ASSAYROUTER: HISTORICAL UTILITY PRIORS FOR FROZEN MOLECULAR PREDICTOR ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Laboratories often face a new molecular assay with 16–64 labels and a bank of predictors whose training data and parameters are unavailable. The practical question is which frozen outputs to include in a small local model. AssayRouter treats completed assays as pseudo-targets and labels each candidate by its post-fit utility: the reduction in held-out discovery loss when the candidate is added to the local target predictor. A shared regressor learns to predict this utility from candidate behavior on the support set, without source identity; on a new assay, one frozen ranking selects four sources and separate labels fit a convex combiner. We train only on completed ChEMBL-MT assays and evaluate 24 external regression assays across six frozen interface families. AssayRouter-C lowers strict four-call negative log-likelihood (NLL) by 0.0409 relative to Support-CV@4. Frozen candidate-label permutations confirm that candidate–utility correspondence carries the transferred information, and leave-one-interface-out training shows that the mapping generalizes to unseen predictor families. Completed assays therefore provide transferable supervision for scarce-label routing through frozen prediction interfaces.

1 INTRODUCTION

Drug discovery programs accumulate predictors assay by assay. When a new continuous endpoint has few measurements, models trained for other assays can already score the same molecules. Their examples, architectures, and parameters may be inaccessible. Reuse then occurs through a frozen contract: each source exposes predictions and metadata, while its training data and parameters remain unavailable. The practical decision is which output channels should enter a small model fitted to the target.

Recent few-shot methods adapt or condition a predictor on a small labeled set: ActFound uses pairwise meta-learning (Feng et al., 2024), while FS-CAP (Eckmann et al., 2024) and UniMatch (Li et al., 2025) encode labeled support. Other systems select or weight source tasks, as in GATE (Lee et al., 2024), or source data, as in AssayMatch (Fan & Barzilay, 2026). These systems select data, tasks, or trainable models without labeling the downstream contribution of a deployed frozen prediction channel.

A source can be accurate in isolation yet redundant after fitting the local target model; a weaker source can correct a consequential residual. The useful supervision is therefore incremental value inside the fitted combination, which completed assays make observable. ASSAYROUTER measures how much each frozen source reduces held-out loss after the same local fit used at deployment. A regressor maps candidate behavior on the support set to this post-fit utility, conditioned on the interface family. Frozen on completed assays, it gives an unseen assay one score per source, one Top-4 set, and separately fitted convex weights.

The evaluation crosses 24 external assays in three collections with six interface families. Matched interventions test whether candidate-specific utility transfers, whether it survives removal of an interface family, and whether frozen ranking can replace subset search on the target. These experiments support three contributions:

- We define post-fit utility, the incremental loss reduction from adding one frozen source to a local target predictor, as transferable supervision for routing.

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.

- AssayRouter amortizes subset search into one frozen score per source, replacing more than 20 million inner nonnegative least-squares (NNLS) fits with zero counterfactual fits on the target.
- Controlled interventions show that candidate-specific utility transfers across external assays and to predictor families absent from router training, improving strict four-call routing over Support-CV@4.

2 PROBLEM SETTING

Let an unseen target assay t provide labeled support $D_t^{\mathrm{sup}} = \{(x_i, y_i)\}_{i=1}^n$ and an untouched confirmation set D_t^{test}. A bank $\mathcal{S}_t$ contains independently trained source predictors f_s. Each interface exposes a prediction $z_s(x)$, optional uncertainty $u_s(x)$, and declared metadata m_s, while its training data and parameters remain inaccessible. We divide each support episode into balanced, disjoint roles: routing observes candidate behavior and metadata on R_t, while fitting observes the selected source outputs on C_t. The two operations receive separate contracts:

$$\mathcal{I}_t^{\mathrm{route}} = (R_t, \{z_s(R_t), u_s(R_t), m_s\}_{s\in\mathcal{S}_t}, \iota_t), \quad \mathcal{I}_t^{\mathrm{fit}}(A_t) = (C_t, \{z_s(C_t)\}_{s\in A_t}), \quad R_t \cap C_t = \varnothing. \tag{1}$$

where ι_t denotes the deployed interface family. In the evaluated contract, m_s includes source training size, while z_s and u_s are the mean and empirical dispersion of each source's frozen members. The confirmation set is absent from both contracts and is opened only for final scoring. The target is also deleted from source construction, router training, model selection, and calibration; historical episodes apply the same deletion.

The task contains two decisions. A selector chooses $A_t \subset \mathcal{S}_t$ with $|A_t| = K = 4$; a normalized NNLS combiner then fits weights on C_t for the local target column and four selected source outputs. This fan-in defines the call budget at confirmation and keeps target adaptation identifiable when labels are scarce. For any legal set A, let $\widehat{h}_{t,A}$ denote that fitted combiner and write $J_t(A) = \mathcal{L}_{D_t^{\mathrm{test}}}(\widehat{h}_{t,A})$. The latent optimal action at confirmation and selection regret are

$$A_t^* = \arg\min_{A\subseteq\mathcal{S}_t,\,|A|=K} J_t(A), \qquad \mathcal{R}_t(\pi) = J_t\big(\pi(\mathcal{I}_t^{\mathrm{route}})\big) - J_t(A_t^*). \tag{2}$$

The oracle action is unavailable at deployment because confirmation labels remain sealed. Since $J_t(A_t^*)$ is shared across selectors, paired differences in confirmation loss equal paired differences in latent regret. AssayRouter learns a policy $\pi : \mathcal{I}_t^{\mathrm{route}} \mapsto A_t$ from completed assays, then fits $\widehat{h}_{t,A_t}$ through the separate fitting contract. The study asks whether post-fit utility learned from completed assays reduces this latent regret on a new label function under a frozen interface family.

3 ASSAYROUTER

3.1 LEARNING A POST-FIT UTILITY PRIOR

Figure 1 connects the three stages: measuring post-fit utility on completed assays, learning one source prior without identity features, and freezing a four-source action before fitting the target.

We measure post-fit utility by mirroring deployment on each completed assay: fit a convex combination of the local predictor and candidate sources, then evaluate it on held-out discovery molecules. Let e index such a historical episode and $A \subseteq \mathcal{S}_e$ a candidate subset. A four-fold out-of-fold Ridge model supplies the local target column $\widehat{y}_{0,e}^{\mathrm{CF}}$. With $\widetilde{X}_{e,A}(x) = [\widehat{y}_{0,e}^{\mathrm{CF}}(x), (z_s(x))_{s\in A}]$, fit $\widetilde{\beta}_{e,A}$ and normalize it as $\widehat{\beta}_{e,A} = \widetilde{\beta}_{e,A}/(\mathbf{1}^\top \widetilde{\beta}_{e,A})$:

$$\widetilde{\beta}_{e,A} = \mathrm{argmin}_{\beta\geq 0} \|\widetilde{y}_{C_e} - \widetilde{X}_{e,A}(C_e)\beta\|_2^2, \qquad \widehat{h}_{e,A}(x) = \mu_{C_e} + \sigma_{C_e}\widetilde{X}_{e,A}(x)\widehat{\beta}_{e,A}. \tag{3}$$

Tildes denote the declared C_e-local robust scaling; $\widehat{h}$ is returned to the raw response scale before evaluation. If every NNLS coefficient vanishes, we use the uniform simplex vector. For discovery set Q_e^{disc}, disjoint from R_e and C_e, write $\mathcal{L}_{e,A} = \mathcal{L}_{Q_e^{\mathrm{disc}}}^{(\sigma_{C_e})}(\widehat{h}_{e,A})$ and define

$$\mathcal{L}_Q^{(\sigma)}(h) = |Q|^{-1}\sum\nolimits_{(x,y)\in Q} \ell_\nu\big((y - h(x))/\sigma\big), \quad \Delta_e(s \mid A) = \mathcal{L}_{e,A} - \mathcal{L}_{e,A\cup\{s\}}, \tag{4}$$

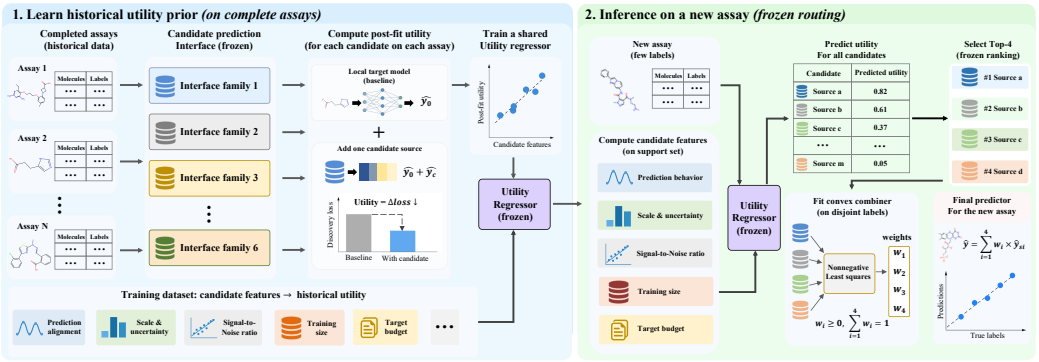

Figure 1: **Completed assays supervise frozen predictor routing.** (1) A historical pseudo-target measures the reduction in discovery loss from adding one candidate. (2) A shared AssayRouter prior maps candidate behavior to utility without source identity. (3) On a new assay, the frozen prior scores the bank once, selects four sources, and fits simplex weights for the target before confirmation.

where the Student-t loss for point predictions with $\nu = 3$ is

$$\ell_\nu(r) = \tfrac{1}{2}\log(\nu\pi) + \log\Gamma(\nu/2) - \log\Gamma((\nu+1)/2) + \tfrac{\nu+1}{2}\log(1 + r^2/\nu). \tag{5}$$

AssayRouter measures the post-fit value $U_e(s) = \Delta_e(s \mid \varnothing)$. Its raw level includes a block-wide offset induced by historical target difficulty and support quality, so we center the candidates within each block:

$$U_e^{\mathrm{C}}(s) = U_e(s) - |\mathcal{S}_e|^{-1} \sum\nolimits_{r \in \mathcal{S}_e} U_e(r) = |\mathcal{S}_e|^{-1} \sum\nolimits_{r \in \mathcal{S}_e} \mathcal{L}_{e,\{r\}} - \mathcal{L}_{e,\{s\}}. \tag{6}$$

Centering removes the shared offset while preserving every candidate ordering (Eq. 9). AssayRouter-Δ retains the uncentered loss reduction as a matched calibration ablation.

A single histogram gradient-boosting (HGB) regressor predicts centered utility from candidate behavior, conditioned on the interface family. Source and assay identity never enter the regressor, and its representation reads neither C_e nor Q_e^{disc}:

$$\widehat{\theta}_\lambda = \mathrm{argmin}_\theta \sum\nolimits_e \sum\nolimits_{s \in \mathcal{S}_e} w_{e,s} \left[q_\theta(\psi_e(s; R_e), \iota_e) - U_e^{\mathrm{C}}(s)\right]^2. \tag{7}$$

We choose $\widehat{\lambda} \in \Lambda$ by cross-validation mean absolute error (MAE) averaged over targets; grouped folds keep every interface of a historical assay on the same side. Appendix A.1.2 gives the configurations and training rule, while Appendix A.1.7 defines the repeated-state weights $w_{e,s} = 13$.

