# GRAPH REACH SETS THE REMOTE-RESPONSE LIMIT OF LOCAL INTERATOMIC POTENTIALS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Force accuracy cannot certify physical fidelity when a response depends on atoms beyond a model's graph reach. For finite-range local-energy message-passing models, we prove an exact architectural law relating interaction depth $L$ to source–target graph distance $d_G$. When $2L - d_G < 0$, attribute-to-force and coordinate force-response cross-derivatives vanish for every parameter setting, so training cannot recover the missing response. In $Au_2$/MgO, increasing MACE depth from two to three blocks crosses this boundary, improves the held-out doped–undoped contrast from density functional theory (DFT), and activates aligned Al-to-Mg sensitivity. By contrast, a parameter-matched width increase leaves graph reach unchanged and does not recover the response. Increasing cutoff closes 57.0% of the predictive gap (95% interval, $55.2$–$58.8\%$), while a SchNet-style model reproduces the depth–width ordering. Ag–Pd exposes the complementary failure: graph-permitted models retain the wrong response direction or scale despite competitive force errors. The graph-reach law therefore provides a pre-training test of representability and separates architectural impossibility from failures of supervision or optimization.

## 1 INTRODUCTION

Low average force error does not tell us whether an interatomic potential has learned the remote response needed for a scientific decision. A weak response has three possible origins: the model graph may exclude the dependency, the training data may not identify it, or optimization may fail to use an available path. Each origin calls for a different remedy. Depth and cutoff change graph access, width changes capacity along existing paths, and response-aligned data change what the fit can identify.

For local-energy message-passing models, one part of this ambiguity is exactly topological. If a force-relevant source and target are separated by more than twice the interaction depth, their cross derivative vanishes for every parameter setting. Width can enrich messages along a path but cannot create a missing path; depth and cutoff expand the support itself. After that boundary is crossed, a held-out response reveals whether supervision has activated the newly available dependency.

We turn this distinction into a response-first analysis. A physical source–target intervention defines the derivative of interest, and the label-free graph reveals whether the architecture can represent it before training. Depth or cutoff crosses the boundary, matched width changes capacity without changing support, and a held-out response measures learned use. A model-only intervention probes elemental identity sensitivity, while response-aligned supervision asks whether the fit can activate an available path (Figure 1).

We study this sequence in two chemically distinct systems. In $Au_2$/MgO, each Au atom is five or six hops from its nearest Al dopant at 3.5 Å, outside two-block support but inside the three-block bound. Depth and cutoff cross this boundary; parameter-matched width and a SchNet-style network isolate capacity and architecture. Ag–Pd paired DFT force constants then show how permission, supervision, response direction, and scale separate after the exact exclusion is removed.

The work contributes an exact graph-margin theorem for remote force response, matched depth–cutoff–width interventions that connect the theorem to $Au_2$/MgO prediction, and paired Ag–Pd force-constant experiments that expose the gap between graph permission and learned fidelity. A



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

separately implemented continuous-filter model shows that the depth–width distinction is not tied to MACE tensor products.

## 2 Related Work

**Hessian and force-response measurements.** Hessian blocks have been used to assess force locality and fragment convergence (Herbold & Behler, 2022); recent work directly supervises sampled force constants (Koker et al., 2026) or predicts molecular Hessians without differentiating an energy model (Burger et al., 2025). These works establish that aggregate energy–force accuracy does not determine curvature. Our additional question is architectural: before training, can a specified block be nonzero in the chosen local-energy function class?

**Local and explicitly long-range interatomic potentials.** Equivariant message-passing potentials improve sample efficiency and local geometric representation (Schütt et al., 2017; Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023), while explicit electrostatics, charge transfer, and learned electronic variables address interactions that finite local cutoffs may miss (Unke et al., 2021; Kosmala et al., 2023; Cheng, 2025; Deng et al., 2023; Ko et al., 2021). We operate at the model-class level: the $2L$ bound asks whether a chosen response lies inside the support of a local-energy architecture before capacity or data are added. Global, reciprocal-space, and charge-equilibration channels lie outside this finite-range support model.

