PROGRESSIVE RESIDUAL SYMBOLIC INTERPRETATION OF MULTI-AGENT REINFORCEMENT LEARNING POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic policy surrogates expose learned decision rules, but a single selected formula can obscure explanatory multiplicity and residual decision structure. We introduce a progressive residual symbolic interpretation framework for multi-agent reinforcement learning policies that characterizes a population of low-complexity symbolic explanations and then allocates residual symbolic corrections path by path using minimum description length (MDL). The resulting population is structurally diverse but functionally organized. After compression to 24 representative paths, MDL selects nonzero residual refinements for 20 and retains the Stage-I form for four. On 200 independently generated nominal episodes, the frozen refinements reduce agent-balanced NMSE by 0.0709 (95% CI: 0.0662–0.0757), showing that the gains persist beyond the discovery trajectories. The discovered symbolic structures transfer without structural rediscovery: on held-out matched states from four independently trained target MATD3 policies, fixed-structure coefficient recalibration achieves an agent-balanced NRMSE of 0.321 and $R^2 = 0.827$. Finally, six selected symbolic controllers jointly replace all neural agents and preserve the principal closed-loop voltage-control profile across nominal and unseen $\pm 10\%$ and $\pm 30\%$ operating shifts. Across 1.29 million raw symbolic-action evaluations, the final controller produces no native action clamps, non-finite evaluations, or rollout failures. Together, this establishes population-level characterization, progressive residual refinement, cross-policy structural reuse, and closed-loop functional validation.

1 INTRODUCTION

Modern reinforcement learning (RL) often relies on neural policies whose decision logic is difficult to inspect directly. This has motivated a growing literature on explainable and interpretable reinforcement learning, spanning explanations of individual decisions, trajectories, internal representations, and policy-level behavior (Bekkemoen, 2024; Glanois et al., 2024; Dazeley et al., 2023). Multi-agent reinforcement learning (MARL) additionally introduces inter-agent dependencies and coordinated behavior (Boggess et al., 2023). In distributed physical control, local decisions are coupled through shared system dynamics and jointly shape future operating states. These challenges arise directly in active voltage control of power distribution networks, where geographically distributed photovoltaic inverters coordinate using local information under spatially coupled and time-varying operating conditions. Wang et al. (2021) formulated this problem as a decentralized partially observable Markov decision process and demonstrated the potential of MARL for distributed voltage regulation, while identifying interpretability and transparency among the challenges facing learned controllers in this setting.

One direct route is to make the learned policy's decision rule explicit and inspectable. Policy distillation showed that policy behavior can be transferred from a neural teacher to a compact student (Rusu et al., 2016), while decision-tree, programmatic, and symbolic approaches make the surrogate itself explicit (Bastani et al., 2018; Verma et al., 2018; Hein et al., 2018; Landajuela et al., 2021). Decision trees support structured inspection and verification, programmatic policies express control logic as readable programs, and symbolic policies provide executable mathematical decision rules.

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.

Together, these approaches show that learned policy behavior can be captured in explicit and executable policy-level representations.

Explicit form alone, however, does not make a representation a useful explanation. Human-subject studies of interpretable model classes show that comprehensibility depends not only on representation format but also on the complexity of the representation (Huysmans et al., 2011; Lage et al., 2019). For symbolic policy interpretation, fidelity and complexity are therefore coupled: an explanation must preserve enough of the learned policy to remain faithful while retaining a sufficiently compact structure for its decision logic to be inspected. Symbolic regression makes this trade-off explicit by exposing the functional relationship while making representation complexity measurable (Makke & Chawla, 2024). The central question is how much symbolic structure is needed to explain a learned policy faithfully while remaining inspectable.

A single returned expression does not reveal whether alternative compact structures achieve comparable fidelity. The Rashomon perspective shows that well-performing models can differ in their reliance on the available variables (Fisher et al., 2019), while TreeFARMS demonstrates structural multiplicity within a near-optimal set of sparse decision trees (Xin et al., 2022). For symbolic policies, the corresponding identification question is whether multiple compact structures can explain the same learned policy with comparable fidelity. Population-level characterization makes such multiplicity observable and reveals the range of symbolic structures consistent with high policy fidelity.

A compact symbolic explanation may still leave structured behavior in its residual. Residual symbolic regression has used successive fitting to improve predictive accuracy (Sipper & Moore, 2021), but symbolic interpretation introduces an additional criterion: gains in fidelity must be weighed against the description complexity required to obtain them. Progressive residual interpretation makes explanation depth a pathwise fidelity–complexity decision, adding symbolic structure only when its explanatory gain justifies the added description complexity.

Cross-policy structural reuse tests whether the discovered symbolic structure persists beyond one numerical realization of policy training. Reusing the same symbolic form after independent retraining through coefficient recalibration probes structural stability without structural rediscovery. Closed-loop validation addresses the distinct consequence of feedback: controller actions alter the future states encountered by the policy (Ross et al., 2011). Fixed-trajectory fidelity and closed-loop behavioral preservation therefore test different properties of a symbolic explanation: how well it reproduces the teacher on a common state distribution, and how it behaves when its own actions generate the state distribution it encounters.

We introduce a progressive residual symbolic interpretation framework that treats interpretability as a problem of identifying how much compact decision structure is needed, whether that structure persists across independent policy training, and whether the resulting symbolic controller preserves behavior under feedback. Canonical population analysis exposes structural alternatives, while complexity-aware residual refinement allocates explanation depth selectively. The resulting structures are evaluated through coefficient-only cross-policy recalibration and joint multi-agent closed-loop replacement.

Our contributions are as follows:

- We reveal structural multiplicity among compact, high-fidelity symbolic explanations of the same learned policy, showing that population-level characterization exposes alternative decision structures beyond any single formula.
- We introduce complexity-aware progressive residual interpretation, making explanation depth a pathwise fidelity–complexity decision and allocating additional symbolic structure only when its explanatory gain justifies the added description complexity.
- We demonstrate cross-policy structural reuse across independently trained MARL policies through coefficient recalibration without structural rediscovery, providing evidence that the extracted symbolic structure persists beyond one numerical realization of policy training.
- We validate the resulting symbolic policies through joint multi-agent closed-loop replacement under nominal and unseen operating conditions, demonstrating behavioral preservation under symbolic feedback.

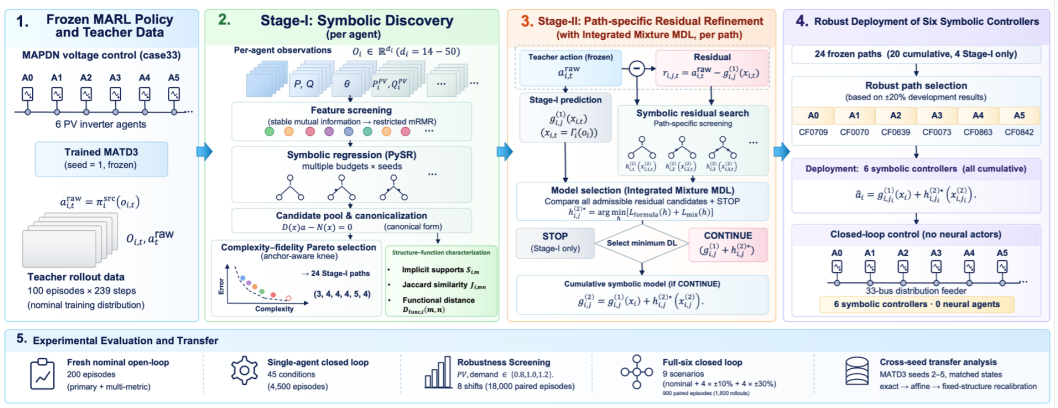

Figure 1: Overview of progressive residual symbolic interpretation. A frozen MARL teacher provides raw-action trajectories for symbolic policy discovery. Stage I constructs a canonical population, characterizes its structure–function relationships, and retains representative paths through anchor-aware Pareto compression. Stage II discovers path-specific symbolic residual corrections, with integrated mixture MDL selecting between STOP and CONTINUE for each path. The resulting frozen symbolic paths support two downstream evaluations: cross-policy structural reuse through coefficient recalibration and source-policy path selection followed by joint six-agent closed-loop control. The evaluation protocols are summarized in the bottom panel.

2 Related Work

We situate symbolic interpretation of trained multi-agent controllers along three lines: executable policy extraction, analysis of structural alternatives, and adaptive symbolic refinement.

Executable policy extraction. VIPER extracts decision trees guided by a neural policy and its Q-function (Bastani et al., 2018); PIRL uses a neural oracle to guide program search (Verma et al., 2018). Genetic-programming methods produce equation-valued controllers from trajectories (Hein et al., 2018), while deep symbolic policy search also distills pretrained neural policies through anchoring (Landajuela et al., 2021). We examine a population of canonical symbolic action rules recovered from the same frozen multi-agent policy, comparing their structures and actions before evaluating representative paths under recalibration and feedback.

Structural alternatives and representative selection. TreeFARMS enumerates near-optimal sparse decision trees and makes alternative model structures available for analysis (Xin et al., 2022). For symbolic action rules, we relate implicit-form structural supports to raw-action distances on common validation states. Complexity–fidelity Pareto selection preserves candidate rules with different trade-offs; hypervolume measures the objective region covered by a representative subset (Guerreiro et al., 2021).

Adaptive discovery and residual depth. i-SISSO screens descriptor spaces using mutual-information-based relevance and redundancy before symbolic modeling (Xu & Qian, 2022). Symbolic-regression boosting adds symbolic predictors in stages to improve predictive fit (Sipper & Moore, 2021). MDL balances the code lengths of a model and its unexplained data (Grünwald & Roos, 2019). Here the screening target changes from the teacher's raw action to each retained path's residual. The pathwise MDL comparison includes the zero correction and codes temporally correlated remaining residuals, making symbolic depth an explicit selection decision.

3 Methodology

Figure 1 provides an overview of the proposed progressive residual symbolic interpretation framework.

3.1 SETTING AND SYMBOLIC POLICY PATHS

We study active voltage control in the MAPDN case33 system with six PV inverter agents indexed by $\mathcal{I} = \{0, \ldots, N-1\}$, where $N = 6$. At time t, agent i receives a local observation $o_{i,t} = \Phi_i(s_t)$, and its frozen source policy produces the raw action $a_{i,t}^{\mathrm{raw}} = \pi_i^{\mathrm{src}}(o_{i,t})$. Let $\Pi^{\mathrm{src}} = (\pi_i^{\mathrm{src}})_{i\in\mathcal{I}}$ denote the joint frozen source policy. For symbolic interpretation, $x_{i,t} = \Gamma_i(o_{i,t})$ collects selected physical variables. Stage I constructs a canonical population $\mathcal{G}_i$; its representative paths satisfy $\{g_{i,j}^{(1)} : j \in \mathcal{J}_i\} \subseteq \mathcal{G}_i$. The selected Stage-II correction for each retained path uses the path-specific physical input $x_{i,j,t}^{(2)} = \Gamma_{i,j}^{(2)}(o_{i,t})$, giving

$$\widetilde{a}_{i,j,t}^{\mathrm{raw}} = g_{i,j}^{(1)}(x_{i,t}) + h_{i,j}^{(2)\star}(x_{i,j,t}^{(2)}), \qquad h_{i,j}^{(2)\star} \equiv 0 \text{ for STOP.} \tag{1}$$

The native action map Λ_i converts neural and symbolic raw actions into environment actions for closed-loop evaluation. The full canonical population $\mathcal{G}_i$ supports structure–function analysis; the retained paths support residual refinement, cross-policy recalibration, and feedback evaluation.

3.2 STAGE I: SYMBOLIC POPULATION AND REPRESENTATIVE PATHS

The shared symbolic-discovery backbone screens physical variables against the target and searches the selected inputs with PySR (Cranmer, 2023) (Appendix A.1). In Stage I, screening uses the frozen teacher's raw actions as targets; Stage II reapplies the backbone to each retained path's residual. Searches span algebraic, rational, and voltage-angle trigonometric operator classes. In Stage I, multiple budgets and seeds yield Hall-of-Fame occurrences pooled into $\mathcal{H}_i^{\mathrm{raw}}$. The deterministic map κ_i consolidates repeated discoveries into

$$\mathcal{G}_i = \{\kappa_i(h) : h \in \mathcal{H}_i^{\mathrm{raw}}\} = \{g_{i,1}, \ldots, g_{i,M_i}\}, \qquad C(g_{i,m}) = \min_{h\in\mathcal{O}(g_{i,m})} C_{\mathrm{PySR}}(h), \tag{2}$$

where $\mathcal{O}(g_{i,m}) = \{h \in \mathcal{H}_i^{\mathrm{raw}} : \kappa_i(h) = g_{i,m}\}$ collects the raw occurrences assigned the same canonical identity. Fidelity on fixed teacher-induced validation states is

$$E_i(g) = \frac{\|g(X_i^{\mathrm{val}}) - \mathbf{a}_i^{\mathrm{val}}\|_2}{\sqrt{N_i^{\mathrm{val}}}\,\sigma_i^{\mathrm{val}}}, \qquad \sigma_i^{\mathrm{val}} = \mathrm{Std}(\mathbf{a}_i^{\mathrm{val}}) > 0. \tag{3}$$

Here N_i^{val} counts the fixed validation transitions, and $E_i(g)$ is the normalized root-mean-square error (NRMSE) of the teacher's raw actions on those transitions.

To compare symbolic explanations in a common structural coordinate system, write each canonical function as $g_{i,m} = N_{i,m}/D_{i,m}$, with $D_{i,m} \equiv 1$ when no variable-dependent denominator remains. Let $\phi_i(x_i) \in \mathbb{R}^{K_i}$ be the shared canonical dictionary, with $N_{i,m} = \theta_{N,i,m}^\top \phi_i$ and $D_{i,m} = \theta_{D,i,m}^\top \phi_i$. The resulting implicit representation is

$$\begin{aligned} D_{i,m}(x_i)a - N_{i,m}(x_i) &= \xi_{i,m}^\top \psi_i(x_i, a) = 0, \\ \xi_{i,m} &= \begin{bmatrix} -\theta_{N,i,m} \\ \theta_{D,i,m} \end{bmatrix}, \qquad \psi_i(x_i, a) = \begin{bmatrix} \phi_i(x_i) \\ a\phi_i(x_i) \end{bmatrix}, \\ S_{i,m} &= \{k : (\xi_{i,m})_k \neq 0\}. \end{aligned} \tag{4}$$

Full support $S_{i,m}$ defines exact structural identity. Let $S_{i,m}^{(-a)} = S_{i,m} \setminus \{k_{a,i}\}$, where $k_{a,i}$ indexes the representation-induced pure-action coordinate. For nonconstant candidates, graded structural similarity and functional distance on the same validation states are

$$J_{i,mn} = \frac{|S_{i,m}^{(-a)} \cap S_{i,n}^{(-a)}|}{|S_{i,m}^{(-a)} \cup S_{i,n}^{(-a)}|}, \quad D_{\mathrm{func},i}(m,n) = \frac{\|g_{i,m}(X_i^{\mathrm{val}}) - g_{i,n}(X_i^{\mathrm{val}})\|_2}{\sqrt{N_i^{\mathrm{val}}}\,\sigma_i^{\mathrm{val}}}. \tag{5}$$

For population-level analysis, we aggregate functional distances over pairs of exact structural supports and test their association with structural similarity using a support-level Spearman QAP (Appendix A.2.2).