The 11 coordinates separate complementarity from reliability. Signed and absolute residual alignment and residual correlation measure how a source tracks what the local model misses; source scale, spread, uncertainty, signal-to-noise ratio (SNR), training size, and routing-support size describe whether that signal is reliable. Let $r_{e,i} = \widetilde{y}_{e,i} - \widehat{y}_{0,e}^{\mathrm{CF}}(x_i)$ and $v_{s,i} = \widetilde{z}_s(x_i)$ on R_e. The primary representation is

$$\begin{aligned}\psi_e(s; R_e) = \big(&\overline{v_s r_e},\ |\overline{v_s r_e}|,\ \mathrm{corr}(v_s, r_e),\ \overline{v_s},\ \mathrm{sd}(v_s),\ \overline{|v_s|},\\ &\overline{u_s},\ \mathrm{sd}(u_s),\ \mathrm{sd}(v_s)/\max(\overline{u_s}, 10^{-3}),\ \log(1+n_s),\ \log_2|R_e|\big).\end{aligned} \tag{8}$$

All coordinates are available under the routing contract; the interface code is appended separately.

Equations 3–7 define the name *AssayRouter*: completed assays convert post-fit candidate contribution into a shared routing prior, and the prior turns a new support set into one frozen source allocation. The centering step changes calibration without changing the supervised action,

$$\mathrm{TopK}_{s \in \mathcal{S}_e} U_e^{\mathrm{C}}(s) = \mathrm{TopK}_{s \in \mathcal{S}_e} U_e(s), \tag{9}$$

while making utility levels comparable across historical blocks.

End-to-end procedure. Learning and deployment follow four operations with separate data roles.

1. For each historical episode, fit the local target Ridge column on R_e by cross-fitting and compute $\psi_e(s; R_e)$ for every frozen candidate.

2. On C_e, fit the target-only model and its candidate-augmented counterpart from Equation 3; on disjoint Q_e^{disc}, convert their loss difference into $U_e^{\mathrm{C}}(s)$.
3. Pool historical episodes and fit $\widehat{q}$ with model selection grouped by assay. Freeze its feature schema, interface encoding, capacity, and parameters.
4. For an unseen target, recompute candidate behavior on R_t, select A_t once, and fit only $\widehat{\beta}_{t,A_t}$ on C_t. Confirmation labels enter the final metric after both the set and weights have been fixed.

Thus R determines the routing representation, C identifies the local convex combiner, and historical Q_{disc} supplies the utility label. No deployment operation reconstructs the discovery query or revisits the selected set.

Equation 4 operationalizes this incremental value under the same convex refit used at deployment. The population target is the component of post-fit value predictable from a new routing support:

$$q^*(\psi, \iota) = \mathbb{E}\big[U_e^{\mathrm{C}}(s) \mid \psi_e(s; R_e) = \psi,\ \iota_e = \iota\big]\,. \tag{10}$$

Leave-one-interface-out training additionally removes ι for the held-out family and tests the behavioral mapping on an unseen representation.

Nested refinements. Let $G_{e,s} = \mathcal{G}_e(s)$ be the 13 frozen partial sets of sizes zero through three that exclude s, and abbreviate $Q_e(s, A) = q_{\mathrm{S}}(\psi_e(s; R_e), g(A), \iota_e)$ and $Q_{e,j-1}(s) = Q_e(s, A_{j-1})$, with $\mathcal{C}_j = \mathcal{S}_e \setminus A_{j-1}$. The marginal-over-states label and state-conditioned action are

$$U_e^{\mathrm{M}}(s) = \mathrm{Avg}_{A \in G_{e,s}}\, \Delta_e(s \mid A), \quad s_j = \mathrm{argmax}_{s \in \mathcal{C}_j}\, Q_{e,j-1}(s), \quad A_j = A_{j-1} \cup \{s_j\}. \tag{11}$$

Here Avg is the arithmetic mean and $g(A)$ contains 14 set-state coordinates. Both refinements retain the historical rows and external replay; they change the supervision or state representation only.

Mechanism interventions. Within each historical target–interface–budget–episode block, a bijection permutes utility among the 19 candidates while preserving feature rows, the label multiset, capacity search, grouped folds, and external assignments. Six leave-one-interface-out routers remove one predictor family and its calibration from historical training.

3.2 Routing once with the frozen utility prior

Frozen action and target-specific fit. At deployment, the frozen router selects four candidates from one score per source:

$$A_t = \mathrm{TopK}_{s \in \mathcal{S}_t}\, \widehat{q}(\psi_t(s; R_t), \iota_t)\,, \qquad |A_t| = K = 4. \tag{12}$$

Routing reads only $\mathcal{I}_t^{\mathrm{route}}$; after A_t is frozen, Equation 3 fits its weights on the separate C_t contract. Writing $z_0(x) = \widehat{y}_{0,t}(x)$, the confirmation predictor is

$$\widehat{y}_t(x) = \sum\nolimits_{j \in \{0\} \cup A_t} \widehat{\beta}_{t,j} z_j(x), \qquad \widehat{\beta}_{t,j} \geq 0, \quad \sum\nolimits_{j \in \{0\} \cup A_t} \widehat{\beta}_{t,j} = 1. \tag{13}$$

Only membership and the fitted simplex weights enter prediction. For $a > 0$ and any constant c, the action is affine-invariant:

$$\mathrm{TopK}(a\widehat{q} + c) = \mathrm{TopK}(\widehat{q}) = A_t. \tag{14}$$

Score calibration is therefore irrelevant until it changes the Top-4 boundary. Each confirmation molecule calls the four selected sources. Across all 24 external targets, the learned prior, feature schema, and interface encoding remain fixed; each target recomputes its local predictor and routing features, then determines A_t and $\widehat{\beta}_{t,A_t}$.

Amortized selection cost. With M sources, AssayRouter performs M frozen score evaluations, one top-K operation, and no counterfactual fit on the target. A Support-CV screen of width L and F inner folds instead uses $B = \min(M, L)$ screened sources, where $\mathrm{Comb}(B, K)$ counts their K-source subsets:

$$N_{\mathrm{score}}^{\mathrm{AR}} = M, \qquad N_{\mathrm{fit}}^{\mathrm{AR}} = 0, \qquad N_{\mathrm{fit}}^{\mathrm{CV}} = F(M\mathbf{1}\{M > L\} + \mathrm{Comb}(B, K))\,. \tag{15}$$

The indicator term is paid only when $M > L$: each fold first scores all M candidates to retain B. The combination term then fits every K-source subset of those B survivors, and the factor F repeats this

selection workload across inner folds. Both methods subsequently fit the same final NNLS combiner with five columns. Across the full evaluation, Equation 15 yields 20,445,696 Support-CV inner fits and zero for AssayRouter; Appendix A.3 gives the ledger by collection and the crossover. This count excludes inference by frozen sources and the common final NNLS fit, isolating the work on the target used to decide membership. Historical utility construction is paid once and reused across external assays, so repeated deployment increases Support-CV search cost while leaving router fitting fixed.

Two evaluation estimands. Let Π_t contain the 16 constituents formed by eight balanced R/C partitions and both role directions, with $\widehat{h}_{t,\pi}$ the frozen four-source predictor for constituent π. On confirmation set D_t^{test}, define

$$\mathcal{L}_{D_t^{\mathrm{test}}}(h) = |D_t^{\mathrm{test}}|^{-1} \sum\nolimits_{(x,y)\in D_t^{\mathrm{test}}} \ell_3\big((y - h(x))/\sigma_{R_t\cup C_t}\big), \tag{16}$$

where $\sigma_{R_t\cup C_t}$ is the robust scale of the complete labeled support. With $\widehat{y}_t^{\mathrm{CF}}(x) = 16^{-1}\sum_{\pi\in\Pi_t}\widehat{h}_{t,\pi}(x)$, the prediction-averaged and strict estimands are

$$\widehat{\mathcal{L}}_t^{\mathrm{CF}} = \mathcal{L}_{D_t^{\mathrm{test}}}(\widehat{y}_t^{\mathrm{CF}}), \qquad \widehat{\mathcal{L}}_t^{\mathrm{strict}} = 16^{-1}\sum\nolimits_{\pi\in\Pi_t}\mathcal{L}_{D_t^{\mathrm{test}}}(\widehat{h}_{t,\pi}). \tag{17}$$

Cross-fit scores one variance-reduced prediction averaged over 16 constituents and can touch 5–21 unique sources per episode. Strict scoring averages 16 scalar losses after each constituent has made exactly four calls; it is therefore the primary deployment estimand. All sparse methods share the assignments, bank, fan-in, and final convex fit. Support-CV@4 searches using support labels before confirmation; Frozen-DES@4 (dynamic ensemble selection) and MINE-WS@4 (Moura et al., 2021) receive query-local access, and all-source convex relaxes fan-in. Random and residual diagnostics remain in the appendix.

Action stability. Let $\widehat{q}_{(K)}$ and $\widehat{q}_{(K+1)}$ bracket the selected set. For any refined scorer q',

$$\gamma_t = \widehat{q}_{(K)} - \widehat{q}_{(K+1)}, \qquad \max_{s\in\mathcal{S}_t}|q'(s) - \widehat{q}(s)| < \gamma_t/2 \implies \mathrm{TopK}(q') = A_t. \tag{18}$$

A refined score can change the selected set only by crossing this rank margin. Whether that change alters prediction depends on the refitted combination function, including the displaced source and all recomputed weights.

4 Results

We fit the router and select its HGB capacity on 20 ChEMBL-MT pseudo-targets, then evaluate a fixed four-source contract on 24 external targets across six interfaces, three label budgets, and 12 support replays. ExpansionRx (OpenADMET Consortium, 2026) provides temporal holdouts, Biogen (Fang et al., 2023) sparse continuous endpoints, and TDC (Huang et al., 2021) scaffold-split regression tasks. Earlier sourcewise studies selected the contract (Appendix C); the primary sparse comparisons share its deployment budget. At 16 labels, each directional fitting split contains eight observations for the local predictor and four selected channels, making reliable subset search especially difficult.

For target t, interface ι, budget b, and replay set $\mathcal{E}_{t\iota b}$, every comparison is first paired at the episode level and then aggregated as

$$d_{t\iota b} = \frac{1}{|\mathcal{E}_{t\iota b}|}\sum_{e\in\mathcal{E}_{t\iota b}}\left(\widehat{\mathcal{L}}_e^{\mathrm{comp}} - \widehat{\mathcal{L}}_e^{\mathrm{AR}}\right). \tag{19}$$

Aggregate intervals independently resample target and interface clusters, so support partitions are not treated as independent scientific replicates; Appendix A.1.10 gives the full unit counts.

