

Highlights

Fidelity Is Not Identification: Validating Robot Digital Twins Under Edits

Suhwan Park, Jihwan Lee, Jiyong Park, Sanghyun Kim

- Fidelity validates a model edit only if the edit preserves the data's null space.
- Fidelity alone never certifies clearing a link's mass on a revolute arm.
- SAPC states a sparse prior and flags edits that the data leave undecided.
- Calibrated unloaded first, a known tool location is the prior to use.
- Gazebo, Isaac Sim and **a blinded physical FR3 session** test the ladder.

Fidelity Is Not Identification: Validating Robot Digital Twins Under Edits

Suhwan Park^a, Jihwan Lee^a, Jiyong Park^a and Sanghyun Kim^{a,*}

^aDepartment of Mechanical Engineering, Kyung Hee University, Yongin, 17104, Republic of Korea

ARTICLE INFO

Keywords:

digital twin
verification and validation
model calibration
identifiability
robot dynamics identification
sparse model selection

ABSTRACT

Simulation models are calibrated to measured behavior and then modified: a component removed, exchanged or added. Parameters that fit the data equally well can predict the modified model differently. For corrections linear in the parameters, a modification is guaranteed by fidelity if and only if it maps the calibration's null space into itself, the estimability condition in operational form. For robot twins, edits by known inertia or of joint friction are guaranteed; removing what a twin attributed to a link is not on revolute arms and rests on a prior. SAPC, a simulator-native calibration, states that prior as sparse change over body and joint groups, writes its correction into the executable model and flags edits the data leave undecided. On a Franka FR3 reference data set, full-library calibrations reproduce held-out motion to 0.05 mm, yet dense and regularized twins put 0.13-0.18 kg of a 0.8 kg tool on the flange, dense twins miss its removal by 12 mm, and SAPC recovers 0.79 kg and 0.1 mm. In a pre-registered 300-pattern tool benchmark, joint-torque payload identification attributes the tool best after unloaded-first calibration, and SAPC flags 88 % of its misattributions at a flag rate of 74 %, leaving 5 % of the fits wrong and unflagged against 20-23 % for Bayesian enumerations. On the Menagerie iiwa model, a near-alias between wrist links leaves the tool mass undecided by these data without a location prior. Gazebo, Isaac Sim and a **blinded physical FR3 session** complete the validation.

1. Introduction

Digital twins of robots are simulators kept in step with a physical machine, so that planning and control can be tested, tuned and certified before they reach hardware [1, 2]. They are calibrated by matching behavior: simulator parameters are tuned until the twin reproduces measured trajectories, torques or disturbance responses [3, 4, 5, 6], or the rigid-body parameters are re-identified from excitation data under physical-consistency constraints [7, 8]. The evidence for their credibility is behavioral as well: validation frameworks [9, 10] and credibility scores [11] compare model outputs with measured ones, scaled to the context of use [12, 13]. And then the twin is edited. A tool is taken off or exchanged, a link or a gearbox replaced, the robot recalibrated before a new payload is mounted, and the twin is expected to follow without a new identification [14, 15].

Behavior does not determine where a correction belongs (Fig. 1). The dynamics of a manipulator depend on its inertial parameters only through base combinations, and the remaining directions of parameter space produce no torque in any motion [16, 17, 18]. That a calibrated model can predict well with wrong physical parameters is familiar from the calibration of computer models [19, 20, 21], from estimability in linear models [22] and from prediction identifiability [23], and change isolation in fault diagnosis and structural model updating rests on priors for the same reason [24, 25, 26]. For an edited twin these classical facts become a practical question: which edits of a calibrated twin does its fidelity guarantee, and which rest on a choice the data did not make? Dense identification, regularization toward nominal values, sparse selection and Bayesian model selection all return one member of the class of equivalent corrections, and whether the edit about to be made depends on that choice is a property of the edit, which a method can only report through its own uncertainty statement; Section 5 measures how well each such statement flags it.

This paper turns the estimability condition into a test on the edit and builds a calibration around it. An edit of the correction is an affine map, and it is certified by fidelity, so that every twin that reproduces the robot still does after the same edit, if and only if its linear part maps the structural null space into itself. Edits by known inertia and edits of joint friction pass; edits that clear what a twin attributed to a link do not on an arm with revolute joints, because a point mass on a joint axis belongs equally to the link and its parent. An attribution-based edit therefore rests on a prior.

*Corresponding author

 dlw1ghks02@khu.ac.kr (J. Lee); qkrwldyd0919@khu.ac.kr (J. Park); kim87@khu.ac.kr (S. Kim)

ORCID(s): 0009-0002-6139-4800 (J. Lee); 0009-0006-7185-9391 (J. Park); 0000-0002-0222-8688 (S. Kim)

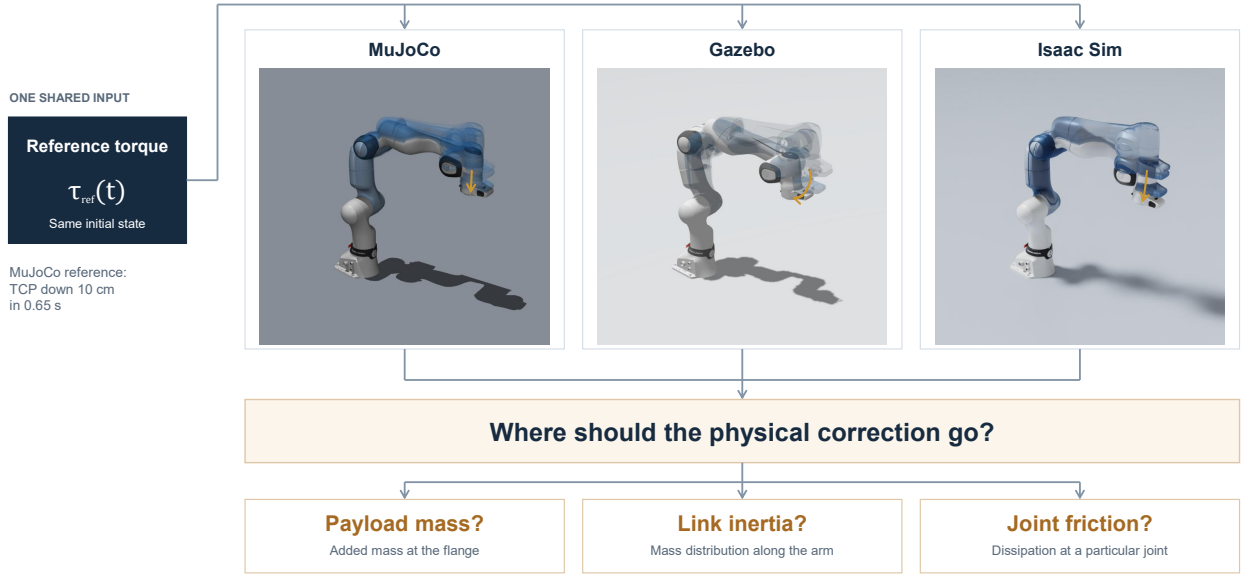


Figure 1: Where should the physical correction go? The reference torque $\tau_{ref}(t)$ moves the tool-center point (TCP) of the MuJoCo FR3 10 cm downward in 0.65 s. Applied from the same initial state to MuJoCo, to Gazebo and to Isaac Sim, the last with a 100 g payload added at the flange, it produces a different motion in each: at 0.65 s the Gazebo TCP is 82 mm and the Isaac Sim TCP 46 mm from the MuJoCo TCP. Ghosts show the robot at 0.000, 0.217 and 0.433 s, the solid pose is at 0.650 s, and the orange line is the TCP path. A calibration that restores the motion must decide whether the correction belongs to the payload, to the mass distribution along the arm, to the friction of a joint or to a property that one simulator does not represent, such as the rotor inertia absent from the Gazebo description; reproducing the motion does not decide it.

When the location of the change is known, as for a tool at the flange, that prior is the location itself, and joint-torque payload identification on a twin calibrated to the unloaded robot is the right method; a wrist force/torque sensor or a weighed part turns the edit into a certified known-inertia edit. Attribution is needed where neither holds: for changes inside the arm, dress packs, unknown grasped objects, wear and unannounced modifications. Structure-Aware Physical Calibration (SAPC) states sparse change over body and joint groups for that case, reads its library off the simulator, selects groups by blocked cross-validation and writes them into the executable model; with the twin it reports the supports of the selected size within its parsimony band, the directions shared inside a support and a flag on the edit the user intends.

The validation follows the ladder of computational verification and validation [27, 28]. Manufactured discrepancies on a Franka FR3 and the MuJoCo Menagerie KUKA LBR iiwa 14 rigid-body model give exact truth, tested in two confirmations fixed before they were run: the first with 640 runs on new trajectories and 820 on the design trajectories with new changes and noise, the second, after a first round of review, a tool benchmark of 300 patterns on new trajectories; Gazebo and Isaac Sim references add model-form discrepancy; a blinded session on a physical FR3 adds real sensors. On the FR3 reference data set, full-library calibrations reproduce held-out motion to 0.05 mm, the dense and regularized ones while placing between 0.13 and 0.18 kg of a 0.8 kg tool on the flange, and removing the tool through each twin’s attribution leaves 12 mm for dense twins and 0.1 mm for SAPC. On the iiwa, a heavier link 3 with the full tool and a heavier link 4 with a lighter tool differ by 0.002 N m on the identification trajectories, so the attribution is set by the prior on these data; SAPC selects the injected support in only 10 of 200 replications, and its parsimony band contains the other support in every run.

The contributions are three. First, the estimability condition as an operational test for edits of executable twins, with the null space read off the simulator’s own regressor, its corollary for revolute chains, the repair that makes certified edits executable, and the leakage interval as a screening statistic before an uncertified edit. Second, a simulator-native calibration that states its prior, writes its correction into the simulator and reports where the prior decided, with an edit flag whose operating characteristics are measured on pre-registered runs against dense, regularized, classical and Bayesian alternatives. Third, a validation ladder of manufactured discrepancies on two robots with two pre-registered

confirmations, two other simulators as references, and a blinded physical session, with code, data and protocols released [29]. Section 2 positions the work, Section 3 develops the criterion, Section 4 presents SAPC, Section 5 reports the validation, and Sections 6 and 7 discuss and conclude.

2. Related work

Calibrating a twin that will be edited touches four bodies of work: dynamic identification, the detection and isolation of changes, the calibration of computer models, and the validation of simulation models. Each supplies part of the answer, and the criterion of Section 3 is a classical one put in operational form for executable twins.

2.1. Dynamic identification and physical consistency

Rigid-body inverse dynamics is linear in the standard inertial parameters, which are fitted to torques measured along optimized excitation trajectories [30, 31, 7]. Only base parameters affect the torques, and a link's mass enters them regrouped with its parent's parameters [16, 17]; the base parameters can be computed numerically from the regressor [32], and which inertial quantities are observable, including the transfer of inertia between neighboring links, has been characterized geometrically [18]. Physical consistency is imposed by linear matrix inequalities on the pseudo-inertia [33, 8] or by a log-det regularizer that pulls each body toward a prior inertia [34]. Payload identification comes in two families. With a wrist force/torque sensor, the payload's ten parameters are estimated from the wrench measured at the tool interface [35, 36, 37], which locates the payload by measurement and turns its removal into a known-inertia edit. From joint torques alone, the payload is identified as the change of the last link's parameters on the premise that nothing else changed [38]; the link-7-only baseline of this paper implements this family. Sparse identification selects few terms from a library [39, 40]; in robotics it learns friction and structural-uncertainty terms for fault-tolerant control [41] and separates unmodeled joint dynamics from the inertia of a grasped object [42]. These methods return a parameter vector that reproduces the torques; which member of the class of equivalent vectors they return is set by their regularizer, their constraints or their premise, and it is this choice that an edit exposes.

2.2. Change detection, isolation and model updating

Fault detection and isolation generates residuals from a model and asks which component a residual points to; a fault is isolable when its residual signature differs from those of the other candidate faults [24], and for manipulators momentum residuals isolate actuator faults and collisions before identification [43]. Structural dynamics updates finite-element models from measured responses, where the choice of updating parameters and the ill-conditioning of their estimation are the central difficulties [25]. Damage localization adds the prior that damage is concentrated in few elements [44], and Bayesian model updating makes the identifiability of the updated parameters explicit, including the case of several equally plausible models [26]. Attribution in a robot twin is isolation in the simulator's parameter space under the same sparse-change prior; the report of Section 4.4 is the analog of an isolability statement, and the edit criterion of Section 3.3 says which consequences of isolation no measurement can check.

2.3. Calibration, identifiability and sim-to-real

Kennedy and O'Hagan calibrate computer models with an explicit discrepancy term [19]; Brynjarsdóttir and O'Hagan showed that a calibration can predict well while returning biased physical parameters unless prior information constrains the discrepancy [20], and Tuo and Wu showed that calibration parameters are defined only through the criterion that selects them [21]. In a linear model, a function of the parameters can be estimated exactly when it is orthogonal to the null space of the design [22]; for dynamic models, prediction identifiability asks whether a prediction, rather than a parameter, is determined by the data [23], and hydrology calls the set of parameter sets that reproduce the observations equally well equifinal [45]. Section 3 applies these ideas to the prediction of an edited model, where they become a test on the edit. Robotic sim-to-real methods adapt simulator parameters until simulated behavior matches the real one, by randomization updated from real rollouts [6], by learned residual tuning [4], by search over friction, inertia and gravity terms [5], or by neural augmentation of a differentiable simulator [46]; hybrid twins add a confidence measure that decides when a learned submodel is trusted [47]. These methods target behavior by construction, which Section 3 shows to be blind to the representative they return, and a neural residual cannot be written back into the simulator's physical parameters.

2.4. Validation of simulation models and digital twins

Simulation credibility rests on conceptual, operational and data validity [13], with evidence scaled to the context of use [12]; computational science separates verification, often by manufactured solutions whose exact answer is known [27], from validation against a hierarchy of references, and a prediction outside the validation domain needs its own justification [28]. Robot simulations have been validated against the real robot as a whole [9], by modular analytical and simulation checks [10], and by evidence-theoretic credibility scores [11]; robot twins in this journal compare simulators [48, 49], validate co-simulation by output trajectories [3], serve as test platforms [2] and follow a drifting physical system by recalibration [14], and predictive twins at scale are updated asset by asset from data [15]. These validate what a twin outputs. An edited twin predicts outside its validation domain, and this paper adds the test that says when its validation still applies, a calibration that states its prior and reports where it decided, and a validation ladder of manufactured discrepancies with known truth, other simulators as references and a physical robot.

3. What fidelity guarantees under edits

This section asks a question about any calibrated simulation model and answers it for models whose correction enters linearly, with the robot twin as the case. The answer separates the edits that the fidelity of a calibrated model guarantees from those that rest on a prior, and Section 3.5 reduces it to two links.

3.1. Setting

A simulation model with parameters θ is calibrated on data D and then modified: a component is removed, exchanged or added, and the modified model is used without new data. Call $C(\hat{\theta})$ the *calibration-equivalence class* of the calibrated parameters, the set of parameters whose outputs on the experimental conditions of D equal those of $\hat{\theta}$. An intended modification is a map E on parameters, and the question is when the fidelity of the calibrated model validates the prediction of the modified one.