**Graph reach and transport.** Under-reaching, over-squashing, and effective-receptive-field analyses distinguish missing paths from weak use or compression along paths that exist (Alon & Yahav, 2021; Topping et al., 2022; Di Giovanni et al., 2023; Finder et al., 2025). Recent completeness theory identifies conditions under which multilayer atomistic GNNs realize their full hop-wise representational range (Ming & Wang, 2026). Our response-first experiment instead asks whether a specified force derivative is structurally available and then learned: it combines label-blind graph analysis, depth–cutoff interventions, a parameter-matched width control, and a held-out physical response.

**Response-aware evaluation.** Phonons, force constants, and charge-transfer phenomena expose derivatives that can be obscured by aggregate metrics or otherwise adequate predictive fits (Herbold & Behler, 2022; Koker et al., 2026; Ko et al., 2021; Rosenbrock et al., 2021; Fu et al., 2025). We treat such responses as primary quantities of interest. The central distinction is between topological permission and statistical identification: a permitted path is a necessary architectural condition, while task-matched supervision determines whether the fitted model uses it.

## 3 From graph reach to learned response

Aggregate energy or force error does not reveal whether an interatomic potential represents a particular remote derivative. Our analysis instead begins with a physically specified source–target response, determines whether the model graph can represent it, and only then compares trained models. Figure 1 summarizes this order of operations.

### 3.1 Measure the response before comparing models

We distinguish a paired coordinate derivative from two $Au_2$/MgO response curves. For target $t$, source $s$, and Cartesian direction $j$, define

$$J^{\mathrm{coord}}_{t\leftarrow s,j}(x;\delta) = \frac{F_t(x+\delta e_{s,j}) - F_t(x-\delta e_{s,j})}{2\delta}.$$

Ag–Pd directly pairs this learned block with the DFT target $J^{\mathrm{DFT}}_{t\leftarrow s,j} = -\Phi^{(:,j)}_{t,s}$. For $Au_2$/MgO, let $S(x)$ denote the projected net-Au force and let $b$ index four held-out non-wetting Au–O bins. The

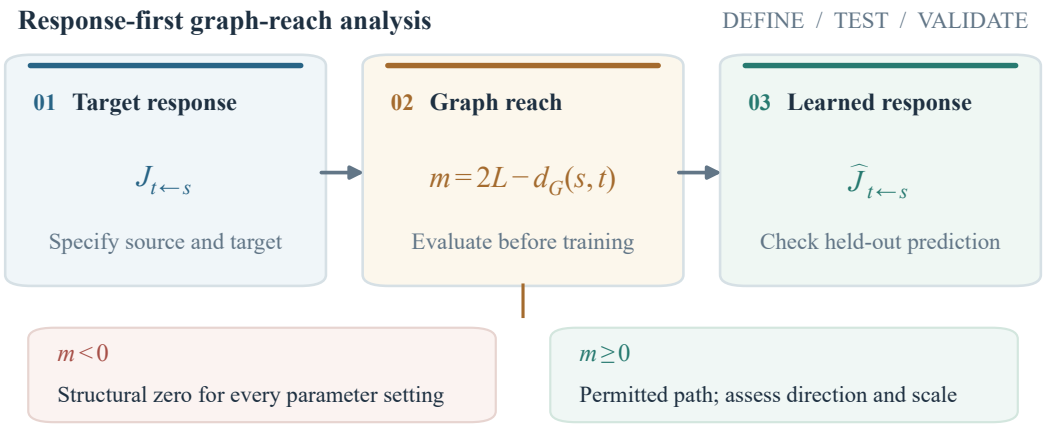


Figure 1: **Graph reach separates exclusion from learned response.** The margin $m = 2L - d_G$ determines structural support; permitted responses are evaluated with held-out prediction.