The comparisons isolate bank reuse, utility supervision, ranking signal, information timing, and action fan-in while retaining the same frozen bank, R/C roles, Top-4 action, and final NNLS fit whenever the tested axis permits. Target-only tests source reuse; source-quality, support statistics, residual compatibility, and permutations test the ranking signal; Support-CV@4 tests search on the target. Query-local and dense controls deliberately relax access or fan-in. Appendix A.2.1 positions every comparator, and Table 2 records its permissions. Support-standardized Student-t NLL is primary; MAE, root mean squared error (RMSE), and Spearman correlation provide complementary error and ranking views.

4.1 ASSAYROUTER TRANSFERS POST-FIT UTILITY

Candidate identity must remain paired with its historical post-fit value for routing to transfer. Five frozen candidate-label permutations each degrade aggregate external NLL under prediction averaging (Fig. 2B). The effect is positive at every budget and across all six interfaces in the uncentered replication (Table 16). The intervention preserves feature rows, the utility multiset, model capacity, grouped folds, and external assignments. Under strict scoring, the same permutation increases AssayRouter-C's NLL by 0.1352 (95% CI $[0.0852, 0.1915]$), favoring the genuine mapping in 90.7% of paired cells. The advantage is largest at 16 labels, when evidence on the target is scarcest (Fig. 2A). It persists with uncentered labels (Table 18) and support-scale normalization (Table 5).

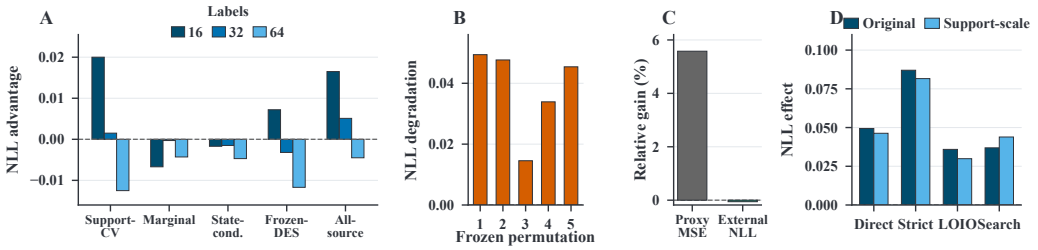

Figure 2: **Post-fit utility transfers across assays.** (A) Comparator-minus-AssayRouter NLL at each target-label budget; positive values favor the utility prior. (B) Five independently frozen candidate-label permutations each degrade external NLL. (C) Direction-normalized proxy MSE and external NLL changes; positive denotes improvement and exposes their divergence. (D) Original and support-scale versions of the direct, strict, unseen-interface, and matched-search comparisons.

The learned correspondence extends beyond the identities available during historical training. The strict intervention retains every interface family; leave-one-interface-out training removes the evaluated family. With AssayRouter-C, permuting the transferred correspondence increases NLL by 0.0497 with a clustered 95% interval of $[0.0135, 0.0963]$ and positive point effects in all six held-out families (Fig. 3). The same unseen-interface routers improve over Support-CV by 0.1116 NLL with interval $[0.0834, 0.1428]$ and win on all 24 targets (Table 20).

A crossed audit excluding both historical target and candidate-source identities retains utility accuracy and Top-4 selection (Fig. 4, right; Appendix A.1.4). Exact-parent removal on TDC preserves both effects in the uncentered replication (Table 21). Table 8 records the evaluation boundary.

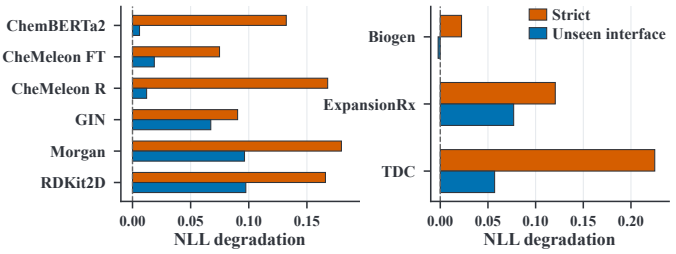

Figure 3: **Candidate–utility transfer spans interfaces and collections.** Mean NLL degradation after label permutation by deployed interface (left) and external collection (right); both interventions use AssayRouter-C, and positive values favor the genuine post-fit labels.

4.2 HISTORICAL UTILITY OUTPERFORMS SELECTION ON THE TARGET UNDER THE SAME ACTION

The strict estimand asks the deployment question directly: after one frozen Top-4 choice, how well does each four-call constituent predict? AssayRouter-C gives the best overall NLL, MAE, and RMSE among the access-matched methods, with its advantage largest at 16 labels (Table 1). The Δ, Z, and marginal-over-states variants remain close, making post-fit candidate value the shared supervision principle across these label constructions.

The overall strict NLL gap to Support-CV is 0.0409 (95% CI $[0.0211, 0.0623]$). AssayRouter-C has lower NLL in 77.8% of target–interface units; in original units, mean MAE falls from 26.24 to 23.39, a 10.9% reduction. Marginal-over-states attains the highest prediction Spearman correlation, distinguishing response ordering from numerical accuracy. Both exhaustive Support-CV and cheaper

Table 1: **Strict four-call deployment.** C, Δ, and Z denote centered, uncentered, and standardized standalone utility; Marginal denotes marginal-over-states utility. Every row uses one-shot constituent scoring; bold and underline mark the two best point estimates.

	NLL by target labels			Overall			
Method	16	32	64	NLL ↓	MAE ↓	RMSE ↓	Pred. ρ ↑
AR-Marginal	2.2711	2.1052	1.9886	2.1216	23.8597	<u>35.3358</u>	**0.3191**
Support-CV	2.3287	2.1309	1.9938	2.1511	26.2446	38.0759	0.3033
Support-greedy	2.3806	2.1695	2.0410	2.1970	28.6550	40.3969	0.2949
Source-quality	2.4467	2.2451	2.0806	2.2575	29.3549	41.0480	0.3111
Residual	2.5339	2.2696	2.0806	2.2947	32.3550	44.8218	0.3053
Target-only	2.4456	2.3084	2.2476	2.3339	29.3636	38.8208	0.1805
AssayRouter-Z	2.2666	2.0989	1.9889	2.1181	23.6652	35.3589	<u>0.3124</u>
AssayRouter-Δ	<u>2.2647</u>	<u>2.0926</u>	**1.9853**	<u>2.1142</u>	<u>23.5806</u>	35.3864	0.2961
AssayRouter-C	**2.2551**	**2.0896**	<u>1.9860</u>	**2.1102**	**23.3872**	**35.0709**	0.3011

support-greedy remain weaker under strict scoring; support-greedy also trails at every budget and interface under prediction averaging (Appendix A.1.13).

Support-CV refits candidate subsets for every support split, whereas AssayRouter scores each source once and reuses the frozen ranking. The search requires 20, 280, and 364 inner NNLS fits per direction on Biogen, ExpansionRx, and TDC; AssayRouter performs none before the shared final combiner. Prediction averaging expands one episode to 64 source calls and more than 20 million inner fits across the evaluation. Strict scoring evaluates each four-call constituent independently. AssayRouter therefore improves the matched deployment action while amortizing its construction cost (Fig. 4, middle; Appendix A.3).

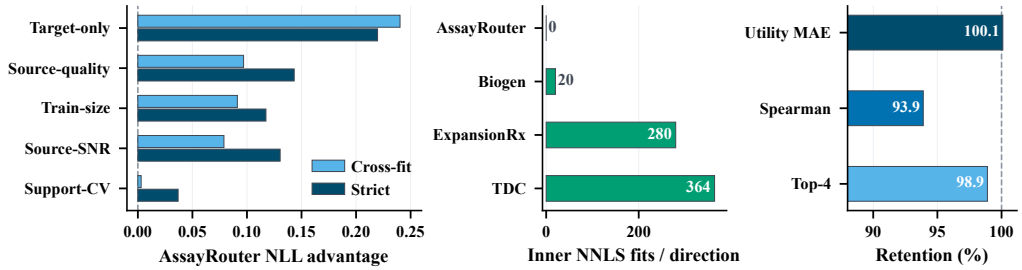

Figure 4: **Historical utility is specific and amortizes search on the target.** Left: matched controls against AssayRouter-Δ under cross-fit and strict scoring; positive values favor the historical prior. Middle: inner subset fits for Support-CV in each collection. AssayRouter performs no counterfactual NNLS fit on the target; both methods retain the same final NNLS combiner. Right: relative performance when both historical target and candidate source identities are held out; 100% is the target-only out-of-fold reference, MAE is inverted, and values above 100% indicate improvement over that reference.

The matched controls test the principal shortcuts when outputs are frozen. Target-only measures whether the bank contributes at all; source quality, training size, and support SNR replace learned utility with metadata or support statistics; residual compatibility uses the local target error profile; and Support-CV supplies subset search on the target. All retain the same Top-4 fan-in and final convex fit. The left panel of Fig. 4 shows the resulting hierarchy: removing source reuse creates the largest gap, generic rules for source strength remain consistently weaker, and Support-CV approaches parity only when its many searched constituents are averaged. Strict scoring exposes the deployed four-call action and separates the frozen historical ranking from this search. Query-local DES and MINE, together with the all-source convex model, answer different questions about information access or fan-in (Appendix A.2.1; Table 15).

4.3 Post-fit supervision survives shortcut and stability tests

Residual compatibility is the closest shortcut based on support profiles under the same four-source contract. AssayRouter-C improves strict NLL by 0.1844 over this selector (95% CI $[0.1138, 0.2609]$).

The advantage appears in all six interfaces and at least 79% of targets at every budget (Table 1; Appendix A.2.2). Residual correlation measures whether a source follows local-model errors; post-fit utility measures whether it remains useful after competing channels enter the normalized NNLS combiner. Their gap localizes the transferred information to constituent value inside the fitted ensemble. The candidate-label intervention, unseen-interface transfer, and strict Support-CV comparison also retain their direction after deleting any one target, interface, or collection (Appendix A.1.11).

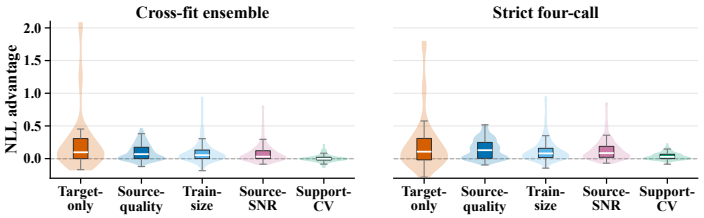

Figure 5: **Matched alternatives underperform.** Distributions aggregate 144 target–interface NLL effects across budgets; positive values favor AssayRouter-Δ. Boxes mark quartiles and medians.

Figure 5 compares matched alternatives across target–interface units. Target-only and generic measures of source strength favor AssayRouter under both estimands. Support-CV reaches parity only through prediction averaging and favors AssayRouter when four-call actions are scored independently. Because all controls share assignments, Top-4 fan-in, and final NNLS, the hierarchy isolates each ranking signal from post-fit supervision. Appendix A.1.13 gives the contracts and point effects.

