a shared RFF state, ridge, and candidate count within each block. Every residual candidate received the same pointwise projection Π_B before audit or high-validation selection, enforcing the bounded-loss premise of Theorem 2; non-expansiveness makes the unprojected linear operator norm a conservative upper bound. Inner validation selected γ among audit-feasible candidates, outer validation then selected α, and the test labels were evaluated once. We repeated this nested selection for $\tau \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ under the nominal gate, fixed margins 0.05 and 0.10, and the finite-family squared-loss margin, using the same selection rule for each setting.

The penalty controls are

$$P_{\mathrm{SSI}} = \frac{\Phi_d^\top \Phi_d}{n_d}, \qquad P_{\mathrm{iso}} = \frac{\mathrm{tr}(P_{\mathrm{SSI}})}{p} I, \qquad P_{\mathrm{train}} = \frac{\mathrm{tr}(P_{\mathrm{SSI}})}{\mathrm{tr}(\Phi_f^\top \Phi_f / n_f)} \frac{\Phi_f^\top \Phi_f}{n_f}. \tag{27}$$

Thus the isotropic and training-covariance controls match the SSI penalty's trace but remove its design-anchor eigenspaces. The residual feature model is shared across these penalties. The fixed-$\gamma = 0$ family keeps the same audit gate and α grid.

Direct spectral mechanism protocol. The spectral-filter comparison uses nested selection with the nominal gate, seven γ values, and five update caps. The squared-loss and fixed-margin comparisons use the quadratic family. These are separate comparisons with their own specified decision criteria below.

The nested spectral runs use the penalty grid listed above, with $\alpha \in \{i/20 : i = 0, \ldots, 20\}$ on QM7b and $\alpha \in \{i/10 : i = 0, \ldots, 10\}$ on MF-PCBA. The stored configurations use $B = 1$ and $\delta = 0.05$. For the finite-family margin, M counts all penalty–step pairs eligible for selection within the fitted family.

Every fitted block compares fixed $\gamma = 0$, the anchor-output quadratic surrogate, direct spectral filtering, an equal-spectrum scrambled-direction filter, and trace-matched isotropic and training-covariance penalties. The scrambled filter uses exactly the direct filter's shrinkage eigenvalues and a seed-frozen random assignment to residual directions. The QM7b anchor comparison uses leverage-selected, in-training random, external random, and external farthest-first inputs. Three update normalizations—training- residual RMS, high-test standard deviation, and pre-update-test standard deviation—are reported for the same selected models. None chooses a model.

B.2 Data and Baseline Implementations

QM7b source and targets. The public QM7b file provides Coulomb matrices $X \in \mathbb{R}^{7211 \times 23 \times 23}$ and 14 target columns (Montavon et al., 2013). The same-property target pairs are specified in Section 5.1.

Features. Each molecule is represented by the upper triangle of its Coulomb matrix plus sorted row-norm summaries, giving 299 inputs. Centering and scaling statistics are fitted on the seed-specific low-fidelity training pool and then applied to high, anchor, validation, and test subsets.

Pooled gate split. For each seed, a single permutation produces 5,000 low-training molecules and 1,500 disjoint test molecules. The high-fidelity subset is nested in the low pool, with budgets 64, 128, 256, and 512. The first 80% of a high budget fits the update and the remaining 20% selects candidates. The next 768 low-pool molecules, disjoint from the high subset, are coarse anchors. All methods use the identical indices for a given seed and budget.

Independent MF-PCBA protocol. MF-PCBA AID 1445 provides a primary-screen SD assay as low fidelity and a DR assay as high fidelity. The latter uses the negative base-10 logarithm of the molar half-maximal inhibitory concentration, denoted $p\mathrm{XC}_{50}$. We use the benchmark's official fixed DR train/test files and exclude every DR compound from the low-only anchor pool. For each of 20 seeds, the high-label budget $n_h \in \{32, 64, 128, 256\}$ is partitioned into four disjoint subsets. The first 25% (at least eight compounds) fits an affine SD-to-DR calibration and thereby defines the pre-update function. Approximately 45% trains the residual, 15% selects γ, and the remainder selects α. The public test labels are used once after selection. The official validation inputs serve as audit anchors, while their high-fidelity labels are not used for selection.

Molecular inputs use simplified molecular-input line-entry system (SMILES) strings, encoded as signed hashed character 1–3-gram features. The low-only pool determines feature standardization. The experiment uses 5,000 low examples, 512 input features, 256 random Fourier features, 256 design anchors, and all 65 disjoint audit inputs from the official validation split without using their DR labels. Every molecule in a DR split is excluded from the low-only fitting and design pools. Because the cheap SD value is assumed queryable for a candidate molecule, the calibrated SD response is available on design, audit, validation, and test inputs. The pre-update predictor is the calibrated SD response, and the residual model learns its remaining discrepancy from the DR target.

Baseline implementation. The paired controls share residual features and training indices. Predictive baselines use their own validation-selected regularization, with the common anchor constraint and blend grid applied after fitting. Quadratic SSI denotes the objective in (23); the direct and scrambled filters use the same base map and differ only in the assignment of spectral weights. Neural continual-learning baselines are specified in Appendix C.3.

In the pooled gate protocol, the RFF dimension is 384. Low ridge is 10^{-5}. High, residual, and random-feature NARGP ridge is 10^{-4}. Exact NARGP uses the validation-selected ridge grid $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ and RBF lengthscales $\{0.5, 1, 2, 4\}$ times the median-distance estimate. These grids tune each predictor on the shared validation split.

B.3 Statistical Evaluation and Computational Resources

The pooled QM7b bootstrap averages each paired contrast over the eight property–budget blocks within a seed, then resamples the 20 seed means. The percentile interval summarizes the resulting distribution. This treats seeds, rather than their dependent blocks, as the replication units. Reported win counts are descriptive.

The gate comparison estimates quadratic-minus-fixed differences in RMSE, normalized update, weighted SSI, selected α, and bound utilization. The predictive comparison uses the baseline chosen by validation RMSE. The frontier comparison uses the validation-selected equal-sized curves defined above. Feasibility is the fraction of selected models whose test update is no larger than the specified cap.