DFT group contrast, model identity sensitivity, and their alignment error are

$$g_b^{\mathrm{DFT}} = \mathbb{E}[S(x) \mid D = 1, b] - \mathbb{E}[S(x) \mid D = 0, b],$$
$$a_{\theta,b}^{\mathrm{id}} = \mathbb{E}[\widehat{S}_\theta(x) - \widehat{S}_\theta(x_{\mathrm{Al}\to\mathrm{Mg}}) \mid D = 1, b],$$
$$R_{\mathrm{align}}(\theta) = \left( \frac{1}{4} \sum_b (a_{\theta,b}^{\mathrm{id}} - g_b^{\mathrm{DFT}})^2 \right)^{1/2}. \tag{1}$$

Here $D$ is the doping indicator and the expectations are empirical held-out means. The identity replacement fixes positions, cell, and graph, making $R_{\mathrm{align}}$ a model-side alignment measure. Its physical reference is the unpaired DFT group contrast rather than a paired causal effect. Pairs, bins, displacements, and sign conventions are chosen before model evaluation.

### 3.2 Declare the exact graph boundary

Let $G_r$ be the periodic neighbor graph at cutoff $r$, and suppose a model represents total energy as $E_\theta = \sum_v \varepsilon_v(h_v^{(L)})$, where $h_v^{(L)}$ depends only on the attributed $L$-hop ball around $v$. Define the response-specific reach margin $m_r(s,t;L) = 2L - d_{G_r}(s,t)$.

**Proposition 1** (Attribute-response reach)**.** *Holding $G_r$ and the geometry fixed, the force on atom $t$ is invariant to an attribute perturbation at atom $s$ whenever* $\mathrm{dist}_{G_r}(s,t) > 2L$.

*Proof.* The force decomposes as $F_t = -\sum_{v:t\in B_L(v)} \partial\varepsilon_v/\partial r_t$. If a summand also contained the perturbed attribute $z_s$, then the triangle inequality would give $\mathrm{dist}_{G_r}(s,t) \leq 2L$. Therefore no contributing term can depend on $s$ beyond the boundary. □

**Corollary 2** (Coordinate force-response reach)**.** *Assume that the complete periodic edge/image set is constant on an open neighborhood of $x$, and that each local energy is twice continuously differentiable there. Then*

$$\mathrm{dist}_{G_r}(s,t) > 2L \quad \Longrightarrow \quad \frac{\partial F_t}{\partial r_s} = -\frac{\partial^2 E_\theta}{\partial r_t\, \partial r_s} = 0.$$

*The centered finite difference is also exactly zero if the entire path $x + \tau e_{s,j}$, $|\tau| \leq \delta$, remains in that fixed-graph neighborhood.*

The same local-energy sum proves the corollary: a mixed derivative requires one center $v$ with both $s, t \in B_L(v)$, which implies graph distance at most $2L$. Appendix A gives the full proof and the corresponding implementation checks.

The proposition separates two questions. A pair beyond $2L$ is structurally excluded; a pair within $2L$ has an available path whose learned use remains an empirical question. The experiment therefore measures the response after the path becomes available.

### 3.3 Cross reach while controlling size and supervision

Each reach matrix contains a base-depth model, a width-expanded model at the same depth, and a deeper base-width model. Depth and cutoff can change the support of the response; width cannot. We therefore choose cutoffs from a label-blind topology analysis so that the same target shell crosses the $2L$ boundary, and we parameter-match the width control to the deeper model. All variants otherwise share the split, structures, optimizer, loss family, and evaluation code.

The evaluation rule combines four pieces of evidence. First, the base and wide models verify the predicted numerical zero when the response is excluded. Second, a graph-permitted model is assessed on the held-out DFT response endpoint and against the zero-response baseline. For $Au_2$/MgO, the separate model-only identity sensitivity is compared with the unpaired DFT group contrast as a measure of alignment; for Ag–Pd, DFT directly supplies the paired force-constant block. Third, the cutoff–depth interaction must be positive, showing that the intervention preferentially helps the model moved across the boundary. Fourth, response evidence must coexist with held-out predictive accuracy, pairing the remote-response measurement with the standard force endpoint.