Let $\mathcal{P}_i^{\mathrm{raw}} \subseteq \mathcal{G}_i$ be the complexity–fidelity Pareto population. Grouping raw Pareto candidates by audited exact-output artifact identity yields $\mathcal{P}_i$; its best-fidelity member $g_i^\star =$

$\arg\min_{g\in\mathcal{P}_i} E_i(g)$ anchors every selected subset. With both objectives normalized and $\mathcal{R} = \{(1.05, 1.05), (1.10, 1.10), (1.20, 1.20)\}$ as the reference-point set, define $H_{i,\varrho}(\mathcal{Q}) = \mathrm{HV}_\varrho(\mathcal{Q})/\,\mathrm{HV}_\varrho(\mathcal{P}_i)$ for $\varrho \in \mathcal{R}$. At cardinality k, select

$$\mathcal{Q}^\star_{i,k} \in \underset{\substack{\mathcal{Q}\subseteq\mathcal{P}_i,\ |\mathcal{Q}|=k \\ g_i^\star\in\mathcal{Q}}}{\arg\max} \min_{\varrho\in\mathcal{R}} H_{i,\varrho}(\mathcal{Q}). \tag{6}$$

The endpoint-chord knee of the robust-retention curve sets k_i^{knee} (Appendix A.2.3). We freeze $\{g_{i,j}^{(1)} : j \in \mathcal{J}_i\} = \mathcal{Q}^\star_{i,k_i^{\mathrm{knee}}}$ as the representative Stage-I paths passed to residual refinement.

3.3 Stage II: path-conditioned residual refinement

Each retained Stage-I path defines its own residual target,

$$r_{i,j,t} = a_{i,t}^{\mathrm{raw}} - g_{i,j}^{(1)}(x_{i,t}), \qquad x_{i,j,t}^{(2)} = \Gamma_{i,j}^{(2)}(o_{i,t}). \tag{7}$$

For each retained Stage-I path, the shared symbolic-discovery backbone constructs candidate explanations of its residual using the path-specific inputs $x_{i,j,t}^{(2)}$. Exact rediscoveries are consolidated into canonical residual expressions within each path, forming $\mathcal{H}_{i,j}^{(2)}$.

Nonzero canonical candidates are grouped by coefficient-invariant symbolic structure. One representative per structure is fixed using development-training data, and the zero correction is included to obtain $\widetilde{\mathcal{H}}_{i,j}^{(2)}$.

For each retained path, a single minimum-description-length (MDL) criterion selects its correction:

$$\mathrm{DL}_{i,j}(h) = L_{\mathrm{formula},i,j}(h) + L_{\mathrm{mix},i,j}(h), \quad h_{i,j}^{(2)\star} = \arg\min_{h\in\widetilde{\mathcal{H}}_{i,j}^{(2)}} \mathrm{DL}_{i,j}(h). \tag{8}$$

$L_{\mathrm{formula},i,j}(h)$ encodes the symbolic correction, while $L_{\mathrm{mix},i,j}(h)$ codes its remaining residual over complete validation episodes under a Matérn-3/2 temporal model calibrated on development-training episodes, with residual scale integrated out (Appendix A.3). The zero correction competes with all nonzero candidates under the same MDL criterion, selecting both the residual correction and the explanation depth for each path. A nonzero selection assigns CONTINUE; a zero selection assigns STOP. The resulting cumulative path is $g_{i,j}^{(2)}(x_{i,t}, x_{i,j,t}^{(2)}) = g_{i,j}^{(1)}(x_{i,t}) + h_{i,j}^{(2)\star}(x_{i,j,t}^{(2)})$.

3.4 Cross-policy structural reuse

We transfer complete source-discovered progressive symbolic paths to independently trained target policies without structural rediscovery or reselection of explanation depth. Each retained path preserves its Stage-I symbolic basis and MDL-selected STOP/CONTINUE decision. CONTINUE paths additionally preserve their selected canonical residual template and Stage-II symbolic basis.

Target-policy adaptation recalibrates the Stage-I outer coefficients against target raw actions. For CONTINUE paths, the Stage-II outer coefficients are subsequently fitted to the remaining target-policy residual.

The resulting complete paths are evaluated on separate matched held-out states against direct source-coefficient reuse and affine output recalibration (Appendix A.4).

3.5 Closed-loop symbolic replacement

For the joint observation $o_t = (o_{i,t})_{i\in\mathcal{I}}$, a symbolic joint policy $\widetilde{\Pi}$ follows the native action pipeline,

$$\widetilde{\mathbf{a}}_t^{\mathrm{raw}} = \widetilde{\Pi}(o_t), \qquad \widetilde{\mathbf{a}}_t^{\mathrm{trans}} = \Lambda(\widetilde{\mathbf{a}}_t^{\mathrm{raw}}), \qquad s_{t+1} = F(s_t, \widetilde{\mathbf{a}}_t^{\mathrm{trans}}). \tag{9}$$

where Λ applies the agent-wise maps Λ_i componentwise. Single-agent substitution evaluates each path against the all-neural source policy Π^{src}. Deployment selection uses the eight shifted conditions specified in Appendix A.5. From paths that complete the selection rollouts with finite raw actions and no native-clamp activation, we select one per agent by lexicographically comparing

worst-case discrepancies in voltage out-of-control rate, average voltage deviation, controllable ratio, and episode return across the eight selection conditions. The resulting six-agent $\widetilde{\Pi}$ is then evaluated jointly against Π^{src} under nominal and held-out operating conditions (Appendix A.5).

4 RESULTS

4.1 STRUCTURAL MULTIPLICITY OF SYMBOLIC POLICY EXPLANATIONS

The Stage-I population exhibits both structural organization and explanatory multiplicity. Structural similarity is associated with smaller functional distances across all six agents (support-level Spearman QAP, Holm-adjusted $p < 0.001$), and all 120 leave-one-episode-out estimates remain negative (Fig. 2a,b).

Across 540 symbolic-regression searches, 8,769 Hall-of-Fame occurrences yield 6,822 distinct canonical functions, spanning 305–762 exact structures per agent. Distinct structures can nevertheless produce closely aligned actions: 52.35% of different-structure pairs for A_0 and 36.56% for A_1 satisfy $D_{\mathrm{func}} \leq 0.05$ (Fig. 2c).

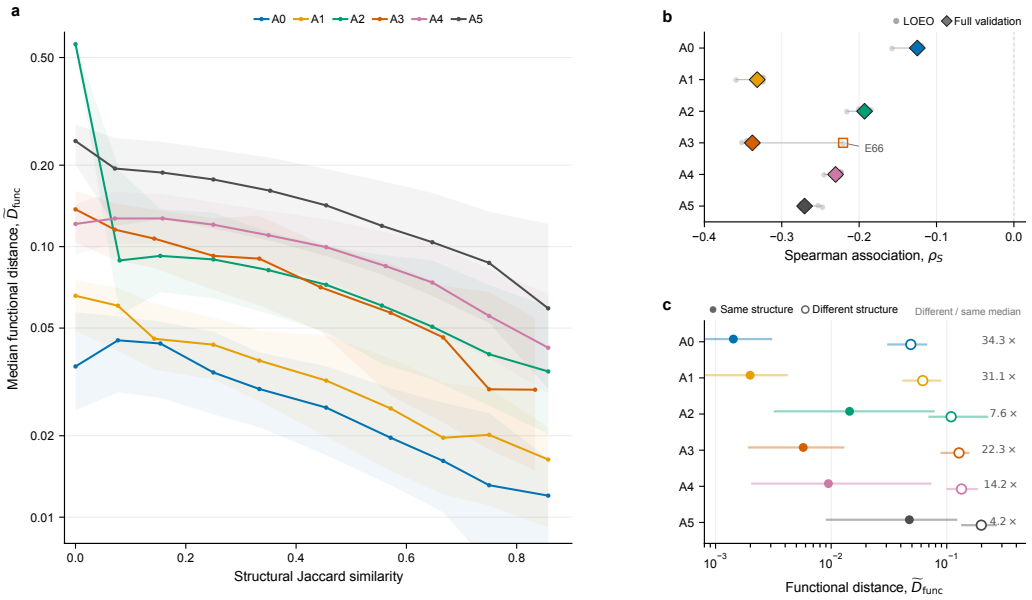

Figure 2: Structural similarity organizes a functionally diverse population of symbolic explanations. **a,** Median functional distance between structure pairs as a function of structural Jaccard similarity; shading indicates interquartile ranges. **b,** Full-validation Spearman QAP coefficients and 20 leave-one-episode-out estimates for each of the six agents. **c,** Functional distances for pairs with identical versus distinct exact structures. Points indicate medians, horizontal segments indicate interquartile ranges, and annotations report the ratio of different-structure to same-structure median functional distances for each agent. Functional-distance axes in **a** and **c** use logarithmic scales.

The anchor-aware Pareto compression retains 24 representative Stage-I paths from 84 candidates, preserving 96.07%–97.73% of the worst-reference hypervolume across agents (Supplementary Table 4).

4.2 PROGRESSIVE RESIDUAL REFINEMENT REVEALS PATH-SPECIFIC EXPLANATION DEPTH

Every retained Stage-I path admits at least one nonzero residual candidate that improves pooled development-validation fidelity. Balancing fidelity against symbolic description cost, MDL selects nonzero corrections for 20 paths and retains the Stage-I representations for four. Explanation depth is thus determined individually for each symbolic path.

On 200 independently generated nominal episodes, the frozen Stage-II corrections yield a mean episode-level, agent-balanced paired NMSE reduction of 0.0709 (95% CI: [0.0662, 0.0757]). Within each episode, improvements are averaged over CONTINUE paths within each agent and then equally across the six agents. For the complete 24-path family, agent-balanced NRMSE decreases from 0.300 to 0.200, while R^2 increases from 0.861 to 0.928.

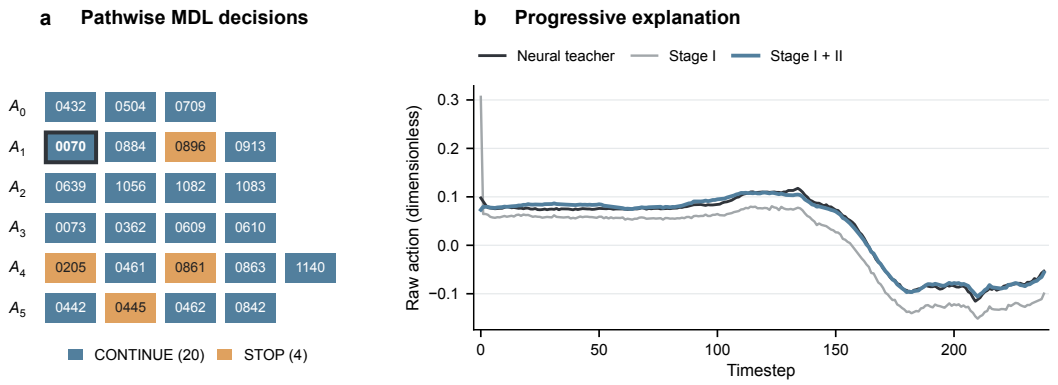

Figure 3: MDL selects path-specific explanation depth through progressive residual refinement. a, MDL decisions for all 24 retained Stage-I paths, grouped by agent. Twenty paths select nonzero residual corrections (CONTINUE), while four retain their Stage-I representations (STOP). The outlined cell identifies A1_CF0070. **b,** Raw-action trajectories of the frozen neural teacher, Stage-I explanation, and complete progressive explanation (Stage I+II) for A1_CF0070 on fresh nominal episode 0, illustrating the effect of path-specific residual refinement.

The pathwise decisions vary both across agents and within individual agents (Fig. 3a). For the illustrated A1_CF0070 path, Stage II corrects the initial Stage-I overshoot and reduces its sustained offset, bringing the cumulative explanation closer to the neural teacher on this fresh nominal episode (Fig. 3b). Training-residual trajectories and fresh-episode raw-action trajectories for all retained paths are provided in Appendix B.3.1 and Appendix B.3.2, respectively.

4.3 Cross-Policy Structural Reuse

Source-discovered symbolic structures transfer across four independently trained target policies through coefficient recalibration, without structural rediscovery. On matched held-out states, fixed-structure recalibration achieves an agent-balanced NRMSE of 0.321, compared with 0.474 for affine output recalibration, a 32.2% reduction (Fig. 4a). The improvement holds across all 24 agent–target-policy combinations after aggregation over the frozen paths within each agent (Fig. 4b).

The same fixed-structure advantage is observed within the 20 CONTINUE paths, where agent-balanced NRMSE decreases from 0.477 under affine recalibration to 0.315 under fixed-structure recalibration.

Residual refinement also contributes to cross-policy fidelity. For the six source-selected deployment paths, the complete progressive explanations achieve lower held-out NRMSE than their recalibrated Stage-I components in 22 of 24 paired agent–target-policy comparisons (Appendix B.4.1).

All 96 Stage-I path–target recalibrations preserve the exact canonical support of the source-discovered structures. Stage-II residual templates and MDL depth decisions also remain fixed throughout transfer.

4.4 Closed-Loop Behavioral Preservation

Under nominal single-agent substitution, Stage-II refinement substantially improves aggregate agreement with the all-neural controller across the 20 matched CONTINUE paths. The magnitudes of the agent-balanced mean discrepancies decrease by 81.6% in episode return, 94.3% in voltage out-of-control rate, 98.1% in controllable ratio, and 98.5% in average voltage deviation (Fig. 5a; Appendix Table 9).

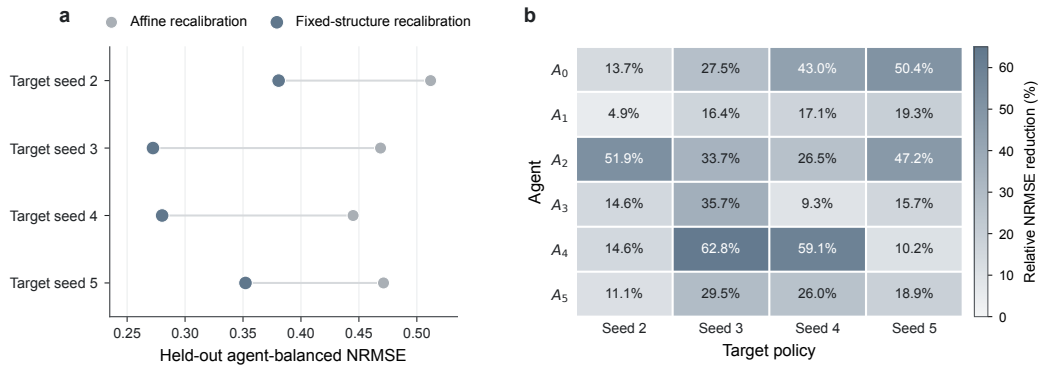

Figure 4: Source-discovered symbolic structures improve fidelity across independently trained target policies. a, Agent-balanced held-out NRMSE across four target policies, comparing affine output recalibration with coefficient recalibration of the source-discovered fixed structures, without structural rediscovery. **b,** Relative NRMSE reduction against affine recalibration for each of the 24 agent–target-policy combinations, using the corresponding agent-level aggregation over frozen source paths. All 24 comparisons favor fixed-structure recalibration.