For the quadratic-control decision tests, the reference setting is the external farthest-first design and nominal gate. Matched-mechanism tests use $\tau = 0.50$. Safety-margin tests use $\tau = 0.50$ on QM7b and $\tau = 0.60$ on MF-PCBA because its official validation audit has only 65 inputs. Four predeclared criteria are evaluated. H1 requires negative upper 95% intervals for both high-RMSE and weighted-SSI differences on QM7b. H4 repeats this criterion on MF-PCBA. H2 requires a positive lower interval for test-feasibility gain on QM7b and a high-RMSE-cost upper interval no larger than 0.02. H5 repeats H2 on MF-PCBA at $\tau = 0.60$. Both safety criteria additionally require the pointwise bound in every block. The other penalties, caps, fixed margins, and anchor strategies are sensitivity analyses. Decision labels below identify the criteria used in the result files.

The direct-filter tests use six specified criteria. H6 requires the design SSI sequence to be non-increasing in every fitted block and the exact-SVD identity gap to be at most 10^{-8}. H7 and H10,

on QM7b and MF-PCBA respectively at the predeclared tight budget $\tau = 0.20$, require direct filtering to reduce disjoint-audit SSI relative to the quadratic surrogate while the upper 95% interval for relative high-RMSE degradation is at most 1%. H8 and H11 repeat that joint operator/non-inferiority requirement against the equal-spectrum scrambled control. H9 asks whether external leverage anchors improve held-out feasibility over random external anchors at $\tau = 0.50$ without more than 1% relative-RMSE degradation. The error margin is assessed jointly with the operator criterion.

Computational resources. Robustness and external comparisons used 64-bit CPU arithmetic. A 501-run batch used four cores and 12 gibibytes (GiB) for 8 h 39 min 27 s; 50 further runs and analysis used eight cores and 24 GiB for 8 min 58 s. Timing comparisons in Tables 17–18 exclude queue waiting and do not represent end-to-end training.

C Additional Results and Stress Tests

C.1 Gate Comparisons and Ablations

The exact numerical values underlying Figure 2, together with the complete direct-spectral sweeps, are retained below.

Table 4: QM7b gate results pooled over 20 seed clusters, two target pairs, and four high-label budgets. Values are cluster mean±standard deviation. Every nontrivial method uses the same anchor cap $\bar{U} \leq 0.50$.

Method	High RMSE	Normalized update
Low-only	1.010 ± 0.015	0.000 ± 0.000
High-only + shared cap	0.829 ± 0.016	0.389 ± 0.029
NARGP/RFF + shared cap	0.829 ± 0.017	0.394 ± 0.023
NARGP/exact RBF + shared cap	0.707 ± 0.018	0.451 ± 0.009
Δ-RFF ($\gamma = 0$)	0.845 ± 0.024	0.436 ± 0.024
Δ-RFF fixed worst-case gate	0.918 ± 0.017	0.189 ± 0.007
SSI-regularized	0.848 ± 0.024	0.458 ± 0.028
SSI worst-case gate	0.909 ± 0.020	0.311 ± 0.035

Pooled gate results by task and budget.

Penalty selection and task-level contrasts. The quadratic gate selects $\gamma = 0$: 28, $\gamma = 0.01$: 37, $\gamma = 0.1$: 53, $\gamma = 1$: 38, $\gamma = 10$: 3, and $\gamma = 100$: 1. Mean bound utilization rises from 0.420 to 0.449, a paired change of 0.029 [0.018, 0.039]. The paired operator-gain change is -0.827 [-0.909, -0.747], and the selected-step change is 0.295 [0.247, 0.344]. Table 5 breaks down the gate comparison by property and budget; Tables 6 and 7 provide the corresponding method-level values.

Table 5: Where the matched-cap mechanism is effective and where held-out feasibility fails. Negative ΔE_h favors the SSI gate. Positive $\Delta\bar{U}$ and $\Delta\alpha$ mean a larger realized and selected update. Cellwise intervals are descriptive and are not multiplicity-adjusted. The pooled contrast is the primary inference.

Pair	n_h	$\Delta\alpha$	ΔE_h	$\Delta\bar{U}$	Gate test-feasible
HOMO	64	0.110 $[0.025, 0.225]$	0.012 $[0.001, 0.027]$	0.014 $[0.004, 0.028]$	20/20
HOMO	128	0.185 $[0.105, 0.285]$	-0.006 $[-0.014, 0.005]$	0.092 $[0.063, 0.127]$	19/20
HOMO	256	0.300 $[0.200, 0.415]$	-0.008 $[-0.016, 0.000]$	0.141 $[0.102, 0.186]$	17/20
HOMO	512	0.485 $[0.355, 0.615]$	-0.033 $[-0.039, -0.028]$	0.215 $[0.167, 0.262]$	12/20

Continued on next page

Table 5 (continued)

Pair	n_h	$\Delta\alpha$	ΔE_h	$\Delta\bar{U}$	Gate test-feasible
LUMO	64	0.170 $[0.075, 0.285]$	0.020 $[0.004, 0.039]$	0.040 $[0.018, 0.067]$	20/20
LUMO	128	0.210 $[0.110, 0.335]$	-0.005 $[-0.014, 0.007]$	0.085 $[0.053, 0.124]$	20/20
LUMO	256	0.340 $[0.230, 0.470]$	-0.010 $[-0.019, 0.001]$	0.141 $[0.102, 0.186]$	17/20
LUMO	512	0.560 $[0.440, 0.680]$	-0.045 $[-0.050, -0.039]$	0.249 $[0.205, 0.292]$	8/20

Table 6: Complete pair–budget high-fidelity RMSE over 20 paired seeds (mean±standard deviation). Low/High are the single-fidelity references. N/RFF and N/exact are random-feature and exact-kernel NARGP. Δ is the matched $\gamma = 0$ control.