State features improve post-fit utility prediction without improving external routing (Fig. 2C). A refined score changes deployment only when it changes Top-4 membership and the ensuing refit changes the combination function. Utility-prediction MSE can therefore improve while preserving either the selected set or its refitted predictor, explaining why proxy regression and deployed decision quality separate. Figure 6 resolves the transferred object across scientific units: candidate value after fitting the local target model, expressed strongly enough to change a sparse action.

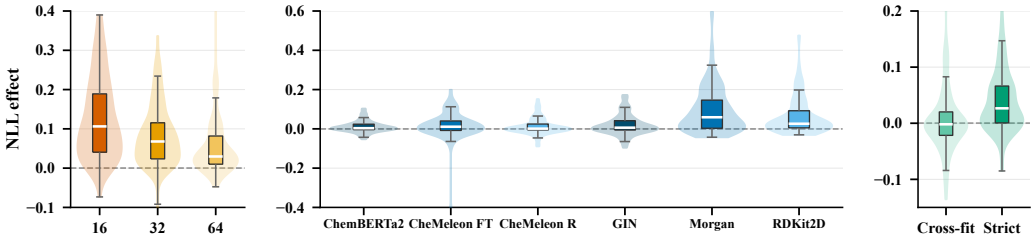

Figure 6: **The central effects span scientific units.** Left: strict candidate-label intervention across 144 target–interface cells; middle: unseen-interface transfer across 72 target–budget cells; right: Support-CV across 144 units under prediction averaging and strict scoring. Positive values favor AssayRouter-Δ; violins contain quartile boxes and medians.

These distributions preserve assay heterogeneity: the strict intervention is broadest at 16 labels, unseen-interface effects are positive in aggregate for every held-out family, and strict Support-CV separates from prediction averaging under the four-call estimand. Agreement between the centered primary result (Table 1) and uncentered Δ replication (Fig. 6) shows that block centering does not create the effect. Historical ordering matters most before a new assay can search the bank reliably.

The deletion analysis comprises 24 leave-one-target, six leave-one-interface, and three leave-one-collection estimates per central comparison. Every deletion preserves the signs of the candidate-label intervention, unseen-interface effect, and strict Support-CV advantage; two-sided exact sign tests over the six interface effects give $p = 0.03125$ for all three. Thus, no single assay, predictor family, or external collection drives the conclusions.

The support-scale sensitivity changes only the historical Q_{disc} loss scale while preserving routing on R, fitting on C, assignments, and deployment; direct, strict, unseen-interface, and strict search effects remain positive (Fig. 2D). The crossed identity audit predicts each historical row without the identity of its pseudo-target or candidate source, yet utility MAE and Top-4 overlap match the reference that holds out the target (Fig. 4, right). Together with leave-one-interface-out deployment,

these results identify a relation from candidate behavior to utility that is shared across assays and prediction contracts.

5 Related Work

Molecular few-shot adaptation. ActFound uses pairwise meta-learning (Feng et al., 2024); FS-CAP (Eckmann et al., 2024) and UniMatch (Li et al., 2025) condition on labeled sets; MolSetRep learns set representations (Boulougouri et al., 2024); and Contrastive KERMT transfers pretrained ADME representations (Xue et al., 2026). Other approaches span multitask benchmarks (Stanley et al., 2021), adaptive kernels (Chen et al., 2023), contextual memory (Schimunek et al., 2023), in-context prediction (Fifty et al., 2023), and learned task transfer (Lee et al., 2024; Zhong & Mottin, 2025). These methods train, fine-tune, or condition a predictor. AssayRouter keeps source contracts frozen and learns which outputs complement a local model. Appendix B reproduces official systems; access-matched routing controls form the main comparison.

Source and task selection. Prediction-profile methods use statistical relations between source outputs and a target: pQSAR transfers dense cross-assay correlations (Martin et al., 2019; Martin & Zhu, 2021), while MINE estimates query-local competence (Moura et al., 2021). Task-level or auxiliary evidence guides transfer weights in GATE (Lee et al., 2024), attribution-supervised assay selection in AssayMatch (Fan & Barzilay, 2026), and imputation in QComp (Yang et al., 2025). TaskWeb (Kim et al., 2023) and MetaGL (Park et al., 2023) likewise transfer task or model relations. These signals describe relatedness or competence before combination; AssayRouter supervises value after the local predictor and source output are fitted together.

Model portfolios and learned routing. Ensemble selection ranks validation libraries (Caruana et al., 2004); zero-shot AutoML ranks pipelines (Öztürk et al., 2022); Model Spider ranks and adapts one model from a pretrained zoo (Zhang et al., 2023); limited-label selection chooses one pretrained classifier (Okanovic et al., 2025); vision-language reuse selects several models per class (Tan et al., 2025); and RouteLLM allocates one frozen language model (Ong et al., 2025). Black-box meta-learning transfers through inaccessible APIs (Hu et al., 2023). AssayRouter selects a fixed-size subset, freezes membership before confirmation, and fits convex weights afterward. Post-fit utility supervision is the contribution; HGB supplies the estimator. Matched comparators retain the bank, R/C roles, fan-in, and final NNLS, while other actions and information budgets provide positioning precedents (Appendix A.2.1; Table 3).

6 Discussion and Conclusion

Historical supervision from completed ChEMBL-MT assays transfers across three external collections and six frozen interfaces. Candidate-label interventions establish that utility must remain paired with candidate behavior; identity and interface holdouts show that this relation extends beyond memorized assays or predictor families. The advantage is strongest at 16 labels and narrows as evidence on the target grows, identifying the scarce-label regime in which amortized historical routing matters most.

The separation between historical utility prediction and external NLL gives a broader evaluation principle. A better proxy matters only when it changes membership near the Top-4 boundary and the refitted combiner assigns consequential weight to that change. Router selection should therefore validate the deployed decision under its exact information and call budget, rather than rely on proxy accuracy alone.

The demonstrated scope is continuous regression with homogeneous interface banks, a fixed four-source action, and 20 historical pseudo-targets. Fixing $K = 4$ makes every selector answer the same sparse decision; learning target-dependent fan-in would extend the action space. Expanding the historical panel would test how utility transfer scales with assay diversity, while learned interface representations could replace the current family code when a deployed contract has no historical counterpart. Classification endpoints require a corresponding post-fit utility loss.

Post-fit routing can also complement few-shot adaptation: AssayRouter selects the frozen channel set, after which an adaptive method can operate within that set. The present evidence establishes this selection step. Completed assays provide reusable supervision for selecting frozen prediction channels, improving strict routing while removing repeated subset search on the target.

Reproducibility Statement

Comparisons reuse frozen source predictions and identical support assignments. Automated checks verify the target–interface–budget grid, delayed confirmation access, and the exact number of source calls under strict scoring. The supplement records the data roles, matched assignments, aggregation units, and statistical analyses used for every reported comparison.

AI Use Statement

Generative AI tools were used to assist with manuscript editing and to provide feedback during discussions of experimental design. All experimental protocols, analyses, implementations, and reported results were independently determined or verified by the authors. The authors reviewed all AI-assisted content and take responsibility for the final manuscript.

References

Matthew Adrian, Yunsie Chung, Kevin Boyd, Saee Paliwal, Srimukh Prasad Veccham, and Alan C. Cheng. Multitask finetuning and acceleration of chemical pretrained models for small molecule drug property prediction. *arXiv preprint arXiv:2510.12719*, 2025. URL https://arxiv.org/abs/2510.12719.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022. URL https://arxiv.org/abs/2209.01712.

Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, 2024. doi: 10.1038/s42256-024-00856-0.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Jackson W. Burns, Akshat Shirish Zalte, Charlles R. A. Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H. Green. Deep learning foundation models from classical molecular descriptors. *arXiv preprint arXiv:2506.15792*, 2025. URL https://arxiv.org/abs/2506.15792.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 18, 2004. doi: 10.1145/1015330.1015432.

Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. Meta-learning adaptive deep kernel gaussian processes for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KXRSh0sdVTP.

Peter Eckmann, Jake Anderson, Rose Yu, and Michael K. Gilson. Ligand-based compound activity prediction via few-shot learning. *Journal of Chemical Information and Modeling*, 64(14):5492–5499, 2024. doi: 10.1021/acs.jcim.4c00485.

Vincent Fan and Regina Barzilay. AssayMatch: Learning to select data for molecular activity models. *Journal of Chemical Information and Modeling*, 66(7):3540–3549, 2026. doi: 10.1021/acs.jcim.5c02858.

Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023. doi: 10.1021/acs.jcim.3c00160.

Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srbuhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, and Sheng Wang. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974, 2024. doi: 10.1038/s42256-024-00876-w.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. In-context learning for few-shot molecular property prediction. *arXiv preprint arXiv:2310.08863*, 2023. URL https://arxiv.org/abs/2310.08863.

Zixuan Hu, Li Shen, Zhenyi Wang, Baoyuan Wu, Chun Yuan, and Dacheng Tao. Learning to learn from APIs: Black-box data-free meta-learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13610–13627. PMLR, 2023. URL https://proceedings.mlr.press/v202/hu23g.html.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4c56ff4ce4aaf9573aa5dff913df997a-Abstract-round1.html.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Joongwon Kim, Akari Asai, Gabriel Ilharco, and Hannaneh Hajishirzi. TaskWeb: Selecting better source tasks for multi-task NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11032–11052, 2023. doi: 10.18653/v1/2023.emnlp-main.680. URL https://aclanthology.org/2023.emnlp-main.680/.

Kerstin Kläser, Błażej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew Fitzgibbon. MiniMol: A parameter-efficient foundation model for molecular learning. In *ICML Workshop on Accessible and Efficient Foundation Models for Biological Discovery*, 2024. URL https://openreview.net/forum?id=uA8zvXmiot.

Greg Landrum and RDKit Contributors. RDKit: Open-source cheminformatics software. Software project, n.d. URL https://www.rdkit.org. Archived releases available at https://doi.org/10.5281/zenodo.591637.

Chanhui Lee, Dae-Woong Jeong, Sung Moon Ko, Sumin Lee, Hyunseung Kim, Soorin Yim, Sehui Han, Sungwoong Kim, and Sungbin Lim. Scalable multi-task transfer learning for molecular property prediction. In *ICML Workshop on AI for Science*, 2024. URL https://openreview.net/forum?id=KE6F41otkx.

Ruifeng Li, Mingqian Li, Wei Liu, Yuhua Zhou, Xiangxin Zhou, Yuan Yao, Qiang Zhang, and Hongyang Chen. UniMatch: Universal matching from atom to task for few-shot drug discovery. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.12453. Spotlight.

Eric J. Martin and Xiang-Wei Zhu. Collaborative profile-QSAR: A natural platform for building collaborative models among competing companies. *Journal of Chemical Information and Modeling*, 61(4):1603–1616, 2021. doi: 10.1021/acs.jcim.0c01342.