Definition 1 (Guaranteed by fidelity). A modification E is *guaranteed by fidelity*, or *certified*, at $\hat{\theta}$ if the output of the modified model $E(\theta)$, on the conditions under which it will be used, is the same for every $\theta \in C(\hat{\theta})$. For the robot twins below these conditions are every state of the arm.

A certified modification inherits the validation of the calibrated model, because every parameter the data admit predicts the same modified behavior. An uncertified one does not: two calibrated models with the same validation record can disagree after it, and the disagreement is decided by whatever chose $\hat{\theta}$ within the class. The definition is the prediction-identifiability question [23] asked of the modified model, and the class is the equifinal set of hydrology [45]. It can be answered when the class has a computable form, which is the case below.

For the robot, let T_0 be the nominal simulator with inverse dynamics $ID_0(q, \dot{q}, \ddot{q})$, the torque that produces the accelerations $\ddot{q}$ at joint positions q and velocities $\dot{q}$, and let R be the reference system that generates the observations. The physical discrepancy is the residual

$$r(q, \dot{q}, \ddot{q}) = \tau - ID_0(q, \dot{q}, \ddot{q}), \quad (1)$$

which calibration expresses through a library of candidate physical corrections, $r \approx A(q, \dot{q}, \ddot{q}) \xi$. The inertial columns of A are extracted from the simulator itself (Section 4), and the friction and rotor-inertia columns are analytic functions of $\dot{q}$ and $\ddot{q}$, so A is the model the twin executes. Its 112 columns are partitioned into 15 groups $\mathcal{G}$: the ten standard inertial parameters of each of the seven links (mass, first mass moments and inertia tensor) [30], five friction terms for each joint, and one rotor-inertia (armature) group for all joints. A calibration returns the selected groups $S \subseteq \mathcal{G}$, their increments ξ_S , and the executable twin T_c obtained by writing ξ_S into the native fields of the simulator. Because the state is measured and the correction enters linearly, the residual is an equation error and, for data that reach the structural rank, $C(\hat{\xi}) = \hat{\xi} + \mathcal{N}$ for the null space $\mathcal{N}$ of Section 3.2. Table 1 maps the robot terms to their generic meaning.

Definition 2 (Two axes of validation). *Behavioral fidelity* compares outputs on held-out data: the torque NRMSE of the correction, $\text{RMS}(r - A\xi)/\text{RMS}(r)$, and of the executable twin; the tool-center-point (TCP) RMSE when T_c replays the reference trajectory in closed loop; and contact quantities of a press task never used for calibration. *Physical attribution* compares the correction with the change of the reference: the precision and recall of S against the changed groups S^* , the error of the reconstructed physical quantities, such as payload mass and center of mass, and the stability of S over measurement-noise realizations.

Table 1

Terms used in this paper, their meaning in the robot case and their generic counterpart for a calibrated simulation model.

Term	Robot case	Generic counterpart
Twin	Simulator of the robot whose parameters are calibrated to it	Calibrated simulation model
Link, joint	Rigid body of the arm; the revolute joint that moves it relative to its parent	Component
Flange, tool, payload	Mounting face of the last link; a body fixed to it	Component at the end of the chain
Armature	Rotor inertia of a joint's motor reflected through its gearbox	Component parameter
Coulomb, viscous, Stribeck friction	Joint friction constant in magnitude, proportional to velocity, and dropping from static to sliding at low velocity	Dissipation terms
Link-side torque	Joint torque measured after the gearbox, by a sensor in the joint	Measured output
Dress pack	Cables and hoses strapped along the arm	Distributed component
MJCF, SDF, USD; Menagerie	Model description formats of MuJoCo, Gazebo and Isaac Sim; MuJoCo's library of robot models	Model description
Residual r	Measured joint torque minus the nominal model's torque	Equation error of the nominal model
Library A , group	Columns of the torque's dependence on the candidate corrections, grouped by body or joint	Sensitivity of the output to the parameters, grouped by component
Null space $\mathcal{N}$	Corrections that change no torque in any state	Directions along which the calibration-equivalence class extends
Edit E	Removing, exchanging or adding a body or joint property	Intended modification of the model
Certified	Guaranteed by fidelity (Definition 1)	Same
Base parameters	The combinations of inertial parameters that the torques determine	Estimable functions
Pseudo-inertia	Matrix whose positive semidefiniteness makes a body physically valid	Physical admissibility constraint

Two anchors fix the scale of both axes: the nominal twin, which carries none of the discrepancy, and the twin with the true parameters of R , whose remaining error is model-form discrepancy that no column of A represents, from integration, friction regularization near zero velocity, contact solvers and simulator-specific representations (Supplementary S3). Attribution can be scored only when S^* is known, which is why the validation of Section 5 starts from manufactured discrepancies.

3.2. Structural null space

Let $\mathcal{N} = \{\xi : A(q, \dot{q}, \ddot{q})\xi = 0 \text{ for every state}\}$ be the structural null space of the library.

Proposition 1 (Structural equivalence). *If two corrections differ by an element of $\mathcal{N}$, the two twins have the same inverse dynamics at every state and, under the same actuation and numerical integration, the same free-space forward dynamics.*

Inverse dynamics is affine in ξ and in the acceleration, $ID = M(q)\ddot{q} + h(q, \dot{q})$, so equality for every $\ddot{q}$ gives equal M and h ; the mass matrix of one twin is nonsingular exactly when that of the other is, and the forward dynamics $M^{-1}(\tau - h)$ follows. Twins whose corrections differ by an element of $\mathcal{N}$ are called *equivalent*. $\mathcal{N}$ is computed numerically, as for base parameters [32]: the library is evaluated on 400 states drawn uniformly inside the joint ranges, its columns are standardized, and the singular vectors with vanishing singular values span $\mathcal{N}$ (the largest torque any of them produces is 10^{-13} of the column scale). For the FR3 and the iiwa, $\mathcal{N}$ has dimension 29 of 112: 27 inertial directions, the difference

between 70 standard and 43 base parameters for seven revolute joints with gravity along the first axis, the second axis perpendicular to the first and no two consecutive axes parallel [16, 17], and two directions that exchange the rotor inertia of joints 1 and 2 with link inertia about the same axes [30]. The dimension depends on the axis geometry and the direction of gravity, so $\mathcal{N}$ is recomputed for every installation. The identification trajectories of this study reach the structural rank, so two corrections that produce the same torques on the identification rows differ by an element of $\mathcal{N}$ and no behavioral measurement in free space can separate them. Data that fall short of the structural rank have a larger null space $\mathcal{N}_D \supseteq \mathcal{N}$, the class becomes $\hat{\xi} + \mathcal{N}_D$, and the test of Proposition 2 becomes $L\mathcal{N}_D \subseteq \mathcal{N}$, so the rank of the excitation is checked before any edit is certified.

3.3. Edits and their certification

A twin is edited when the robot is: a tool is removed or exchanged, a link or a gearbox replaced, a payload of known inertia added. In parameter space an edit is an affine map of the correction,

$$E(\xi) = L\xi + c, \quad (2)$$

with L linear and c known. Subtracting a tool of known inertia π_t from link 7 is $L = I$, $c = -P_7^\top \pi_t$; removing what the twin attributed to link 7 is $L = I - P_7$, $c = 0$, with P_g the coordinate projector onto group g ; replacing link b by a part of known parameters π_{new} is $L = I - P_b$, $c = P_b^\top (\pi_{\text{new}} - \pi_b)$.

Proposition 2 (Edit certification). *The edit (2) is certified if and only if $L\mathcal{N} \subseteq \mathcal{N}$.*

The edited corrections of ξ and ξ' differ by $L(\xi - \xi')$, which lies in $\mathcal{N}$ for every $\xi - \xi' \in \mathcal{N}$ exactly when $L\mathcal{N} \subseteq \mathcal{N}$; taking $\xi' = \xi + n$ for an n with $Ln \notin \mathcal{N}$ shows the converse.

Remark 1 (Lineage). Proposition 2 is the estimability condition of linear models in operator form: a linear function $w^\top \xi$ is estimable when w is orthogonal to the null space of the design [22], and the edited twin's torque at every state is a family of such functions. In robotics it is base-parameter regrouping [16, 17]: $L\mathcal{N} \subseteq \mathcal{N}$ says that L induces a map on the quotient $\mathbb{R}^{112}/\mathcal{N}$, that is, on the base parameters, as known-payload identification and its regrouped estimates presuppose [38, 18]. What this paper adds is operational: the test for executable twins with $\mathcal{N}$ read off the simulator's own regressor, its consequence for revolute chains (Corollary 2), the repair of an executable edit (Corollary 1), and the report, the leakage bound and the validation ladder of the following sections.

Remark 2 (Beyond linear corrections). Proposition 2 needs the correction to enter the model linearly and the state to be observed, so that the residual is an equation error and the class $\mathcal{C}$ is the affine set $\hat{\xi} + \mathcal{N}$. For a model whose output depends nonlinearly on its parameters, let $\mathcal{N}_D(\theta)$ be the null space of the output sensitivity on the calibration data at θ (the Jacobian of those outputs with respect to θ), the tangent of the class at $\hat{\theta}$, and $\mathcal{N}_U(\theta)$ that of the outputs on the conditions of use. The modified output is then constant to first order along the class if and only if $L\mathcal{N}_D(\hat{\theta}) \subseteq \mathcal{N}_U(E(\hat{\theta}))$, with L the Jacobian of E at $\hat{\theta}$: the sensitivity is taken at the edited point, which in a nonlinear model differs from the calibrated one. For $y = \theta_1\theta_2$ at $\hat{\theta} = (1, 1)$ and the known offset $E(\theta) = (\theta_1, \theta_2 + 1)$, $L = I$ but the edited output $1 + t$ changes along the class $(t, 1/t)$, as the condition predicts. Globally the class is a manifold that can curve away from its tangent space, so a modification certified to first order can still be decided by the prior far from $\hat{\theta}$; Definition 1 remains the test.

The guarantee of a certified edit needs two hypotheses beyond Proposition 2. First, the physical change is the edit applied to the reference's own correction ξ^* : the removed tool is the one whose inertia c describes, and nothing else changes with it. Then, when $\hat{\xi} - \xi^* \in \mathcal{N}$, $E(\hat{\xi}) - E(\xi^*) \in \mathcal{N}$, and the edited twin predicts the edited reference exactly where the unedited twin predicted the unedited one; the estimation error of the estimable part, $\hat{\xi} - \xi^*$ modulo $\mathcal{N}$, passes through L unchanged in size for a known-inertia edit. Second, the executable twin must carry a physically valid correction after the edit, which the corrected one need not.

Corollary 1 (Certified edits are repairable). *If the edit is certified, the reference correction ξ^* and its edit $E(\xi^*)$ are physically valid, and $\hat{\xi} - \xi^* \in \mathcal{N}$, then some $n \in \mathcal{N}$ makes both $\hat{\xi} + n$ and $E(\hat{\xi} + n)$ physically valid.*

Take $n = \xi^* - \hat{\xi} \in \mathcal{N}$; then $\hat{\xi} + n = \xi^*$ and $E(\hat{\xi} + n) = E(\xi^*)$ are valid, and $E(\hat{\xi} + n) - E(\hat{\xi}) = Ln \in \mathcal{N}$ leaves every torque unchanged. The *null-space repair* finds such an n without knowing ξ^* : the smallest n for which the edited and

the unedited corrections satisfy the pseudo-inertia inequalities of every link [8] and non-negative friction and armature totals, a semidefinite program; when no n reaches the floors, the shift that minimizes the total shortfall. It changes no torque before or after the edit and is used for every known-inertia edit of Section 5.

Two kinds of edit are certified for every twin. *Known-inertia edits*: with $L = I$ the condition holds for every $\mathcal{N}$, so subtracting a known tool, adding a known payload or exchanging one known tool for another transfers the fidelity of the unedited twin to the edited one, whichever correction the twin carries. *Joint-friction edits*: a friction law that depends only on the joint's own velocity produces torque only at that joint and as a function of that velocity alone, which no inertial, rotor-inertia or other friction column reproduces, so $\mathcal{N}$ has no component in it and clearing or replacing the friction of a joint is certified. Load- or temperature-dependent friction couples a joint's friction with the load and falls outside this statement. Edits that rely on where the twin placed inertia behave differently.

Corollary 2 (Clearing link groups on revolute chains). *Let joint $j \geq 2$ of a serial chain be revolute, and let its axis contain a point whose position depends on the joints $1, \dots, j-1$ in some motion. Clearing a set B of link groups that contains exactly one of the links $j-1$ and j is not certified.*

Let p be such a point and $m > 0$. A point mass m at p moves identically whether it is attached to link j or to link $j-1$, because rotation about the axis of joint j leaves p fixed relative to link $j-1$, so the difference n_j of the two attachments, a mass m on one link and $-m$ on the other at the same point, lies in $\mathcal{N}$. Clearing B maps n_j to a single point mass $\pm m$ at p on the one link outside B . That point moves in some motion of the joints $1, \dots, j-1$, so it produces a torque, $(I - P_B)n_j \notin \mathcal{N}$, and Proposition 2 fails. On an all-revolute arm such as the FR3 or the iiwa the hypothesis holds for every $j \geq 2$, so clearing link groups is certified only for all links at once, and with them the armature, which shares the rotor-inertia directions of joints 1 and 2. Computed group by group on both robots (Supplementary S8), clearing any single link group or the armature group is not certified, clearing all links with the armature is, and clearing the friction of any joint is. On a chain of prismatic joints the obstruction does not arise, because no point of a translating link is fixed relative to its parent; the masses of a Cartesian gantry are separately determined by the axis forces. Removing a tool by clearing what the twin attributed to link 7 is therefore not certified on a revolute arm, nor is replacing link 3 by clearing its correction: fidelity alone guarantees friction edits, known-inertia edits and the clearing of the whole arm, and no edit that relies on how the twin distributed inertia among the links.

3.4. Leakage interval

For an uncertified edit, the edited behavior depends on which of the equivalent corrections the calibration returned. The data cannot narrow this choice, but physics bounds it: of the corrections equivalent to a fitted $\hat{\xi}$, only those whose links have positive semidefinite pseudo-inertia [8] and whose joint friction and armature totals are non-negative describe a robot. For a linear quantity φ that the edit acts on, the *leakage interval* is

$$\Lambda_{\varphi}(\hat{\xi}) = \{\varphi(\hat{\xi} + n) : n \in \mathcal{N}, J(\pi_b + (\hat{\xi} + n)_b) \geq 0 \text{ for every link } b, \text{ friction and armature totals } \geq 0\}, \quad (3)$$

an interval whose ends are two semidefinite programs. It uses no ground truth, and it adds to Proposition 2 a size: how far physically valid twins that reproduce every motion can disagree on the quantity an uncertified edit changes. For a certified edit the edited behavior is the same everywhere in the interval. With positive semidefiniteness as the only constraint the interval is wide for link-attribution edits on revolute arms: for the removal of a 0.8 kg tool it spans a median of 2.7 kg on the FR3 and 4.9 kg on the iiwa over the 20 replication trajectory sets of Section 5 (Table 5), nearly the same around correct and wrong attributions. It is therefore a screening statistic computed before an edit, which says that the attribution, not the data, decides that edit, and not a confidence interval that becomes narrow with more data; tighter physical bounds, such as CAD tolerances on each link, would narrow it.