Pair	Method	$n_h = 64$	$n_h = 128$	$n_h = 256$	$n_h = 512$
HOMO	Low	1.077 ± 0.020	1.077 ± 0.020	1.077 ± 0.020	1.077 ± 0.020
HOMO	High	0.856 ± 0.044	0.879 ± 0.032	0.945 ± 0.033	1.060 ± 0.029
HOMO	N/RFF	0.851 ± 0.037	0.886 ± 0.028	0.948 ± 0.026	1.048 ± 0.029
HOMO	N/exact	0.805 ± 0.043	0.797 ± 0.025	0.803 ± 0.020	0.799 ± 0.021
HOMO	Δ	0.941 ± 0.051	0.932 ± 0.036	0.928 ± 0.027	0.898 ± 0.025
HOMO	Δ-gate	0.977 ± 0.034	0.995 ± 0.024	0.989 ± 0.023	0.982 ± 0.022
HOMO	SSI	0.951 ± 0.054	0.936 ± 0.038	0.929 ± 0.027	0.901 ± 0.028
HOMO	Gate	0.989 ± 0.040	0.989 ± 0.029	0.980 ± 0.032	0.948 ± 0.024
LUMO	Low	0.943 ± 0.020	0.943 ± 0.020	0.943 ± 0.020	0.943 ± 0.020
LUMO	High	0.640 ± 0.039	0.662 ± 0.026	0.711 ± 0.034	0.877 ± 0.026
LUMO	N/RFF	0.648 ± 0.038	0.659 ± 0.034	0.709 ± 0.035	0.884 ± 0.034
LUMO	N/exact	0.617 ± 0.050	0.628 ± 0.041	0.606 ± 0.029	0.604 ± 0.033
LUMO	Δ	0.786 ± 0.037	0.776 ± 0.033	0.761 ± 0.029	0.738 ± 0.025
LUMO	Δ-gate	0.850 ± 0.024	0.859 ± 0.021	0.849 ± 0.022	0.844 ± 0.021
LUMO	SSI	0.789 ± 0.037	0.785 ± 0.054	0.763 ± 0.036	0.733 ± 0.025
LUMO	Gate	0.869 ± 0.048	0.855 ± 0.036	0.840 ± 0.034	0.799 ± 0.024

Table 7: Complete pair–budget normalized functional update over 20 paired seeds (mean±standard deviation). Low/High are the single-fidelity references. N/RFF and N/exact are random-feature and exact-kernel NARGP. Δ is the matched $\gamma = 0$ control.

Pair	Method	$n_h = 64$	$n_h = 128$	$n_h = 256$	$n_h = 512$
HOMO	Low	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
HOMO	High	0.438 ± 0.076	0.431 ± 0.049	0.380 ± 0.085	0.199 ± 0.148
HOMO	N/RFF	0.433 ± 0.030	0.419 ± 0.048	0.406 ± 0.080	0.252 ± 0.110
HOMO	N/exact	0.456 ± 0.028	0.461 ± 0.030	0.448 ± 0.022	0.445 ± 0.017
HOMO	Δ	0.383 ± 0.121	0.421 ± 0.054	0.437 ± 0.053	0.461 ± 0.013
HOMO	Δ-gate	0.222 ± 0.042	0.178 ± 0.009	0.190 ± 0.009	0.184 ± 0.005
HOMO	SSI	0.391 ± 0.122	0.430 ± 0.048	0.451 ± 0.057	0.493 ± 0.035
HOMO	Gate	0.236 ± 0.054	0.270 ± 0.075	0.331 ± 0.095	0.400 ± 0.109
LUMO	Low	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
LUMO	High	0.452 ± 0.031	0.441 ± 0.032	0.449 ± 0.045	0.323 ± 0.099
LUMO	N/RFF	0.432 ± 0.035	0.443 ± 0.037	0.453 ± 0.042	0.315 ± 0.122
LUMO	N/exact	0.447 ± 0.038	0.440 ± 0.037	0.457 ± 0.033	0.455 ± 0.040
LUMO	Δ	0.467 ± 0.021	0.426 ± 0.061	0.446 ± 0.026	0.449 ± 0.022
LUMO	Δ-gate	0.209 ± 0.019	0.169 ± 0.011	0.182 ± 0.007	0.178 ± 0.008
LUMO	SSI	0.479 ± 0.014	0.444 ± 0.067	0.472 ± 0.035	0.506 ± 0.032
LUMO	Gate	0.249 ± 0.060	0.254 ± 0.083	0.323 ± 0.099	0.427 ± 0.105

Quadratic-penalty ablations. Table 8 gives the absolute outcomes for the nested quadratic-control comparison at $\tau = 0.50$.

Table 8: Absolute performance under the nominal $\tau = 0.50$ gate.

Benchmark	Method	High-fidelity RMSE	Test feasibility
QM7b	fixed $\gamma = 0$	0.912 [0.904, 0.920]	1.000 [1.000, 1.000]
QM7b	SSI	0.913 [0.904, 0.922]	1.000 [1.000, 1.000]
QM7b	isotropic	0.834 [0.825, 0.842]	1.000 [1.000, 1.000]
QM7b	training covariance	0.905 [0.896, 0.913]	1.000 [1.000, 1.000]
QM7b	low only	0.981 [0.974, 0.987]	1.000 [1.000, 1.000]
QM7b	high-only RFF	0.794 [0.786, 0.802]	0.869 [0.812, 0.919]
QM7b	NARGP/RFF	0.797 [0.790, 0.804]	0.881 [0.831, 0.931]
MF-PCBA	fixed $\gamma = 0$	0.252 [0.247, 0.259]	1.000 [1.000, 1.000]
MF-PCBA	SSI	0.253 [0.248, 0.260]	1.000 [1.000, 1.000]
MF-PCBA	isotropic	0.251 [0.246, 0.258]	1.000 [1.000, 1.000]
MF-PCBA	training covariance	0.252 [0.247, 0.259]	1.000 [1.000, 1.000]
MF-PCBA	low only	0.254 [0.249, 0.262]	1.000 [1.000, 1.000]
MF-PCBA	high-only RFF	0.254 [0.249, 0.260]	1.000 [1.000, 1.000]
MF-PCBA	NARGP/RFF	0.255 [0.249, 0.262]	0.988 [0.963, 1.000]

Thresholds, anchor design, and margins. Tables 9, 10, and 11 vary the update cap, design anchors, and safety margin, respectively. All paired effects use the same selection rule within each setting.