Eric J. Martin, Valery R. Polyakov, Xiang-Wei Zhu, Li Tian, Prasenjit Mukherjee, and Xin Liu. All-assay-max2 pQSAR: Activity predictions as accurate as four-concentration IC50s for 8558 novartis assays. *Journal of Chemical Information and Modeling*, 59(10):4450–4459, 2019. doi: 10.1021/acs.jcim.9b00375.

Thiago J. M. Moura, George D. C. Cavalcanti, and Luiz S. Oliveira. MINE: A framework for dynamic regressor selection. *Information Sciences*, 543:157–179, 2021. doi: 10.1016/j.ins.2020.07.056. URL https://doi.org/10.1016/j.ins.2020.07.056.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings*

of the 28th International Conference on Artificial Intelligence and Statistics, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043. PMLR, 2025. URL https://proceedings.mlr.press/v258/okanovic25a.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

OpenADMET Consortium. OpenADMET–ExpansionRx blind challenge dataset. Hugging Face Datasets, 2026. URL https://huggingface.co/datasets/openadmet/openadmet-expansionrx-challenge-data. Official post-challenge data release.

Ekrem Öztürk, Fabio Ferreira, Hadi Jomaa, Lars Schmidt-Thieme, Josif Grabocka, and Frank Hutter. Zero-shot AutoML with pretrained models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17138–17155. PMLR, 2022. URL https://proceedings.mlr.press/v162/ozturk22a.html.

Namyong Park, Ryan A. Rossi, Nesreen Ahmed, and Christos Faloutsos. MetaGL: Evaluation-free selection of graph learning models via meta-learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=C1ns08q9jZ.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XrMWUuEevr.

Megan Stanley, John F. Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-Mol: A few-shot learning dataset of molecules. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=701FtuyLlAd.

Hao-Zhe Tan, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Vision-language model selection and reuse for downstream adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58726–58738, 2025. URL https://proceedings.mlr.press/v267/tan25i.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Yifan Xue, Srimukh Prasad Veccham, Saee Paliwal, Tyler Shimko, and Micha Livne. Probabilistic contrastive pretraining for multi-task ADME property prediction. *arXiv preprint arXiv:2606.11508*, 2026. URL https://arxiv.org/abs/2606.11508.

Bingjia Yang, Yunsie Chung, Archer Y. Yang, Bo Yuan, Tianchi Chen, and Xiang Yu. QComp: A QSAR-based imputation framework for drug discovery. *Journal of Chemical Information and Modeling*, 65(15):7862–7873, 2025. doi: 10.1021/acs.jcim.5c00059.

Yi-Kai Zhang, Ting-Ji Huang, Yao-Xiang Ding, De-Chuan Zhan, and Han-Jia Ye. Model spider: Learning to rank pre-trained models efficiently. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0604. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2c71b14637802ed08eaa3cf50342b2b9-Abstract-Conference.html.

Zhiqiang Zhong and Davide Mottin. Automatic auxiliary task selection and adaptive weighting boost molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/61c2975281d60d3b1ce4cefc157d99df-Abstract-Conference.html.

A PROTOCOL AND DECISION CONTROLS

This appendix specifies the training protocol and maps each central claim to its matched control and statistical unit. Tables 2–4 define access, alternative explanations, and the central evidence chain; Tables 5–21 report the interventions and sensitivity analyses; Appendix A.2 contains the matched comparators; and Appendix A.3 analyzes selection cost. Appendix C records the development boundary, and Appendix B separates reproductions that use full source training from the routing comparison at matched support budgets.

A.1 TRAINING, INTERVENTIONS, AND AUDITS

A.1.1 HISTORICAL MATRIX

AssayRouter is trained only on continuous ChEMBL-MT (Adrian et al., 2025) pseudo-targets. Each of 20 targets is crossed with six predictor interfaces, support budgets 16, 32, and 64, and four historical episodes. Each support is divided deterministically into R and C; a third role Q_{disc} is disjoint from both. For support budget n, the two fitting roles contain $n/2$ molecules each (8, 16, or 32), while Q_{disc} draws up to 128 molecules from an independent internal discovery split (25–128 across the 20 historical assays). The primary target is the reduction in discovery loss from adding one candidate to the local target predictor. This gives 27,360 distinct candidate contexts. To match the refinements' state frequency, each candidate context is repeated over the same grid of 13 states, yielding 355,680 rows. The router is a histogram gradient-boosting regressor conditioned on the interface and selected by grouped cross-validation over target identity.

A.1.2 FROZEN FITTING CONFIGURATION

The local target learner is Ridge regression with $\alpha = 1$ on molecular fingerprints. Four shuffled folds produce its out-of-fold support column; the fit on all C rows produces query predictions. Regression labels are centered by the median and divided by 1.4826 times the median absolute deviation, falling back to the sample standard deviation and then 10^{-6}. The local column and selected source columns are fitted by NNLS and normalized to the simplex. HGB model selection compares three fixed configurations (η, T, L, m, λ): $(0.05, 200, 15, 20, 1)$, $(0.05, 300, 31, 30, 1)$, and $(0.08, 220, 31, 40, 3)$, denoting learning rate, iterations, maximum leaves, minimum leaf samples, and ℓ_2 regularization. Early stopping is disabled. Five folds grouped by target select MAE averaged over targets, after which the winning configuration is refitted on the complete historical matrix.

Table 2: Information available to the principal selectors. Query-local methods may change the selected set across confirmation molecules.

Selector	History	Full-bank query	Query-local
AssayRouter	Utility	No	No
Support-CV	None	No	No
Frozen-DES	None	Yes	Yes
MINE-WS (Moura et al., 2021)	Competence	Yes	Yes
All-source	None	Yes	No

Source-quality is the capacity-matched learned selector: it retains all 11 candidate coordinates, HGB capacity search, grouped folds, the matched expansion over 13 states, Top-4 action, external assignments, and final NNLS, changing only the historical supervision label. It therefore tests whether learning a selector from generic source strength can explain the result. Additional model selection methods with the same inputs and action would repeat this scientific role; a new numerical comparator is relevant only when it instantiates an uncovered supervision, access, timing, or action axis.

A.1.3 SUPPORT-SCALE Q_{disc} UTILITY-LOSS SENSITIVITY

For each frozen historical episode, we recompute only the utility loss as $\ell_{\mathrm{NLL}}(e_C \sigma_C / \sigma_{R \cup C})$, where e_C is the original discovery residual standardized on C. Routing features remain local to R, the local target and normalized NNLS fits remain local to C, and external deployment is unchanged. The

Table 3: Coverage of alternative explanations relevant to the decision across five contract axes. Each control changes one scientific axis while preserving the frozen prediction contract wherever that axis permits.

Alternative explanation	Matched test	Changed axis
No value from bank reuse	Target-only	Bank reuse
Generic source strength suffices	Source-quality	Utility supervision
Metadata/support statistics suffice	Size / SNR	Ranking signal
Profile compatibility suffices	Residual (strict)	Ranking signal
Candidate–label pairing is arbitrary	Frozen permutations	Ranking signal
Target-side search suffices	Support-CV@4	Access/timing
Query-local competence suffices	DES@4 / MINE-WS@4	Access/timing
Sparse fan-in creates the gain	All-source convex	Action/fan-in
Calibration or state creates the gain	Δ / C / Z / Marginal / State	Utility supervision

Table 4: Evidence-chain index. Effects favor AssayRouter-C; the final column reports the scientific unit appropriate to each test.

Test	NLL effect	Positive breadth
Candidate-label intervention	0.1352	90.7% of cells
Unseen-interface intervention	0.0497	6/6 held-out families
Support-CV@4 vs. C	0.0409	77.8% of target–interface units
Residual compatibility vs. C	0.1844	83.3% of targets

sensitivity reuses the original direct and leave-one-interface-out donor mappings and every external assignment. Each paired decision uses the same 24 targets, six interfaces, three budgets, and 12 support episodes as its parent experiment, with the target–interface bootstrap of 10,000 resamples defined in Appendix A.1.10. The prespecified criteria require positive NLL interval lower bounds for the direct, strict, unseen-interface, and strict-search comparisons, plus positive unseen-interface effects in at least four of six held-out families. The direct, strict, unseen-interface, and matched-search comparisons preserve the candidate-specific utility advantage (Table 5).

Table 5: Support-scale Q_{disc} sensitivity. Values are paired point effects; all five prespecified criteria are met.

Comparison	Effect	Value
Direct intervention	Permuted − genuine NLL	$+0.0463$
Strict intervention	Permuted − genuine NLL	$+0.0816$
Unseen interface	Permuted − genuine NLL	$+0.0299$
Unseen interfaces	Positive held-out families	$5/6$
Strict search	Support-CV − genuine NLL	$+0.0439$

A.1.4 SOURCE-IDENTITY HOLDOUT

The 20 historical candidate-source identities are disjoint from the 37 source-assay identities in the external prediction banks. We additionally cross-fit the historical matrix over five target folds and five candidate-source folds, training each of 25 HGB models only on rows whose pseudo-target and candidate source both lie outside the held folds. Every one of the 355,680 historical rows is predicted exactly once, with source identity absent from the features. Relative to the original target-held-out predictions, the double-held-out model has matched utility MAE (-0.00006, 95% interval $[-0.00184, 0.00178]$) and matched Top-4 overlap (-0.00382, $[-0.02205, 0.01493]$); fine-grained Spearman correlation decreases by 0.0197 ($[-0.0355, -0.0025]$). The deployment candidate set therefore transfers across historical source identities; the observed decrease is confined to fine-grained ordering beyond the selection boundary.

A.1.5 MATCHED POST-FIT LABEL CONTROLS

These controls retain the historical rows, $R/C/Q_{\text{disc}}$ roles, routing features, interface code, grouped HGB search, external assignments, Top-4 rule, and simplex fit. They replace only the historical label with negative post-fit loss after adding the candidate, its percentile rank within context, its value centered within the block, or its value standardized within the block. Raw loss and rank are weaker than standalone utility. AssayRouter-C and AssayRouter-Z remain matched to AssayRouter-Δ in overall strict NLL; their control-minus-Δ effects are -0.0039 with interval $[-0.0102, 0.0017]$ and $+0.0039$ with interval $[-0.0044, 0.0130]$. This identifies post-fit candidate value relative to its historical block as the transferable supervision shared by the successful constructions.

Table 6: Strict matched post-fit controls. Point estimates are averaged after each four-call constituent is scored.

Method	NLL ↓	MAE ↓	RMSE ↓	Spearman ↑
AssayRouter-Δ	2.1142	23.5806	35.3864	0.2961
AssayRouter-C	2.1102	23.3872	35.0709	0.3011
AssayRouter-Z	2.1181	23.6652	35.3589	0.3124
Post-fit loss	2.1684	25.9818	37.2848	0.3002
Post-fit rank	2.1515	24.9610	36.3301	0.3228

A.1.6 DIRECT ASSAYROUTER-C CORRESPONDENCE INTERVENTION

The primary centered label receives the same frozen candidate bijection used by the standalone-utility intervention. Permutation and block centering commute to numerical precision, so the manipulation preserves every block label multiset while changing only which candidate receives each value. The same HGB search, external assignments, Top-4 rule, and simplex fit are replayed. Table 7 shows that the candidate–label pairing remains decisive for the selected deployment variant under both estimands.