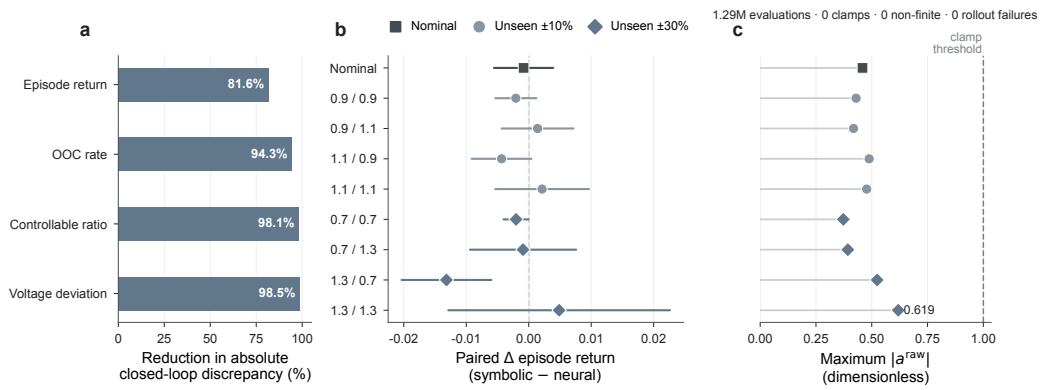

Figure 5: Residual refinement improves closed-loop agreement and supports six-agent symbolic execution. a, Reduction in the absolute agent-balanced mean discrepancy from the all-neural controller after Stage-II refinement, evaluated through nominal single-agent substitution on the same 20 CONTINUE paths. **b,** Mean paired episode-return differences between the joint six-agent symbolic controller and the all-neural controller under nominal operation and eight unseen PV–demand shifts. Differences are symbolic minus neural; horizontal intervals indicate paired 95% confidence intervals. **c,** Maximum absolute raw symbolic action under each operating condition, relative to the native action-clamp threshold.

All six paths selected for joint deployment contain MDL-selected Stage-II corrections. Together, they form a complete six-agent symbolic controller (Appendix Table 10).

Under nominal operation, its mean paired episode-return difference from the all-neural controller is $\Delta R = -8.24 \times 10^{-4}$ (95% CI $[-0.005566, 0.003917]$).

Across eight held-out PV–demand shifts, the largest absolute mean paired episode-return difference is 0.0132 (Fig. 5b). At $(s_{\mathrm{PV}}, s_{\mathrm{D}}) = (1.3, 0.7)$, the mean paired return difference is $\Delta R = -0.013171$, while voltage out-of-control rate decreases by 6.09×10^{-4}, controllability increases by 4.73×10^{-3}, and average voltage deviation decreases by 2.10×10^{-4} (Appendix Table 11).

Across these nine conditions and $1{,}290{,}600$ raw symbolic-action evaluations, the joint controller produces no native action clamps, non-finite evaluations, environment-destroy events, or premature terminations; its largest observed raw-action magnitude is 0.619 (Fig. 5c). The selected symbolic paths therefore form an explicit, numerically stable six-agent controller across the evaluated conditions.

5 DISCUSSION

Our results establish a population-based view of symbolic policy interpretation in multi-agent reinforcement learning. Structural similarity is systematically associated with functional proximity, while distinct exact structures can produce closely aligned actions. Together, these findings reveal that learned policies admit multiple symbolic descriptions within an organized structure–function landscape. Preserving representative explanations along the fidelity–complexity frontier provides a basis for exploring this multiplicity rather than reducing the policy to a single symbolic expression.

Progressive residual refinement extends this population-based representation by selecting explanation depth for each path. MDL balances the fidelity gained from residual correction against its additional symbolic description cost, retaining Stage I where further complexity is not selected. The resulting refinements achieve aggregate fidelity improvements on independently generated nominal episodes, connecting path-specific complexity selection to predictive performance beyond the discovery data.

The downstream evaluations establish two complementary properties of the discovered symbolic representations. Across independently trained target policies, source-discovered structures can be reused through coefficient recalibration without structural rediscovery. The additional gains from complete progressive explanations over their Stage-I components further demonstrate the contribution of residual refinement during transfer. Closed-loop evaluation then establishes a distinct property: the symbolic representations remain executable when their actions influence subsequent observations and interact through the shared network dynamics. Together, structural reuse and feedback execution connect symbolic policy interpretation to both independent policy realizations and the behavior induced by the resulting symbolic controllers.

6 CONCLUSION

We introduce progressive residual symbolic interpretation as a population-based framework for representing learned multi-agent policies through executable symbolic paths with individually selected explanation depth. On MAPDN case33, MDL selects residual refinements for 20 of 24 representative paths, achieving a paired agent-balanced NMSE reduction of 0.0709 on 200 independent nominal episodes. The source-discovered structures also support cross-policy reuse, reducing held-out agent-balanced NRMSE by 32.2% relative to affine recalibration across four independently trained target policies. Six selected paths execute jointly across nine evaluated operating conditions, with no native action clamps or non-finite evaluations among $1{,}290{,}600$ raw symbolic actions. Together, these results establish a framework that preserves alternative symbolic explanations, allocates expression complexity by path, and connects interpretable policy representations to multi-agent feedback execution.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including improving grammar, wording, and clarity of expression. The authors reviewed the AI-assisted material and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The experimental setup and methodology are detailed in Section 3. The appendices provide the protocols for Stage-I symbolic discovery and candidate screening (Appendix A.1), MDL-based residual selection (Appendix A.3), cross-policy coefficient recalibration (Appendix A.4), and closed-loop evaluation (Appendix A.5). Symbolic discovery and model selection are separated from independent evaluation, with source-discovered structures and MDL decisions held fixed during cross-policy transfer.

REFERENCES

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e6d8545daa42d5ced125a4bf747b3688-Paper.pdf.

Yanzhe Bekkemoen. Explainable reinforcement learning (xrl): a systematic literature review and taxonomy. *Machine Learning*, 113:355–441, 2024. doi: 10.1007/s10994-023-06479-7. URL https://link.springer.com/article/10.1007/s10994-023-06479-7.

Kayla Boggess, Sarit Kraus, and Lu Feng. Explainable multi-agent reinforcement learning for temporal queries. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 55–63. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/7. URL https://doi.org/10.24963/ijcai.2023/7. Main Track.

Miles Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, May 2023. URL http://arxiv.org/abs/2305.01582. arXiv:2305.01582 [astro-ph, physics:physics].

Richard Dazeley, Peter Vamplew, and Francisco Cruz. Explainable reinforcement learning for broad-xai: a conceptual framework and survey. *Neural Computing and Applications*, 35:16893–16916, 2023. doi: 10.1007/s00521-023-08423-1. URL https://link.springer.com/article/10.1007/s00521-023-08423-1.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *Machine Learning*, 113:5847–5890, 2024. doi: 10.1007/s10994-024-06543-w. URL https://link.springer.com/article/10.1007/s10994-024-06543-w.

Peter Grünwald and Teemu Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, 2019. doi: 10.1142/S2661335219300018. URL https://doi.org/10.1142/S2661335219300018.

Andreia P. Guerreiro, Vasco Manquinho, and José Rui Figueira. Exact hypervolume subset selection through incremental computations. *Computers & Operations Research*, 136:105471, 2021. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2021.105471. URL https://www.sciencedirect.com/science/article/pii/S0305054821002215.

Daniel Hein, Steffen Udluft, and Thomas A. Runkler. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76:158–169, 2018. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2018.09.007. URL https://www.sciencedirect.com/science/article/pii/S0952197618301933.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2010.12.003. URL https://www.sciencedirect.com/science/article/pii/S0167923610002368.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pp. 59–67, 2019. doi: 10.1609/hcomp.v7i1.5280. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5280.

Mikel Landajuela, Brenden K Petersen, Sookyung Kim, Claudio P Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F Pettit, and Daniel Faissol. Discovering symbolic policies with deep reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5979–5989. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/landajuela21a.html.

Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-023-10622-0. URL https://link.springer.com/article/10.1007/s10462-023-10622-0. Article 2.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation, 2016. URL https://arxiv.org/abs/1511.06295.

Moshe Sipper and Jason H. Moore. Symbolic-regression boosting. *Genetic Programming and Evolvable Machines*, 22:357–381, 2021. doi: 10.1007/s10710-021-09400-0. URL https://link.springer.com/article/10.1007/s10710-021-09400-0.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5045–5054. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/verma18a.html.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3271–3284. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1a6727711b84fd1efbb87fc565199d13-Paper.pdf.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14071–14084. Curran Associates, Inc., 2022. doi: 10.52202/068431-1023. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5afaa8b4dd18eb1eed055d2d821b58ae-Paper-Conference.pdf.

Yuqin Xu and Quan Qian. i-sisso: Mutual information-based improved sure independent screening and sparsifying operator algorithm. *Engineering Applications of Artificial Intelligence*, 116:105442, 2022. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105442. URL https://www.sciencedirect.com/science/article/pii/S0952197622004328.

APPENDIX CONTENTS

A Method details and evaluation protocols **12**

A.1 Shared symbolic-discovery backbone 12

A.2 Stage I: population analysis and path selection 13

A.3 Stage II: residual candidates and MDL depth selection 15

A.4 Cross-policy recalibration and structure checks 17

A.5 Single-agent selection and joint feedback evaluation 18

B Research hypotheses and additional results **19**

B.1 Formal statements of the research hypotheses 19

B.2 Additional Results for Structural Multiplicity of Symbolic Policy Explanations . . 20

B.3 Additional Results for Progressive Residual Refinement 20

B.4 Additional Results for Cross-Policy Structural Reuse 25

B.5 Additional Results for Closed-Loop Behavioral Preservation 26

C Selected symbolic expressions **28**

C.1 Selected Stage-I expressions and validation fidelity 28

C.2 MDL-Selected Stage-II Residual Corrections 30

A METHOD DETAILS AND EVALUATION PROTOCOLS

A.1 SHARED SYMBOLIC-DISCOVERY BACKBONE

The frozen teacher dataset contains 100 complete episodes, with 239 usable transitions per episode. Stage-I discovery uses episodes E0–E59 for training and E60–E79 for development validation. Episodes E80–E99 form a separate held-out split. The independent nominal evaluation of the frozen refinements uses newly generated episodes. Each agent's symbolic target is the teacher raw action $a_{i,t}^{\mathrm{raw}} = \pi_i^{\mathrm{src}}(o_{i,t})$, before native environment-side action translation.

Target-adaptive variable screening. For each candidate physical variable $o_i^{(k)}$, estimate its mutual information $I(o_i^{(k)}; a_i^{\mathrm{raw}})$ with the teacher raw action. Rank variables by this relevance within each episode-preserving resample and use the deterministic elbow to identify the relevance-supported subset. Five-fold episode-level cross-validation over five repetitions produces the resamples; variables selected in more than half form the stable marginal pool. Within the relevance-supported set, mutual-information relevance and redundancy characterize complementary and overlapping variables.

Episode-block permutation evidence supplies a conditional rescue pool. For candidate $o_i^{(k)}$, the mean increase in prediction error is

$$\Delta_{i,k}^{\mathrm{perm}} = \mathrm{MAE}_{\mathrm{perm}(k)} - \mathrm{MAE}_{\mathrm{base}}. \tag{10}$$

Variables with finite, positive mean $\Delta_{i,k}^{\mathrm{perm}}$ are ranked in descending order. A deterministic maximum-distance elbow identifies the conditional rescue set. The union of this set and the stable marginal pool determines the coordinates retained by Γ_i, giving $x_{i,t} = \Gamma_i(o_{i,t})$. The same target-adaptive screening principle is applied to each frozen Stage-I path's residual $r_{i,j,t} = a_{i,t}^{\mathrm{raw}} - g_{i,j}^{(1)}(x_{i,t})$, giving the path-specific Stage-II input $x_{i,j,t}^{(2)} = \Gamma_{i,j}^{(2)}(o_{i,t})$.

Constrained search. PySR searches expressions in three operator classes:

$$\begin{aligned}\mathcal{F}_{\mathrm{ALG}} &= \{+, -, \times\},\\ \mathcal{F}_{\mathrm{RAT}} &= \mathcal{F}_{\mathrm{ALG}} \cup \{\div\},\\ \mathcal{F}_{\mathrm{TRIG}} &= \mathcal{F}_{\mathrm{ALG}} \cup \{\sin, \cos\}.\end{aligned} \tag{11}$$

Write $\mathcal{C}_{\mathrm{SR}} = \{\mathrm{ALG}, \mathrm{RAT}, \mathrm{TRIG}\}$. Division is restricted to denominator subtrees of complexity at most three, with no nested division. Rational candidates undergo a numerical denominator audit on the observed trajectory support and interpolations between adjacent observations. Trigonometric operators are restricted to observed voltage-angle variables. Where screening retains no angle variable, the TRIG class reduces to the ALG operator family.

For observations $\{(z_k, y_k)\}_{k=1}^{n}$ and expressions admissible under the selected operator class, the search objective is

$$\min_{h} \frac{1}{n} \sum_{k=1}^{n} \big(y_k - h(z_k)\big)^2 \quad \text{subject to} \quad C_{\mathrm{PySR}}(h) \le C_{\max}. \tag{12}$$

The input–target pairs instantiate this objective as

$$(z_t, y_t) = \begin{cases} (x_{i,t}, a_{i,t}^{\mathrm{raw}}), & \text{Stage I,} \\ (x_{i,j,t}^{(2)}, r_{i,j,t}), & \text{Stage II for path } j. \end{cases} \tag{13}$$

For Stage I, each of the six agents is searched over all three classes, iteration budgets $\mathcal{B}_{\mathrm{SR}} = \{50, 100, 200, 400, 800, 1600\}$, and PySR seeds $\Omega_{\mathrm{SR}} = \{0, 1, 2, 3, 4\}$, giving 90 searches per agent. These seeds control symbolic search while the source teacher remains frozen.

The recorded Stage-I PySR settings are maximum expression size 30, 31 populations of 27 expressions each, 380 cycles per iteration, squared-error loss, zero explicit parsimony, float32 precision, enabled annealing, crossover probability 0.2, deterministic serial execution, disabled warm start and denoising, and enabled simplification.

A.2 Stage I: population analysis and path selection

A.2.1 Canonical consolidation and implicit structural support

Let $\mathcal{H}_{i,c,b,\omega}$ be the complete Hall-of-Fame occurrence collection returned for class c, iteration budget b, and search seed ω. The pooled occurrence collection is the multiset union

$$\mathcal{H}_i^{\mathrm{raw}} = \biguplus_{(c,b,\omega) \in \mathcal{C}_{\mathrm{SR}} \times \mathcal{B}_{\mathrm{SR}} \times \Omega_{\mathrm{SR}}} \mathcal{H}_{i,c,b,\omega} = \{h_{i,r}\}_{r=1}^{R_i}. \tag{14}$$

A deterministic symbolic normalization κ_i applies algebraic and trigonometric simplification and rational normalization. Canonical identity uses exact equality of the resulting representations and retains unrounded coefficients. Define

$$\begin{aligned}\mathcal{G}_i &= \{\kappa_i(h) : h \in \mathcal{H}_i^{\mathrm{raw}}\} = \{g_{i,1}, \ldots, g_{i,M_i}\},\\ \mathcal{O}(g_{i,m}) &= \{h \in \mathcal{H}_i^{\mathrm{raw}} : \kappa_i(h) = g_{i,m}\},\\ C(g_{i,m}) &= \min_{h \in \mathcal{O}(g_{i,m})} C_{\mathrm{PySR}}(h).\end{aligned} \tag{15}$$

The occurrence collection $\mathcal{O}(g_{i,m})$ preserves the search provenance associated with canonical function $g_{i,m}$.