Table 9: Complete threshold ablation for the quadratic SSI surrogate minus the matched fixed-$\gamma = 0$ control. Negative ΔE_h favors the surrogate. Intervals resample seeds with all within-seed blocks attached.

Benchmark	τ	ΔE_h	$\Delta \bar{U}$	$\Delta \alpha$	Δ SSI
QM7b	0.20	-0.0023 $[-0.0035, -0.0009]$	0.013 [0.011, 0.015]	0.032 [0.028, 0.036]	-0.438 $[-0.495, -0.387]$
QM7b	0.30	0.0020 [0.0005, 0.0037]	0.007 [0.004, 0.009]	0.021 [0.016, 0.027]	-0.233 $[-0.286, -0.184]$
QM7b	0.40	0.0008 $[-0.0014, 0.0034]$	0.012 [0.009, 0.015]	0.032 [0.026, 0.038]	-0.243 $[-0.296, -0.194]$
QM7b	0.50	0.0006 $[-0.0022, 0.0041]$	0.017 [0.012, 0.021]	0.043 [0.033, 0.052]	-0.298 $[-0.347, -0.252]$
QM7b	0.60	0.0034 [0.0005, 0.0071]	0.014 [0.009, 0.019]	0.043 [0.032, 0.054]	-0.307 $[-0.354, -0.262]$
MF-PCBA	0.20	-0.0001 $[-0.0004, 0.0002]$	0.016 [0.011, 0.020]	0.036 [0.025, 0.046]	-0.505 $[-0.568, -0.443]$
MF-PCBA	0.30	0.0007 [0.0001, 0.0014]	0.006 [0.001, 0.013]	0.028 [0.011, 0.045]	-0.390 $[-0.486, -0.291]$
MF-PCBA	0.40	0.0007 [0.0001, 0.0013]	0.009 [0.000, 0.018]	0.044 [0.020, 0.066]	-0.443 $[-0.529, -0.350]$
MF-PCBA	0.50	0.0008 [0.0002, 0.0014]	0.012 [0.000, 0.025]	0.055 [0.026, 0.084]	-0.444 $[-0.520, -0.365]$
MF-PCBA	0.60	0.0013 [0.0005, 0.0020]	0.004 $[-0.007, 0.015]$	0.042 [0.014, 0.071]	-0.411 $[-0.494, -0.325]$

Table 10: Anchor-design sensitivity at $\tau = 0.50$: SSI minus fixed $\gamma = 0$.

Design anchors	ΔE_h	$\Delta \bar{U}$	$\Delta \alpha$	Δ SSI
QM7b				
external farthest-first	0.0006 $[-0.0022, 0.0041]$	0.017 $[0.012, 0.021]$	0.043 $[0.033, 0.052]$	-0.298 $[-0.347, -0.252]$
external random	0.0003 $[-0.0021, 0.0030]$	0.017 $[0.013, 0.021]$	0.047 $[0.038, 0.057]$	-0.294 $[-0.342, -0.245]$
in-pool random	0.0045 $[0.0008, 0.0085]$	0.012 $[0.006, 0.017]$	0.039 $[0.030, 0.048]$	-0.298 $[-0.335, -0.257]$
MF-PCBA				
external farthest-first	0.0008 $[0.0002, 0.0014]$	0.012 $[0.000, 0.025]$	0.055 $[0.026, 0.084]$	-0.444 $[-0.520, -0.365]$
external random	0.0011 $[0.0005, 0.0017]$	0.004 $[-0.008, 0.016]$	0.034 $[0.005, 0.064]$	-0.432 $[-0.516, -0.345]$
in-pool random	0.0009 $[0.0003, 0.0016]$	0.001 $[-0.011, 0.012]$	0.026 $[0.000, 0.053]$	-0.402 $[-0.500, -0.305]$

Table 11: Safety-policy ablation relative to the nominal SSI gate. Positive feasibility change is better. Positive RMSE cost is worse.

Benchmark	Policy	Feasibility change	RMSE cost	Bound-valid blocks
QM7b	fixed margin 0.05	0.000 $[0.000, 0.000]$	0.0073 $[0.0057, 0.0087]$	100.0%
QM7b	fixed margin 0.10	0.000 $[0.000, 0.000]$	0.0122 $[0.0105, 0.0137]$	100.0%
QM7b	finite-family margin	0.000 $[0.000, 0.000]$	0.0151 $[0.0134, 0.0167]$	100.0%
MF-PCBA	fixed margin 0.05	0.000 $[0.000, 0.000]$	-0.0001 $[-0.0003, 0.0000]$	100.0%
MF-PCBA	fixed margin 0.10	0.000 $[0.000, 0.000]$	-0.0002 $[-0.0005, 0.0000]$	100.0%
MF-PCBA	finite-family margin	0.000 $[0.000, 0.000]$	-0.0001 $[-0.0006, 0.0004]$	100.0%

The specified quadratic-control and margin criteria H1, H2, H4, and H5 are not met. Their definitions are given in Appendix B.3.

C.2 Spectral Robustness and External Validity

Spectral-filter sensitivity. Table 12 rescales the same selected predictions using alternative denominators. Table 13 reports the threshold sweep, complementing the tight-cap comparison in Section 5.3.

Table 12: Normalization sensitivity for direct spectral filtering minus the anchor-output quadratic surrogate at $\tau = 0.20$. High-std and Pre-std use the standard deviations of high-fidelity labels and pre-update predictions, respectively.