Table 7: Direct AssayRouter-C candidate-label intervention. NLL effects are permuted minus genuine; win rate is the fraction of 432 target–interface–budget cells favoring the genuine mapping. Intervals use target–interface clustered resampling.

Estimand	NLL effect	95% interval	Cell win rate
Cross-fit	0.0814	$[0.0427, 0.1246]$	0.7222
Strict	0.1352	$[0.0852, 0.1915]$	0.9074

Tables 8 and 9 give the complete training/evaluation inventory and the six fixed interface contracts used by the reported AssayRouter experiments.

Table 8: Training and frozen external evaluation. External target identities and confirmation labels are absent from router training and model selection.

Role	Collection	Targets	Interfaces
Historical supervision	ChEMBL-MT (Adrian et al., 2025)	20	6
Temporal holdout	ExpansionRx (OpenADMET Consortium, 2026)	9	6
Sparse-panel holdout	Biogen (Fang et al., 2023)	6	6
Scaffold holdout	TDC regression (Huang et al., 2021)	9	6

Within each source, Morgan/Ridge and CheMeleon/Ridge use 12 fixed members, RDKit2D/LightGBM and ChemBERTa2/LightGBM use eight, and GIN and the fine-tuned CheMeleon head use three. The interface returns the member mean as $z_s(x)$ and their empirical standard deviation as $u_s(x)$.

The feature inventory below separates the primary 11-coordinate AssayRouter input from the 14 state coordinates used only by the extension.

Table 9: The six formal interfaces to frozen predictors. Abbreviations keep each cell on one line: FP, emb., repr., and GIN denote fingerprint, embedding, representation, and graph-isomorphism network. A bank contains independently trained source-assay predictors from one interface family.

Molecular representation	Source predictor	Reuse mode
Morgan FP (Rogers & Hahn, 2010)	Ridge	Contract
RDKit2D (Landrum & RDKit Contributors, n.d.)	LightGBM (Ke et al., 2017)	Contract
CheMeleon emb. (Burns et al., 2025)	Ridge	Frozen repr.
ChemBERTa2 emb. (Ahmad et al., 2022)	LightGBM (Ke et al., 2017)	Frozen repr.
Molecular graph	GIN (Xu et al., 2019)	Per-source
CheMeleon repr. (Burns et al., 2025)	Prediction head	Fine-tuned

A.1.7 Feature inventory

The primary 11-coordinate HGB utility prior uses signed and absolute residual alignment, residual correlation, source mean and standard deviation, absolute source response, uncertainty mean and spread, signal-to-noise ratio, log size of the routing support, and declared training size. The state-conditioned extension adds signed and absolute remaining alignment, remaining correlation, mean and maximum redundancy among selected sources, set size, residual scale, variance reduction, sum of source weights, number of active sources, current and augmented condition numbers, and effective rank. The six interface indicators are appended separately to every design. All coordinates are available before confirmation labels are read; coordinates derived from predictions use R, and training size is interface metadata.

The three routers share the 11 coordinates for each candidate used by AssayRouter and both refinements. The remaining 14 coordinates are available only to the state-conditioned extension. The primary router represents each candidate context by 13 equally weighted copies with identical candidate features and standalone label. Marginal-over-states uses the 13 labels specific to each state, while the state-conditioned extension additionally exposes the partial set. This schema of 11 candidate and 14 state coordinates is separate from the earlier pairwise feature design summarized in Appendix C.

All three regressors append the same interface code to those coordinates. The full historical model uses six one-hot indicators. Each leave-one-interface-out model fits an encoder over the five observed interfaces and maps the held-out family to a five-dimensional zero vector, with no calibration specific to that interface.

Table 10 compares standalone and marginal-over-states supervision with the same estimator, rows, and external replay.

Table 10: Matched supervision refinements. Differences are marginal-over-states utility NLL minus standalone-utility NLL; positive values favor the primary utility prior.

Scope	Mean difference
Overall	−0.0038
16 labels	−0.0067
32 labels	−0.0002
64 labels	−0.0043
Biogen	−0.0035
ExpansionRx	0.0008
TDC	−0.0085

Table 11 aligns the supervision, representation, and external decision quality of the three nested routers.

Table 12 then isolates prediction refinement on the shared marginal target before any external routing decision is made.

Table 11: Prediction refinement and external decision quality. Historical out-of-fold (OOF) mean-squared error (MSE) is defined only for the shared marginal-over-states label; lower is better.

Router	Historical target	State features	External NLL
AssayRouter (standalone-utility prior)	$\Delta(s \mid \varnothing)$	No	2.0418
Marginal-over-states utility	$\mathbb{E}_A \Delta(s \mid A)$	No	2.0380
State-conditioned extension	$\Delta(s \mid A)$	Yes	2.0392

Table 12: Nested post-fit utility-prediction audit. Both columns use frozen target-grouped out-of-fold predictions for the same marginal labels.

Scope	Marginal utility MSE	State-conditioned MSE	Relative reduction
Overall	0.02064	0.01949	5.58%
16 labels	0.02584	0.02419	6.41%
32 labels	0.02061	0.01923	6.72%
64 labels	0.01546	0.01504	2.66%

A.1.8 Historical-label permutations

The primary intervention uses one prespecified bijection, independent of labels, to permute standalone utility among 19 candidates inside each block defined by target, interface, support budget, and episode. Candidate features and the label multiset remain fixed, and the same permuted label is repeated over the matched grid of 13 states. The frozen mapping moves candidates in every block and retains 5.32% fixed points overall. The clustered interval estimates the external effect of this frozen intervention; no Monte Carlo permutation p-value is inferred. A supporting intervention permutes the *complete utility trajectory over 13 states* for marginal-over-states utility. It preserves all 355,680 labels, coverage of set sizes, and trajectory structure. Both interventions retain estimator capacity, grouped validation, external assignments, $K = 4$ combination, and the boundary on confirmation labels.

The main bijection is replicate 1 in Figure 2B. Four additional fixed mappings repeat the intervention without changing the training or replay protocol. Table 13 reports their external point effects; every replicate has a positive clustered lower bound.

Table 13: Independent candidate-label interventions under the cross-fit prediction-average estimand. Each row uses a distinct frozen bijection; effects are permuted minus genuine NLL.

Frozen mapping	Mean NLL effect	Cell win rate
1	0.0494	73.38%
2	0.0477	70.83%
3	0.0146	60.65%
4	0.0339	72.69%
5	0.0454	71.53%

A.1.9 External replay

The three external collections by six interfaces form 18 evaluated deployments. Each target–interface–budget cell has 12 support episodes. Every episode produces eight balanced R/C partitions and two role directions. The prediction-averaged estimand averages the 16 raw-scale predictions before scoring; the strict estimand scores each four-source constituent and then averages its scalar losses. Sparse methods select exactly four source inputs per constituent, and all methods fit the same no-prior simplex NNLS on C. The target predictor is deleted from the bank, external targets never enter router training, and confirmation labels enter only after all predictions have been frozen.

A.1.10 STATISTICAL UNIT AND DIRECTION

Support repetitions, partitions, and role directions are averaged before scientific comparison. A cell is a collection–target–interface–budget combination. Aggregate uncertainty uses a two-way pigeonhole bootstrap that independently resamples target and interface clusters 10,000 times. Unless stated otherwise, positive paired differences mean comparator NLL minus the stated AssayRouter variant.

A.1.11 SMALL-CLUSTER SENSITIVITY

We recompute the three headline NLL effects after leaving out each target, interface, or collection in turn, with frozen predictions and no model refitting. The strict candidate-label intervention, leave-one-interface-out transfer, and strict Support-CV comparison remain positive in every deletion. Exact sign-flip tests over the six interface effects give $p = 0.03125$ for all three comparisons; sign-flip tests over the 24 target effects also remain positive. This sensitivity treats the small number of clusters directly and agrees with the target–interface bootstrap used for the primary analysis.

A.1.12 EVIDENCE DESIGN

Five controlled changes answer distinct questions while preserving frozen assignments and the boundary on confirmation labels. Candidate-label permutation tests whether post-fit utility remains informative when paired with the correct source; leave-one-interface-out training tests transfer to an unseen family; source-quality supervision tests whether generic source strength suffices; the target-only learner tests whether the frozen bank adds value; and strict constituent scoring tests whether the result survives the deployed call budget.

A.1.13 MATCHED ALTERNATIVE EXPLANATIONS

The target-only control keeps the same cross-fitted local learner, support roles, and confirmation scoring while deleting every source call. The source-quality prior keeps the complete AssayRouter training and replay pipeline but replaces post-fit utility with the improvement on the discovery set of an affine-calibrated source over a constant predictor; its label never fits or evaluates the local target model. Two deterministic rankings select the four sources with the largest declared training sets or the strongest signal-to-noise ratio on the routing set. Support-greedy uses routing-support labels to add one source at a time, then freezes the set before the same final fit and confirmation scoring. These controls share the assignments and are evaluated under both prediction-averaged and strict constituent estimands. Table 14 reports point effects; Fig. 5 shows their target–interface distributions.

Table 14: **Matched alternatives.** Paired mean NLL differences; positive values favor AssayRouter-Δ.

Alternative	Cross-fit	Strict
Target-only	0.2403	0.2197
Source-quality	0.0969	0.1433
Train-size	0.0912	0.1174
Source-SNR	0.0789	0.1304
Support-CV	0.0030	0.0369

Table 15 preserves the exact values behind Figure 2A. Rows are grouped by scientific role because refinements, information-rich controls, interventions, and classical diagnostics do not form a single access-matched leaderboard.

Table 16 reports the direct intervention at the aggregate, budget, collection, and interface levels.

The direct intervention moves all four matched metrics in the same scientific direction. Table 17 keeps errors in the original units separate from support-standardized NLL and reverses the Spearman subtraction so that positive entries always favor genuine post-fit labels.

Tables 18 and 19 replay the prespecified primary mapping under the strict deployment estimand. Each constituent is scored before aggregation, so every confirmation prediction uses exactly four source calls and no averaged prediction vector.

Table 15: **Stability-focused cross-fit estimand.** Absolute support-standardized robust NLL on 24 external targets excluded from router fitting and six prediction interfaces; lower is better. All sparse constituent models have four-source combiner fan-in, but their information permissions differ as described in the text.