Each canonical function is written as $a = g_{i,m}(x_i) = N_{i,m}(x_i)/D_{i,m}(x_i)$, with $D_{i,m}(x_i) \equiv 1$ when no variable-dependent denominator remains. Let $\phi_i(x_i) = [\phi_{i,1}(x_i), \ldots, \phi_{i,K_i}(x_i)]^\top$ be the shared canonical dictionary, and write $N_{i,m} = \theta_{N,i,m}^\top \phi_i$ and $D_{i,m} = \theta_{D,i,m}^\top \phi_i$. The implicit representation is

$$\begin{aligned}D_{i,m}(x_i)a - N_{i,m}(x_i) &= \xi_{i,m}^\top \psi_i(x_i, a) = 0,\\ \psi_i(x_i, a) &= \begin{bmatrix} \phi_i(x_i) \\ a\phi_i(x_i) \end{bmatrix}, \qquad \xi_{i,m} = \begin{bmatrix} -\theta_{N,i,m} \\ \theta_{D,i,m} \end{bmatrix},\\ S_{i,m} &= \{k \in \{1, \ldots, 2K_i\} : (\xi_{i,m})_k \neq 0\}.\end{aligned} \tag{16}$$

The full support $S_{i,m}$ defines exact structural identity; graded similarity is defined below after masking the pure-action coordinate.

A.2.2 STRUCTURAL COMPARISON AND PERMUTATION INFERENCE

Let $k_{a,i}$ index the dictionary's pure-action term $a \cdot 1$. The reduced support used for graded similarity is $S_{i,m}^{(-a)} = S_{i,m} \setminus \{k_{a,i}\}$. For two nonconstant canonical functions $g_{i,m}, g_{i,n}$, define structural similarity and fixed-validation-state functional distance by

$$\begin{aligned} J_{i,mn} &= \frac{|S_{i,m}^{(-a)} \cap S_{i,n}^{(-a)}|}{|S_{i,m}^{(-a)} \cup S_{i,n}^{(-a)}|}, \\ D_{\mathrm{func},i}(m,n) &= \frac{\|g_{i,m}(X_i^{\mathrm{val}}) - g_{i,n}(X_i^{\mathrm{val}})\|_2}{\sqrt{N_i^{\mathrm{val}}}\,\sigma_i^{\mathrm{val}}}. \end{aligned} \tag{17}$$

The formal association uses nonconstant expressions passing the validation evaluation audit. For each full exact support u, let $\mathcal{G}_i(u) = \{g_{i,m} \in \mathcal{G}_i : S_{i,m} = u\}$. For distinct supports u, v, use their reduced-support Jaccard score $J_i(u,v)$ and aggregate over the corresponding coefficient-level expressions:

$$\begin{aligned} \widetilde{D}_i(u,v) &= \operatorname*{median}_{g_{i,m} \in \mathcal{G}_i(u),\, g_{i,n} \in \mathcal{G}_i(v)} D_{\mathrm{func},i}(m,n), \\ T_i^{\mathrm{obs}} &= \rho_{\mathrm{S}}\big(\{J_i(u,v)\}_{u<v}, \{\widetilde{D}_i(u,v)\}_{u<v}\big). \end{aligned} \tag{18}$$

Each unordered pair of exact structures contributes once to this Spearman statistic.

For each of $B_{\mathrm{QAP}} = 10{,}000$ permutations (seed 20260908), one permutation of the exact-support node labels is applied to both endpoints of every functional-distance dyad; the structural-similarity matrix is held fixed. Recomputing the statistic yields $T_i^{(b)}$. The two-sided Monte Carlo value is

$$p_i^{\mathrm{QAP}} = \frac{1 + \sum_{b=1}^{B_{\mathrm{QAP}}} \mathbf{1}\{|T_i^{(b)}| \geq |T_i^{\mathrm{obs}}|\}}{B_{\mathrm{QAP}} + 1}. \tag{19}$$

The six agent-wise primary Spearman p-values form one family and are adjusted with Holm's step-down procedure. As an episode-level stability analysis, the structure-pair statistic is also recomputed after removing each complete development-validation episode in turn (20 removals per agent). The QAP and leave-one-episode-out analyses characterize the canonical population. Representative selection is determined by complexity and teacher fidelity.

A.2.3 ANCHORED ROBUST-HYPERVOLUME COMPRESSION

Teacher fidelity on the development-validation inputs is

$$E_i(g) = \frac{\|g(X_i^{\mathrm{val}}) - \mathbf{a}_i^{\mathrm{val}}\|_2}{\sqrt{N_i^{\mathrm{val}}}\,\sigma_i^{\mathrm{val}}}, \qquad \sigma_i^{\mathrm{val}} = \mathrm{Std}(\mathbf{a}_i^{\mathrm{val}}) > 0. \tag{20}$$

Let $\mathcal{P}_i^{\mathrm{raw}} \subseteq \mathcal{G}_i$ be the functions nondominated under joint minimization of $(C(g), E_i(g))$. Candidates in $\mathcal{P}_i^{\mathrm{raw}}$ are grouped by their audited exact-output artifact identity. Artifact-group membership is obtained from the audited output-identity map indexed by canonical-model ID. This selection-level grouping is distinct from the canonical normalization κ_i. Within each group, we retain the candidate with lowest Stage-I complexity, breaking ties by validation NRMSE and then canonical-model ID. The retained representatives form $\mathcal{P}_i$. The compulsory fidelity anchor is $g_i^\star = \arg\min_{g \in \mathcal{P}_i} E_i(g)$.

The two objectives are min–max normalized over $\mathcal{P}_i$:

$$\widetilde{C}_i(g) = \frac{C(g) - C_{i,\min}}{C_{i,\max} - C_{i,\min}}, \qquad \widetilde{E}_i(g) = \frac{E_i(g) - E_{i,\min}}{E_{i,\max} - E_{i,\min}}. \tag{21}$$

For the fixed reference points

$$\mathcal{R} = \{(1.05, 1.05), (1.10, 1.10), (1.20, 1.20)\}, \tag{22}$$

let $\mathrm{HV}_\varrho(\mathcal{Q})$ denote the two-dimensional dominated hypervolume of the normalized objective pairs for $\mathcal{Q} \subseteq \mathcal{P}_i$. The anchored subset selection is

$$\begin{aligned} H_{i,\varrho}(\mathcal{Q}) &= \frac{\mathrm{HV}_\varrho(\mathcal{Q})}{\mathrm{HV}_\varrho(\mathcal{P}_i)}, \\ \mathcal{Q}^\star_{i,k} &\in \operatorname*{arg\,max}_{\substack{\mathcal{Q}\subseteq\mathcal{P}_i,\,|\mathcal{Q}|=k \\ g_i^\star\in\mathcal{Q}}} \min_{\varrho\in\mathcal{R}} H_{i,\varrho}(\mathcal{Q}), \\ R_i(k) &= \min_{\varrho\in\mathcal{R}} H_{i,\varrho}(\mathcal{Q}^\star_{i,k}). \end{aligned} \tag{23}$$

The endpoint-chord knee is computed over $k_{\min} = 2$ through $k_{\max} = |\mathcal{P}_i|$. Define

$$\begin{aligned} \widetilde{k} &= \frac{k - k_{\min}}{k_{\max} - k_{\min}}, \\ \widetilde{R}_i(k) &= \frac{R_i(k) - R_i(k_{\min})}{R_i(k_{\max}) - R_i(k_{\min})}, \\ d_i(k) &= \frac{\widetilde{R}_i(k) - \widetilde{k}}{\sqrt{2}}, \qquad k_i^{\mathrm{knee}} \in \operatorname*{arg\,max}_{k_{\min}\le k\le k_{\max}} d_i(k). \end{aligned} \tag{24}$$

The frozen representative paths are $\{g_{i,j}^{(1)} : j \in \mathcal{J}_i\} = \mathcal{Q}^\star_{i,k_i^{\mathrm{knee}}}$. The knee calculation begins at $k_{\min} = 2$; $k = 1$ is the anchor-only diagnostic state.

Algorithm 1 Population-Level Symbolic Structure Discovery

Require: Frozen source policies $\{\pi_i^{\mathrm{src}}\}_{i\in\mathcal{I}}$, teacher-induced development data, selected inputs $\{\Gamma_i\}_{i\in\mathcal{I}}$, symbolic classes $\mathcal{C}_{\mathrm{SR}}$, budgets $\mathcal{B}_{\mathrm{SR}}$, and seeds Ω_{SR}

Ensure: Stage-I symbolic paths $\{g_{i,j}^{(1)} : j \in \mathcal{J}_i\}$ for each agent i

for each agent $i \in \mathcal{I}$ **do**
  $\mathcal{H}_i^{\mathrm{raw}} \leftarrow$
  for each $(c, b, \omega) \in \mathcal{C}_{\mathrm{SR}} \times \mathcal{B}_{\mathrm{SR}} \times \Omega_{\mathrm{SR}}$ **do**
    Run constrained symbolic search under (c, b, ω)
    Pool the complete Hall-of-Fame into $\mathcal{H}_i^{\mathrm{raw}}$
  end for
  Canonicalize $\mathcal{H}_i^{\mathrm{raw}} \xrightarrow{\kappa_i} \mathcal{G}_i$ and assign canonical complexities
  Construct the raw complexity–fidelity Pareto population $\mathcal{P}_i^{\mathrm{raw}}$
  Retain one representative per audited exact-output artifact group to obtain $\mathcal{P}_i$
  Set the fidelity anchor $g_i^\star \leftarrow \arg\min_{g\in\mathcal{P}_i} E_i(g)$
  for $k = 2, \ldots, |\mathcal{P}_i|$ **do**
    Compute the robust anchor-constrained subset $\mathcal{Q}^\star_{i,k}$
    $R_i(k) \leftarrow \min_{\varrho\in\mathcal{R}} H_{i,\varrho}(\mathcal{Q}^\star_{i,k})$
  end for
  Normalize the retention curve and compute $d_i(k) = (\widetilde{R}_i(k) - \widetilde{k})/\sqrt{2}$
  $k_i^{\mathrm{knee}} \leftarrow \arg\max_{2\le k\le|\mathcal{P}_i|} d_i(k)$
  $\{g_{i,j}^{(1)} : j \in \mathcal{J}_i\} \leftarrow \mathcal{Q}^\star_{i,k_i^{\mathrm{knee}}}$
end for

A.3 Stage II: residual candidates and MDL depth selection

Path-conditioned candidate construction. For each retained $g_{i,j}^{(1)}$, apply the shared symbolic-discovery backbone of Appendix A.1 with target $r_{i,j,t} = a_{i,t}^{\mathrm{raw}} - g_{i,j}^{(1)}(x_{i,t})$ and input $x_{i,j,t}^{(2)} = \Gamma_{i,j}^{(2)}(o_{i,t})$, as specified in equation 13. Canonicalize the resulting Hall-of-Fame occurrences within each path (i, j). Identical symbolic corrections arising from different Stage-I paths retain their path identity. After symbolic-class admissibility and finite-prediction checks, add the unique zero correction to obtain

$$0 \in \mathcal{H}_{i,j}^{(2)}, \qquad C_2(0) = 0, \qquad C_2(h) = \min_{\widetilde{h}\in\mathcal{O}_{i,j}^{(2)}(h)} C_{\mathrm{PySR}}(\widetilde{h}) \quad (h \neq 0), \tag{25}$$

where $\mathcal{O}_{i,j}^{(2)}(h)$ contains the raw occurrences consolidated into correction h.

Group nonzero canonical corrections by coefficient-ignoring symbolic structure. Within each structure, fix one coefficient realization using development-training loss, then $C_2(h)$, then deterministic candidate identity as the tie-break order. The selected representatives and zero form

$$0 \in \widetilde{\mathcal{H}}_{i,j}^{(2)} \subseteq \mathcal{H}_{i,j}^{(2)}. \tag{26}$$

Formula code and temporal residual code. For $h \in \widetilde{\mathcal{H}}_{i,j}^{(2)}$, define

$$\mathrm{DL}_{i,j}(h) = L_{\mathrm{formula},i,j}(h) + L_{\mathrm{mix},i,j}(h). \tag{27}$$

The formula code begins with a one-bit indicator for STOP or CONTINUE. A nonzero correction is encoded by prefix-AST tokens under the frozen path-specific grammar, with tokens encoded uniformly. Numeric constants use canonical decimal literals, a self-delimiting literal-length code, and a fixed numeric-character alphabet. Operator arities make the prefix expression self-delimiting. $C_2(h)$ remains the symbolic search-complexity measure; the model cost in equation 27 is the actual formula code length.

For validation episode e, let $\mathbf{a}_{i,e}^{\mathrm{raw}}$, $\mathbf{X}_{i,e}$, and $\mathbf{X}_{i,j,e}^{(2)}$ denote its episode-stacked teacher raw actions, Stage-I inputs, and Stage-II inputs. The remaining action residual is

$$\mathbf{u}_{i,j,e}(h) = \mathbf{a}_{i,e}^{\mathrm{raw}} - g_{i,j}^{(1)}(\mathbf{X}_{i,e}) - h(\mathbf{X}_{i,j,e}^{(2)}). \tag{28}$$

The path-specific validation data are

$$\mathcal{D}_{i,j}^{\mathrm{val}} = \left\{ \left(\mathbf{X}_{i,e}, \mathbf{X}_{i,j,e}^{(2)}, \mathbf{a}_{i,e}^{\mathrm{raw}} \right) : e \in \mathcal{E}_{\mathrm{val}} \right\}. \tag{29}$$

Complete episodes define the temporal blocks. Conditional on residual scale, the block model is

$$\mathbf{u}_{i,j,e}(h) \mid \sigma^2 \sim \mathcal{N}\big(0, \sigma^2 \mathbf{K}_{i,j}^{\mathrm{train}}\big), \tag{30}$$

where $\mathbf{K}_{i,j}^{\mathrm{train}}$ is the path-specific Matérn-3/2 temporal correlation matrix. Its length scale and nugget are calibrated from development-training episodes. For the validation episodes,

$$Q_{i,j}(h) = \sum_{e \in \mathcal{E}_{\mathrm{val}}} \mathbf{u}_{i,j,e}(h)^{\top} \big(\mathbf{K}_{i,j}^{\mathrm{train}}\big)^{-1} \mathbf{u}_{i,j,e}(h). \tag{31}$$

The train-calibrated residual-scale prior is

$$\sigma^2 \sim \mathrm{IG}(a_0, b_{0,i,j}), \qquad a_0 = 2, \tag{32}$$

with $b_{0,i,j}$ determined from the development-training residual scale. Let N_{val} count all validation transitions and $E_{\mathrm{val}} = |\mathcal{E}_{\mathrm{val}}|$ count complete validation episodes. Integrating out σ^2 yields

$$\begin{aligned} \log p\big(\mathcal{D}_{i,j}^{\mathrm{val}} \mid h, \mathbf{K}_{i,j}^{\mathrm{train}}\big) = & -\frac{N_{\mathrm{val}}}{2}\log(2\pi) - \frac{E_{\mathrm{val}}}{2}\log\big|\mathbf{K}_{i,j}^{\mathrm{train}}\big| \\ & + a_0 \log b_{0,i,j} - \log\Gamma(a_0) + \log\Gamma\left(a_0 + \frac{N_{\mathrm{val}}}{2}\right) \\ & - \left(a_0 + \frac{N_{\mathrm{val}}}{2}\right)\log\left(b_{0,i,j} + \frac{Q_{i,j}(h)}{2}\right). \end{aligned} \tag{33}$$

The validation residual code length is

$$L_{\mathrm{mix},i,j}(h) = -\log_2 p\big(\mathcal{D}_{i,j}^{\mathrm{val}} \mid h, \mathbf{K}_{i,j}^{\mathrm{train}}\big). \tag{34}$$

Pathwise depth selection. For every path, select and freeze

$$\begin{aligned} h_{i,j}^{(2)\star} &= \arg\min_{h \in \widetilde{\mathcal{H}}_{i,j}^{(2)}} \mathrm{DL}_{i,j}(h), \\ g_{i,j}^{(2)}(x_{i,t}, x_{i,j,t}^{(2)}) &= g_{i,j}^{(1)}(x_{i,t}) + h_{i,j}^{(2)\star}(x_{i,j,t}^{(2)}). \end{aligned} \tag{35}$$

The choice $h_{i,j}^{(2)\star} = 0$ assigns STOP; a nonzero choice assigns CONTINUE.