3.5. A two-link example

Figure 2 reduces the argument to two links, and later sections return to it. A 0.5 kg clamp sits on the elbow axis; attached to the upper or to the lower link it produces the same joint torques in every state, so every twin on the line $m_1 + m_2 = 0.5$ is equivalent; this line is the one-dimensional slice of $\mathcal{N}$ that the clamp spans, the direction n_j of Corollary 2 with $j = 2$. Subtracting the clamp's known mass from the lower link translates that line and keeps the twins equivalent. Clearing the lower link's increment projects the line onto the m_1 axis and splits it: the twin that placed the clamp on the lower link becomes exact, the minimum-norm twin keeps half of it, and the twin that placed it on the upper link keeps all of it.

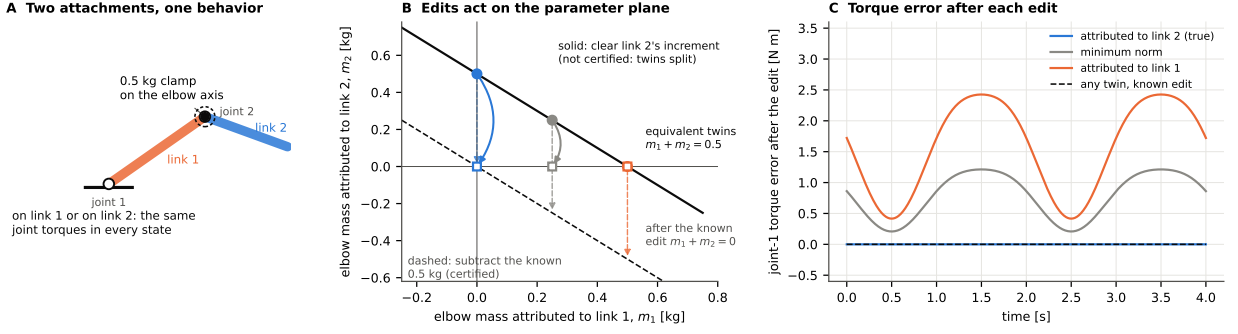


Figure 2: Edit certification on two links, a one-dimensional slice of the null space. (A) A 0.5 kg clamp on the elbow axis produces the same joint torques attached to either link. (B) Equivalent twins lie on the line $m_1 + m_2 = 0.5$; subtracting the known 0.5 kg (dashed) keeps them equivalent, clearing the lower link's increment (solid) does not. (C) Joint-1 torque error of the edited twins against the reference without the clamp: zero for every twin after the known edit, and zero, half or all of the clamp's torque after clearing, depending on the attribution.

3.6. Consequence: an attribution needs a stated prior

An uncertified edit cannot be validated by fidelity, because every twin in the equivalence class has the same fidelity and the class contains twins that the edit sends anywhere in the leakage interval. Any calibration returns one member of the class, and the choice is a prior: a minimum-norm fit chooses the member closest to the nominal twin in its metric, a physically consistent fit one that satisfies the consistency inequalities, a log-det fit the one closest to the nominal bodies [34], joint-torque payload identification the one that places every change on the last link, and a sparse fit the one supported on the fewest groups. When the location of a change is known, the location prior is the right one and needs no selection; when it is not, as for changes inside the arm, a dress pack, wear or an unannounced modification, the prior has to be chosen and its influence reported. Section 4 states the prior used here, sparse change over body and joint groups, and reports where it was decisive; Section 5 tests it where it holds and where it does not.

4. Structure-aware calibration with a report

SAPC states the prior that Section 3.6 shows to be unavoidable when the location of a change is unknown, sparse change over body and joint groups, and reports where that prior decided the answer. It reads the library off the simulator (Section 4.1), bounds every fit and writes the result back into native fields (Section 4.2), selects groups by blocked cross-validation (Section 4.3) and returns the selected twin together with a report (Section 4.4); Fig. 3 summarizes the pipeline.

4.1. Simulator-native library

For a rigid body b with standard parameters $\pi_b = (m, mc, \mathbf{I}_o)$, inverse dynamics is affine, $\text{ID}(\pi_b + \delta) = \text{ID}(\pi_b) + Y_b(q, \dot{q}, \ddot{q}) \delta$ [30]. Classical identification derives Y_b symbolically from the kinematic chain [30, 7]; SAPC reads it off the simulator, so the library is by construction the model the twin executes.

Remark (π -basis extraction). With ten physically valid bodies $\pi_1, \dots, \pi_{10}$ whose parameter matrix P is invertible, $Y_b = [\text{ID}(\pi_b + \pi_1) - \text{ID}(\pi_b), \dots, \text{ID}(\pi_b + \pi_{10}) - \text{ID}(\pi_b)] P^{-1}$ at every state. The sum of two valid bodies is valid, so the simulator accepts every call, and a link's regressor costs eleven inverse-dynamics calls of any simulator that exposes inverse dynamics and writable inertial fields. The extracted regressor reproduces the torque change of the injected payload and link-3 mass to 1.4×10^{-5} N m over a 9.6 N m range; a payload rigidly attached to the flange is a change of link 7's ten parameters.

Each joint adds five friction columns acting on its own equation, Coulomb $\tanh(\dot{q}_j/0.02)$, viscous $\dot{q}_j$, Stribeck $\tanh(\dot{q}_j/0.02) \exp(-(\dot{q}_j/0.1)^2)$, quadratic $\dot{q}_j |\dot{q}_j|$ and a constant offset, and one armature column $\dot{q}_j$. The Stribeck column is weakly excited above the velocity mask below and is kept for low-speed segments; the seven armature columns form one group with a diagonal rotor model, because a reference description without rotor inertia, such as SDF, changes all of them at once. The simulator's Coulomb friction is a constraint whose force is linear in velocity

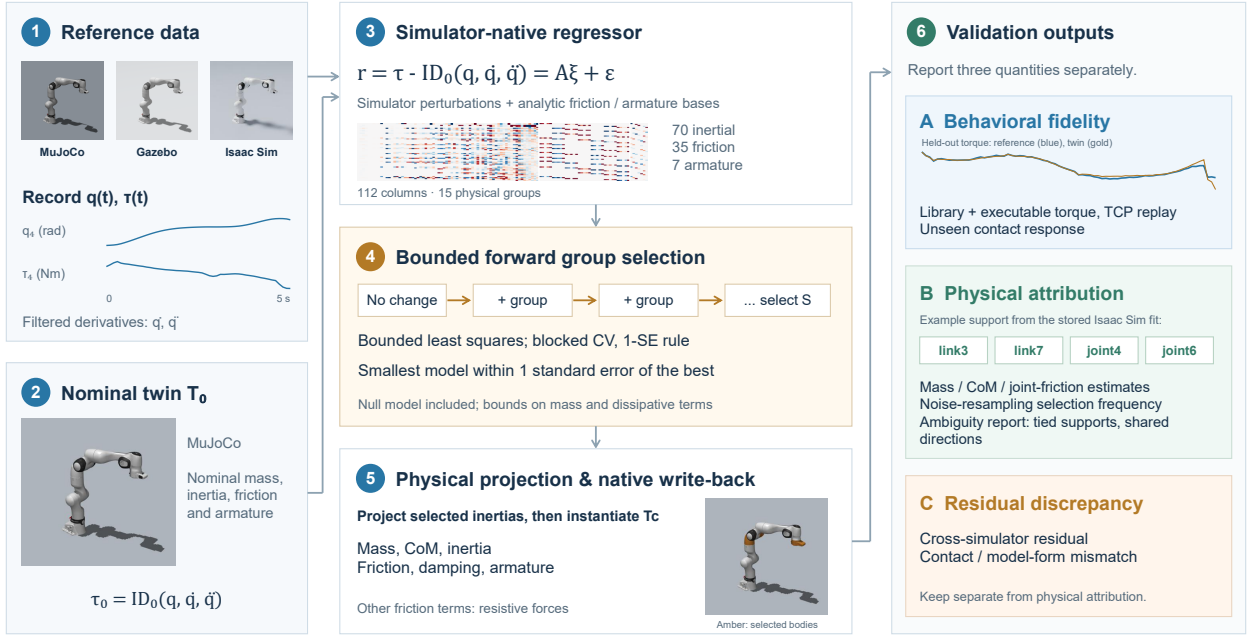


Figure 3: Overview of structure-aware physical calibration (SAPC). (1) Joint positions and torques are recorded on the reference, and velocities and accelerations are obtained by filtered differentiation. (2) The nominal twin T_0 gives $\tau_0 = ID_0(q, \dot{q}, \ddot{q})$. (3) The residual $r = \tau - ID_0$ is expressed in a simulator-native library of 70 inertial, 35 friction and 7 armature columns in 15 physical groups. (4) Bounded forward selection from the uncorrected twin, a backward pass and one-for-one exchanges choose the smallest support within one standard error of the best blocked cross-validation score. (5) The selected corrections are written into the native fields of the executable twin T_c . (6) Behavioral fidelity, physical attribution and the report, with the supports of the selected size that the data cannot reject and the edit flag, are returned separately.

below a knee of about 0.03 rad/s, so the equation of joint j is dropped wherever $|\dot{q}_j| < 0.1$ rad/s, as in identification practice [7].

4.2. Bounds and executable write-back

Every candidate fit is bounded least squares on standardized columns with a ridge of 10^{-4} : the total Coulomb, viscous, Stribeck and quadratic coefficients and the total armature of each joint stay non-negative, and no body loses more than half its nominal mass. Without these bounds a sparse fit on a short record can return non-passive friction, and the twin built from it diverges. A selected body whose pseudo-inertia

$$J(\pi) = \begin{bmatrix} \frac{1}{2} \text{tr}(\mathbf{I}_o) \mathbf{1} - \mathbf{I}_o & m\mathbf{c} \\ m\mathbf{c}^\top & m \end{bmatrix} \quad (4)$$

is not positive semidefinite [8] is replaced by the nearest consistent body, and the projection distance is reported; projection follows selection, because projecting an aliased dense solution changes the torque function it was fitted to.

The identified increments become a simulator file, not a residual function. Inertial increments are converted to the native mass, center of mass, principal inertia and inertia frame of each body; Coulomb and viscous increments are added to the joint's friction-loss and damping fields and armature increments to its rotor inertia. Stribeck, quadratic and offset terms have no native field in MuJoCo and act as resistive generalized forces that depend only on the current velocity. Inertial changes are never applied as external forces, because an inertial residual computed from a lagged acceleration closes an algebraic loop through the mass matrix. The executable twin inherits the simulator's friction law, contact solver and integrator, and its correction can be read, edited and transferred in the simulator's own parameter space, unlike a learned residual [47, 46]. Known-inertia edits are written back after the null-space repair of Corollary 1.

4.3. Group selection

A group is the unit at which a physical change occurs, a tool, a worn joint, a replaced link, and the unit at which a twin drifts from its counterpart in service [14]. SAPC scores a support S by blocked cross-validation [50]: each training trajectory is cut into four contiguous blocks, each block is left out in turn, the columns of S are fitted on the remaining blocks, and the score is the mean held-out torque RMSE over the twelve folds of three trajectories. Contiguous blocks keep the filtered, serially correlated rows of one segment out of its own training set; neighboring blocks still share the correlation at their boundaries, so the fold scores are not independent and the standard error below is a working measure of spread.

The search starts from the uncorrected twin, whose score is the RMS of the residual itself. Forward steps add the group that lowers the score most until no addition lowers it or the support reaches eight groups; among the models on the path the smallest within one standard error of the best is chosen, the one-standard-error rule [51] with the standard error of the fold scores of the best model. A backward pass then removes a group whenever the score without it stays within one standard error of the best score seen, because a forward path cannot undo an early choice, such as a neighboring group that mimics most of a change before the group that carries it enters. Finally, one-for-one exchanges of a selected for an unselected group are taken while they lower the score, so the result does not depend on the order of the path; the representative is the best-scoring support of the selected size. The cap of eight groups bounds the search on a nominal model far from the robot; it was reached in 5 of the 1460 runs of the first confirmation, all random iiwa patterns, and it is lifted for the first stage of the physical session, whose nominal model differs from the robot at every joint. The rule and its constants were chosen on development data with known ground truth (Supplementary S5) and frozen by hash before the confirmation runs; the group lasso [52] is compared in Section 5.2.

4.4. Report

The selected twin comes with three statements of what the data did not decide. *Parsimony band*: every support of the selected size that the search scored, forward, backward and in the exchanges, whose score lies within one standard error of the best, with the paired fold-score gap to the selected support and its standard error, and with the physical quantities each attributes. The band is not an enumeration of all supports of that size, and its width reflects the heterogeneity of the folds as well as the difference between supports; the paired gap separates the two. *Edit flag*: for the edit the user intends, the edited quantity, such as the link-7 mass that a tool removal clears, is compared across the supports of the band whose paired gap is at most two paired standard errors, the supports that a paired comparison of the folds cannot reject; the edit is flagged when one of them differs from the selected value by more than a declared tolerance, here $\max(0.02 \text{ kg}, 20 \%)$. *Shared directions*: for a support S , $o(S) = \sum_{g \in S} \text{rank } A_g - \text{rank } A_S$ counts the directions its groups share; if $o(S) > 0$, the split of the correction along them is set by the ridge, not by the data, and the groups that share them are reported. Within a support whose restricted regressor has full column rank the fit is unique, so these statements cover what the data leave open inside and between the scored supports. The report is *non-empty* when the band holds a support other than the selected one or the selected support has shared directions, and *silent* otherwise; comparing the band's supports on where they place mass, rather than on the edited quantity, gives the *location flag* used for the physical session (Section 5.7). The leakage interval of Section 3.4 complements them before an edit: it does not depend on the selection and says how far the edited quantity can move along $\mathcal{N}$.

A flagged edit should use a known inertia, which is certified (Section 3.3), or be preceded by an excitation that separates the supports that disagree. An unflagged edit is decided by the data under the stated prior among the supports the search scored; a support larger than the selected one lies outside the band, and Section 5.3 shows a case in which it carries the truth. SAPC returns the best-scoring support in every case, because a twin must be produced, and the report doubles as a diagnosis of where the robot changed.

The prior and its alternatives. Every calibration chooses a representative of the equivalence class of Section 3.2. Dense least squares chooses the one of minimum norm in its regularizer, a physically consistent fit [33, 8] one inside the consistency cone, a log-det regularized fit [34] the one closest to the nominal bodies, joint-torque payload identification [38] assumes that only link 7 changed, and Bayesian model selection over supports [53] places a prior on sparse supports and reports a posterior probability. Section 5 compares all of them on the same rows, folds and groups.

5. Experimental validation

The validation follows the ladder of computational science [27, 28]. Manufactured discrepancies, injected into a simulator, give the exact truth for every attribution and every edit; another simulator as the reference adds model-form

Table 2

Validation ladder: what each stage can and cannot test. Manufactured discrepancies give exact truth for attribution and edits; other simulators add model-form discrepancy with a known injected change; the physical robot adds real sensors and friction with a weighed block at a location withheld from the analyst.