Benchmark	Residual-RMS $\Delta \bar{U}$	High-std ΔU	Pre-std ΔU
QM7b	0.011 $[0.009, 0.014]$	0.019 $[0.014, 0.024]$	0.021 $[0.015, 0.026]$
MF-PCBA	0.004 $[-0.001, 0.010]$	0.001 $[-0.003, 0.006]$	0.005 $[-0.002, 0.011]$

Table 13: Direct spectral SSI against matched controls. The scrambled control has the same shrinkage eigenvalues but assigns them to mismatched residual directions.

Benchmark	τ	Control	Relative ΔE_h (%)	Δ SSI	$\Delta \bar{U}$
			-0.828	-0.778	0.025
QM7b	0.20	fixed $\gamma = 0$	$[-0.950, -0.703]$	$[-0.869, -0.682]$	$[0.022, 0.028]$
			-0.645	-0.616	0.025
QM7b	0.30	fixed $\gamma = 0$	$[-0.814, -0.461]$	$[-0.699, -0.529]$	$[0.020, 0.029]$
			-0.844	-0.382	0.018
QM7b	0.30	quadratic	$[-1.077, -0.622]$	$[-0.465, -0.301]$	$[0.014, 0.022]$
			-0.556	-0.419	0.021
QM7b	0.30	scrambled	$[-0.711, -0.400]$	$[-0.516, -0.323]$	$[0.016, 0.026]$
			-0.853	-0.600	0.032
QM7b	0.40	fixed $\gamma = 0$	$[-1.053, -0.632]$	$[-0.681, -0.514]$	$[0.028, 0.037]$
			-0.928	-0.357	0.021
QM7b	0.40	quadratic	$[-1.212, -0.686]$	$[-0.433, -0.280]$	$[0.017, 0.024]$
			-0.615	-0.396	0.024
QM7b	0.40	scrambled	$[-0.797, -0.410]$	$[-0.473, -0.316]$	$[0.019, 0.030]$
			-1.006	-0.606	0.040
QM7b	0.50	fixed $\gamma = 0$	$[-1.239, -0.744]$	$[-0.686, -0.526]$	$[0.035, 0.045]$
			-1.059	-0.309	0.024
QM7b	0.50	quadratic	$[-1.441, -0.727]$	$[-0.378, -0.238]$	$[0.019, 0.029]$
			-0.694	-0.389	0.030
QM7b	0.50	scrambled	$[-0.879, -0.486]$	$[-0.451, -0.325]$	$[0.025, 0.035]$
			-0.937	-0.587	0.045
QM7b	0.60	fixed $\gamma = 0$	$[-1.111, -0.761]$	$[-0.673, -0.494]$	$[0.038, 0.051]$
			-1.279	-0.280	0.031
QM7b	0.60	quadratic	$[-1.700, -0.931]$	$[-0.362, -0.195]$	$[0.025, 0.036]$
			-0.772	-0.370	0.036
QM7b	0.60	scrambled	$[-1.030, -0.534]$	$[-0.453, -0.281]$	$[0.028, 0.044]$
			-0.283	-0.910	0.020
MF-PCBA	0.20	fixed $\gamma = 0$	$[-0.435, -0.154]$	$[-1.036, -0.767]$	$[0.015, 0.027]$
			-0.202	-0.766	0.018
MF-PCBA	0.30	fixed $\gamma = 0$	$[-0.355, -0.055]$	$[-0.932, -0.596]$	$[0.011, 0.025]$
			-0.451	-0.376	0.012
MF-PCBA	0.30	quadratic	$[-0.709, -0.227]$	$[-0.551, -0.191]$	$[0.002, 0.021]$
			-0.157	-0.339	0.014
MF-PCBA	0.30	scrambled	$[-0.296, -0.018]$	$[-0.542, -0.127]$	$[0.006, 0.022]$
			-0.143	-0.769	0.015
MF-PCBA	0.40	fixed $\gamma = 0$	$[-0.288, -0.004]$	$[-0.915, -0.619]$	$[0.007, 0.022]$
			-0.409	-0.326	0.006
MF-PCBA	0.40	quadratic	$[-0.638, -0.194]$	$[-0.477, -0.158]$	$[-0.006, 0.018]$
			-0.078	-0.325	0.014
MF-PCBA	0.40	scrambled	$[-0.233, 0.091]$	$[-0.512, -0.122]$	$[0.004, 0.024]$
			-0.103	-0.739	0.017
MF-PCBA	0.50	fixed $\gamma = 0$	$[-0.267, 0.052]$	$[-0.886, -0.590]$	$[0.008, 0.024]$
			-0.400	-0.294	0.004
MF-PCBA	0.50	quadratic	$[-0.637, -0.184]$	$[-0.450, -0.127]$	$[-0.011, 0.018]$
			-0.110	-0.215	0.019
MF-PCBA	0.50	scrambled	$[-0.316, 0.103]$	$[-0.376, -0.041]$	$[0.008, 0.030]$
			-0.013	-0.738	0.011
MF-PCBA	0.60	fixed $\gamma = 0$	$[-0.205, 0.173]$	$[-0.895, -0.579]$	$[0.001, 0.021]$
			-0.489	-0.327	0.008
MF-PCBA	0.60	quadratic	$[-0.734, -0.266]$	$[-0.477, -0.161]$	$[-0.007, 0.021]$
			-0.076	-0.294	0.017
MF-PCBA	0.60	scrambled	$[-0.289, 0.150]$	$[-0.478, -0.093]$	$[0.005, 0.030]$

Robustness, external validity, and scaling. The eleven configurations in Section 5.1 consist of the nine (p, ρ) pairs at $n_d = 256$, plus $n_d = 64, 128$ at $p = 256$, $\rho = 10^{-4}$. External families share the stored predictor, high-label splits, audit inputs, and blend grid with the direct-filter comparison.

Exact NARGP uses the augmented input/low-prediction RBF kernel. AR1 fits a training-only affine low-to-high relation followed by an exact RBF discrepancy model. This is a co-kriging-style predictive baseline, not joint covariance maximum-likelihood estimation. Both select ridge and lengthscale on inner validation. The frozen-low MFNN has a linear low-prediction branch and two width-64 tanh layers; inner validation selects among three weight decays and early-stopped checkpoints (400-epoch cap, patience 50).