Ag–Pd additionally separates three training signals while retaining the same topology and evaluation set: aggregate energy–force labels, a mixture of aggregate and harmonic-displacement labels, and harmonic-only force supervision. This staged design tests a second necessary condition after graph permission: whether the training signal identifies the response being evaluated. Seed-specific endpoints quantify optimization variation; 10,000 structure-bootstrap draws quantify held-out response uncertainty without treating Cartesian components as independent samples. Complete definitions and proof extensions are provided in Appendix A. These bootstrap intervals condition on the fitted ensemble and do not represent uncertainty over a population of training runs.

## 4 Results

### 4.1 Label-blind graph analysis identifies two exact boundaries

No trained-model score is used to choose a source–target pair or cutoff. In $Au_2$/MgO, each Au atom's nearest Al dopant is 5–6 hops away at 3.5 Å and 3–4 hops away at 4.5 Å. In Ag–Pd, the fixed 120-pair shell moves from exactly three hops at 3.0 Å to exactly two at 5.0 Å. At 3.0 Å, all 720 source/axis/sign checks preserve the complete periodic edge/image set, with a minimum cutoff margin of 0.06596 Å. At 5.0 Å, hop classes and local balls remain fixed but some periodic edge images change, so the strict coordinate-null result is applied only at 3.0 Å. These boundaries support separate tests of depth, cutoff, parameter-matched width, and training signal rather than an observational comparison of differently sized models.

### 4.2 Depth crosses the boundary; matched width does not

The 5,000-structure $Au_2$/MgO corpus is split into 4,000/500/500 train/validation/test structures before model evaluation. At the fixed 3.5 Å cutoff, every Au atom in all 2,500 doped structures is five or six graph hops from its nearest Al dopant. Proposition 1 excludes an Al-identity contribution to the Au force for two interaction blocks but does not exclude one for three.

D2-base has 102,320 parameters; D2-wide and D3-base have 165,222 and 165,072, respectively, a 0.091% gap between the width and depth controls. Across architecture variants, each structure's cell, cutoff, and complete atomic input are unchanged; the target is a nonlocal charge-transfer response rather than finite-size electrostatic convergence. All variants are trained from scratch with the same structures, loss, cutoff, batch size four, and a maximum of 200,000 optimizer updates, with early-stopping patience 30. The test set is evaluated once for each of three planned training runs.

The held-out DFT group contrast is negative in all four distance bins, and every 95% stratified structure-bootstrap interval excludes zero. These bins contain 148 of the 500 test structures (77 doped and 71 undoped); the model-only paired curve uses the 77 doped structures, whereas ordinary

force metrics use all 500. This reference target is shown directly in both response figures below and is tabulated in Appendix B.