Stage	Reference and truth	Scored	Nuisance	Not tested
Main change	MuJoCo FR3 and iiwa with injected tool, link mass and friction; exact	support, magnitudes, edits, leakage interval, replication (20 trajectory sets $\times$ 10 noise draws)	sensor noise $\times 1$, $\times 3$	model-form discrepancy, real sensors
Detection power	tool mass 0.02-1.2 kg $\times$ offset 0-50 mm; friction ± 10 -150 %; no change; exact	sensitivity, false alarms	sensor noise $\times 1$, $\times 3$	the same
Random benchmark	120 sparse patterns per robot, 1-4 groups, pre-registered distribution; exact for the pattern	support, tool mass, edits, report	background mass error ± 2 %, friction drift, colored noise, offsets	model-form discrepancy
Prior violated	dress pack over links 3-6; on-axis tool; background error 5 and 10 %	what is returned and whether the report flags it	sensor noise $\times 1$	the same
Tool benchmark	150 patterns per robot, a tool and 0-3 other changes; single stage and unloaded first	tool attribution, removal edit, edit flag of every method	as the random benchmark	the same
Gazebo	another simulator, same model description, injected change; injected part exact	attribution of the injected change, terms absent from the description	friction law, integrator, rotor inertia	real sensors
Isaac Sim	another simulator; injected change and a no-change control	attribution, false alarms	contact and solver differences	real sensors
Physical FR3	weighed blocks at five placements and none, withheld from the analyst; blocks exact, robot unknown	location, mass, report on an on-axis block, false alarms, edits by attribution and by known inertia	sensors, friction, cabling, offsets	changes other than the blocks

discrepancy that no column of the library represents; a physical robot adds real sensors, friction and cabling (Table 2). Two manipulators are studied in the first stage, the Franka FR3 and the KUKA LBR iiwa 14, and the method code was frozen before any run whose data had not been seen during its design.

5.1. Design

Robots, references and injected changes Reference and twin are the MuJoCo Menagerie models of the two robots [54, 55], driven by joint torques with contacts disabled in free space. The reference carries the changes of Table 3: a 0.8 kg tool at the flange with an eccentric center of mass, Coulomb and viscous friction increments on joint 4 and a Coulomb increment on joint 6, and a 12 % increase of the mass and inertia of link 3. A tool change and a friction drift are the archetypal changes a twin must follow over the life of the system [14]; the friction increments are large, and Section 5.4 shows that increments of 10 % are detected as well. The reference integrates with MuJoCo's implicit integrator at 0.5 ms and the twin with its implicit-fast variant, which drops the velocity derivatives of the Coriolis forces, at 1 ms; the twin with the true parameters replays the FR3 reference to 0.027 mm, against 14.43 mm for the nominal twin. The iiwa is the MuJoCo Menagerie iiwa 14 rigid-body model, used unmodified: it carries no joint friction or rotor inertia, so its joint changes are the FR3 increments added to zero. Its library has the same friction and armature columns as the FR3's, which do not depend on the nominal values, so its null space has the same 29 directions; realistic friction and rotor inertia would change the nominal twin, not the structure the edit test reads. Its spherical wrist, whose joint 5-7 axes meet in one point, is where the two robots differ for attribution (Fig. 4).

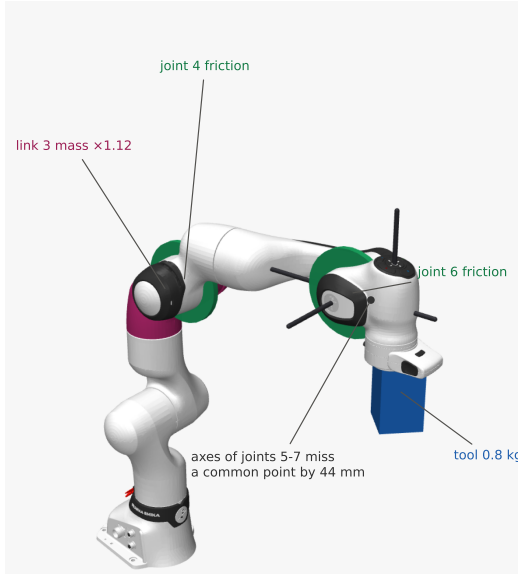
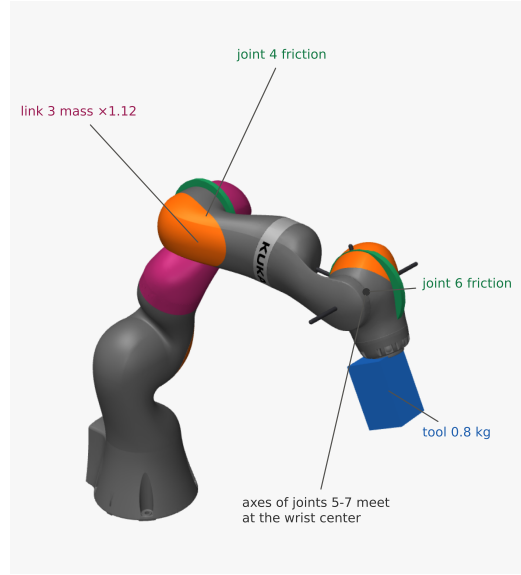
A Franka FR3**B KUKA LBR iiwa 14**

Figure 4: The two robots of the manufactured-discrepancy stage, MuJoCo Menagerie models at their home poses. Marked are the injected changes, the tool at the flange (a box with the tool's mass and principal moments), the link whose mass and inertia increase and the two joints whose friction increases, and the axes of joints 5-7, which meet at the wrist center on the iiwa and miss a common point by 44 mm on the FR3.

Table 3

Injected changes of the main study, with the nominal values of the two robots. The iiwa model has no nominal joint friction, so its friction increments are the FR3 values added to zero.

Change	FR3 (nominal → reference)	iiwa (nominal → reference)	Library group
Tool at the flange	0.8 kg at (0.02, -0.015, 0.06) m; 1.28 times the link-7 mass	the same; 0.67 times the link-7 mass	link 7
Joint-4 Coulomb friction	1.14 → 2.80 N m (+146 %)	0 → 2.80 N m	joint 4
Joint-4 viscous friction	0.21 → 0.55 N m s (+162 %)	0 → 0.55 N m s	joint 4
Joint-6 Coulomb friction	0.44 → 1.10 N m (+150 %)	0 → 1.10 N m	joint 6
Link-3 mass and inertia	$\times 1.12$ (2.25 → 2.51 kg)	$\times 1.12$ (3.50 → 3.92 kg)	link 3
Library	112 columns, 70 inertial, 35 friction and 7 rotor-inertia, in 15 groups		

Excitation, noise and derivatives Excitation trajectories are five-harmonic Fourier series with a 0.1 Hz fundamental [31, 30], scaled to 35 % of each joint range and capped at 1.6 rad/s and 8 rad/s²; each lasts 20 s, three identify and two held-out ones score every behavioral metric. A nominal-feedforward PD controller tracks them, so the residual torque is the controller's own effort to compensate the change, and positions and torques are logged at 1 kHz. The sensor model adds encoder quantization of 10^{-4} rad, jitter of 10^{-5} rad and torque noise of 0.05 N m, all scaled by a common factor of 1 (nominal) or 3 (stress). Velocities and accelerations come from a Savitzky-Golay filter of 0.1 s and order 3, the rows are decimated to 200 Hz without filtering the torques, and joint equations below 0.1 rad/s are dropped (Section 4.1). On these trajectories the standardized base regressor has a condition number of 109 on the FR3 and 80 on the iiwa.

Methods compared The calibrations run on the same rows, folds and groups. *Joint-torque payload identification* [38] fits the ten link-7 parameters (link 7 only), and in the tool benchmark also each joint's Coulomb and viscous friction (link 7 + friction), the classical comparator for a tool. *Dense least squares* fits all 112 columns with a ridge of 10^{-6} in standardized units, the minimum-norm correction; *dense LMI* uses SAPC's bounded fit on all groups with the pseudo-inertia of every link constrained positive definite inside the fit [33, 8]; *dense log-det* adds the log-det divergence to the nominal bodies [34], with the largest weight whose blocked cross-validation score lies within one standard error of the best. *Group lasso* [52] scores every support on its regularization path by SAPC's refit and cross-validation.

Bayesian selection enumerates every support of at most five groups with a closed-form evidence under an i.i.d. Gaussian likelihood whose noise variance is taken from the dense fit, a log-uniform prior scale and a beta-binomial prior over support size [53], without the bounds of the other fits. It is run in three variants: with the noise variance inflated by the residual autocorrelation time (Bayes), without inflation (Bayes iid), and as a generalized posterior whose learning rate is chosen, in the spirit of the SafeBayes repair for misspecified models [56] but by blocked cross-validation of its posterior-expected prediction error instead of SafeBayes's sequential criterion (Bayes η -CV). Its flag compares the 95 % credible interval of the edited quantity, averaged over supports, with the same tolerance as SAPC's. An exhaustive search over all supports of at most four groups verifies SAPC's search on the reference data set.

Protocol The reference data set, training seeds 1-3 and noise draw 17, is the one on which the selection rule was designed (Supplementary S5), and its numbers are labeled as such. The first confirmation fixed the method code by hash and 1460 runs in a protocol before they were made: 20 training sets of three new trajectories times 10 noise draws per robot and 240 random change patterns on new trajectories, 640 runs that test the method out of sample, and 740 detection-power runs and 80 runs with a violated prior that reuse the design trajectories with new changes and noise draws. Its four primary endpoints are exact support recovery, the error of the edit made through the attribution, the support error with and without a non-empty report, and the link-7 mass error; an amendment made before any *iiwa* run set the joint changes of the frictionless *iiwa* model as increments (Supplementary S5). After the first round of review, a second protocol fixed the edit flag of Section 4.4, the Bayesian variants and a tool benchmark of 300 new patterns on new trajectories before they were run, with the harness and analysis scripts frozen by hash; the tie rule of the flag was chosen on a 40-pattern pilot outside that benchmark. The flag statistics of the first confirmation's suites are recomputed on the same data and are exploratory.

5.2. Behavior versus attribution

Every calibration that uses the full library reproduces the held-out motion (Table 4, Fig. 5). On the FR3 reference data set the library torque NRMSE is 0.041-0.042 for all of them and the replay error 0.024-0.046 mm, against 0.027 mm for the true parameters and 14.43 mm for the nominal twin; link 7 only, which ignores the other changes, replays at 2.65 mm. A validation that stops at output agreement [9, 3] accepts all of them.

The attribution columns separate them. Dense least squares, the dense LMI fit and the dense log-det fit place 0.130-0.180 kg of the 0.8 kg tool on link 7, with centers of mass 0.12-0.37 m from the true one, and group lasso places 0.138 kg; the rest of the tool is spread over links 3-6. Neither physical consistency nor a prior toward the nominal bodies helps: over the whole grid of log-det weights the link-7 increment stays between -0.31 and 0.49 kg, and the weight that cross-validation selects keeps 0.180 kg. SAPC recovers 0.794 kg with a center-of-mass error of 1.3 mm and the four injected groups, as does the exhaustive search; the inflated Bayesian selection returns the same support with 0.800 kg. Over the 200 replications SAPC selects exactly the injected groups in 187 and misattributes the tool in 9, all in two trajectory sets; the dense LMI, log-det and group-lasso fits never select the injected groups and misattribute the tool in every run; the inflated Bayesian selection is exact in every run and the uninflated one in 165.

The *iiwa* repeats the behavioral picture and changes the attribution one. Its spherical wrist lets a heavier link 4 and a lighter tool reproduce the torques of a heavier link 3 and the full tool to within 0.002 N m RMS on these trajectories, a twenty-fifth of the torque noise, so two sparse supports of the same size explain the data almost equally well, and which one wins depends on the excitation and the derivative filter. On the reference data set SAPC selects the injected support with link 4 in its band, and the inflated Bayesian selection returns the other support with 0.350 kg. Over the 200 replications SAPC selects the injected groups in 10, eight of them in one trajectory set, the uninflated Bayesian selection in 15, the inflated one and the dense fits in none; SAPC prefers the support with link 4 in the other 190, a systematic preference of these data rather than a coin toss. The methods differ in what they say about it. SAPC's report is non-empty in every run; the Bayesian selection gives a wrong support a posterior probability of at least 0.95 in 198 of the 200 runs with the inflated variance and in 138 without it, and never has the injected support in its 95 % credible set with the inflation, the overconfidence expected of an i.i.d. likelihood under misspecification [56]. How the edit flag sorts these runs is the subject of Section 5.4.

5.3. Edits

Proposition 2 predicts which edits separate the twins, and Table 6 tests the prediction on the reference data set (Fig. 6). Without an edit the twins agree also beyond the identification data: on faster and larger motions the dense

Table 4

Behavior versus attribution, noise $\times 1$. Reference data set (the development data of the selection rule): library and executable-twin torque NRMSE and replay TCP RMSE on the held-out trajectories, payload mass (true 0.8 kg) and its center-of-mass error. Replications (20 trajectory sets $\times$ 10 noise draws): runs with exactly the injected groups (a primary endpoint of the first protocol), and runs with the tool misattributed (link-7 mass outside max(0.02 kg, 20%) of 0.8 kg; exploratory, defined after the replications). The *iiwa* reference data set is one on which SAPC happens to select the injected support; the replications show how often it does.

Calibration	Reference data set					Replications [of 200]	
	NRMSE lib.	NRMSE twin	Replay [mm]	Payload [kg]	CoM err. [mm]	Injected groups	Tool misattributed
<i>FR3</i>							
Nominal twin	1.000	1.000	14.43	-	-	-	-
Link 7 only	0.326	0.329	2.65	0.861	37	0	0
Dense least squares	0.041	0.025	0.027	0.130	368	-	-
Dense LMI	0.041	0.024	0.030	0.130	369	0	200
Dense log-det	0.041	0.025	0.030	0.180	122	0	200
Group lasso	0.042	0.025	0.046	0.138	340	0	200
Bayesian selection (inflated)	0.041	0.024	0.024	0.800	0.1	200	0
SAPC	0.041	0.024	0.030	0.794	1.3	187	9
True parameters	0.042	0.024	0.027	0.800	0	-	-
<i>iiwa</i>							
Nominal twin	1.000	1.000	16.38	-	-	-	-
Link 7 only	0.423	0.429	6.48	0.947	73	0	30
Dense least squares	0.030	0.019	0.066	0.141	407	-	-
Dense LMI	0.031	0.019	0.053	0.140	407	0	200
Dense log-det	0.030	0.019	0.053	0.181	138	0	198
Group lasso	0.031	0.021	0.119	0.252	408	0	200
Bayesian selection (inflated)	0.030	0.020	0.078	0.350	241	0	200
SAPC	0.031	0.019	0.046	0.797	0.8	10	190

twins replay to 0.022-0.026 mm and SAPC to 0.030 mm, against 0.025 mm for the true parameters and 13.5 mm for the nominal twin.

Edits by attribution are not certified and separate the twins. Removing the tool by clearing each twin's link-7 increment leaves SAPC 0.099 mm from the unloaded reference and every dense variant 11.7-12.5 mm, worse than the uncalibrated twin at 2.60 mm, because the tool mass the dense fits placed on links 3-6 stays behind; replacing the increment by a known 0.5 kg tool gives 0.100 and 11.7-12.5 mm. On the *iiwa* the two supports of the band reproduce the loaded robot alike, and the removal leaves the support with link 3 at 0.046 mm and the one with link 4 at 5.18 mm, against 9.5 mm for the dense twins: the edit outcome depends on the choice that the report exposes.