Algorithm 2 Progressive Residual Symbolic Refinement

Require: Frozen Stage-I paths $\{g_{i,j}^{(1)}\}$, development data, symbolic-discovery backbone, and MDL specification
Ensure: Refined paths $\{g_{i,j}^{(2)}\}$ and STOP/CONTINUE decisions
for each $i \in \mathcal{I}$ **do**
  for each $j \in \mathcal{J}_i$ **do**
    Construct residual targets $r_{i,j,t}$ and path-specific inputs $x_{i,j,t}^{(2)}$
    Discover and canonicalize admissible residual corrections to obtain $\mathcal{H}_{i,j}^{(2)}$, including 0
    Consolidate symbolic structures to obtain $\widetilde{\mathcal{H}}_{i,j}^{(2)}$
    Calibrate the path-specific temporal codebook and scale prior
    $h_{i,j}^{(2)\star} \leftarrow \arg\min_{h \in \widetilde{\mathcal{H}}_{i,j}^{(2)}} \mathrm{DL}_{i,j}(h)$
    Set $g_{i,j}^{(2)} \leftarrow g_{i,j}^{(1)} + h_{i,j}^{(2)\star}$; STOP if $h_{i,j}^{(2)\star} = 0$, otherwise CONTINUE
  end for
end for

A.4 Cross-policy recalibration and structure checks

The source-selected symbolic bases and MDL depth decisions are fixed before target-policy calibration. A shared matched-state bank of 200 complete episodes is divided equally between calibration and held-out evaluation.

For agent i and retained path $j \in \mathcal{J}_i$, let $\mathbf{B}_{i,j}^{(1),\mathrm{cal}}$ denote the frozen Stage-I basis evaluated on calibration states. CONTINUE paths additionally use $\mathbf{B}_{i,j}^{(2),\mathrm{cal}}$. The superscript ho denotes the corresponding held-out basis evaluation.

Algorithm 3 Progressive Cross-Policy Recalibration

Require: Frozen progressive paths and matched-state target-policy data
Ensure: Held-out raw-action predictions $\widehat{\mathbf{a}}_{i,j}^{\mathrm{tar,raw,ho}}$
for each target policy, agent i, and path $j \in \mathcal{J}_i$ **do**
  $\widehat{\mathbf{w}}_{i,j}^{(1)} \leftarrow \arg\min_{\mathbf{w}} \|\mathbf{a}_i^{\mathrm{tar,raw,cal}} - \mathbf{B}_{i,j}^{(1),\mathrm{cal}}\mathbf{w}\|_2^2$
  if CONTINUE **then**
    $\mathbf{r}_{i,j}^{\mathrm{tar,cal}} \leftarrow \mathbf{a}_i^{\mathrm{tar,raw,cal}} - \mathbf{B}_{i,j}^{(1),\mathrm{cal}}\widehat{\mathbf{w}}_{i,j}^{(1)}$
    $\widehat{\mathbf{w}}_{i,j}^{(2)} \leftarrow \arg\min_{\mathbf{w}} \|\mathbf{r}_{i,j}^{\mathrm{tar,cal}} - \mathbf{B}_{i,j}^{(2),\mathrm{cal}}\mathbf{w}\|_2^2$
    $\widehat{\mathbf{a}}_{i,j}^{\mathrm{tar,raw,ho}} \leftarrow \mathbf{B}_{i,j}^{(1),\mathrm{ho}}\widehat{\mathbf{w}}_{i,j}^{(1)} + \mathbf{B}_{i,j}^{(2),\mathrm{ho}}\widehat{\mathbf{w}}_{i,j}^{(2)}$
  else
    $\widehat{\mathbf{a}}_{i,j}^{\mathrm{tar,raw,ho}} \leftarrow \mathbf{B}_{i,j}^{(1),\mathrm{ho}}\widehat{\mathbf{w}}_{i,j}^{(1)}$
  end if
end for

We evaluate held-out cumulative-path NRMSE against direct source-coefficient reuse and affine recalibration of the frozen source mapping on the same matched-state evaluation split.

Progressive path structure. For each retained path $j \in \mathcal{J}_i$, let $S_{i,j}^{(1,\mathrm{src})}$ denote the source Stage-I canonical implicit support. This is the support $S_{i,m(j)}$ of the canonical Stage-I function underlying path j. Let $d_{i,j} \in \{0,1\}$ denote the source-selected MDL depth decision, where $d_{i,j} = 0$ represents STOP and $d_{i,j} = 1$ represents CONTINUE. For each CONTINUE path, let $\mathcal{T}_{i,j}^{(2)}$ denote the selected coefficient-invariant residual structure and its frozen additive basis. Let $q \in \{\mathrm{src}\} \cup \mathcal{L}_{\mathrm{tar}}$ index the source or a target-policy realization. For each target policy ℓ, the coefficients $\mathbf{w}_{i,j}^{(2,\ell)}$ are those fitted in Algorithm 3; the target-policy index is suppressed within the algorithm's outer loop. Define the

Stage-II outer-coefficient support by

$$U_{i,j}^{(2,q)} = \begin{cases} \operatorname{supp}\left(\mathbf{w}_{i,j}^{(2,q)}\right), & d_{i,j} = 1, \\ \emptyset, & d_{i,j} = 0. \end{cases} \tag{36}$$

For STOP paths, the residual template is also defined as empty.

The progressive path structure is defined by

$$\mathcal{S}_{i,j}^{(q)} = \left(S_{i,j}^{(1,q)}, d_{i,j}, \mathcal{T}_{i,j}^{(2)}, U_{i,j}^{(2,q)}\right). \tag{37}$$

This representation retains both stages of the progressive explanation. Stage-I structure is specified by its canonical implicit support, while the selected Stage-II residual structure is specified by its canonical template and outer-coefficient support.

Progressive structure verification. We reconstruct the source and recalibrated Stage-I expressions in the original agent-specific canonical implicit dictionary and compare their full structural supports. Source reconstruction is checked against the saved canonical representation.

For CONTINUE paths, we additionally compare the source and target Stage-II outer-coefficient supports under the frozen residual template. The selected template and MDL depth decision are inherited from the source policy.

Held-out fidelity is evaluated for the complete progressive path. Stage-I-only predictions are also retained to characterize the contribution of the residual correction.

A.5 Single-agent selection and joint feedback evaluation

Native action and physical execution. The symbolic target is the teacher's post-tanh raw action, before environment-side translation. For the evaluated MAPDN case33 configuration, the native translation is

$$a_{i,t}^{\mathrm{trans}} = \Lambda_i(a_{i,t}^{\mathrm{raw}}) = 0.8\ \mathrm{clip}(a_{i,t}^{\mathrm{raw}}, -1, 1). \tag{38}$$

The translated action scales the inverter's instantaneous reactive-power capability:

$$q_{i,t}^{\mathrm{PV}} = a_{i,t}^{\mathrm{trans}} \sqrt{(s_i^{\max})^2 - (p_{i,t}^{\mathrm{PV}})^2}. \tag{39}$$

Symbolic deployment uses the native action translation and environment transition in equation 9. In each single-agent substitution, one neural agent is replaced by a symbolic path and the remaining agents follow Π^{src}. Nominal substitutability uses 100 paired episodes per controller condition. Deployment selection uses 50 paired episodes per condition for the eight shifts

$$(s_{\mathrm{PV}}, s_{\mathrm{D}}) \in \{0.8, 1.0, 1.2\}^2 \setminus \{(1, 1)\}. \tag{40}$$

Candidates must complete these rollouts with finite raw symbolic actions and zero native-clamp activation. For each agent, one path is selected lexicographically by its worst-case closed-loop discrepancy in voltage out-of-control rate, average voltage deviation, controllable ratio, and episode return, in that order.

The six selected paths form $\widetilde{\Pi}$. Joint evaluation compares it with the all-neural Π^{src} on the nominal condition and on held-out operating shifts

$$(s_{\mathrm{PV}}, s_{\mathrm{D}}) \in \{0.9, 1.1\}^2 \cup \{0.7, 1.3\}^2, \tag{41}$$

using 100 paired episodes per scenario. The evaluation records paired system-level performance, the maximum absolute raw symbolic action, native-clamp counts, non-finite evaluations, environment-destroy events, and premature terminations.

B RESEARCH HYPOTHESES AND ADDITIONAL RESULTS

B.1 FORMAL STATEMENTS OF THE RESEARCH HYPOTHESES

Multiplicity of symbolic structures. Let $\varepsilon_i > 0$ denote the teacher-fidelity criterion and $\delta_i \in \mathbb{N}$ the complexity criterion. The admissible set is

$$\mathcal{A}_i(\varepsilon_i, \delta_i) = \{g_{i,m} \in \mathcal{G}_i : E_i(g_{i,m}) \le \varepsilon_i,\ C(g_{i,m}) \le \delta_i\}. \tag{42}$$

Hypothesis 1 (Structural multiplicity). For agent i, the admissible symbolic representations are structurally non-unique:

$$|\{S_{i,m} : g_{i,m} \in \mathcal{A}_i(\varepsilon_i, \delta_i)\}| \ge 2. \tag{43}$$

Equivalently, there exist $m \neq n$ such that $g_{i,m}, g_{i,n} \in \mathcal{A}_i(\varepsilon_i, \delta_i)$ and $S_{i,m} \neq S_{i,n}$.

Symbolic recovery of residual behavior. Let $\mathcal{J}_i$ index the retained Stage-I paths. The residual target for $g_{i,j}^{(1)}$ is $r_{i,j,t} = a_{i,t}^{\mathrm{raw}} - g_{i,j}^{(1)}(x_{i,t})$. Let $x_{i,j}^{(2)} \in \mathbb{R}^{q_{i,j}}$ be the physical input for a residual correction, and let $\mathcal{H}_{i,j}^{(2)}$ contain candidate corrections $h : \mathbb{R}^{q_{i,j}} \to \mathbb{R}$, including $0 \in \mathcal{H}_{i,j}^{(2)}$. On the fixed validation sample, define

$$E_{i,j}^{\mathrm{res}}(h) = \frac{\|\mathbf{r}_{i,j}^{\mathrm{val}} - h(X_{i,j}^{(2),\mathrm{val}})\|_2}{\sqrt{N_i^{\mathrm{val}}}\,\sigma_i^{\mathrm{val}}}, \quad \Delta_{i,j}(h) = E_{i,j}^{\mathrm{res}}(0) - E_{i,j}^{\mathrm{res}}(h). \tag{44}$$

Let $C_2(h)$ denote correction complexity, with $C_2(0) = 0$. For fixed $\eta > 0$ and $\delta^{(2)} \ge 0$, a residual is recoverable when some $h \in \mathcal{H}_{i,j}^{(2)}$ satisfies $\Delta_{i,j}(h) \ge \eta$ and $C_2(h) \le \delta^{(2)}$.

Hypothesis 2 (Residual symbolic recovery). At least one retained Stage-I path has a residual correction meeting both criteria:

$$\exists i \in \mathcal{I},\ \exists j \in \mathcal{J}_i,\ \exists h \in \mathcal{H}_{i,j}^{(2)} : \quad \Delta_{i,j}(h) \ge \eta, \quad C_2(h) \le \delta^{(2)}. \tag{45}$$

Cross-policy reuse of progressive symbolic structure. Let $\mathcal{L}_{\mathrm{tar}}$ denote the set of independently trained target policies considered in the transfer evaluation.

For each retained path $j \in \mathcal{J}_i$, let $\widehat{g}_{i,j}^{(2;\ell)}$ denote the complete progressive explanation following sequential coefficient recalibration for target policy $\ell \in \mathcal{L}_{\mathrm{tar}}$. For STOP paths, the Stage-II correction is zero.

Let $\mathcal{S}_{i,j}^{(\mathrm{src})}$ and $\mathcal{S}_{i,j}^{(\ell)}$ denote the corresponding source and recalibrated progressive path structures defined by equation 37. Let $E_i^{(\ell)}$ denote held-out NRMSE relative to target policy ℓ, and let $\varepsilon_{\mathrm{tr}} > 0$ denote the fidelity tolerance.

Hypothesis 3 (Cross-policy progressive structural reuse). For every agent, a common source-discovered progressive path admits structurally consistent recalibration across the evaluated target policies:

$$\begin{aligned} &\forall i \in \mathcal{I},\ \exists j \in \mathcal{J}_i,\ \forall \ell \in \mathcal{L}_{\mathrm{tar}} : \\ &\qquad \mathcal{S}_{i,j}^{(\mathrm{src})} = \mathcal{S}_{i,j}^{(\ell)}, \\ &\qquad E_i^{(\ell)}\big(\widehat{g}_{i,j}^{(2;\ell)}\big) \le \varepsilon_{\mathrm{tr}}. \end{aligned} \tag{46}$$

Closed-loop behavioral preservation. Let $\Pi^{\mathrm{src}} = (\pi_i^{\mathrm{src}})_{i\in\mathcal{I}}$ denote the source multi-agent teacher and $\widetilde{\Pi}$ a symbolic replacement policy. Open-loop fidelity $E_{\mathrm{ol}}(\widetilde{\Pi}, \Pi^{\mathrm{src}})$ is measured on fixed teacher-induced observations. In closed loop, the symbolic joint raw action is $\widetilde{\mathbf{a}}_t^{\mathrm{raw}} = \widetilde{\Pi}(o_t)$, and the environment evolves as

$$s_{t+1} = F\big(s_t, \Lambda(\widetilde{\Pi}(o_t))\big). \tag{47}$$

Let $D_{\mathrm{cl}}(\widetilde{\Pi}, \Pi^{\mathrm{src}})$ denote the closed-loop functional discrepancy, with criteria $\varepsilon_{\mathrm{ol}} > 0$ and $\varepsilon_{\mathrm{cl}} > 0$.