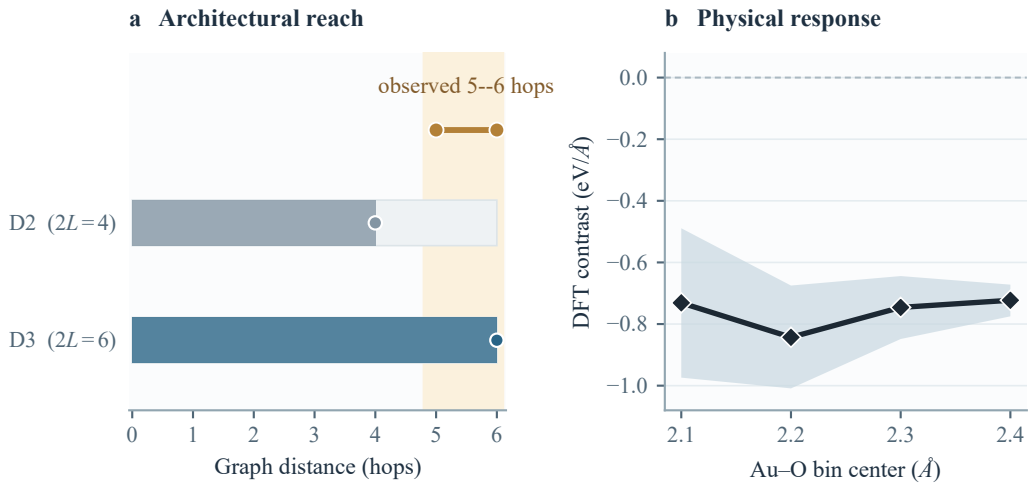


Figure 2: **The target crosses the D2–D3 reach boundary.** The 5–6-hop shell lies outside D2 support and inside D3 support. Held-out DFT contrasts are shown by distance bin; the shaded region is a 95% structure-bootstrap interval.

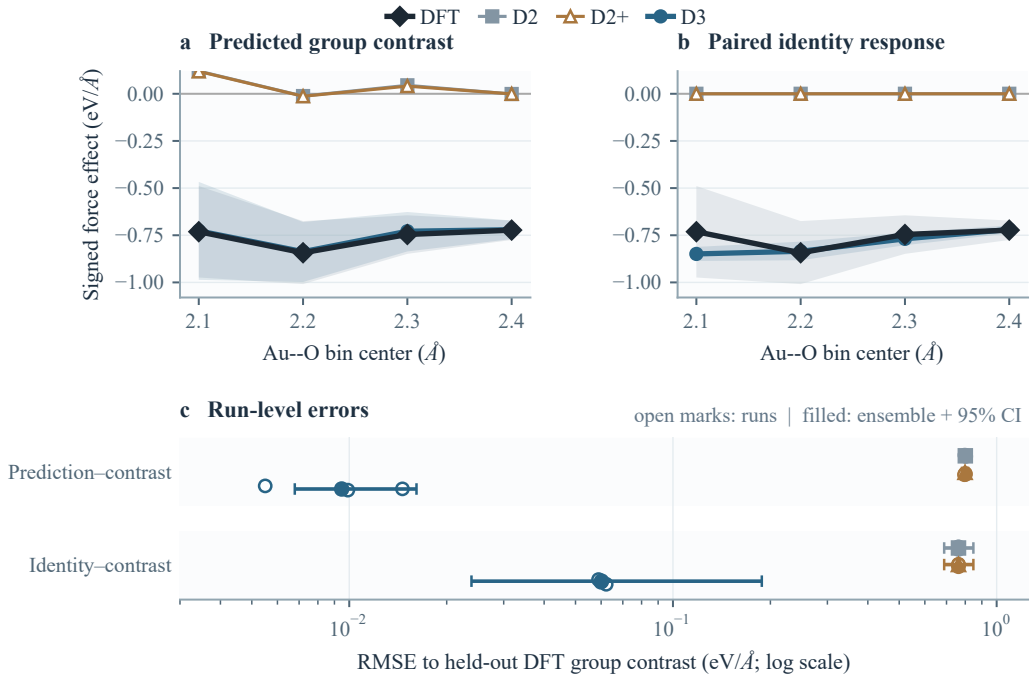


Figure 3: **Prediction and identity sensitivity across graph reach.** D3 follows the held-out DFT contrast and aligned Al-to-Mg counterfactual; both D2 controls remain near zero across three runs. In panel c, open marks show individual runs and filled marks show fitted-ensemble endpoints with 95% structure-bootstrap intervals.

### 4.3 DEPTH IMPROVES PREDICTION AND ACTIVATES ALIGNED IDENTITY SENSITIVITY

On the primary four-bin contrast, D3-base attains 0.0095 eV/Å RMSE (95% structure-bootstrap interval [0.0068, 0.0161]), compared with 0.8002 [0.7833, 0.8176] for D2-base and 0.7985 [0.7817, 0.8159] for D2-wide. The D2-base-minus-D3 and D2-wide-minus-D3 improvements are 0.7907 [0.7726, 0.8055] and 0.7891 [0.7712, 0.8038] eV/Å. All 10,000 structure-bootstrap differences are positive, and the ordering holds in every matched training run. D2-wide differs from D2-base by only 0.0017 eV/Å despite matching D3-base in parameter count. Within this matched design, parameter count alone therefore does not explain the improvement.