Edits by known inertia are certified and work for every twin, once the twin is physically valid. Subtracting the known tool returns every twin to the reference in the linear library, as the proposition requires. In the executable twin the dense link 7, which carries 0.13 kg of the 0.8 kg tool, receives a negative mass, and projecting it back to a valid body costs 1.20-1.68 mm. The null-space repair of Corollary 1, which changes no torque, keeps every link valid after the edit, and the known removal then replays to 0.025-0.029 mm for every twin, the level of the true parameters at 0.026 mm; the exchange for a known tool needs no repair and replays to 0.029-0.037 mm.

The leakage interval says before the edit that the attribution decides it. Around every fit of the replications the interval of the link-7 mass increment contains the true 0.8 kg and spans 1.9-3.2 kg on the *FR3* and 2.9-5.0 kg on the *iiwa*, around correct and wrong attributions alike (Table 5). It never certifies a tool removal, as Corollary 2 predicts, and it does not separate good from bad fits; what it gives is the size of the disagreement between valid twins with the same torques, a screening statistic that tells the maintainer before the edit to use a known inertia.

Calibrating the unloaded robot first moves the attribution problem, it does not remove it. When the unloaded robot is calibrated first and the tool fitted alone afterward, every method finds 0.800 kg and the removal leaves 0.021-0.024 mm, as payload identification assumes [38, 42]. The robot's own change keeps its ambiguity: the dense stage one splits the link-3 change between links 3 and 4, so replacing link 3 through its own correction leaves 0.37 mm against 0.71 mm without the edit, whereas SAPC's twin returns to 0.026 mm, the level of the true parameters. The tool benchmark of Section 5.5 tests this route on 300 random patterns.

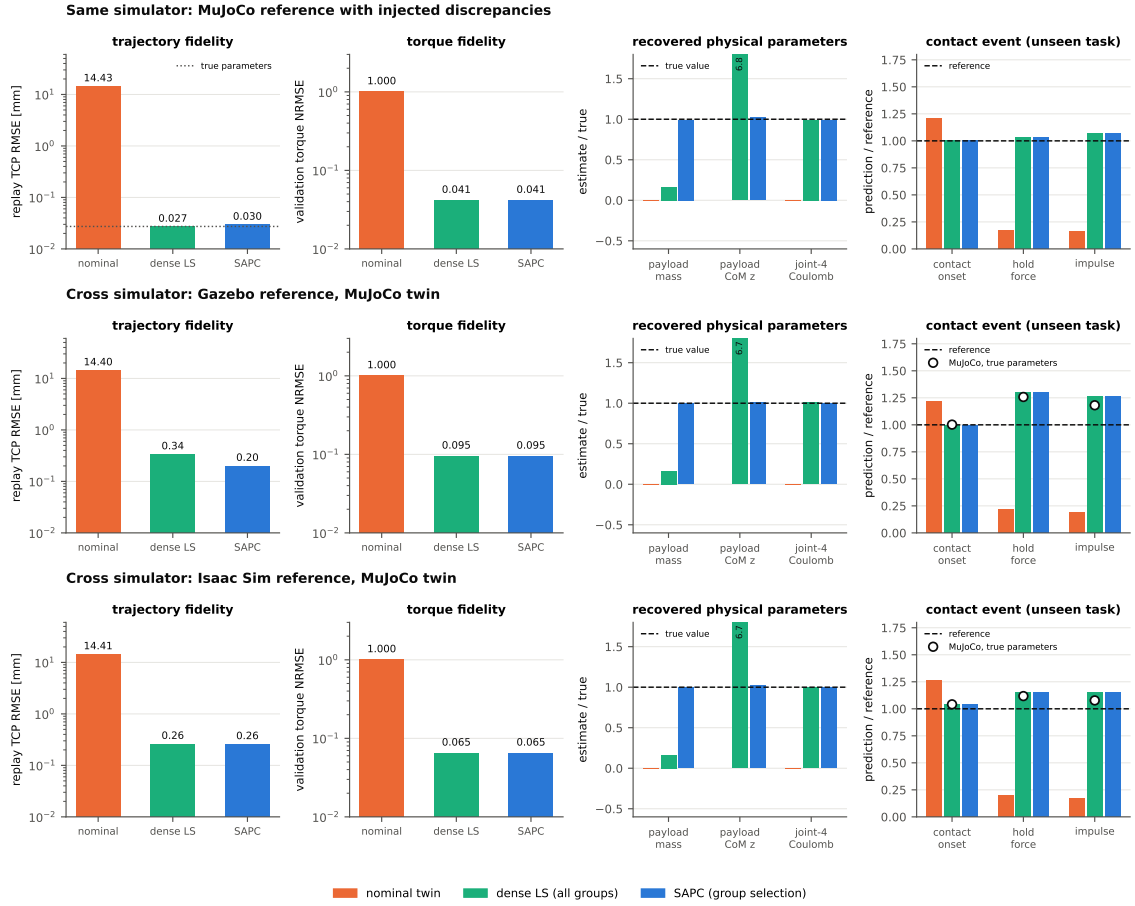


Figure 5: Same behavior, different physics. Columns compare the nominal twin, dense least squares and SAPC on trajectory fidelity, torque fidelity, physical attribution and an unseen press task; rows: MuJoCo, Gazebo and Isaac Sim references, with a MuJoCo twin in every row. The first two panels of each row do not separate dense from structure-aware calibration, the third does, and the fourth shows that both predict the quasi-static press alike.

Table 5

Leakage interval of the link-7 mass increment over the 20 replication trajectory sets (one draw each), noise $\times 1$, true increment 0.8 kg: median fitted increment, median ends of the interval, median width with its 10th-90th percentile range, and the sets whose interval contains the truth.

Robot	Fit	Fitted [kg]	Lower [kg]	Upper [kg]	Width [kg]	Contains truth
FR3	SAPC	0.79	0.03	2.72	2.69 (2.26-3.09)	20/20
FR3	Dense LMI	0.13	-0.15	2.73	2.89 (2.57-3.14)	20/20
iiwa	SAPC	0.34	-1.11	3.80	4.92 (4.72-4.97)	20/20
iiwa	Dense LMI	0.14	-0.45	3.46	3.83 (3.06-4.47)	20/20

Where the prior fails, the edit fails with it. In case C5 of Supplementary S6 the tool sits on the joint-7 axis next to a heavier link 6; a point mass on that axis is a null direction between links 6 and 7, so SAPC and the Bayesian selection merge the tool into link 6, and clearing link 7 leaves 17.5 mm. The report stays silent on the FR3, because the injected support is one group larger than the selected one and lies outside the band; the edit is not certified, and a known-inertia edit recovers it.

Table 6

Edits on the reference data set, noise $\times 1$, replay TCP RMSE against the reference after the same physical change [mm]. By attribution: the twin's own link-7 increment is cleared (removal) or replaced by a known 0.5 kg tool (exchange). By known inertia: the known tool is subtracted, as projected after the edit or with the null-space repair before it. Unloaded first: stage one calibrates the robot without the tool and stage two fits the tool alone; link 3 replaced: the twin's own link-3 increment is cleared and replayed against the reference with a nominal link 3, which every twin misses by 0.71 mm without the edit.

Twin	Link-7 increment [kg]	By attribution		By known inertia		Faster motions
		Removal	Exchange	Projected	Repaired	
<i>FR3, single stage</i>						
Nominal twin	0	2.60	2.61	2.60	2.60	13.49
Dense least squares	0.130	11.73	11.73	1.20	0.025	0.022
Dense LMI	0.130	11.74	11.73	1.20	0.029	0.026
Dense, raw-unit minimum norm	0.090	12.45	12.45	1.68	0.028	0.023
SAPC	0.794	0.099	0.100	0.027	0.028	0.030
True parameters	0.800	0.026	0.027	0.026	0.026	0.025
<i>iiwa, single stage</i>						
Dense LMI	0.140	9.51	9.54	-	-	0.049
SAPC, support with link 3	0.797	0.046	0.048	-	-	0.043
SAPC, support with link 4	0.350	5.18	5.19	-	-	0.075
True parameters	0.800	0.029	0.031	-	-	0.025
	Tool [kg]	Removal	Link 3 replaced			
<i>FR3, unloaded first</i>						
Dense least squares	0.800	0.023	0.37			
Dense LMI	0.800	0.024	0.37			
SAPC	0.800	0.021	0.026			
True parameters	0.800	0.026	0.027			

5.4. Detection, random patterns and violated priors

Detection power and false alarms Fig. 7 varies one change at a time on both robots, with five noise draws per condition, on the design trajectories. A tool 25 mm or more off the flange axis is attributed to link 7 in every run at noise $\times 1$ from 20 g upward, and from 50 g upward at 10 mm; a 10 % increase of the joint-4 friction is found in every run at both noise levels. A tool on the joint-7 axis is a near-null direction between links 6 and 7 and is attributed to link 7 in only some runs, but the report names link 7 in every one. Where nothing changed, SAPC returns the uncorrected twin in all 80 runs of the two robots, whereas the inflated Bayesian selection reports a change in 80 % (FR3) and 100 % (iiwa) of the runs at noise $\times 1$ and the uninflated one in all of them: their likelihood treats the small, consistent residual of the integrator difference between reference and twin as signal.

Primary endpoints of the first confirmation On the FR3 replications SAPC selected exactly the injected groups in 187 of 200 runs (94 %; 82-100 % clustered by trajectory set), with all 13 misses in two of the 20 sets, and its link-7 mass error had a median of 0.007 kg and a 90th percentile of 0.021 kg. Removing the tool through its attribution left a torque NRMSE of 0.10 (90th percentile 0.13), where 1.0 is the uncalibrated twin, against 2.41 for the dense LMI fit. A non-empty report went with a wrong support in 13 of 115 runs and an empty one in none of 85. On the iiwa replications the injected support was selected in 10 of 200 runs, eight of them in one trajectory set, the report was non-empty in every run, the edit NRMSE had a median of 0.94 (90th percentile 0.97) and the link-7 mass error a median of 0.46 kg (0.47 kg). On the random patterns with a tool the edit NRMSE had medians of 0.36 and 0.66 and 90th percentiles of 2.34 and 2.15 on the two robots. The tool misattribution of Tables 4 and 7 is not one of these endpoints; it was defined after the replications and is exploratory, and on the FR3 replications its rate of 4.5 % has a 95 % interval of 0-13 % by cluster bootstrap over the 20 trajectory sets. Table 7 gives every compared method on every suite.

Payload identification and the edit On the tool mass alone, joint-torque payload identification matches or beats SAPC in every suite with a tool (both misattribute 8 of the 18 FR3 random patterns with a tool): its location prior is

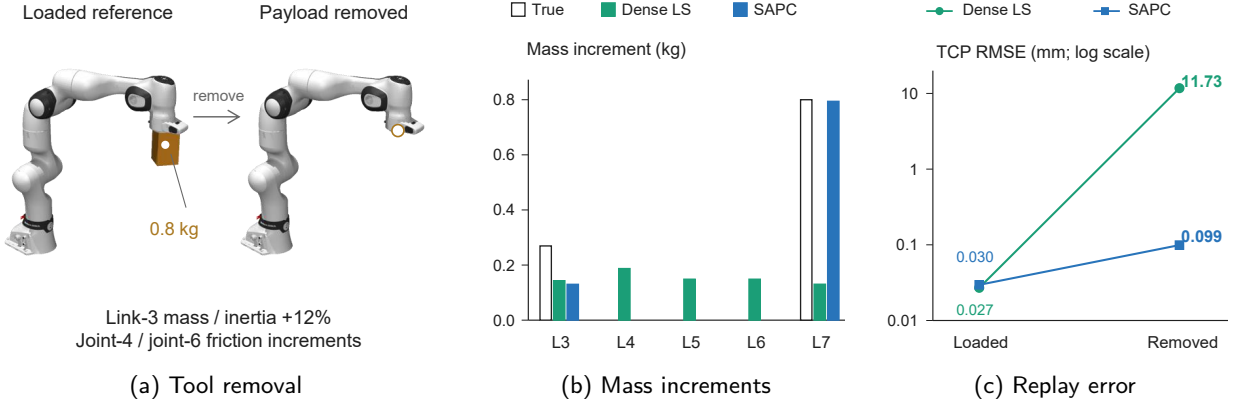


Figure 6: An edit by attribution, FR3, noise $\times 1$, held-out trajectories. (a) The 0.8 kg tool is removed while the link-3, joint-4 and joint-6 changes remain; each twin clears its own link-7 increment without refitting. (b) True and identified mass increments per link; the link-3 mass has a structural null direction along which SAPC keeps the nominal value, so its smaller increment there carries no torque error. (c) Replay TCP RMSE before and after the edit; the uncalibrated twin reaches 2.60 mm after the removal.

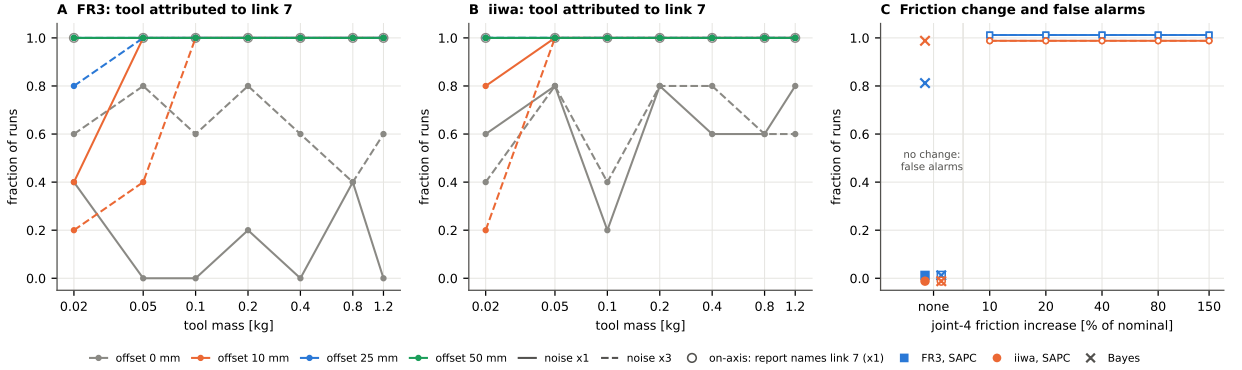


Figure 7: Detection power, five noise draws per condition, design trajectories with new noise. (A, B) Fraction of runs in which SAPC attributes a tool to link 7, against its mass, for four lateral offsets from the joint-7 axis and two noise levels; hollow markers, for the tool on the axis, the fraction whose report names link 7. (C) Fraction of runs in which joint 4 is selected, against its friction increase, and the false-alarm rate where nothing changed, for SAPC and the inflated Bayesian selection.

right for every tool of the first confirmation. It pays for that prior in the edit: removing its link-7 increment returns the uncalibrated twin, an edit NRMSE of 1.00, because it represents none of the other changes, whereas SAPC's removal reaches 0.10 on the FR3 replications and 0.36 and 0.66 on the random patterns. The dense and regularized fits misattribute the tool in every suite; the Bayesian variants never misattribute it on the FR3 replications and are mixed against SAPC elsewhere.