Method	16 labels	32 labels	64 labels
AssayRouter (standalone-utility prior)	2.1606	2.0300	1.9348
Matched refinements			
Marginal-over-states utility	2.1539	2.0297	1.9305
State-conditioned extension	2.1589	2.0285	1.9301
Search and information-rich controls			
Support-CV@4 ensemble	2.1806	2.0315	1.9223
Frozen-DES@4	2.1678	2.0268	1.9231
MINE-WS@4 (Moura et al., 2021)	2.2058	2.0611	1.9309
All-source convex	2.1771	2.0351	1.9303
Mechanism intervention and classical diagnostics			
AssayRouter, utility labels permuted	2.2419	2.0714	1.9602
Residual compatibility	2.3634	2.1490	1.9949
Support-greedy selection (Caruana et al., 2004)	2.2133	2.0530	1.9572
50 random four-source sets	2.2212	2.0619	1.9515

Table 16: Direct standalone-utility permutation. Differences are permuted-router NLL minus genuine AssayRouter NLL; positive values favor genuine post-fit labels.

Scope	Mean difference	Win rate
Overall	0.0494	73.38%
16 labels	0.0813	84.03%
32 labels	0.0415	72.22%
64 labels	0.0254	63.89%
Biogen	0.0030	50.00%
ExpansionRx	0.0865	85.19%
TDC	0.0433	77.16%
ChemBERTa2 / LightGBM	0.0838	80.56%
Fine-tuned CheMeleon	0.0321	68.06%
Frozen CheMeleon / Ridge	0.0397	66.67%
GIN from scratch	0.0316	70.83%
Morgan / Ridge	0.0327	75.00%
RDKit2D / LightGBM	0.0766	79.17%

Table 20 asks whether the primary candidate–utility mapping centered within each block transfers when an entire predictor family is absent from historical supervision. Each router trains on the other five families, maps the held-out interface to a five-dimensional zero vector, forbids calibration specific to that interface, and reuses the same 5,184 external episode pairs.

A.1.14 Compound provenance and nonoverlap sensitivity

Canonical parents are the largest sanitized fragments represented by canonical isomeric SMILES. The analysis detects no leakage of target labels. Biogen and ExpansionRx have zero exact-parent overlap with training data for the historical router or source predictors; TDC contains compound exposure across assays. For TDC, H-clean removes confirmation parents used to train the historical router, S-clean removes parents used to train any other source predictor, and Union-clean applies both filters. Stored assignments, selections, weights, and predictions are rescored without retraining or reselection. All six interface replays reproduce metrics on the full set to numerical precision before filtering, and confirmation labels are read only after cell predictions are frozen. A target enters clustered inference only when at least 16 confirmation molecules remain; the clearance microsome AZ endpoint retains five Union-clean rows and is excluded, leaving eight TDC targets. This sensitivity

Table 17: Secondary metrics for the direct standalone-utility permutation. NLL, MAE, and RMSE are permuted minus genuine; Spearman is genuine minus permuted.

Metric	Mean effect
Robust NLL	0.0494
MAE	3.5805
RMSE	3.1850
Spearman correlation	0.0327

Table 18: Strict four-call standalone-utility intervention. Differences are permuted minus genuine NLL; positive values favor AssayRouter.

Scope	Mean difference
Overall	0.0870
16 labels	0.1260
32 labels	0.0816
64 labels	0.0533
Biogen	0.0263
ExpansionRx	0.1081
TDC	0.1063
ChemBERTa2 / LightGBM	0.1295
Fine-tuned CheMeleon	0.0681
Frozen CheMeleon / Ridge	0.0940
GIN from scratch	0.0494
Morgan / Ridge	0.0675
RDKit2D / LightGBM	0.1133

measures exact-parent overlap; scaffold independence lies outside its scope. Table 21 reports each frozen filtering scope.

A.2 Routing comparators and search cost

The following inventory separates AssayRouter from the closest paradigms for reusing models and tasks by the supervision available before deployment and the action taken on a new task.

A.2.1 Closest reuse paradigms

All-Assay-Max2 and collaborative pQSAR use prediction profiles and target–assay correlation to filter channels and fit target PLS without a fixed K (2019; 2021). Model Label Learning uses capability labels for each concept to select top-k VLMs per class and ensemble zero-shot predictions (2025). Model Spider uses approximate historical model rankings to rank a model zoo and adapt its highest-ranked model (2023). AssayMatch uses assay compatibility supervised by TRAK to rank assays for an unlabeled target and train on selected data (2026). AssayRouter uses standalone post-fit loss reduction to select one global Top-4 set from the support data and then fits a convex combiner.

The cross-fit Support-CV comparison uses AssayRouter-Δ, whose utility is uncentered; the primary strict comparison uses AssayRouter-C (Table 23). Table 22 reports cross-fit accuracy, and the following cost analysis records the subset fits and source calls on the target required to construct it.

The closest reuse methods above are positioning precedents rather than numerical baselines with matched access. Residual compatibility is the profile correlation control within the contract for the explanation suggested by pQSAR. The other original contracts require dense historical prediction profiles, rankings over a model zoo, assay descriptions, training on selected data, or auxiliary experiments, and their actions range from choosing one model to training a new model. Adapting them to frozen heterogeneous outputs, the declared support roles, a fixed four-source action, and the shared NNLS would define a new router. The alternatives within the contract are therefore tested by

Table 19: Overall secondary effects for the strict four-call intervention. Losses are permuted minus genuine; Spearman is genuine minus permuted.

Metric	Mean effect
Robust NLL	0.0870
MAE	4.6262
RMSE	4.9810
Spearman correlation	0.0405

Table 20: AssayRouter-C leave-one-interface-out evaluation. Differences are comparator minus genuine-router NLL after the named family is removed from historical supervision; positive values favor AssayRouter-C.

Held-out scope	Permuted	Support-CV
Overall	0.0497	0.1116
16 labels	0.0697	0.1769
32 labels	0.0492	0.1029
64 labels	0.0301	0.0550
Biogen	−0.0023	0.0367
ExpansionRx	0.0770	0.1594
TDC	0.0570	0.1138
ChemBERTa2 / LightGBM	0.0060	0.1117
Fine-tuned CheMeleon	0.0188	0.0908
Frozen CheMeleon / Ridge	0.0121	0.1228
GIN from scratch	0.0673	0.0955
Morgan / Ridge	0.0963	0.1298
RDKit2D / LightGBM	0.0975	0.1190

the coverage matrix in Table 3; Support-CV@4 is the primary matched search comparator, while Frozen-DES@4 and MINE-WS@4 deliberately receive richer query access.

A.2.2 Strict residual compatibility

For each directional partition, a four-fold out-of-fold local target predictor defines $r_{t,i} = \widetilde{y}_{t,i} - \widehat{y}^{\mathrm{CF}}_{0,t}(x_i)$ on R_t. The profile correlation score is

$$c_t(s) = \mathrm{corr}_{i \in R_t}(\widetilde{z}_s(x_i), r_{t,i}), \qquad A_t^{\mathrm{res}} = \mathrm{TopK}_{s \in \mathcal{S}_t}\, c_t(s). \tag{20}$$

Candidates are ranked by decreasing signed Pearson correlation; the canonical bank order breaks exact ties. The score and ranking use R_t only, while C_t fits the same simplex NNLS with the target predictor and four sources as AssayRouter-C. On the identical 5,184 strict episodes, its NLL is 2.5339, 2.2696, and 2.0806 at 16, 32, and 64 labels, respectively; the overall NLL is 2.2947. The NLL difference between Residual and AssayRouter-C is 0.1844 across the 432 paired target–interface–budget cells, with a two-way target–interface bootstrap interval of $[0.1138, 0.2609]$. The corresponding MAE, RMSE, and Spearman values are 32.3550, 44.8218, and 0.3053. This control shares the call budget, access timing, assignments, and final fit while replacing historical post-fit supervision with one correlation rule fitted to the target.

Support-CV@4 averages 16 $K = 4$ directional constituents per episode. It therefore makes 64 source calls, uses 5–21 unique sources per episode (mean 10.45), and has zero episodes with a strict four-source union. The full evaluation performs 20,445,696 inner NNLS fits: 414,720 on Biogen, 8,709,120 on ExpansionRx, and 11,321,856 on TDC. These counts arise during selection to construct the prediction-averaged ensemble; Table 23 separately evaluates accuracy under the strict four-call estimand.

The one-shot replay evaluates both methods under the same strict four-call estimand. Support-CV@4 and AssayRouter-C are each scored as 16 independent four-call constituents on the same 5,184

Table 21: TDC exact-parent nonoverlap sensitivity. Differences are permuted minus genuine NLL for the AssayRouter-Δ replication; positive values favor historical candidate–utility correspondence. Retention is the mean confirmation fraction.

Router	Scope	Targets	Retention	Mean difference
Direct	Full	9	100.00%	0.0433
Direct	H-clean	9	62.38%	0.0419
Direct	S-clean	8	55.97%	0.0403
Direct	Union-clean	8	42.71%	0.0385
LOIO	Full	9	100.00%	0.0525
LOIO	H-clean	9	62.38%	0.0512
LOIO	S-clean	8	55.97%	0.0464
LOIO	Union-clean	8	42.71%	0.0468

Table 22: Matched Support-CV@4 cross-fit comparison. NLL is computed after averaging the same 16 directional prediction vectors. Differences are Support-CV minus AssayRouter-Δ.

Scope	Support-CV NLL	Difference
Overall	2.0448	0.0030
16 labels	2.1806	0.0200
32 labels	2.0315	0.0015
64 labels	1.9223	-0.0125

episodes. Table 23 gives the paired robust-NLL comparison at every prespecified scope; the complete replay retains the secondary metrics under the same assignments.

Table 23: Strict four-call Support-CV@4 comparison. Differences are Support-CV NLL minus AssayRouter-C NLL; positive values favor the frozen historical prior.

Scope	Mean difference
Overall	0.0409
16 labels	0.0736
32 labels	0.0413
64 labels	0.0078
Biogen	0.0005
ExpansionRx	0.0716
TDC	0.0371
ChemBERTa2 / LightGBM	0.0444
Fine-tuned CheMeleon	0.0238
Frozen CheMeleon / Ridge	0.0403
GIN from scratch	0.0348
Morgan / Ridge	0.0513
RDKit2D / LightGBM	0.0505

The right panel of Figure 6 shows the target–interface distributions for the matched AssayRouter-Δ comparison. For AssayRouter-C, the overall difference is 0.0409 with interval $[0.0211, 0.0623]$; the 16- and 32-label effects and the ExpansionRx and TDC effects are resolved, whereas the 64-label and Biogen intervals cross zero. AssayRouter-C has lower NLL in 77.8% of the 144 target–interface units and 63.3% of the 5,184 support episodes. These rates use the same 16 independently scored four-call constituents as the strict mean.

A.3 AMORTIZED SELECTION COST

A.3.1 SELECTION-WORK PROPOSITION

Suppose M candidate predictions and their routing features have already been constructed. AssayRouter selects K sources using M evaluations of a frozen scalar score and a top-K operation, with zero counterfactual fits on the target. A sequential exhaustive selector that adds one candidate at a time and fits every remaining candidate requires

$$\sum_{j=0}^{K-1} (M - j) = KM - \frac{K(K-1)}{2} \tag{21}$$

counterfactual fits on the target per routing split.