The Cu comparison uses two classical potentials. The ten-seed neural audit uses full-parameter Jacobians from a width-64 network, with 450 low-fidelity training steps and 300 fine-tuning steps. Cu and public-benchmark intervals use 20 seed clusters; the Jacobian intervals use 10. Each interval uses 10,000 bootstrap draws.

Table 14: Complete exploratory RFF-dimension, ridge, and design-anchor grid at $\tau = 0.20$, with 20 seeds per row. Negative relative RMSE differences favor the direct filter.

Benchmark	(p, ρ, n_d)	Direct–quadratic RMSE (%)	Direct–scrambled RMSE (%)
QM7b	128, 1e-4, 256	-0.431 $[-0.515, -0.343]$	-1.112 $[-1.211, -1.013]$
QM7b	128, 3e-4, 256	-0.356 $[-0.454, -0.252]$	-0.246 $[-0.362, -0.133]$
QM7b	128, 3e-5, 256	-0.472 $[-0.565, -0.378]$	-1.267 $[-1.342, -1.187]$
QM7b	256, 1e-4, 128	-0.266 $[-0.397, -0.140]$	-0.189 $[-0.315, -0.062]$
QM7b	256, 1e-4, 256	-0.590 $[-0.766, -0.435]$	-0.575 $[-0.689, -0.460]$
QM7b	256, 1e-4, 64	-0.434 $[-0.544, -0.330]$	-0.326 $[-0.407, -0.243]$
QM7b	256, 3e-4, 256	-0.397 $[-0.606, -0.216]$	-0.151 $[-0.268, -0.025]$
QM7b	256, 3e-5, 256	-0.670 $[-0.839, -0.511]$	-0.518 $[-0.657, -0.378]$
QM7b	512, 1e-4, 256	-0.438 $[-0.542, -0.342]$	-0.282 $[-0.378, -0.188]$
QM7b	512, 3e-4, 256	-0.726 $[-0.890, -0.570]$	-0.530 $[-0.635, -0.421]$
QM7b	512, 3e-5, 256	-0.690 $[-0.798, -0.589]$	-0.416 $[-0.504, -0.338]$
MF-PCBA	128, 1e-4, 256	-0.240 $[-0.435, -0.081]$	-0.200 $[-0.365, -0.063]$
MF-PCBA	128, 3e-4, 256	-0.219 $[-0.396, -0.070]$	-0.190 $[-0.339, -0.072]$
MF-PCBA	128, 3e-5, 256	-0.217 $[-0.394, -0.051]$	-0.136 $[-0.297, 0.010]$
MF-PCBA	256, 1e-4, 128	-0.323 $[-0.530, -0.144]$	-0.161 $[-0.331, -0.018]$
MF-PCBA	256, 1e-4, 256	-0.241 $[-0.390, -0.100]$	-0.131 $[-0.251, -0.015]$
MF-PCBA	256, 1e-4, 64	-0.175 $[-0.299, -0.066]$	-0.080 $[-0.213, 0.029]$
MF-PCBA	256, 3e-4, 256	-0.243 $[-0.379, -0.109]$	-0.096 $[-0.226, 0.034]$
MF-PCBA	256, 3e-5, 256	-0.278 $[-0.448, -0.119]$	-0.216 $[-0.371, -0.080]$

Continued on next page

Table 14 (continued)

Benchmark	(p, ρ, n_d)	Direct–quadratic RMSE (%)	Direct–scrambled RMSE (%)
MF-PCBA	512, 1e-4, 256	-0.217 $[-0.332, -0.109]$	-0.032 $[-0.121, 0.060]$
MF-PCBA	512, 3e-4, 256	-0.194 $[-0.322, -0.055]$	0.031 $[-0.066, 0.138]$
MF-PCBA	512, 3e-5, 256	-0.231 $[-0.337, -0.126]$	-0.077 $[-0.155, 0.004]$

Table 15: Transfer contrasts, defined as direct filter minus control. Cu uses 20 seed clusters and the neural Jacobian audit uses 10. Public-benchmark contrasts appear in Table 3. Values below 0.0005 in magnitude round to zero.

Audit	Control	Relative ΔE_h (%)	$\Delta \bar{U}$
Cu potentials	quadratic SSI	3.681 $[2.253, 5.290]$	-0.042 $[-0.055, -0.029]$
Cu potentials	scrambled directions	-5.875 $[-6.556, -5.160]$	0.066 $[0.058, 0.073]$
Neural Jacobian	quadratic SSI	0.088 $[0.000, 0.263]$	-0.001 $[-0.003, 0.000]$
Neural Jacobian	scrambled directions	0.091 $[0.000, 0.274]$	-0.001 $[-0.003, 0.000]$

Table 16: Native ($\alpha = 1$) and shared-cap external predictions. Cells show seed-cluster means and 95% intervals. RMSE is pooled within, not across, benchmarks.

External method	Native RMSE	Native $\bar{U}$	Capped RMSE	Capped $\bar{U}$
QM7b				
exact NARGP	0.600 [0.594, 0.606]	0.916 [0.901, 0.931]	0.851 [0.845, 0.857]	0.177 [0.173, 0.180]
AR1 discrepancy	0.595 [0.590, 0.601]	0.832 [0.820, 0.844]	0.849 [0.843, 0.855]	0.181 [0.178, 0.184]
composite MFNN	2.257 [1.936, 2.574]	2.590 [2.270, 2.902]	0.897 [0.889, 0.906]	0.139 [0.127, 0.151]
MF-PCBA				
exact NARGP	0.355 [0.351, 0.359]	1.152 [1.100, 1.207]	0.253 [0.247, 0.259]	0.079 [0.063, 0.093]
AR1 discrepancy	0.247 [0.244, 0.251]	0.428 [0.399, 0.460]	0.248 [0.243, 0.255]	0.128 [0.111, 0.143]
composite MFNN	2.267 [1.920, 2.615]	9.820 [8.149, 11.664]	0.254 [0.248, 0.261]	0.030 [0.017, 0.046]