Random patterns and violated priors The random patterns change one to four of the 15 groups by amounts drawn from a pre-registered distribution, 2-50 % for link masses and 10-150 % for friction, and add nuisances that no sparse support explains: a mass and inertia error of up to 2 % on every link, a friction drift of up to 5 % between training trajectories, colored torque noise and per-trajectory torque offsets. With these nuisances no method recovers the injected groups exactly, because the background error is a real change of every link; only 18 of the 120 patterns per robot contain a tool, too few to rank the methods, which the tool benchmark of Section 5.5 does with 150 per robot. A background mass error of 5-10 % on every link defeats the sparse prior relative to the CAD twin: SAPC misattributes the tool in 70-100 % of those runs, the inflated Bayesian selection in 80-100 % and the uninflated one in 20-100 %, whereas joint-torque payload identification, whose premise still holds, misattributes it in 20-60 %; calibrating the unloaded robot

Table 7

First confirmation, every compared method. Rows: robot (runs / runs with a tool). Exact: SAPC's fraction of runs with exactly the injected groups, with its 95 % interval (Wilson; for the replications, clustered by trajectory set). Tool misattributed (exploratory, defined after the replications): link-7 mass outside $\max(0.02 \text{ kg}, 20 \%)$ of its true change, the tool mass plus the link-7 part of the background error, over the runs with a tool, in percent; Bayes η -CV from the exploratory refit of the same data. Edit: median torque NRMSE after removing the tool by attribution, SAPC and dense LMI, where 1.0 is the uncalibrated twin; link 7 only returns the uncalibrated twin, 1.00, in every suite; undefined for a tool alone, whose removal leaves no injected change. Log-det: dense log-det; lasso: group lasso.

Robot	Exact [%]	Tool misattributed [%]								Edit	
		SAPC	Link 7 only	Dense LMI	Log-det	Lasso	Bayes	Bayes iid	Bayes η -CV	SAPC	LMI
<i>Replication of the main change (20 trajectory sets $\times$ 10 draws)</i>											
FR3 (200/200)	94 [82,100]	4	0	100	100	100	0	0	0	0.10	2.41
iiwa (200/200)	5 [0,14]	95	15	100	99	100	100	82	85	0.94	1.46
<i>Random sparse patterns with nuisances</i>											
FR3 (120/18)	4 [2,9]	44	44	100	100	100	61	33	39	0.36	1.86
iiwa (120/18)	0 [0,3]	72	56	100	78	100	89	89	89	0.66	1.28
<i>Dress pack over links 3-6 (no tool)</i>											
FR3 (10/0)	0 [0,28]	-	-	-	-	-	-	-	-	-	-
iiwa (10/0)	0 [0,28]	-	-	-	-	-	-	-	-	-	-
<i>Tool on the joint-7 axis</i>											
FR3 (10/10)	40 [17,69]	60	0	100	100	100	80	70	70	-	-
iiwa (10/10)	90 [60,98]	10	0	100	100	100	10	10	10	-	-
<i>Main change, CAD error 5 %</i>											
FR3 (10/10)	0 [0,28]	100	20	100	100	100	100	20	20	1.76	2.30
iiwa (10/10)	0 [0,28]	100	60	100	100	100	100	100	100	0.95	1.38
<i>Main change, CAD error 10 %</i>											
FR3 (10/10)	0 [0,28]	70	50	100	100	100	80	50	50	1.59	2.16
iiwa (10/10)	10 [2,40]	100	60	100	100	100	100	100	100	1.01	1.33

first restores the sparse prior, because the background error then belongs to stage one. A dress pack of 0.15 kg on each of links 3-6 is reproduced by fewer groups on both robots: on the FR3 by three link groups, which the report flags in every run, and on the iiwa by link 5 alone, which explains it to a torque NRMSE of 0.002-0.003 on these trajectories because the wrist center exchanges mass among links 4-7, so that no report computed from these data can tell it from a single change.

The edit flag on the first confirmation (exploratory) The first protocol scored the report by whether it was non-empty; its link-7 flag was fixed after the replications and is exploratory. That flag fired in 9 of the 200 FR3 replications, exactly the nine misattributions, but in 92 of 120 FR3 random patterns, of which 74 had a correct link-7 mass, so it said little where tools were rare. Refitted on the same data with the flag of the second protocol, SAPC flagged 9 of the 200 FR3 replications, 8 of its 9 misattributions and one correct run, and caught 16 of 20 and 26 of 36 misattributions on the random patterns at flag rates of 45 % and 52 %, counting runs without a tool (Supplementary S9). On the iiwa replications it flagged 148 of its 190 misattributions: in 42 runs the paired comparison rejected the other wrist support, which differs from the selected one by a fraction of a percent of the score, and the whole parsimony band, the secondary rule, flags all 190. There the Bayesian intervals without inflation did better, flagging every one of their misattributions with a specificity of 80-83 %, because the posterior keeps a few percent on the other support; on the random patterns they caught 10-38 % of theirs.

Fixed cases and stability Seven further patterns of the earlier study, fixed before any fit, and an eighth for the iiwa, each over 20 noise draws, are reported in Supplementary S6. Single changes of the tool, of a joint's friction and of a link's mass, and a light tool with two friction changes, are recovered exactly in every draw on both robots, and removing a recovered tool leaves 0.025-0.029 mm against 4.9-17.6 mm for the dense twins. On fixed trajectories, the four injected groups of the main change are selected in all 20 MuJoCo draws at noise $\times 1$, all 10 at $\times 3$ and all 20 Isaac Sim draws, with a payload of $0.794 \pm 0.001 \text{ kg}$ in MuJoCo (Fig. 8). The audit explains the pattern of the ties: the tool's torque is 99.9 % reproducible by link 6 on both robots, and the link-3 change 99.5 % by link 2 on the FR3 and 99.995 % by link 4 on the

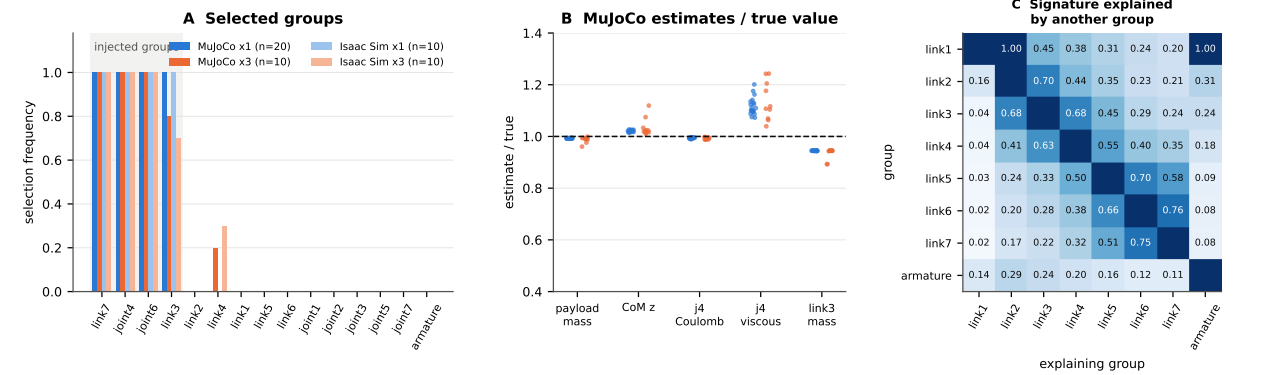


Figure 8: Selection on fixed trajectories. (A) Selection frequency of each group over 20 (MuJoCo, $\times 1$), 10 (MuJoCo, $\times 3$) and 10 (Isaac Sim, $\times 1$ and $\times 3$) noise draws. (B) MuJoCo estimates normalized by the true value; the link-3 mass has a structural null direction. (C) Fraction of a group’s standardized torque signature explained by another group’s column space on the identification rows.

iiwa, a remainder of 0.002 N m (Supplementary S8). A point mass at the intersection of the first three joint axes is a null direction of link 3 alone: on the FR3 it is, per kilogram, $n = (1, 0, 0, -0.316, 0.0999, 0, 0, 0.0999, 0, 0)$ in the order mass, first moments and inertia (xx, xy, xz, yy, yz, zz) of link 3, a point at (0, 0, -0.316) m in link-3 coordinates, and it produces at most 10^{-13} N m. SAPC leaves that direction at its nominal value, so its link-3 mass increment of 0.13 kg differs from the injected 0.27 kg while its torque effect is right.

5.5. Tool benchmark: single stage and unloaded first

The second protocol tests the edit that motivates the paper, removing a tool through its attribution, on 150 new patterns per robot, each with its own new trajectories. Every pattern carries a tool at the flange, drawn from the distribution of the random benchmark, and zero to three other changes with the same nuisances. In the single stage, the loaded runs are calibrated against the nominal twin; in the unloaded-first route, as in the physical session, SAPC first calibrates three unloaded runs and every method then fits three loaded runs relative to that stage-one twin. A tool is misattributed when its link-7 mass is outside $\max(0.02 \text{ kg}, 20 \%)$ of the truth, and the edit error is the torque error of the twin after clearing its link-7 increment, divided by the tool’s own torque, so that 1 is the cost of leaving the tool in the twin (Table 8).

Attribution In the single stage, joint-torque payload identification misattributes the tool in 31 % of the FR3 patterns and 42 % of the iiwa ones, with or without the friction terms, SAPC in 38 % and 66 %, the Bayesian variants in 37-51 % and 67-75 %, and the dense LMI fit in 99 % on both robots. The other changes of a pattern are what the location prior cannot absorb: link 7 alone leaves an edit error of 0.17 and 0.20 at the median, where 1 is the cost of leaving the tool in, and adding friction brings it to 0.11 and 0.14; SAPC reaches 0.11 on the FR3 and 0.48 on the iiwa, whose wrist lets links 4-6 take part of the tool, and the dense fit 0.82 and 0.79. Calibrating the unloaded robot first makes the tool the only change, and the location prior then decides it exactly: link 7 alone and link 7 with friction misattribute no tool on either robot and edit to 0.03-0.04. Stage two then carries the tool and the per-trajectory friction drift of the nuisances; a global friction change between the stages is part of the dry-run stress conditions of the physical session (Supplementary S7). The unloaded stage lowers SAPC’s misattribution to 25 % and 50 %, significantly on both robots (exact McNemar, 28 against 9 and 43 against 19 discordant patterns, $p = 0.003$), but not that of the Bayesian variants ($p \geq 0.06$); with the per-joint offsets removed as in the physical session, SAPC misattributes 25 % and 44 %.

The edit flag SAPC's flag is informative on both robots and in both arms (one-sided Fisher test, $p \leq 0.008$): it catches 81 % and 87 % of the misattributions in the single stage and 87 % and 96 % after the unloaded stage, at flag rates of 65-83 %. Over the 600 fits it flags 237 of its 269 misattributions and leaves 32, 5.3 % of the fits, wrong and quiet; the price is specificity, 37 %, so that a flag sends many correct attributions to a known-inertia edit or to new excitation. Pooled over robots and arms, the Bayesian variants are more specific, 76-96 %, and catch 56-67 % of their misattributions, leaving 20-23 % of the fits wrong with a credible interval inside the tolerance; SAPC has fewer such

Table 8

Tool benchmark of the second protocol, 150 patterns per robot. Misattributed: link-7 mass outside $\max(0.02 \text{ kg}, 20\%)$ of the tool, in percent with the Wilson 95 % interval. Edit: median (90th percentile) torque error after removing the tool by attribution, relative to the tool's torque; 1 is the cost of leaving the tool in. Flag: sensitivity (misattributions flagged) and flag rate, in percent; SAPC with the tie rule of two paired standard errors, the Bayesian variants with their 95 % credible interval; full operating characteristics in Supplementary S9.

Method	Single stage				Unloaded first			
	Misattr. [%]	Edit	Sens.	Flags	Misattr. [%]	Edit	Sens.	Flags
<i>FR3, 150 patterns</i>								
SAPC	38 [31,46]	0.11 (1.00)	81	65	25 [19,33]	0.06 (1.00)	87	68
Link 7 only	31 [24,38]	0.17 (1.09)	-	-	0 [0,2]	0.03 (0.13)	-	-
Link 7 + friction	31 [24,38]	0.11 (0.73)	-	-	0 [0,2]	0.04 (0.14)	-	-
Dense LMI	99 [96,100]	0.82 (0.94)	-	-	100 [98,100]	0.82 (0.90)	-	-
Bayes	51 [43,59]	0.23 (1.00)	54	41	49 [41,57]	0.19 (1.01)	48	32
Bayes iid	37 [30,45]	0.13 (1.00)	27	11	32 [25,40]	0.11 (1.00)	23	11
Bayes η -CV	42 [34,50]	0.17 (1.00)	33	15	37 [30,45]	0.12 (1.00)	29	13
<i>iiwa, 150 patterns</i>								
SAPC	66 [58,73]	0.48 (1.00)	87	81	50 [42,58]	0.29 (1.00)	96	83
Link 7 only	42 [34,50]	0.20 (1.41)	-	-	0 [0,2]	0.04 (0.13)	-	-
Link 7 + friction	42 [34,50]	0.14 (1.14)	-	-	0 [0,2]	0.04 (0.14)	-	-
Dense LMI	99 [95,100]	0.79 (0.91)	-	-	100 [98,100]	0.80 (0.90)	-	-
Bayes	75 [67,81]	0.51 (1.00)	81	67	64 [56,71]	0.51 (1.00)	75	58
Bayes iid	67 [59,74]	0.49 (1.00)	70	49	65 [57,72]	0.51 (0.99)	74	51
Bayes η -CV	68 [60,75]	0.49 (1.00)	71	50	63 [55,71]	0.50 (0.99)	72	48

fits than every variant in both arms on both robots (exact McNemar, $p < 0.005$) except the inflated variant in the iiwa single stage ($p = 0.15$). Choosing the learning rate by blocked cross-validation kept the untempered posterior in 65-80 % of the patterns and never went below 2^{-5} , so the tempered variant behaves like the uninflated one. Where the location of the tool is known and the unloaded robot has been calibrated first, the location should decide the attribution; where it is not known, the flag is what tells the user that the data did not decide it.

5.6. Other simulators as references

Simulators disagree even at identical parameters [48], so a reference from another simulator adds discrepancy that is not of the robot. The twin stays in MuJoCo and two references are calibrated with it (Table 9; implementation in Supplementary S3). Gazebo Harmonic [57] receives SDF worlds generated from the MJCF with the inertias, damping and Coulomb friction copied and the changes of Table 3 injected; SDF has no rotor-inertia field, so the reference lacks the rotor inertia that the MuJoCo model carries, as deployed Gazebo FR3 models do. Isaac Sim 4.5 [58] imports the same description as a native articulation that keeps the rotor inertia. Neither exposes accelerations, so both sets use filtered derivatives, and the torque metrics include their error.

Against Gazebo the simulator-only replay error is 0.52 mm and the nominal twin misses the changed reference by 14.40 mm. Dense calibration reaches 0.34 mm and SAPC 0.20 mm at the same torque NRMSE of 0.095, and the attribution repeats the controlled study: dense calibration places 0.131 kg of the tool on link 7, SAPC 0.797 kg. SAPC also finds a change that was never injected: it selects the armature group with corrections that match the rotor inertia missing from SDF on all seven joints, -0.194 to -0.190 kg m^2 on joints 1-4 and -0.070 to -0.074 on joints 5-7. The values on joints 1 and 2 share null directions with link inertia (Section 3.2), so they follow from the sparse prior; the other five follow from the data. It also selects joint 7, a model-form correction that absorbs Gazebo's low-speed friction law, not a physical change of the robot. Calibrated on unchanged Gazebo data alone, it selects these two groups and nothing else and reduces the simulator-only error from 0.52 to 0.11 mm; removing the armature from the calibrated twin raises its error back to 0.52 mm. The two-stage variant makes this a procedure: stage one absorbs the armature and joint 7, and stage two selects the injected groups with joint 7 and recovers 0.798 kg at 0.20 mm, while dense calibration on the same stage-one twin returns 0.131 kg.