A.3.2 PROOF AND BOUNDARY

At step j, the sequential selector has already chosen j sources and must evaluate each of the $M - j$ remaining candidates to certify its next greedy choice. Summing over $j = 0, \ldots, K-1$ gives the expression above. AssayRouter evaluates its historical score independently for all M candidates and obtains the set by top-K. The count isolates exact selection work after frozen predictions and candidate features have been constructed; downstream accuracy, the final NNLS combiner, approximate search, caching, and specialized algebraic updates have separate costs. Support-CV@4 is a stronger screened exhaustive implementation. With F cross-fitting folds, screen width L, bank size M, and subset size K, its selection stage evaluates

$$F\left[M\,\mathbf{1}\{M > L\} + \binom{\min(M, L)}{K}\right] \tag{22}$$

counterfactual support fits per direction: the screening term applies when the bank exceeds L, and the second exhaustively scores the surviving subsets of size K. For the formal comparator, $L = 8$ and $K = 4$. Both the singleton screen and subset search use only labels and predictions from the routing support, and the selected set is frozen before confirmation scoring; neither post-fit utility nor confirmation labels enter the search. The implementation repeats this operation across targets, interfaces, budgets, episodes, partitions, and role directions. Table 22 reports the resulting total of 20,445,696 inner NNLS fits. For TDC, this workload is the same screened procedure over the $L = 8$ survivors.

Table 24: Comparison scopes. Selection fit counts measure subset construction on the target; strict confirmation calls define the accuracy budget. Counts exclude total wall-clock latency.

Quantity	AssayRouter-C	Support-CV@4
Historical label NNLS fits	28,800	0
Selection NNLS fits per direction	0	246.5
Selection NNLS fits, full evaluation	0	20,445,696
Strict confirmation calls per constituent	4	4
Constituents per cross-fit episode	16	16

A.3.3 AUDITED FIT-COUNT CROSSOVER

Building the complete historical supervision matrix used 711,360 inner NNLS fits; constructing only the standalone labels used by the primary router requires 28,800, followed by 16 HGB fits for model selection and the final model. Support-CV@4 averages 246.5 inner NNLS fits per routing direction, or 3,944 per cross-fit evaluation episode with 16 directions. Measured by this shared NNLS primitive, its repeated search reaches the cost of the primary standalone labels after 116.8 routing directions, equivalent to 7.3 such cross-fit episodes, and the full historical development cost after 2,885.8 directions, or 180.4 episodes. These crossover counts exclude feature construction, HGB execution, frozen source prediction, and the final combiner; they quantify the shared NNLS primitive rather than wall-clock time.

Table 25: **Cross-fit greedy and information-rich diagnostics by collection.** Comparator NLL minus AssayRouter-Δ NLL, averaged over interfaces and support budgets; positive values favor the post-fit utility prior.

Comparator	Biogen	ExpansionRx	TDC
50 random four-source sets	−0.0026	0.0713	0.0275
Support-greedy selection (Caruana et al., 2004)	−0.0047	0.0547	0.0356
Frozen-DES@4	−0.0055	0.0013	−0.0044
MINE-WS@4 (Moura et al., 2021)	−0.0010	0.0421	0.0229
All-source convex	−0.0019	0.0440	−0.0274

Table 26: **Cross-fit greedy and information-rich diagnostics by interface.** Comparator NLL minus AssayRouter-Δ NLL, averaged over external targets and budgets; positive values favor the post-fit utility prior.

Interface	Random	Greedy (Caruana et al., 2004)	DES@4	All-source
Morgan / Ridge	0.0333	0.0292	−0.0010	0.0078
RDKit2D / LGBM	0.0445	0.0413	−0.0005	0.0140
CheMeleon / Ridge	0.0311	0.0304	−0.0096	−0.0023
ChemBERTa2 / LGBM	0.0474	0.0459	0.0070	0.0100
GIN	0.0367	0.0351	0.0026	0.0126
CheMeleon / FT	0.0254	0.0141	−0.0138	−0.0078

Table 25 localizes the matched primary comparators by external collection; Table 26 gives the complementary view across frozen predictor families.

Table 27 preserves the marginal-over-states intervention as an independent replication in the same direction as the direct standalone-utility result.

Table 27: Matched marginal-over-states mechanism tests. Differences are comparator NLL minus genuine marginal-over-states NLL; positive values favor the genuine labels.

Comparator and scope	Mean difference
State-conditioned extension, overall	0.0011
State-conditioned, 16 labels	0.0051
State-conditioned, 32 labels	−0.0012
State-conditioned, 64 labels	−0.0004
Marginal labels permuted, overall	0.0532
Marginal labels permuted, 16 labels	0.0824
Marginal labels permuted, 32 labels	0.0428
Marginal labels permuted, 64 labels	0.0344
Marginal labels permuted, Biogen	0.0006
Marginal labels permuted, ExpansionRx	0.0716
Marginal labels permuted, TDC	0.0698

B External Checkpoint and Benchmark Reproductions

B.1 Checkpoint and benchmark reproductions

B.1.1 Official few-shot checkpoints

We downloaded the no-FEP ChEMBL checkpoints released with ActFound (Feng et al., 2024) and evaluated five official systems (ActFound, ActFound-transfer, MAML, ProtoNet, and transfer-QSAR) on the nine TDC regression targets. Every method receives the identical 16/32/64 support molecules used by the corresponding AssayRouter episode and predicts the same official test split.

The official 2,048-dimensional count-Morgan input and five-step adaptation are preserved. Across means computed over targets, the strongest official checkpoint is ProtoNet: the NLL difference between AssayRouter and ProtoNet is -0.0190, 0.2182, and 0.0125 at the three budgets, where positive values favor ProtoNet. Bootstrap intervals over targets cross zero in all three cases; the complete matrix for all five methods accompanies the submission.

B.1.2 Published Biogen predictors

We also reproduce methods evaluated on the exact public Biogen (Fang et al., 2023) split. The original Random Forest (Breiman, 2001) and LightGBM (Ke et al., 2017) code gives mean Pearson correlations of 0.6683 and 0.7069, close to the published 0.6700 and 0.7000. The official MolSetRep (Boulougouri et al., 2024) GINE implementation completes three runs for each of six endpoints and reaches mean Pearson $r = 0.6373$; its SR-GINE counterpart reaches $r = 0.7259$ in another 18 runs. Across both neural models, endpoint values reproduce the published table within 0.029. These are full-training predictors and are reported separately from support-budget-matched routing methods. Table 28 gives the resulting macro correlations and run counts.

Table 28: Official full-training reproduction on all six public Biogen endpoints. Pearson r is macro-averaged across endpoints.

Method	Runs	Macro Pearson r
Random Forest (Breiman, 2001)	6	0.6683
LightGBM (Ke et al., 2017)	6	0.7069
MolSetRep GINE (Boulougouri et al., 2024)	18	0.6373
MolSetRep SR-GINE (Boulougouri et al., 2024)	18	**0.7259**

B.1.3 MiniMol reproduction

The official MiniMol (Kläser et al., 2024) protocol uses all 22 TDC ADMET tasks, train/validation splits from five author-provided seeds, five fold heads per split, and 25 epochs per head. This yields 550 fitted heads above one frozen official checkpoint. All 22 tasks complete; the mean and maximum absolute deviations from the released metrics are 0.0122 and 0.0830, respectively. The complete task table and released split definitions accompany the submission.

B.1.4 Contrastive KERMT reproduction

We fine-tune the official Contrastive KERMT (Xue et al., 2026) checkpoint for 100 epochs on each of five seeds and retain the authors' multitask readout for each task. The reproduction completes four Biogen endpoints with high coverage and all nine ExpansionRx endpoints, giving mean absolute error (MAE) averaged over endpoints of 0.3515 and 0.2836, respectively. Table 29 separates these public-checkpoint reproductions from the manuscript's reported macro means; both use full training and remain separate from routing at matched support budgets.

Table 29: Full-training KERMT comparison (macro MAE; lower is better). Author values are transcribed from the official manuscript source; reproductions use the released Contrastive KERMT checkpoint, five seeds, and the frozen public splits.

Protocol	Biogen (4)	ExpansionRx (9)
Author KERMT task-specific	0.3320	0.3750
Author best Contrastive KERMT	**0.3210**	0.3590
Official-checkpoint reproduction	0.3515	**0.2836**

Target-only and all-source fit the same regularized combiner on the local prediction or the full bank. Compatibility, training-size, and signal-to-uncertainty selectors rank the frozen bank by their named support statistic. Local-kernel estimates residual compatibility among molecular neighbors. QComp-style (Yang et al., 2025) applies a residual-covariance update through the same interface;

original QComp can observe auxiliary experimental values for a query and therefore has a different information budget.

B.1.5 STATISTICAL AND REPRODUCIBILITY DETAILS

Let $r = (y - \hat{y})/\sigma_{\text{sup}}$, where σ_{sup} is the median absolute deviation of the complete episode support $R \cup C$, multiplied by 1.4826. The sample standard deviation is used when that quantity vanishes, with a 10^{-6} floor. Robust NLL is the Student-t loss for point predictions with $\nu = 3$,

$$\ell_{\text{NLL}}(r) = \tfrac{1}{2}\log(\nu\pi) + \log\Gamma(\nu/2) - \log\Gamma((\nu+1)/2) + \tfrac{\nu+1}{2}\log(1 + r^2/\nu). \tag{23}$$

The methods do not emit predictive variances for this metric. We average support repetitions before comparison. Aggregate bootstraps independently resample predictor interfaces and target tasks, preserving their crossed structure. Each analysis uses at least 4,000 replicates, and complete intervals accompany the submission.

All controls reuse identical supports, predictions, combiners, and evaluation rows. Random subsets are fixed before scoring, and confirmation labels never participate in selection. The supplementary records preserve metrics, selections, access checks, and the evaluation configuration for every deployment.

C RETROSPECTIVE DEVELOPMENT BOUNDARY

Earlier sourcewise studies used ExpansionRx, Biogen, and TDC to develop the abstraction of prediction contracts and the fixed four-source action. They established three design choices carried into the reported protocol: residual behavior is sufficient to expose portable structure between source and target; a common convex combiner is necessary to compare selectors without confounding from coefficient fitting; and a fan-in of four channels balances complementary corrections against interference from the full bank. Their scientific role is selection of the deployment contract.

AssayRouter in the main comparisons is the HGB regressor without identity features defined by Equations 4–7, trained on standalone post-fit utility centered within each block and frozen before the action in Equation 12. Marginal-over-states and state-conditioned regressors are matched refinements. Earlier pairwise, soft-prior, graph, classification, and semantic systems remain separate development studies.

The external claim is model-held-out transfer under this fixed contract: all 24 evaluation targets are excluded from router fitting and grouped capacity selection. Appendix A reports the evidence through frozen candidate-label interventions, leave-one-interface-out routers, strict matched comparisons, identity holdouts, provenance checks, and leave-one-cluster sensitivities. These tests carry the paper's conclusions; the retrospective studies explain how the final question and action budget were chosen.