Table 17: Complete compute grid, part 1. M is candidate count. Time sums separately measured median setup and all-candidate costs. It is not end-to-end training time. Storage is in mebibytes (MiB) and counts named matrices only, excluding allocator/workspace overhead.

n_f	n_c	p	M	Direct ms	Quadratic ms	Direct / quadratic MiB
64	64	128	7	4.214	1.970	0.281 / 0.626
64	64	128	31	90.425	6.404	0.281 / 0.626
64	64	256	7	1.853	8.174	0.406 / 2.252
64	64	256	31	3.062	31.476	0.406 / 2.252
64	256	128	7	3.997	1.868	0.656 / 0.813
64	256	128	31	5.166	6.870	0.656 / 0.813
64	256	256	7	4.423	9.584	0.969 / 2.627

Continued on next page

Table 17 (continued)

n_f	n_c	p	M	Direct ms	Quadratic ms	Direct / quadratic MiB
64	256	256	31	6.512	36.138	0.969 / 2.627
64	1024	128	7	36.472	2.483	2.156 / 1.563
64	1024	128	31	33.078	7.776	2.156 / 1.563
64	1024	256	7	30.922	13.480	3.219 / 4.127
64	1024	256	31	31.316	70.807	3.219 / 4.127
128	64	128	7	18.914	2.102	0.688 / 0.688
128	64	128	31	3.936	6.212	0.688 / 0.688
128	64	256	7	2.739	7.362	0.875 / 2.377
128	64	256	31	4.060	113.179	0.875 / 2.377
128	256	128	7	10.650	2.196	1.250 / 0.876
128	256	128	31	10.653	7.643	1.250 / 0.876
128	256	256	7	9.755	13.938	1.625 / 2.752
128	256	256	31	11.694	33.947	1.625 / 2.752
128	1024	128	7	37.072	2.458	3.500 / 1.626
128	1024	128	31	33.529	8.383	3.500 / 1.626
128	1024	256	7	27.897	9.086	4.625 / 4.252
128	1024	256	31	35.928	36.102	4.625 / 4.252

Table 18: Complete compute grid, part 2. M is candidate count. Time sums separately measured median setup and all-candidate costs. It is not end-to-end training time. Storage counts named matrices only, excluding allocator/workspace overhead.

n_f	n_c	p	M	Direct ms	Quadratic ms	Direct / quadratic MiB
256	64	128	7	6.889	1.806	2.062 / 0.813
256	64	128	31	7.978	7.347	2.062 / 0.813
256	64	256	7	6.459	8.200	2.375 / 2.627
256	64	256	31	8.777	38.013	2.375 / 2.627
256	256	128	7	18.573	1.892	3.000 / 1.001
256	256	128	31	27.727	6.352	3.000 / 1.001
256	256	256	7	32.855	8.312	3.500 / 3.002
256	256	256	31	35.656	39.136	3.500 / 3.002
256	1024	128	7	72.051	2.315	6.750 / 1.751
256	1024	128	31	82.445	8.067	6.750 / 1.751
256	1024	256	7	63.179	9.065	8.000 / 4.502
256	1024	256	31	57.125	40.967	8.000 / 4.502
512	64	128	7	19.853	2.009	7.062 / 1.063
512	64	128	31	15.459	6.252	7.062 / 1.063
512	64	256	7	15.209	9.112	7.625 / 3.127
512	64	256	31	14.221	32.591	7.625 / 3.127
512	256	128	7	53.378	2.007	8.750 / 1.251
512	256	128	31	45.051	6.723	8.750 / 1.251
512	256	256	7	50.838	10.715	9.500 / 3.502
512	256	256	31	50.453	33.975	9.500 / 3.502
512	1024	128	7	180.427	2.647	15.500 / 2.001
512	1024	128	31	180.223	7.798	15.500 / 2.001
512	1024	256	7	203.103	9.437	17.000 / 5.002
512	1024	256	31	167.955	31.669	17.000 / 5.002

C.3 Neural Updates and Prediction Profiles

Neural continual-learning audit. The 20-seed HOMO-256 audit compares fine-tuning, elastic weight consolidation (EWC) (Kirkpatrick et al., 2017), replay (Lopez-Paz & Ranzato, 2017), orthogonal gradient descent (OGD) (Farajtabar et al., 2020), A-GEM (Chaudhry et al., 2019), PCGrad (Yu et al., 2020), learning without forgetting (LwF) (Li & Hoiem, 2018), and soft-orthogonal residual learning. Each update is measured against the method's own stored predictor. This experiment measures actual neural updates and is distinct from the ten-seed experiment that applies the filter to a trained-network Jacobian.

Table 19: Twenty-seed QM7b HOMO-256 neural continual-learning audit. Update is measured against each method's stored pre-update predictor.

Method	High RMSE	Update RMSE	Normalized update
Fine-tune	0.771 ± 0.012	0.869 ± 0.026	0.902 ± 0.042
EWC	0.771 ± 0.013	0.868 ± 0.028	0.901 ± 0.044
Replay	1.094 ± 0.022	0.361 ± 0.019	0.374 ± 0.019
OGD	0.816 ± 0.033	0.851 ± 0.023	0.884 ± 0.046
A-GEM	0.881 ± 0.060	0.615 ± 0.079	0.639 ± 0.087
PCGrad	1.074 ± 0.028	0.386 ± 0.017	0.401 ± 0.018
LwF	0.789 ± 0.015	0.779 ± 0.028	0.809 ± 0.049
Soft orthogonal	0.947 ± 0.182	0.827 ± 0.190	0.807 ± 0.180

The neural methods make different accuracy–stability trade-offs. Fine-tuning and EWC attain the lowest high-RMSE but change their stored predictors by nearly one residual scale. Replay and PCGrad reduce that change substantially while paying a large accuracy cost; OGD, A-GEM, and LwF occupy intermediate positions. The soft-orthogonal update has wide seed variability and does not establish a stable improvement.