Against Isaac Sim the simulator-only error is 0.19 mm, and dense calibration and SAPC both reach 0.26 mm, the level of the twin with the true parameters, at a torque NRMSE of 0.065. Dense calibration places 0.131 kg on link 7, and SAPC 0.795 kg with exactly the four injected groups, in all 10 noise draws at $\times 1$ and all 10 at $\times 3$. On unchanged

Table 9

Other simulators as references, MuJoCo twin, noise $\times 1$: groups, torque NRMSE (linear library, filtered derivatives), replay TCP RMSE on the held-out trajectories and payload mass (true 0.8 kg).

Level and twin	Groups	NRMSE	Replay [mm]	Payload [kg]
<i>Gazebo reference</i>				
Unchanged reference, nominal twin	-	-	0.52	-
Unchanged reference, SAPC	armature, joint 7	0.41	0.11	0
Changed reference, nominal twin	-	1.000	14.40	-
Dense least squares	all 15	0.095	0.34	0.131
SAPC	injected, armature, joint 7	0.095	0.20	0.797
Injected groups and armature	5	0.095	0.20	0.794
Injected groups without armature	4	0.100	0.54	0.784
Two stages, SAPC after the unchanged calibration	injected, joint 7	0.104*	0.20	0.798
<i>Isaac Sim reference</i>				
Unchanged reference, nominal twin	-	-	0.19	-
Unchanged reference, SAPC (10 draws)	none	1.000	0.19	0
Changed reference, nominal twin	-	1.000	14.41	-
Dense least squares	all 15	0.065	0.26	0.131
SAPC (20 draws, $\times 1$ and $\times 3$)	injected	0.065	0.26	0.795
True parameters	-	-	0.26	0.800

*Relative to the stage-one residual.

Isaac Sim data SAPC returns the uncorrected twin in all 10 draws. Where the two descriptions agree, the calibration finds only what was injected; where they differ, it finds the difference and names it.

5.7. A physical FR3

A physical robot has one change whose truth is exact, a weighed block, and the session tests the whole protocol on it, blinded and with criteria fixed in advance (HARDWARE.md in the data archive [29]; amended after the first round of review, before the session). Two steel blocks of about 0.5 and 1.0 kg, weighed to 1 g, are mounted on an FR3 without its hand in seven coded runs whose order a person who does not analyze the data draws and seals: block 1 off the joint-7 axis at the flange, mounted twice with a remount between, block 2 at the same place, block 1 on the joint-7 axis, block 1 on a bracket on the link-6 housing, block 1 on a bracket on the link-4 housing, and nothing. The on-axis run is a positive control of the report: a block on that axis lies on a near-null direction of links 6 and 7, so the report must name both. Every housing mount keeps the block's center of mass at least 80 mm from both joint axes that bound its link, because a block near the distal axis would lie on a null direction with the child link, and a pre-session check computes for every mount the part of the block's torque signature that the best competing group cannot reproduce, which must reach one torque-noise standard deviation; the check moved the first link-6 stand-in, which sat 2 mm from the joint-7 axis. Unloaded runs before and after the coded ones provide the stage-one twin and the drift check; five Fourier trajectories at 60 % of the simulation's velocity and acceleration caps identify and a sixth is held out, recorded as joint positions and link-side joint torques at 1 kHz [7]. Stage one calibrates the nominal twin to the unloaded robot without the eight-group cap, because a link-side torque differs from the twin's actuator torque at every joint; stage two identifies each coded run relative to the complete stage-one twin with all five methods, with the per-joint torque offset of each run removed, because torque-sensor offsets drift between runs; a single-stage variant fits every run against the nominal twin.

The criteria and their thresholds are fixed in advance. H1a: every off-axis block lies exactly on its mount's link in SAPC's support. H1b: otherwise the location report, the inertial groups of the parsimony band, names the mount's link among at most two groups and is flagged; for the on-axis block it is flagged and names links 6 or 7 only. H2: on every block run, whatever H1, the mass increment on the mount's link is within 10 % of the weighed block (on-axis: links 6 and 7 together). H3: at the flange the center of mass is within 10 mm of the CAD value. H4: on the empty run and on the final unloaded run no link group is selected, and no friction group whose held-out torque exceeds 0.10 N m. H5: the ratio of the two flange masses is within 5 % of the weighed ratio, a linearity check of the whole chain. H6: edits from both flange mountings (removal, exchange for block 2, relocation to the link-6 housing) and from both housing runs (removal, relocation to the other housing) predict the held-out torques of the target run within 1.2 times

Table 10

Blinded session on a physical FR3, stage two relative to the unloaded stage-one twin. Top: mass increment on the link that carries the block [kg], for every method (link 7 only: its link-7 increment), and SAPC's location report. Bottom: held-out torque RMSE on the target run [N m] of the twin edited through its own attribution, and of SAPC fitted on the target run itself; last row: the largest ratio of a known-inertia edit to the method's own fit on the target (criterion H6b).

Coded run (revealed after the analysis)	SAPC (report)	Link 7 only	Dense LMI	Log-det	Bayes	Weighed
Block 1, flange (first mounting)	0.50 (7)	0.50	0.10	0.43	0.50	0.50
Block 1, flange (remounted)	0.50 (7)	0.50	0.10	0.43	0.50	0.50
Block 2, flange	1.00 (7)	1.00	0.24	0.24	1.00	1.00
Block 1, on the joint-7 axis	0.50 (6,7; flagged)	0.50	0.19	0.50	0.50	0.50
Block 1, link-6 housing	0.50 (6)	0.40	0.10	0.51	0.50	0.50
Block 1, link-4 housing	0.50 (4)	0.08	0.14	-0.05	0.50	0.50
Nothing	no link	-	-	-	-	-
Edit by attribution [N m]	SAPC	Link 7 only	Dense LMI	Log-det	Bayes	SAPC on target
Flange removal	0.07	0.07	0.22	0.21	0.07	0.07
Flange exchange	0.08	0.08	0.22	0.21	0.08	0.08
Flange relocation to link 6	0.07	0.07	0.22	0.21	0.07	0.07
Link 6 removal	0.07	0.21	0.18	0.08	0.07	0.07
Link 6 relocation to link 4	0.07	0.21	0.18	0.08	0.07	0.07
Link 4 removal	0.07	0.10	0.10	0.13	0.07	0.07
Link 4 relocation to link 6	0.07	0.10	0.10	0.13	0.07	0.07
Known inertia, largest ratio to own fit	1.00	1.64	1.00	1.00	1.00	-

the method fitted on that run itself, by SAPC's attribution (H6a) and, for every full-library method, by known inertia (H6b); a criterion on the flange placement passes only if both mountings pass. Joint-torque payload identification is reported beside H6b: it cannot represent a housing block, so its twin does not reproduce those runs, which Proposition 2 presupposes. For every edit the torque difference between SAPC's and joint-torque payload identification's edited twins is also computed, so that an edit for which the two cannot differ is known before the session.

A dry run of the whole session on a simulated stand-in, with a tenth of the nominal friction, no rotor inertia and the stand-in mounts, under five conditions of increasing difficulty, is reported in Supplementary S7. In every condition SAPC met all eight criteria: it located the off-axis blocks exactly, within 1.8 % of their mass and 2.7 mm of their center of mass, reported the on-axis block on links 6 and 7 with both flags raised, and kept the edit flag quiet on every other run. The link-6 edits are the discriminating ones, with 0.19 N m between the twins edited through SAPC's and through joint-torque payload identification's attribution; the flange and link-4 edits, with 0.00 and 0.06 N m, are labeled non-discriminating in advance. In the single-stage variant, removing the flange block leaves 0.11-0.28 N m for SAPC against 0.88-0.92 N m for joint-torque payload identification, which absorbs the robot's own changes into link 7.

On the robot, SAPC placed every off-axis block exactly on the link that carried it, within 2 % of its weighed mass and, at the flange, within 3 mm of its center of mass, in both flange mountings, and selected no link where nothing was mounted. It reported the on-axis block on links 6 and 7 with the location flag and the edit flag raised and the total mass within 1 %, the outcome the report is built to give. The ratio of the two flange masses matched the weighed ratio to 0.1 %. Joint-torque payload identification put 0.40 kg of the link-6 block and 0.08 kg of the link-4 block on link 7, and the dense LMI fit spread every block over links 3-7 while reproducing the torques as well as SAPC. Edited through their own attribution, the SAPC and Bayesian twins predicted the removed and relocated housing blocks as well as twins fitted on the target runs, whereas joint-torque payload identification missed the link-6 edits by a factor of three; edited by known inertia, every full-library twin matched its own fit on every target.

6. Discussion

6.1. A protocol for edited twins

The results reduce to a protocol for any twin whose correction enters the simulator's parameters linearly, and to first order otherwise (Remark 2); Table 11 condenses it. First, calibrate the unloaded robot, so that a later change is

Table 11

Decision table for an edit of a calibrated twin; the rates are those of the tool benchmark (Section 5.5) and the violated-prior runs (Section 5.4).

Situation	Action
Edit by known inertia, or of a joint's friction	Certified: edit any calibrated twin, after the null-space repair
Change at a known location (tool at the flange)	Calibrate the unloaded robot first, then identify the change with its location as the prior
Location unknown, held-out residual above the noise floor	Read the whole parsimony band: the paired rule misses most misattributions under systematic misfit
Location unknown, edit flagged	Use a known inertia, or re-excite to separate the disagreeing supports
Location unknown, edit not flagged	Edit through the attribution, knowing that 10-21 % (FR3) and 11-45 % (iiwa) of unflagged tools were still wrong in the benchmark
No inertia known and re-excitation impossible	Treat the edited behavior as a prior assumption; the leakage interval gives its range

sparse relative to the stage-one twin, and check that the held-out residual is near the noise floor of repeated trajectories and free of structure over time and joints before trusting any attribution; recompute $\mathcal{N}$ for the installation and check that the excitation reaches its rank. Second, decide whether the intended edit is certified, by Proposition 2: an edit by a known inertia, or of a joint's friction, is guaranteed by fidelity and needs no attribution, and a wrist force/torque sensor or a weighed part turns a tool edit into one. Third, if the edit acts on where the twin placed inertia and the location of the change is known, as for a tool at the flange, identify it with that location as the prior, by joint-torque payload identification on the stage-one twin; the tool benchmark shows that this is the most accurate route once the tool is the only change, and that in a single stage the robot's other changes cost it 31-42 % misattributions. Fourth, if the location is not known, calibrate with a stated prior and read the report: the leakage interval says before the edit that the attribution, not the data, will decide it; a flagged edit should use a known inertia or wait for an excitation that separates the supports that disagree; an unflagged edit is decided by the data among the supports the search scored, which is strong but not complete evidence: in the tool benchmark 10-21 % of the unflagged FR3 tools and 11-45 % of the iiwa ones were still wrong, and case C5 shows a failure outside the band. When the residual of the first step stays above the noise floor, the whole parsimony band, not the paired rule, is the flag to read (Section 6.2). None of these steps needs ground truth; the validation of Section 5 is what justifies relying on them. In the two-link example of Section 3.5, the second step subtracts the known clamp and is done; without the clamp's mass, the leakage interval spans every split of the clamp between the links, and re-excitation cannot help, because the clamp lies on the elbow axis and no motion separates the two attachments.

6.2. The prior and its alternatives

Sparse change over body and joint groups is the right default for events whose location is unknown and local, a worn joint, a replaced part, a mounted bracket, as damage is in structural health monitoring [44]. It is wrong for distributed changes and for a CAD error on every link, and Section 5.4 shows what SAPC returns then: a twin that still reproduces the motion, with a report that depends on the tie rule. With a CAD error of 5-10 % on every link, the edit flag of two paired standard errors catches only 12 of the 37 misattributed tools, because the systematic misfit makes the paired comparison reject the supports that disagree, whereas the whole parsimony band flags 36 of them; case C5 on the FR3, whose injected support lies outside the band, stays silent under both. The adequacy check of the first step therefore also selects the tie rule: near the noise floor the paired rule is the specific one, above it the band is the safe one, which is also why the location criterion of the physical session uses the band. Single-stage SAPC is valid while the change is sparse relative to the twin and its signature exceeds the noise by more than the systematic residual of the unmodeled effects; when the sensors are noise-free, the integrator difference between reference and twin becomes the largest residual and SAPC adds a group for it (Supplementary S1), and calibrating the unloaded robot first moves both effects into stage one. The Bayesian enumeration tested here, with an i.i.d. Gaussian likelihood, a plug-in noise variance and no bounds, returns the same representative where the data separate the supports, as on the FR3 replications, where the inflated variant is exact in every run and the uninflated one in 165 of 200. Where they do not, its posterior concentrates on one support: on the iiwa replications 138 of its 200 runs put a posterior of at least 0.95 on a wrong support without inflation and 198 with the autocorrelation inflation, and in the tool benchmark its credible interval of the tool mass

flags 56-67 % of its misattributions, more specifically than SAPC's flag but with four times as many wrong and quiet fits. Where two supports nearly alias, as on the iiwa replications, the uninflated posterior keeps a few percent on the other one, and its interval then flagged every misattribution, more than SAPC's paired rule. Choosing the learning rate by blocked cross-validation, a variant of the SafeBayes repair [56], kept the untempered posterior in 65-80 % of the patterns, because the held-out prediction error barely depends on which of the near-equivalent supports carries the weight. Its reports of a change where nothing changed come from the integrator difference between reference and twin, a small, consistent residual that its likelihood treats as signal. These results concern the implementations tested; a hierarchical model with a learned noise structure may do better, and the flag statistics of Section 5.5 are the measure any such variant should be compared on.

6.3. Limitations and outlook

The certification criterion holds exactly for corrections that enter the model linearly with the state observed, and to first order otherwise (Remark 2); the robot case shows how the test reads, and the same steps apply to any calibrated simulation model whose parameter sensitivities can be computed. The edit flag's operating characteristics are measured here for the edit that motivates the paper, the tool removal; the rule applies unchanged to the mass of any link, and the housing runs of the physical session test it on inner links. How often the prior decides an attribution depends on the kinematics and the excitation. On the iiwa the two supports that the data cannot separate differ by 0.002 N m on these trajectories; the report itself points to the remedy, because the disagreeing supports are rival models, and trajectories on which their predicted torques differ most can be chosen from the report alone, as in optimal excitation [31] and model-discrimination design [59], before the robot is excited again. The leakage interval is bounded by physical consistency alone, which keeps it wide for link attributions; bounds on how far a link may move from its CAD values would turn it into a usable tolerance. Contact lies outside Proposition 1, and extending the edit test to contact parameters, with a force sensor in the loop, is the next step.