The paired identity intervention supplies the support-sensitive check. Across all 250 doped test structures and every training run, the largest absolute D2 model sensitivity is only $1.15 \times 10^{-5}$ eV/Å for D2-base and $1.91 \times 10^{-5}$ eV/Å for D2-wide. Their ensemble identity–contrast alignment RMSE is 0.7622 and 0.7622 eV/Å, essentially the magnitude of the held-out DFT group contrast. D3-base instead returns the same negative sign as that contrast in all four bins and reduces the alignment RMSE to 0.0602 [0.0239, 0.1882] eV/Å. Every conditional structure-bootstrap difference and all three matched runs favor D3.

Secondary controls agree with the primary endpoints. Mean per-Au force-component MAE is 0.0105 eV/Å for D3-base versus 0.0944 and 0.0940 for the two D2 models; within-group relative-energy RMSE is 0.445 versus 0.694 and 0.827 meV/atom. Together, prediction and the paired intervention identify the same mechanism: the architecture with a permitted path improves the held-out group contrast and expresses an aligned model identity sensitivity, whereas additional width at the excluded depth does not.

The remote-response difference does not arise from a collapsed short-range force map. On 250 undoped test structures, both D2 controls retain Pearson correlation at least 0.9971 after geometry-wise centering. Their centered RMSEs are 0.0665 and 0.0725 eV/Å, while D3 reduces the same descriptive error to 0.0327 eV/Å.

Table 1: **All nine MACE fits.** Contrast RMSE compares each predicted doped-minus-undoped curve with the held-out DFT group contrast. Identity–contrast RMSE compares the model-only fixed-geometry Al→Mg sensitivity with that unpaired DFT contrast; it is not error against a paired DFT target. Force quantities are in eV/Å and relative-energy RMSE is in meV/atom.

| Variant | Seed | Contrast RMSE | Identity–contrast RMSE | Au-force MAE | Rel.-energy RMSE |
|---|---|---|---|---|---|
| D2-base | S1 | 0.8005 | 0.7622 | 0.0949 | 0.721 |
| D2-base | S2 | 0.7975 | 0.7622 | 0.0940 | 0.671 |
| D2-base | S3 | 0.8027 | 0.7622 | 0.0944 | 0.688 |
| D2-wide | S1 | 0.8005 | 0.7622 | 0.0940 | 0.835 |
| D2-wide | S2 | 0.8000 | 0.7622 | 0.0938 | 0.823 |
| D2-wide | S3 | 0.7952 | 0.7622 | 0.0941 | 0.824 |
| D3-base | S1 | 0.0146 | 0.0589 | 0.0107 | 0.434 |
| D3-base | S2 | 0.0055 | 0.0622 | 0.0108 | 0.433 |
| D3-base | S3 | 0.0099 | 0.0599 | 0.0099 | 0.467 |

### 4.4 CUTOFF CROSSING PROVIDES A SECOND INTERVENTION

The fixed-cutoff comparison establishes that depth crosses the boundary. To test whether crossing the same support boundary through a distinct architectural intervention yields the predicted improvement, we trained a second matrix with cutoffs 3.5 and 4.5 Å, the same D2-base, D2-wide, and D3-base variants, and three new matched runs per cell. The larger cutoff moves each Au atom's nearest Al dopant within D2 reach without changing depth or channel width. It also changes graph density, neighbor and radial features, and potentially optimization, so we interpret it together with the cleaner fixed-cutoff depth–width comparison.