Hypothesis 4 (Closed-loop behavioral preservation). There exists a symbolic replacement satisfying both criteria:

$$\exists \widetilde{\Pi} : \quad E_{\mathrm{ol}}(\widetilde{\Pi}, \Pi^{\mathrm{src}}) \leq \varepsilon_{\mathrm{ol}}, \qquad D_{\mathrm{cl}}(\widetilde{\Pi}, \Pi^{\mathrm{src}}) \leq \varepsilon_{\mathrm{cl}}. \tag{48}$$

B.2 Additional Results for Structural Multiplicity of Symbolic Policy Explanations

The agent-specific canonical populations contain between 746 and 1,324 distinct functions. These yield 305–762 exact implicit structures per agent, of which 304–761 are nonconstant and enter the formal structural analysis.

Table 1: Structural geometry of the Stage-I symbolic populations. "Structures" denotes the nonconstant exact implicit structures used in the formal Jaccard analysis.

Agent	Structures	Mean J	Median J	Fraction $J = 0$	NN mean J	NN median J
A_0	761	0.130804	0.105263	0.158763	0.660894	0.700000
A_1	702	0.132286	0.117647	0.140999	0.652131	0.666667
A_2	344	0.269842	0.250000	0.004560	0.818663	0.833333
A_3	304	0.124123	0.090909	0.315529	0.684459	0.750000
A_4	413	0.232432	0.214286	0.058358	0.779341	0.818182
A_5	372	0.243004	0.222222	0.073834	0.794344	0.833333

Table 2: Largest connected-component fraction across structural-similarity thresholds. Entries give the fraction of nonconstant exact structures contained in the largest connected component at each Jaccard threshold.

Agent	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
A_0	0.8265	0.6912	0.4849	0.1340	0.0499	0.0013
A_1	0.8376	0.7009	0.2593	0.0826	0.0641	0.0043
A_2	0.9942	0.9884	0.9767	0.6366	0.2587	0.0320
A_3	0.7533	0.7007	0.5428	0.1316	0.0855	0.0066
A_4	0.9685	0.9249	0.7845	0.5932	0.1792	0.0073
A_5	0.9892	0.9462	0.8280	0.6452	0.3118	0.0323

Table 3: Stage-I structure–function association and episode-level robustness. The same- and different-structure columns report median functional distance. "Diff. ≤ 0.05" is the fraction of different-exact-structure function pairs satisfying $D_{\mathrm{func}} \leq 0.05$. All six raw QAP values are $1/10{,}001$, with $p_{\mathrm{Holm}} = 6/10{,}001 < 0.001$ after adjustment.

Agent	Same-structure	Different-structure	Diff. ≤ 0.05	ρ_S	LOEO range
A_0	0.001425	0.048854	52.35%	-0.1248	$[-0.1576, -0.1192]$
A_1	0.001990	0.061938	36.56%	-0.3317	$[-0.3591, -0.3236]$
A_2	0.014431	0.109166	13.46%	-0.1930	$[-0.2159, -0.1846]$
A_3	0.005729	0.127653	7.38%	-0.3380	$[-0.3517, -0.2208]$
A_4	0.009464	0.134210	4.51%	-0.2305	$[-0.2453, -0.2235]$
A_5	0.047466	0.199009	2.53%	-0.2705	$[-0.2751, -0.2475]$

The raw agent-wise Pareto counts are $13, 14, 14, 14, 15$, and 15. Selection-layer identity consolidation merges one duplicate identity for A_3, producing the effective selection pools reported in Table 4.

B.3 Additional Results for Progressive Residual Refinement

The common MDL criterion selects nonzero residual corrections for 20 of the 24 retained paths, with a median selected correction complexity of $C_2 = 12$. The four STOP paths are A1_CF0896, A4_CF0205, A4_CF0861, and A5_CF0445.

Table 4: Stage-I robust hypervolume retention at the selected anchor-aware knees. Retention is reported under each of the three fixed hypervolume reference points. The robust retention is the minimum across reference points and is the quantity optimized by the maximin selection.

Agent	Pareto pool	Selected	$R_{1.05}$	$R_{1.10}$	$R_{1.20}$	Robust R
A_0	13	3	96.7896%	96.7534%	96.7364%	96.7364%
A_1	14	4	97.9592%	97.8332%	97.6700%	97.6700%
A_2	14	4	98.4029%	98.1090%	97.7003%	97.7003%
A_3	13	4	97.9960%	97.8624%	97.6899%	97.6899%
A_4	15	5	98.0630%	97.9183%	97.7314%	97.7314%
A_5	15	4	96.0671%	96.1298%	96.2675%	96.0671%
Total	84	24	–	–	–	–

Table 5: Residual recoverability across the 24 frozen Stage-I paths. Candidate-space summaries are reported under two residual-complexity caps.

Residual cap	Median MSE reduction	Median fraction of path-wise maximum	Paths reaching $\geq 90\%$ of maximum	Paths improving $\geq 15/20$ episodes	Paths improving $\geq 18/20$ episodes
$C_2 \leq 15$	47.59%	84.44%	10/24	24/24	21/24
$C_2 \leq 20$	52.64%	95.61%	19/24	24/24	23/24

B.3.1 PATH-SPECIFIC STAGE-II RESIDUAL TRAJECTORIES

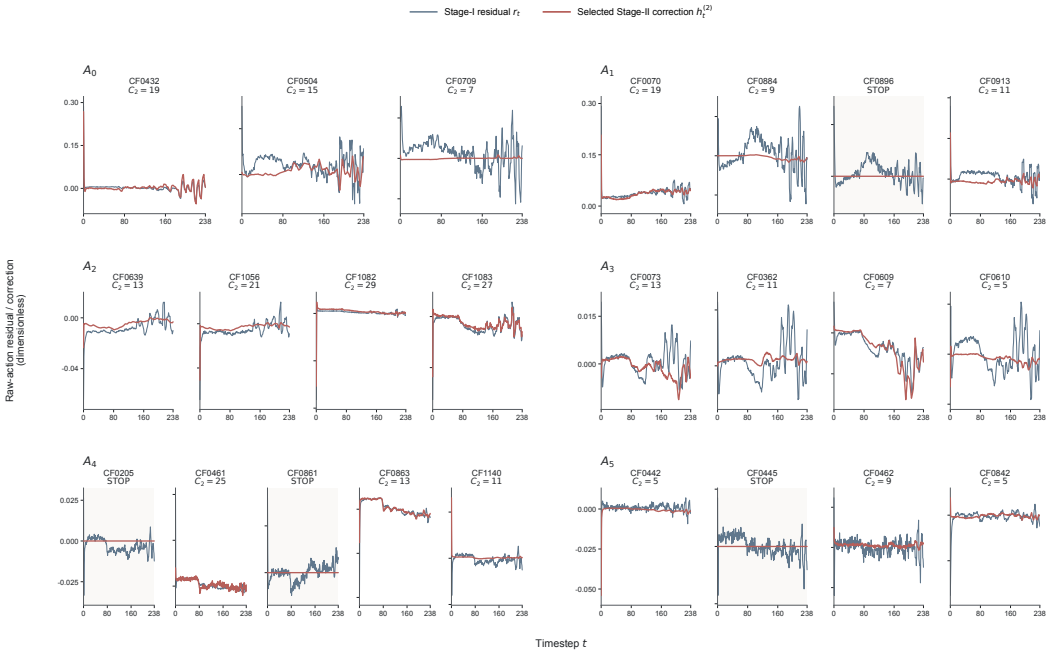

Figure 6: Path-specific Stage-II residual refinements across the 24 frozen Stage-I symbolic explanations. Paths are grouped by agent. Each panel shows the Stage-I raw-action residual r_t and the MDL-selected correction $h_t^{(2)}$ on a common Stage-II training trajectory. CONTINUE paths overlay the nonzero correction; STOP paths select zero. Axes show the dimensionless tanh-squashed raw-action space before native action translation, with path-specific vertical scales.

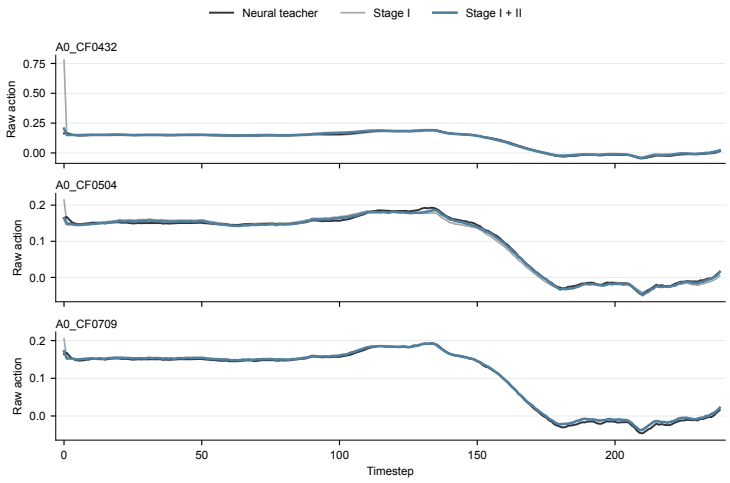

Figure 7: Fresh nominal raw-action trajectories for agent A_0. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

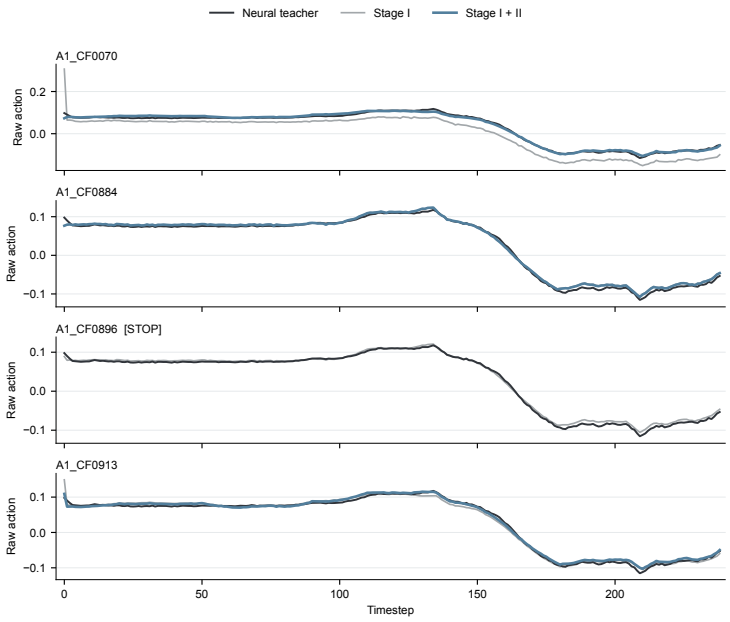

Figure 8: Fresh nominal raw-action trajectories for agent A_1. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

B.3.2 Fresh-episode raw-action trajectories of retained paths

The following figures show all 24 retained paths on fresh nominal episode 0. Within each agent, the paths are evaluated on the same teacher-induced observations. Each trajectory contains timesteps 0–238, including the initial transient.

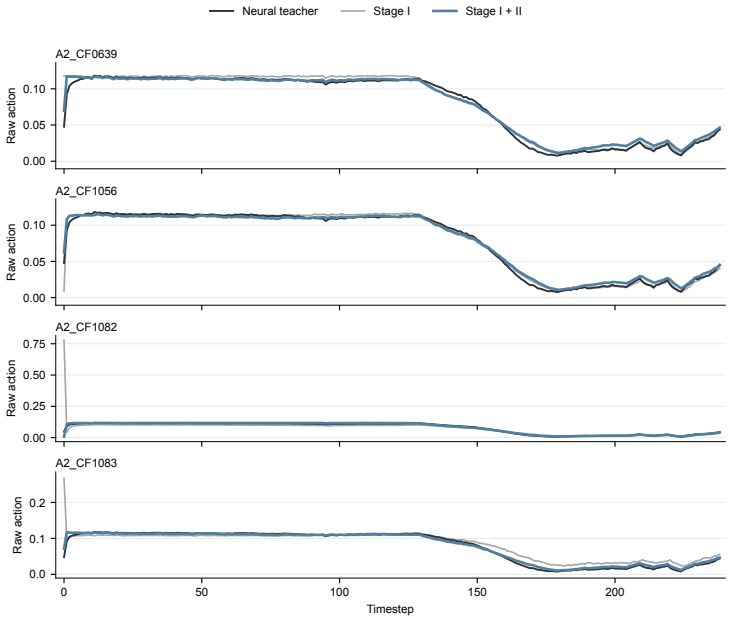

Figure 9: Fresh nominal raw-action trajectories for agent A_2. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

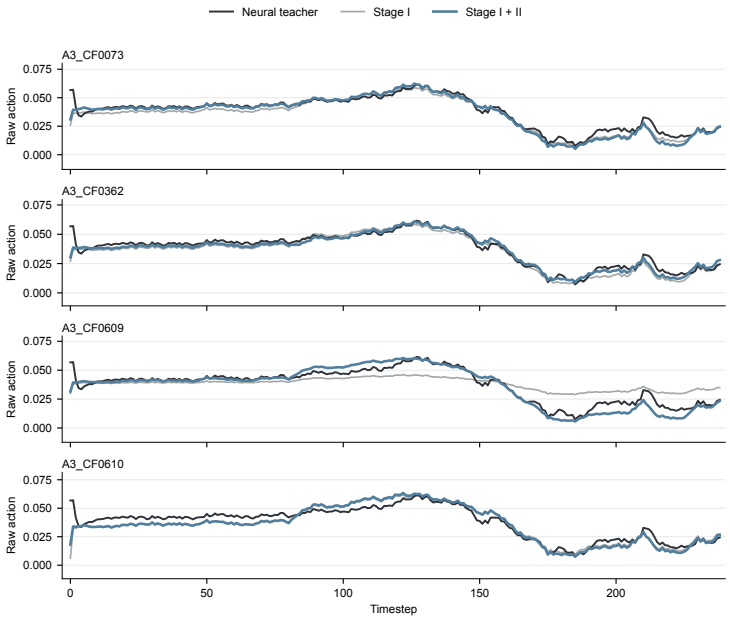

Figure 10: Fresh nominal raw-action trajectories for agent A_3. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

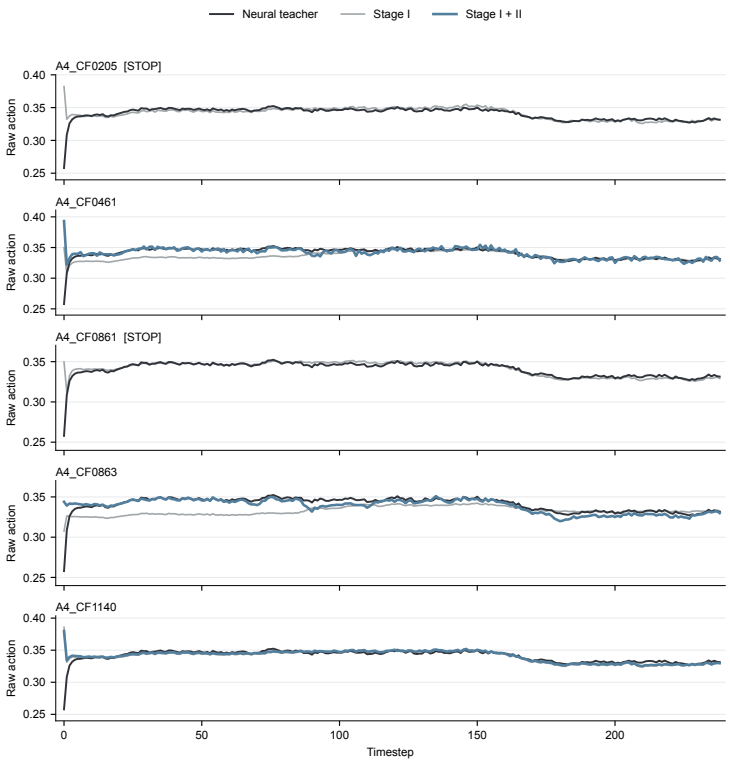

Figure 11: Fresh nominal raw-action trajectories for agent A_4. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