Anchor density and neural interference. Section 5.2 reports the anchor-density experiment. The full precision coefficients of variation are 0.005673 for weighted SSI and 0.478427 for the unweighted norm. In the neural stress comparison, negative spatial gap, unweighted SSI, and whitened overlap have Spearman correlations -0.123, 0.028, and -0.023 with measured interference, respectively.

Sketch scaling. Random projections can reduce full Jacobian or feature dimensions to rank r. At feature dimension 32,768, relative error and time are 17.16%/0.046 s, 1.69%/0.086 s, 25.48%/0.148 s, and 147.93%/0.284 s for ranks 16, 32, 64, and 128, respectively. The approximation error is non-monotone in this experiment, so rank alone is insufficient to choose an accurate sketch (Heckel, 2022).

When anchor guarantees fail. Table 20 links observable failures to specific checks or changes in the update policy. It complements the statistical conditions in the theorems and the measured transfer behavior in Section 5.7.

Table 20: Interpreting an anchor certificate requires a specified input measure and a response when held-out updates exceed the cap.

Failure mode	Observable symptom	Required response
Anchor–test mismatch	Anchor $\bar{U} \leq \tau$ but test $\bar{U} > \tau$	Add coverage-representative anchors or treat the certificate as local only.
Feature learning	Empirical NTK or SSI changes during training	Re-estimate the operator. Do not reuse an initialization certificate.
Biased low model	Small update preserves known low-fidelity error	Relax or remove the update budget in that region.
Label-mechanism shift	Input coverage remains high while residual law changes	Use labeled validation probes. Input coverage alone is insufficient.

Continued on next page

Table 20 (continued)

Failure mode	Observable symptom	Required response
Ill-conditioned fine kernel	SSI and predictions are ridge-sensitive	Report ridge sensitivity and avoid interpreting an unstable coefficient.
Unrepresentative weighting	Duplicates or dense regions dominate anchors	Use normalized quadrature/empirical weights and report the anchor construction.

Reconstructed prediction profiles. Figures 3–4 visualize the existing QM7b seed-0 blocks at budget 256 and $\tau = 0.20$, using external farthest-first anchors and the direct filter. Both properties use the same 1,500 test molecules. These fixed-block views illustrate predictions and their limitations; they are not additional benchmark trials.

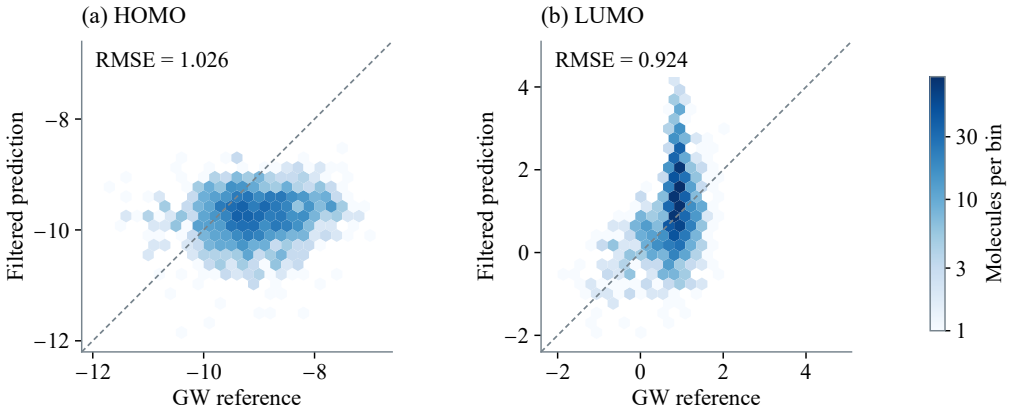

Figure 3: **Reconstructed high-fidelity predictions.** HOMO and LUMO predictions against GW references in native dataset units. Hexagonal bins retain all 1,500 test molecules per panel and share a logarithmic count scale; dashed lines denote exact agreement. The remaining systematic errors are visible despite the small allowed update.

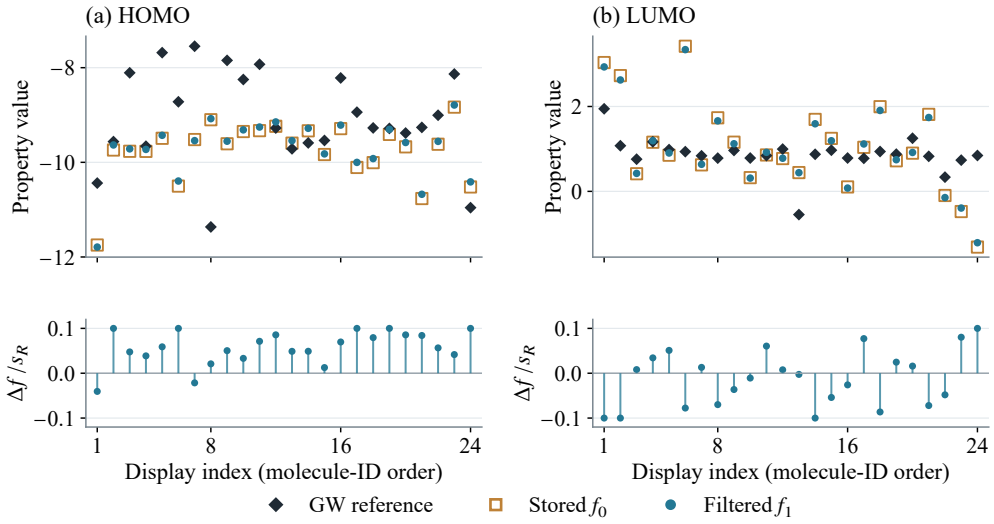

Figure 4: **Stored and updated prediction profiles.** Top: GW reference, stored predictor, and filtered predictor for 24 molecules selected at evenly spaced ranks in test-set molecule-ID order, independently of errors. Bottom: signed update divided by the training-residual RMS s_R, the scale s in (5). The display index is categorical, not a spatial coordinate. Both properties select $\gamma = 1$ and $\alpha = 0.10$; small changes can leave substantial error.