7. Conclusion

A calibrated simulation model that reproduces every measured behavior can still carry the wrong physics, and the difference matters when the model is modified. For corrections that enter linearly, a modification is guaranteed by fidelity exactly when it maps the null space of the calibration into itself; on robot twins, edits by known inertia and of joint friction are, and edits that clear what the twin attributed to a link are not, on any arm with revolute joints. Such edits rest on a prior: the location of the change when it is known, and otherwise a stated one, which SAPC takes to be sparse change over body and joint groups, with a report that flags an edit when supports the data cannot reject disagree on it. On two manipulators with manufactured discrepancies, full-library calibrations reproduced held-out motion while dense and structure-aware ones placed the FR3 tool on different links; in a pre-registered tool benchmark, joint-torque payload identification was the most accurate once the unloaded robot had been calibrated first, and SAPC flagged 88 % of its misattributions at a flag rate of 74 %, leaving 5 % of the tools wrong and unflagged where Bayesian enumerations left 20-23 %. The same separation held against Gazebo and Isaac Sim references **and on a physical FR3 with blocks at locations the analyst did not know**. Validating a twin should therefore include the edits it will undergo: certify them where the structure allows, use the location prior where it is known, and where it is not, report the prior that decides them.

CRedit authorship contribution statement

Suhwan Park: Methodology, Software, Validation, Investigation, Writing - original draft. **Jihwan Lee:** Investigation, Validation. **Jiyong Park:** Investigation, Validation. **Sanghyun Kim:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Every result in this paper can be reproduced from public data [29]. The project page, <https://rcilab.khu.ac.kr/sapc>, provides the archives: the result files behind every table and figure, including the 1460 confirmation runs; the recorded Gazebo and Isaac Sim reference logs for identification, validation and the press task, **together with the joint logs of the physical FR3 session**; and the robot models, the MuJoCo Menagerie FR3 and KUKA LBR iiwa 14, and generated simulator worlds. It also provides the source code of SAPC, of the comparison methods and of every experiment, the protocols fixed before the confirmation runs and before the physical session, the MJCF-to-SDF and MJCF-to-USD conversions, the simulator bridges and the fixed-seed scripts that regenerate the MuJoCo trajectories.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) in order to draft and edit the manuscript text, to write and debug the simulation and analysis code, and to produce the data plots from the result files by scripts. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments

This work was supported by the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (RS-2025-11082970), and by the Industrial Strategic Technology Development Program funded by MOTIE (RS-2024-00419641). This research was also supported by a Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Science and ICT (RS-2025-00564593). The funders had no role in the study design, in the collection, analysis and interpretation of data, in the writing of the article or in the decision to submit it for publication.

References

- [1] G. L. Stavrinos, H. D. Karatza, Cyber-physical systems, digital twins and Industry 4.0: The role of modeling and simulation, *Simulation Modelling Practice and Theory* 124 (2023) 102727. doi:10.1016/j.simpat.2023.102727.
- [2] G. Wang, Z. Zhang, A. Nazarova, A digital twin-based platform for testing and optimization of path planning algorithms, *Simulation Modelling Practice and Theory* 145 (2025) 103206. doi:10.1016/j.simpat.2025.103206.
- [3] I. Ahmad, B. Siddique, M. A. Islam, Z. U. Haq, Z. Niu, M. M. Waqas, Q. Yang, Z. Qiu, Development and evaluation of digital twin model for rack and pinion drive vegetable seedling transmission device using Adams/MATLAB co-simulation, *Simulation Modelling Practice and Theory* 142 (2025) 103132. doi:10.1016/j.simpat.2025.103132.
- [4] A. Allevato, E. S. Short, M. Pryor, A. Thomaz, TuneNet: One-shot residual tuning for system identification and sim-to-real robot task transfer, in: L. P. Kaelbling, D. Kragic, K. Sugiura (Eds.), *Proceedings of the Conference on Robot Learning (CoRL 2019)*, Vol. 100 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 445–455, arXiv:1907.11200.
- [5] D. Bargellini, A. Pasquali, A. Govoni, R. Zanella, G. Palli, Enhancing Sim2Real transfer for torque-controlled robots through Real2Sim dynamics estimation and reinforcement learning, arXiv:2608.22629, presented at the 2026 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM) (2026).
- [6] Y. Chebotar, A. Handa, V. Makovychuk, M. Macklin, J. Issac, N. Ratliff, D. Fox, Closing the sim-to-real loop: Adapting simulation randomization with real world experience, in: *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8973–8979. doi:10.1109/icra.2019.8793789.
- [7] C. Gaz, M. Cognetti, A. Oliva, P. Robuffo Giordano, A. De Luca, Dynamic identification of the Franka Emika Panda robot with retrieval of feasible parameters using penalty-based optimization, *IEEE Robotics and Automation Letters* 4 (4) (2019) 4147–4154. doi:10.1109/LRA.2019.2931248.
- [8] P. M. Wensing, S. Kim, J.-J. E. Slotine, Linear matrix inequalities for physically consistent inertial parameter identification: A statistical perspective on the mass distribution, *IEEE Robotics and Automation Letters* 3 (1) (2018) 60–67. doi:10.1109/LRA.2017.2729659.
- [9] P. J. Durst, D. McInnis, J. Davis, C. T. Goodin, A novel framework for verification and validation of simulations of autonomous robots, *Simulation Modelling Practice and Theory* 117 (2022) 102515. doi:10.1016/j.simpat.2022.102515.
- [10] S. Xie, W. Zhang, M. Shen, F. Xue, A modular configuration-based rapid verification and validation framework combining analytical and simulation methods for intelligent manufacturing systems, *Simulation Modelling Practice and Theory* 142 (2025) 103136. doi:10.1016/j.simpat.2025.103136.
- [11] J. Shi, Q. Zhou, C. Pan, A credibility evaluation method for digital twin based on improved evidence theory, *Simulation Modelling Practice and Theory* 143 (2025) 103152. doi:10.1016/j.simpat.2025.103152.
- [12] ASME, Assessing credibility of computational modeling through verification and validation: Application to medical devices, Tech. Rep. V&V 40-2018, American Society of Mechanical Engineers (2018).

- [13] R. G. Sargent, Verification and validation of simulation models, *Journal of Simulation* 7 (1) (2013) 12–24. doi:10.1057/jos.2012.20.
- [14] T. Xia, J. Zuo, J. Ding, Y. Pan, Degradation-adaptive digital twin for long-term consistency, *Simulation Modelling Practice and Theory* 149 (2026) 103276. doi:10.1016/j.simpat.2026.103276.
- [15] M. G. Kapteyn, J. V. R. Pretorius, K. E. Willcox, A probabilistic graphical model foundation for enabling predictive digital twins at scale, *Nature Computational Science* 1 (5) (2021) 337–347. doi:10.1038/s43588-021-00069-0.
- [16] M. Gautier, W. Khalil, Direct calculation of minimum set of inertial parameters of serial robots, *IEEE Transactions on Robotics and Automation* 6 (3) (1990) 368–373. doi:10.1109/70.56655.
- [17] H. Mayeda, K. Yoshida, K. Osuka, Base parameters of manipulator dynamic models, *IEEE Transactions on Robotics and Automation* 6 (3) (1990) 312–321. doi:10.1109/70.56663.
- [18] P. M. Wensing, G. Niemeyer, J.-J. E. Slotine, A geometric characterization of observability in inertial parameter identification, *The International Journal of Robotics Research* 43 (14) (2024) 2274–2302. doi:10.1177/02783649241258215.
- [19] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3) (2001) 425–464. doi:10.1111/1467-9868.00294.
- [20] J. Brynjarsdóttir, A. O'Hagan, Learning about physical parameters: The importance of model discrepancy, *Inverse Problems* 30 (11) (2014) 114007. doi:10.1088/0266-5611/30/11/114007.
- [21] R. Tuo, C. F. J. Wu, Efficient calibration for imperfect computer models, *The Annals of Statistics* 43 (6) (2015) 2331–2352. doi:10.1214/15-aos1314.
- [22] S. R. Searle, *Linear Models*, Wiley, 1997. doi:10.1002/9781118491782.
- [23] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics* 25 (15) (2009) 1923–1929. doi:10.1093/bioinformatics/btp358.
- [24] R. Isermann, Model-based fault-detection and diagnosis - status and applications, *Annual Reviews in Control* 29 (1) (2005) 71–85. doi:10.1016/j.arcontrol.2004.12.002.
- [25] M. I. Friswell, J. E. Motershead, *Finite Element Model Updating in Structural Dynamics*, Solid Mechanics and its Applications, Springer Netherlands, 1995. doi:10.1007/978-94-015-8508-8.
- [26] L. S. Katafygiotis, J. L. Beck, Updating models and their uncertainties. ii: Model identifiability, *Journal of Engineering Mechanics* 124 (4) (1998) 463–467. doi:10.1061/(asce)0733-9399(1998)124:4(463).
- [27] P. J. Roache, Code verification by the method of manufactured solutions, *Journal of Fluids Engineering* 124 (1) (2002) 4–10. doi:10.1115/1.1436090.
- [28] W. L. Oberkampf, C. J. Roy, *Verification and Validation in Scientific Computing*, Cambridge University Press, 2010. doi:10.1017/cbo9780511760396.
- [29] S. Park, J. Lee, J. Park, S. Kim, SAPC research artifacts: result files, reference logs, robot models and code [dataset], GitHub release research-artifacts-v2, <https://github.com/RCILab/sapc/releases/tag/research-artifacts-v2> (2026).
- [30] W. Khalil, E. Dombre, *Modeling, Identification and Control of Robots*, Butterworth-Heinemann, Oxford, UK, 2004, first published by Hermes Penton Science, London, 2002.
- [31] J. Swevers, C. Ganseman, D. Tukul, J. de Schutter, H. Van Brussel, Optimal robot excitation and identification, *IEEE Transactions on Robotics and Automation* 13 (5) (1997) 730–740. doi:10.1109/70.631234.
- [32] M. Gautier, Numerical calculation of the base inertial parameters of robots, *Journal of Robotic Systems* 8 (4) (1991) 485–506. doi:10.1002/rob.4620080405.
- [33] C. D. Sousa, R. Cortesão, Physical feasibility of robot base inertial parameter identification: A linear matrix inequality approach, *The International Journal of Robotics Research* 33 (6) (2014) 931–944. doi:10.1177/0278364913514870.
- [34] T. Lee, P. M. Wensing, F. C. Park, Geometric robot dynamic identification: A convex programming approach, *IEEE Transactions on Robotics* 36 (2) (2020) 348–365. doi:10.1109/tro.2019.2926491.
- [35] C. G. Atkeson, C. H. An, J. M. Hollerbach, Estimation of inertial parameters of manipulator loads and links, *The International Journal of Robotics Research* 5 (3) (1986) 101–119. doi:10.1177/027836498600500306.
- [36] D. Kubus, T. Kroger, F. Wahl, On-line estimation of inertial parameters using a recursive total least-squares approach, in: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008, pp. 3845–3852. doi:10.1109/iro.2008.4650672.
- [37] N. Mavrikis, R. Stolkin, Estimation and exploitation of objects' inertial parameters in robotic grasping and manipulation: A survey, *Robotics and Autonomous Systems* 124 (2020) 103374. doi:10.1016/j.robot.2019.103374.
- [38] W. Khalil, M. Gautier, P. Lemoine, Identification of the payload inertial parameters of industrial manipulators, in: Proceedings 2007 IEEE International Conference on Robotics and Automation, 2007, pp. 4943–4948. doi:10.1109/robot.2007.364241.
- [39] S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proceedings of the National Academy of Sciences* 113 (15) (2016) 3932–3937. doi:10.1073/pnas.1517384113.
- [40] Z. Chen, Y. Liu, H. Sun, Physics-informed learning of governing equations from scarce data, *Nature Communications* 12 (1) (2021) 6136. doi:10.1038/s41467-021-26434-1.
- [41] T. N. Truong, S. Kim, High-performance fixed-time active fault-tolerant control for robotic manipulators via sparse physics-informed dynamics learning, *Information Sciences* 758 (2027) 124001. doi:10.1016/j.ins.2026.124001.
- [42] H. Kim, S. Kim, J. Park, Enhanced online parameter estimation of unknown objects via sparse identification of unmodeled dynamics, *Journal of Mechanical Science and Technology* 39 (5) (2025) 2343–2354. doi:10.1007/s12206-025-2201-2.
- [43] S. Haddadin, A. De Luca, A. Albu-Schaffer, Robot collisions: A survey on detection, isolation, and identification, *IEEE Transactions on Robotics* 33 (6) (2017) 1292–1312. doi:10.1109/tro.2017.2723903.
- [44] E. M. Hernandez, Identification of isolated structural damage from incomplete spectrum information using l1-norm minimization, *Mechanical Systems and Signal Processing* 46 (1) (2014) 59–69. doi:10.1016/j.ymssp.2013.12.009.

- [45] K. Beven, A manifesto for the equifinality thesis, *Journal of Hydrology* 320 (1-2) (2006) 18–36. doi:10.1016/j.jhydrol.2005.07.007.
- [46] E. Heiden, D. Millard, E. Coumans, G. S. Sukhatme, Augmenting differentiable simulators with neural networks to close the sim2real gap, arXiv:2007.06045 (2020).
- [47] S. Chabanet, H. Bril El-Haouzi, P. Thomas, Active learning confidence measures for coupling strategies in digital twins integrating simulation and data-driven submodels, *Simulation Modelling Practice and Theory* 140 (2025) 103092. doi:10.1016/j.simpat.2025.103092.
- [48] A. Farley, J. Wang, J. A. Marshall, How to pick a mobile robot simulator: A quantitative comparison of CoppeliaSim, Gazebo, MORSE and Webots with a focus on accuracy of motion, *Simulation Modelling Practice and Theory* 120 (2022) 102629. doi:10.1016/j.simpat.2022.102629.
- [49] G. D. Wijaya, W. Caesarendra, M. I. Petra, G. Królczyk, A. Glowacz, Comparative study of Gazebo and Unity 3D in performing a virtual pick and place of Universal Robot UR3 for assembly process in manufacturing, *Simulation Modelling Practice and Theory* 132 (2024) 102895. doi:10.1016/j.simpat.2024.102895.
- [50] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecography* 40 (8) (2017) 913–929. doi:10.1111/ecog.02881.
- [51] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, 2009. doi:10.1007/978-0-387-84858-7.
- [52] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (1) (2006) 49–67. doi:10.1111/j.1467-9868.2005.00532.x.
- [53] J. G. Scott, J. O. Berger, Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem, *The Annals of Statistics* 38 (5) (2010) 2587–2619. doi:10.1214/10-aos792.
- [54] K. Zakka, Y. Tassa, MuJoCo Menagerie Contributors, MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, http://github.com/google-deepmind/mujoco_menagerie (2022).
- [55] E. Todorov, T. Erez, Y. Tassa, MuJoCo: A physics engine for model-based control, in: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033. doi:10.1109/IRoS.2012.6386109.
- [56] P. Grünwald, T. van Ommen, Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it, *Bayesian Analysis* 12 (4) (2017) 1069–1103. doi:10.1214/17-ba1085.
- [57] N. Koenig, A. Howard, Design and use paradigms for Gazebo, an open-source multi-robot simulator, in: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE Cat. No.04CH37566), Vol. 3, 2004, pp. 2149–2154. doi:10.1109/iros.2004.1389727.
- [58] NVIDIA, Isaac Sim 4.5 documentation, <https://docs.isaacsim.omniverse.nvidia.com/4.5.0/> (2025).
- [59] G. E. Box, W. J. Hill, Discrimination among mechanistic models, *Technometrics* 9 (1) (1967) 57–71. doi:10.1080/00401706.1967.10490441.