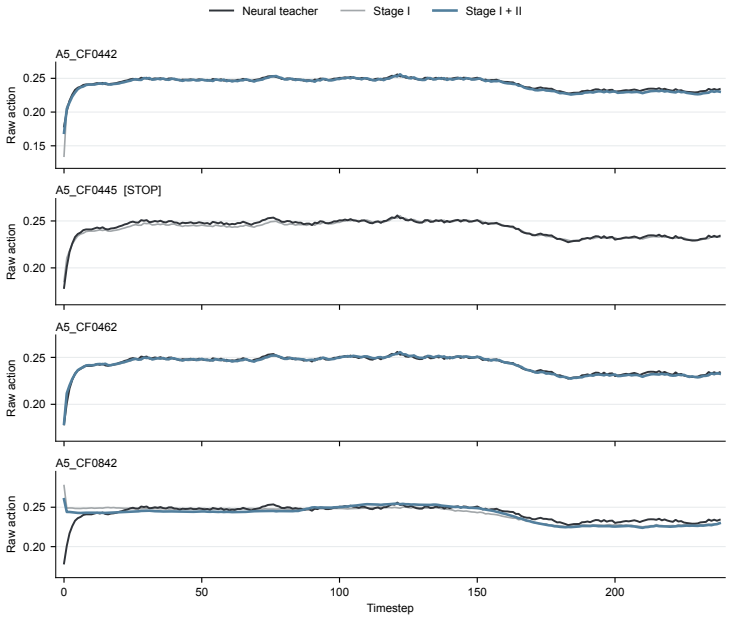

Figure 12: Fresh nominal raw-action trajectories for agent A_5. Black shows the frozen neural teacher, gray the Stage-I path, and blue the cumulative Stage-I+II path on teacher-induced episode 0. For STOP paths, the cumulative output equals Stage I, so the coincident blue line is omitted. Vertical scales are set separately for each path.

B.4 Additional Results for Cross-Policy Structural Reuse

Cross-policy reuse is evaluated on four independently trained target policies with training seeds 2–5. The symbolic bases, path identities, and Stage-II STOP/CONTINUE decisions discovered from the source policy are retained. Target-policy episodes 0–99 are used for coefficient recalibration and episodes 100–199 for held-out matched-state evaluation.

Three adaptation levels are considered. EXACT retains both the source-discovered symbolic structure and its source-policy coefficients. AFFINE retains the complete source symbolic mapping and adapts only an output scale and bias. FIXED_STRUCTURE retains the source-discovered symbolic basis and recalibrates its outer coefficients. For CONTINUE paths, Stage-I coefficients are recalibrated before fitting the Stage-II residual coefficients; STOP paths remain Stage-I explanations.

Table 6: Held-out cross-policy fidelity under the three recalibration levels. Results are averaged over target training seeds 2–5. Within each target policy, frozen paths are first averaged within agent and the six agents are then equally weighted.

Scope	Adaptation	NRMSE	SD across target seeds	R^2
ALL24	EXACT	4.4488	2.5642	-51.8053
	AFFINE	0.4742	0.0278	0.6443
	FIXED_STRUCTURE	**0.3213**	0.0535	**0.8265**
CONTINUE20	EXACT	4.4468	2.5630	-51.7426
	AFFINE	0.4771	0.0301	0.6410
	FIXED_STRUCTURE	**0.3149**	0.0575	**0.8348**
FINAL6	EXACT	4.4317	2.5597	-51.5174
	AFFINE	0.4774	0.0373	0.6377
	FIXED_STRUCTURE	**0.3308**	0.1078	**0.8126**

Table 7: Agent-balanced held-out NRMSE for each independently trained target policy. Values are reported for the complete 24-path symbolic family.

Target policy	EXACT	AFFINE	FIXED_STRUCTURE
Seed 2	5.4367	0.5119	**0.3807**
Seed 3	7.5204	0.4687	**0.2723**
Seed 4	1.6859	0.4450	**0.2802**
Seed 5	3.1523	0.4712	**0.3521**

Table 8: Agent-wise held-out NRMSE across independently trained target policies. Each entry reports AFFINE $\rightarrow$ FIXED_STRUCTURE. Values are averaged across the frozen symbolic paths belonging to the corresponding agent.

Agent	Target seed 2	Target seed 3	Target seed 4	Target seed 5
A_0	$0.0702 \rightarrow \mathbf{0.0606}$	$0.0702 \rightarrow \mathbf{0.0509}$	$0.1032 \rightarrow \mathbf{0.0588}$	$0.1140 \rightarrow \mathbf{0.0565}$
A_1	$0.0658 \rightarrow \mathbf{0.0625}$	$0.0605 \rightarrow \mathbf{0.0506}$	$0.0851 \rightarrow \mathbf{0.0705}$	$0.0869 \rightarrow \mathbf{0.0701}$
A_2	$1.0019 \rightarrow \mathbf{0.4821}$	$0.2054 \rightarrow \mathbf{0.1361}$	$0.2741 \rightarrow \mathbf{0.2015}$	$0.6960 \rightarrow \mathbf{0.3671}$
A_3	$0.3938 \rightarrow \mathbf{0.3362}$	$0.6355 \rightarrow \mathbf{0.4087}$	$0.2306 \rightarrow \mathbf{0.2091}$	$0.5441 \rightarrow \mathbf{0.4585}$
A_4	$0.7265 \rightarrow \mathbf{0.6205}$	$0.9337 \rightarrow \mathbf{0.3477}$	$0.9708 \rightarrow \mathbf{0.3970}$	$0.4141 \rightarrow \mathbf{0.3717}$
A_5	$0.8130 \rightarrow \mathbf{0.7224}$	$0.9069 \rightarrow \mathbf{0.6398}$	$1.0063 \rightarrow \mathbf{0.7443}$	$0.9723 \rightarrow \mathbf{0.7888}$

Across target policies, fixed-structure recalibration improves the agent-balanced held-out NRMSE relative to affine recalibration for each of the four independently trained policy instances (Table 7).

The same direction holds after resolving the aggregate by agent. All 24 agent–target-policy combinations exhibit lower held-out NRMSE under fixed-structure recalibration than under affine output recalibration (Table 8). Together with the family-level results in Table 6, this shows that the observed structural reuse extends across target-policy retraining, agent identity, and symbolic-path subsets.

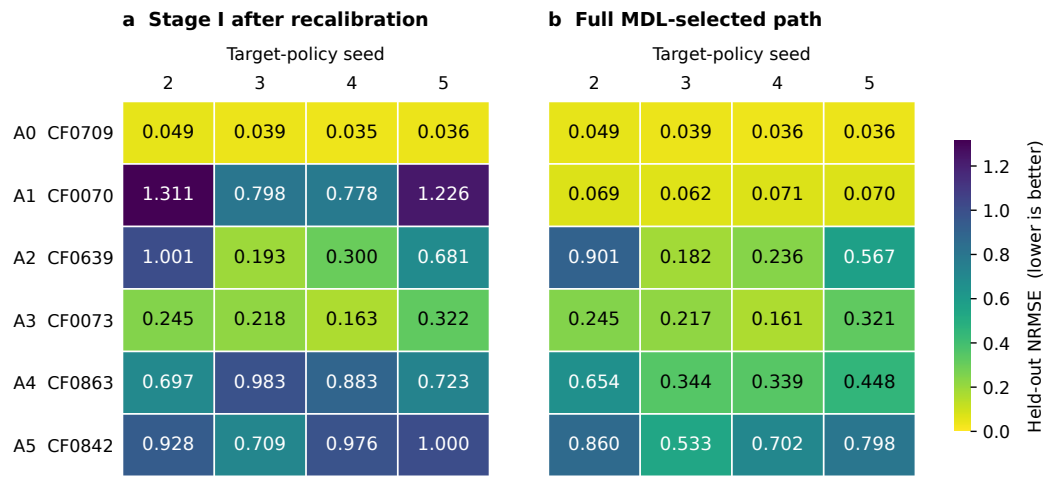

Figure 13: Cross-policy fidelity of the six source-selected progressive symbolic paths. Held-out matched-state NRMSE for the same six frozen deployment paths under four independently trained target policies. **a,** Stage-I component after coefficient recalibration. **b,** Full MDL-selected Stage-I+II path after Progressive coefficient recalibration. Both panels use a common color scale; lower values indicate better target-policy fidelity. The six paths were selected by source-policy robustness screening, without target-side path reselection.

B.4.1 Structure preservation and transfer of the final six paths

Progressive structural preservation. All 24 retained Stage-I paths preserve their exact canonical implicit supports across four independently trained target policies, yielding 96 matching path–target recalibrations. For the 20 CONTINUE paths, the selected Stage-II residual templates and MDL depth decisions remain fixed. Their nonzero outer-coefficient supports also match across all 80 target recalibrations. Together, these checks verify Stage-I canonical support preservation and Stage-II active-term retention under the frozen progressive representation defined in equation 37.

Transfer of the final six progressive paths. We evaluate the same six deployment paths selected using source-policy development data (Table 10). All six have MDL-selected Stage-II corrections, and their identities remain unchanged across the four target policies. Figure B.4.1 compares each path's recalibrated Stage-I component with its complete progressive explanation on matched held-out states. The corresponding aggregate transfer results are reported in Table 6.

The complete progressive explanations achieve lower held-out NRMSE than their recalibrated Stage-I components in 22 of the 24 paired FINAL6 comparisons. The contribution of residual refinement is particularly pronounced for A_1: across the four target policies, cumulative NRMSE ranges from 0.062 to 0.071, compared with 0.778 to 1.311 for the Stage-I component. These results demonstrate the cross-policy explanatory contribution of the frozen MDL-selected residual structures following sequential coefficient recalibration.

B.5 Additional Results for Closed-Loop Behavioral Preservation

This section provides the additional closed-loop evidence underlying Sec. 4.4. We distinguish three evaluation levels: nominal single-agent substitution, development-time composition selection, and final joint replacement of all six neural agents.

B.5.1 Matched Closed-Loop Effect of Stage-II Refinement

The functional contribution of Stage-II refinement is evaluated on the same 20 CONTINUE paths at both symbolic depths. For each metric, the matched Stage-I discrepancy from the neural controller is obtained as

$$\Delta_{\mathrm{S1-N}} = \Delta_{\mathrm{Cum-N}} - \Delta_{\mathrm{Cum-S1}},$$

where all three quantities use the same agent-balanced paired-episode aggregation.

Table 9 reports the resulting closed-loop differences. Progressive refinement substantially reduces the discrepancy from the neural controller across return, voltage-regulation, and loss-related metrics.

Table 9: Matched nominal closed-loop effect of Stage-II refinement. Results use the 20 CONTINUE paths only. Differences are symbolic minus neural. The final column reports the reduction in absolute discrepancy after adding the selected Stage-II correction.

Metric	Stage-I − neural	Stage-I+II − neural	Absolute discrepancy reduction
Episode return	$+9.9841 \times 10^{-3}$	-1.8364×10^{-3}	81.6%
Voltage out-of-control rate	$+8.2566 \times 10^{-4}$	-4.6789×10^{-5}	94.3%
Controllable ratio	-4.2631×10^{-3}	$+8.1358 \times 10^{-5}$	98.1%
Average voltage deviation	$+2.4919 \times 10^{-4}$	$+3.6877 \times 10^{-6}$	98.5%
Total line loss	-1.3167×10^{-4}	$+1.4230 \times 10^{-5}$	89.2%
Reactive-power loss	-2.9096×10^{-3}	$+3.9960 \times 10^{-5}$	98.6%

B.5.2 Frozen Six-Controller Composition

The development robustness screen evaluates the frozen symbolic candidates over eight shifted conditions,

$$s_{\mathrm{PV}}, s_{\mathrm{D}} \in \{0.8, 1.0, 1.2\}, \qquad (s_{\mathrm{PV}}, s_{\mathrm{D}}) \neq (1.0, 1.0),$$

using single-agent symbolic substitution.

A candidate is eligible for final composition only if it produces no non-finite symbolic evaluations, environment-destroy events, premature terminations, or native raw-action clamps over the development grid. Among eligible candidates, selection prioritizes the worst-case absolute paired discrepancy in voltage out-of-control rate, followed by average voltage deviation, controllability, and mean absolute episode-return difference.

The resulting composition contains one controller per agent (Table 10). All six selected controllers use the cumulative Stage-I+II representation.

Table 10: Final six symbolic controllers selected for joint deployment. The final column reports the maximum absolute raw symbolic action observed for that path over the $\pm 20\%$ development screen.

Agent	Path	Variant	Dev. max $\lvert a^{\mathrm{raw}} \rvert$
A_0	A0_CF0709	Stage-I+II	0.468619
A_1	A1_CF0070	Stage-I+II	0.464420
A_2	A2_CF0639	Stage-I+II	0.151707
A_3	A3_CF0073	Stage-I+II	0.179369
A_4	A4_CF0863	Stage-I+II	0.409086
A_5	A5_CF0842	Stage-I+II	0.282541

B.5.3 Full Six-Agent Closed-Loop Results

After composition selection, all six neural agents are replaced simultaneously by the frozen symbolic controllers. The final evaluation contains the nominal condition, four previously unseen $\pm 10\%$ PV–demand shifts, and four previously unseen $\pm 30\%$ shifts. Each controller–scenario pair is evaluated over 100 paired episodes.

Under nominal operation, the symbolic controller attains a mean episode return of -6.749642, compared with -6.748818 for the all-neural controller.

Table 11 reports the paired symbolic-minus-neural differences in return and the principal voltage-regulation metrics.

Table 12 completes the system-level evaluation with line-loss and reactive-power-loss differences, together with the raw-action range of the explicit symbolic controller.

Across the nine final operating conditions, the symbolic controller executes 1,290,600 raw symbolic actions. No raw action reaches the native clamp threshold, no symbolic evaluation becomes non-

Table 11: Full-six symbolic closed-loop behavior across the final operating conditions. All differences are full-symbolic minus all-neural. Return intervals are paired 95% confidence intervals.

PV / Demand	ΔR [95% CI]	ΔOOC	ΔControllable	ΔV_{dev}
1.0 / 1.0	-0.000824 $[-0.005566,\ 0.003917]$	-0.000015	-0.000335	$+0.000031$
0.9 / 0.9	-0.002061 $[-0.005340,\ 0.001217]$	$+0.000056$	-0.000544	$+0.000088$
0.9 / 1.1	$+0.001414$ $[-0.004345,\ 0.007173]$	$+0.000048$	-0.000711	$+0.000052$
1.1 / 0.9	-0.004345 $[-0.009129,\ 0.000439]$	-0.000076	$+0.000586$	-0.000015
1.1 / 1.1	$+0.002139$ $[-0.005375,\ 0.009653]$	-0.000086	-0.000126	-0.000042
0.7 / 0.7	-0.002027 $[-0.004041,\ -0.000012]$	$+0.000055$	-0.000711	$+0.000150$
0.7 / 1.3	-0.000909 $[-0.009406,\ 0.007588]$	$+0.000185$	-0.000962	$+0.000071$
1.3 / 0.7	-0.013171 $[-0.020359,\ -0.005984]$	-0.000609	$+0.004728$	-0.000210
1.3 / 1.3	$+0.004868$ $[-0.012886,\ 0.022622]$	-0.000654	$+0.001339$	-0.000226

Table 12: Loss-related differences and raw symbolic-action diagnostics in the final full-six evaluation. Differences are full-symbolic minus all-neural. All nine scenarios have zero native action clamps, zero non-finite symbolic evaluations, zero environment-destroy events, and zero premature terminations.

PV / Demand	Δ line loss	Δq-loss	Max $\lvert a^{\mathrm{raw}}\rvert$
1.0 / 1.0	-0.000162	-0.000275	0.458530
0.9 / 0.9	-0.000216	-0.000794	0.429836
0.9 / 1.1	-0.000281	-0.000582	0.418302
1.1 / 0.9	$+0.000222$	$+0.000330$	0.488328
1.1 / 1.1	-0.000060	$+0.000328$	0.477188
0.7 / 0.7	-0.000166	-0.001417	0.372462
0.7 / 1.3	-0.000149	-0.000671	0.392815
1.3 / 0.7	$+0.002200$	$+0.002653$	0.524078
1.3 / 1.3	-0.000065	$+0.002052$	0.618812

finite, and no episode ends through an environment-destroy event or premature termination. The largest observed raw-action magnitude is 0.618812, under $(s_{\mathrm{PV}}, s_{\mathrm{D}}) = (1.3, 1.3)$.

Taken together, the supplementary closed-loop results show that the Stage-II correction improves agreement with the neural controller at the feedback-system level, that the selected symbolic paths compose into a complete six-agent controller, and that the resulting explicit control law remains numerically well behaved over the final evaluated operating conditions.

C Selected symbolic expressions

C.1 Selected Stage-I expressions and validation fidelity

The 24 Stage-I representatives retained by the robust Pareto compression are listed below with their symbolic expressions, minimum PySR discovery complexity, and development-validation teacher NRMSE.

Table 13: Primary Stage-I symbolic representatives. The displayed expression is a minimum-complexity PySR realization associated with each retained canonical function, written using physical-variable notation.

Path	ID	Symbolic expression	$C(g)$	Val. NRMSE
$g^{(1)}_{0,1}$	CF0432	$0.54967165\, Q_{12}^{PV}$	3	0.272534
$g^{(1)}_{0,2}$	CF0504	$-1.9523535 V_{15} - 1.9523535\theta_{16} + 2.0410025$	7	0.063675
$g^{(1)}_{0,3}$	CF0709	$\dfrac{-P_{12}^{PV}\theta_{15} + \theta_{15} + 1.2492019}{V_{15}} \times \Big[-0.19124326 P_{12}^{PV} + 0.19124326 P_6 + 0.126185307791163 P_7 - V_{15}\,(V_{15}\theta_{13} + 0.46198922) + 0.546461034659358 \Big]$	27	0.023245

Continued on next page

Path	ID	Symbolic expression	$C(g)$	Val. NRMSE
$g_{1,1}^{(1)}$	CF0070	$-3.7888513\theta_{11}$	3	0.257059
$g_{1,2}^{(1)}$	CF0913	$-1.8670233V_{12} - 1.8670233\theta_{16} + 1.88339840125731$	7	0.079696
$g_{1,3}^{(1)}$	CF0884	$-(V_{12} - 1.4339235)\left(-P_{17}^{PV} + 0.70401776P_6 + 0.70401776P_7\right) + 0.052868456$ $-\theta_{11}\left[V_{12} + \theta_{17}\left(2Q_{17}^{PV} - \theta_{17} + 1.7178397\right)\right]$	13	0.041224
$g_{1,4}^{(1)}$	CF0896	$-(V_{12} - 1.3502903)\left(-P_{17}^{PV} + 0.7526653P_6 + 0.7526653P_7 + 0.7526653\theta_{14} + 0.130073238723652\right)$	29	0.033642
$g_{2,1}^{(1)}$	CF1082	$0.5148161Q_{21}^{PV}$	3	0.558179
$g_{2,2}^{(1)}$	CF1083	$0.080296814 - 10.2150160\theta_{19}$	5	0.251350
$g_{2,3}^{(1)}$	CF0639	$\frac{0.0967776399192 - 0.14094095P_{21}^{PV}}{P_{21}^{PV} + 0.818864}$	7	0.131248
$g_{2,4}^{(1)}$	CF1056	$\left(0.6770084 - P_{21}^{PV}\right)\left[\left(\left[P_{21}^{PV} + 1.3044355(Q_{21}^{PV})^2\right] \times \left(Q_{21}^{PV} + 4.5015388\right) - 10.950373\right)\sin(\theta_{19}) + 0.14034784\right]$	22	0.097697
$g_{3,1}^{(1)}$	CF0609	$1.0317682 - V_{24}$	3	0.696916
$g_{3,2}^{(1)}$	CF0610	$-4.026491 + \frac{4.034883}{V_{23}}$	5	0.210228
$g_{3,3}^{(1)}$	CF0362	$\frac{V_{24}\left[Q_{24}^{PV}\left(P_{23}(P_{23} - 0.33459392) + 0.04951871\right) - 2.7062352\right] + 2.7203705}{V_{24}} - 2.6656835V_{24}$	15	0.151343
$g_{3,4}^{(1)}$	CF0073	$+\left[0.09918596Q_{24}^{PV} + \left(1.2540977P_{24}^{PV} - V_{22}\right) \times (9.386009P_{23} - 0.93096031323364) \times (V_{24} - 0.9847905)\right] \times (P_{23} + 0.19637279) + 2.6798484089823$	29	0.124353
$g_{4,1}^{(1)}$	CF0863	$1.3040671 - V_{29}$	3	0.640473
$g_{4,2}^{(1)}$	CF0461	$Q_{28}^{PV} - V_{29} + 1.2163559$	5	0.415776
$g_{4,3}^{(1)}$	CF0205	$Q_{28}^{PV}Q_{29} - \theta_{25} + 0.30299$	7	0.239369
$g_{4,4}^{(1)}$	CF1140	$-\theta_{25} + \left(Q_{28}^{PV} + 0.6965253\right) \times (0.08827932Q_{29} + 0.39499265)$	11	0.203710
$g_{4,5}^{(1)}$	CF0861	$-Q_{28}^{PV}\left(Q_{28}^{PV} + 0.08790261\right) \times \left(94.83461(Q_{28}^{PV})^2V_{29} - 3.6847553\right) \times \left[-55.52055Q_{28}^{PV}\left(P_{28}^{PV} + Q_{28}^{PV}\right) + Q_{29} + 3.1192632\right] + 0.20012584$	29	0.145031
$g_{5,1}^{(1)}$	CF0842	$0.24359299 - \theta_{25}$	3	0.553044
$g_{5,2}^{(1)}$	CF0442	$\frac{0.15069222Q_{32}^{PV} - 0.016602970181245}{Q_{32}^{PV} - 0.09306475}$	7	0.338248

Continued on next page

Path	ID	Symbolic expression	$C(g)$	Val. NRMSE
$g_{5,3}^{(1)}$	CF0445	$-\frac{V_{25}\left(Q_{32}^{PV}+0.6813544\right)}{Q_{32}^{PV}-0.09530889}\times\left[\left(Q_{32}^{PV}-0.09530889\right)\times\left(V_{25}-1.219321\right)+0.0028812147\right]$	15	0.253540
$g_{5,4}^{(1)}$	CF0462	$\frac{V_{25}^2}{Q_{32}^{PV}-0.09479328}\Bigg[-V_{25}\Bigg(Q_{32}^{PV}\Bigg(Q_{32}^{PV}\Big(Q_{32}^{PV}\left(Q_{32}^{PV}-0.09479328\right)+V_{25}-0.7293505\Big)-0.55627155Q_{32}^{PV}+0.052730804795184\Bigg)+0.0969075\theta_{25}\left(Q_{32}^{PV}-0.09479328\right)\Bigg)-\left(Q_{32}^{PV}-0.09479328\right)\times\left(V_{25}-1.1646181\right)\Bigg]$	29	0.216102

C.2 MDL-Selected Stage-II Residual Corrections

Table 14 reports the frozen Stage-II outcome for every Stage-I path. A zero correction denotes a path for which the integrated MDL criterion retains the Stage-I representation without an additional symbolic layer.

Table 14: MDL-selected Stage-II residual refinements. For each frozen Stage-I path, the table reports the correction selected by exact minimization of the conditional integrated mixture MDL. C_2 is the minimum original PySR complexity of the selected correction, with $C_2(0) = 0$.

Stage-I path	Stage-II candidate	Selected correction $h_{i,j}^{(2)\star}$	C_2	Decision
A0_CF0432	A0_CF0432 H2_0057	$-0.293834703485869P_{12}^{PV}-0.53171617Q_{12}^{PV}+0.293834703485869P_6+0.53171617P_9-0.53171617V_{15}-\sin(\theta_{11})+0.53171617\cos(\theta_9)+0.101714825873354$	19	CONTINUE
A0_CF0504	A0_CF0504 H2_0091	$-3.1626995P_{14}P_6+3.1626995P_{14}\theta_{17}+P_6^2-P_6\theta_{17}-23.5311569861836P_6\sin^2(\theta_{15})+23.5311569861836\theta_{17}\sin^2(\theta_{15})$	15	CONTINUE
A0_CF0709	A0_CF0709 H2_0077	$0.44301963\left(Q_{12}^{PV}\right)^2\theta_6$	7	CONTINUE
A1_CF0070	A1_CF0070 H2_0024	$-1.7900722V_{12}+55.423985\theta_{14}\theta_{16}\theta_9-8.3708884148148\theta_{16}\theta_9+\theta_9+\frac{1.8111011}{V_{17}}$	19	CONTINUE
A1_CF0913	A1_CF0913 H2_0013	$-0.06162032941281P_{17}^{PV}-0.06162032941281Q_{17}^{PV}-0.1835378P_{13}+0.1835378P_6$	11	CONTINUE
A1_CF0884	A1_CF0884 H2_0173	$P_{17}^{PV}\left(Q_{17}^{PV}\right)^2P_9+0.49453956P_{17}^{PV}Q_{17}^{PV}P_9$	9	CONTINUE

Continued on next page

Stage-I path	Stage-II candidate	Selected correction $h_{i,j}^{(2)\star}$	C_2	Decision
A1_CF0896	STOP	0	0	STOP
A2_CF1082	A2_CF1082 H2_0059	$-0.0774062\left(P_{21}^{PV}\right)^3 + 0.21760482\left(P_{21}^{PV}\right)^2 - 0.25947756P_{21}^{PV} - 0.05143357\left(Q_{21}^{PV}\right)^3 V_{20}^2 V_{21} - 0.4253996Q_{21}^{PV} V_{20}^2 V_{21} + 0.10053369$	29	CONTINUE
A2_CF1083	A2_CF1083 H2_0100	$-4.45765207553672\left(P_{21}^{PV}\right)^2\theta_{19} + 0.0951736467891366\left(P_{21}^{PV}\right)^2 - 0.145989295232816P_{21}^{PV}P_{18} + 37.7922551284811P_{21}^{PV}\theta_{19}\theta_{21} + 11.6838775359616P_{21}^{PV}\theta_{19} - 2.38756290532084P_{21}^{PV}\theta_{21} - 0.255901838734675P_{21}^{PV} + 0.382650107760192P_{18} - 99.0566498341672\theta_{19}\theta_{21} + 6.258001325016958\theta_{21} + 0.0168896840566848$	27	CONTINUE
A2_CF0639	A2_CF0639 H2_0152	$Q_{21}^{PV}\sin(\theta_{21}) + P_{18}P_{19} + 0.28121987P_{18} + 0.15816922\sin(\theta_{21}) - 0.015308235$	13	CONTINUE
A2_CF1056	A2_CF1056 H2_0208	$69.061185813636\left(Q_{21}^{PV}\right)^3\sin^2(\theta_{21}) + 84.6806782418017\left(Q_{21}^{PV}\right)^2\sin^2(\theta_{21}) + 2.5213127681454\left(Q_{21}^{PV}\right)^2\sin(\theta_{21}) + 8.310306Q_{21}^{PV}P_{18}\sin(\theta_{21}) + 2.763163872066Q_{21}^{PV}\sin(\theta_{21}) + 0.3033959P_{18} - 0.0119890427481194$	21	CONTINUE
A3_CF0609	A3_CF0609 H2_0074	$-1.871402V_{23} - \theta_{23} + 1.8577629$	7	CONTINUE
A3_CF0610	A3_CF0610 H2_0119	$0.02566137Q_{24}^{PV} - 0.002889395$	5	CONTINUE
A3_CF0362	A3_CF0362 H2_0094	$P_{22}P_{23} - 0.12590893P_{22} + 2.22640866211897P_{23}\theta_{24} - 0.047362328P_{23} - 0.2803247323901320\theta_{24} + 0.00596334004078904$	11	CONTINUE
A3_CF0073	A3_CF0073 H2_0142	$P_{22}^2V_{24} + 2.0V_{22}^2 - 1.9914901V_{24}$	13	CONTINUE
A4_CF0863	A4_CF0863 H2_0154	$0.268879802954016Q_{28}^{PV} - 0.268879802954016P_{30} - 3.20590494212640\theta_{25} + 2.569744\theta_{29} - 0.22095065P_{30} - 0.4419013\theta_{25}$	13	CONTINUE
A4_CF0461	A4_CF0461 H2_0111	$-21.6447599140410\theta_{26} + 79.5622998280820\theta_{28} - 21.6447599140410\theta_{29} - 36.27278\theta_{30} + 0.0274367734381662$	25	CONTINUE
A4_CF0205	STOP	0	0	STOP
A4_CF1140	A4_CF1140 H2_0177	$-6.769354\left(Q_{28}^{PV}\right)^3 - 6.769354\left(Q_{28}^{PV}\right)^2 P_{27} + 0.07944733Q_{28}^{PV}$	11	CONTINUE
A4_CF0861	STOP	0	0	STOP
A5_CF0842	A5_CF0842 H2_0015	$-0.6631183V_{29} + 0.643453301353278$	5	CONTINUE
A5_CF0442	A5_CF0442 H2_0049	$0.18381794Q_{32}^{PV} - 0.0117957093387434$	5	CONTINUE
A5_CF0445	STOP	0	0	STOP

Continued on next page

Stage-I path	Stage-II candidate	Selected correction $h_{i,j}^{(2)\star}$	C_2	Decision
A5_CF0462	A5_CF0462 H2_0117	$-2.1179085\theta_{28} + 2.0\sin(\theta_{32}) + 0.0013468722$	9	CONTINUE