Table 2 shows the resulting intervention. At 3.5 Å, both D2 variants retain numerical-zero model identity sensitivity, whereas D3 has the same sign as the held-out DFT group contrast in all four bins. At 4.5 Å, D2-base and D2-wide produce that group-aligned sign in every bin and reduce

predictive RMSE to 0.3500 and 0.3712 eV/Å, respectively. The response interaction (0.414, 0.457, 0.472 eV/Å) and predictive interaction (0.415, 0.465, 0.459 eV/Å) are positive in every new run. Expanding the D2 cutoff closes 57.0% of the original D2-to-D3 predictive gap, with a 95% structure-bootstrap interval of $[55.2, 58.8]\%$. The width control remains ineffective below the boundary and is no better than D2-base after the boundary is crossed.

Table 2: **$Au_2$/MgO matched interventions.** Depth and cutoff admit the remote dependency; parameter-matched width below the boundary does not. The identity column reports model-only fixed-geometry Al→Mg sensitivity. Its alignment RMSE is measured against the held-out DFT doped-minus-undoped group contrast, not a paired DFT target. Predictive RMSE is in eV/Å.

| Model | cutoff | parameters | reach | model-only identity | predictive RMSE |
|---|---|---|---|---|---|
| D2-base | 3.5 Å | 102,320 | excluded | max $1.15 \times 10^{-5}$ | 0.8002 |
| D2-wide | 3.5 Å | 165,222 | excluded | max $1.91 \times 10^{-5}$ | 0.7985 |
| D3-base | 3.5 Å | 165,072 | permitted | aligned sign, 4/4 bins | **0.0095** |
| D2-base | 4.5 Å | 102,320 | permitted | aligned sign, 4/4 bins | **0.3500** |
| D2-wide | 4.5 Å | 165,222 | permitted | align. RMSE 0.3450 | 0.3712 |
| D3-base | 4.5 Å | 165,072 | permitted | align. RMSE 0.0613 | 0.0062 |
| D2 cutoff-induced gap closure | | | | **57.0%** $[55.2, 58.8]\%$ | |

The D2 endpoints after cutoff expansion are stable across runs. At 4.5 Å, D2-base has predictive RMSE $0.3500 \pm 0.0292$ and identity–contrast alignment RMSE $0.3234 \pm 0.0248$ eV/Å; the corresponding D2-wide values are $0.3712 \pm 0.0064$ and $0.3450 \pm 0.0077$. D2-base remains within 5% of D2-wide in all three matched runs, while D3 retains predictive RMSE $0.0062 \pm 0.0016$ eV/Å.

### 4.5 The same exclusion reappears in a continuous-filter architecture

A separately implemented SchNet-style continuous-filter model (Schütt et al., 2017) tests whether the boundary survives a change of architecture. It uses scalar one-hop interaction blocks and a node-wise atomic-energy readout, without MACE's equivariant tensor products or product basis. S2-base has 10,145 parameters, S2-wide has 14,664, and S3-base has 14,881; the width–depth parameter gap is 1.46%. All nine runs use the same 3.5 Å graph, fixed 4,000/500/500 split, loss, and endpoints as the primary $Au_2$/MgO study.

Both S2 variants return an exactly zero float64 model identity sensitivity in every test structure. S3 returns the same negative sign as the held-out DFT group contrast in all four bins for 3/3 seeds. Its fitted-ensemble identity–contrast alignment RMSE is 0.128 eV/Å, compared with 0.762 for S2-base and 0.762 for S2-wide; its predictive RMSE is 0.071, compared with 0.806 and 0.805 eV/Å. Every matched-seed comparison favors S3. Conditional structure-bootstrap intervals for the S2-minus-S3 improvement are $[0.444, 0.787]$ eV/Å for the alignment endpoint and at least $[0.695, 0.772]$ for prediction. Ordinary force RMSE also falls from 0.066/0.065 to 0.030 eV/Å. Across the three runs, S3 identity–contrast alignment RMSE spans 0.085–0.164 and predictive RMSE spans 0.038–0.110 eV/Å; the S2 alignment error remains 0.762 in every run.

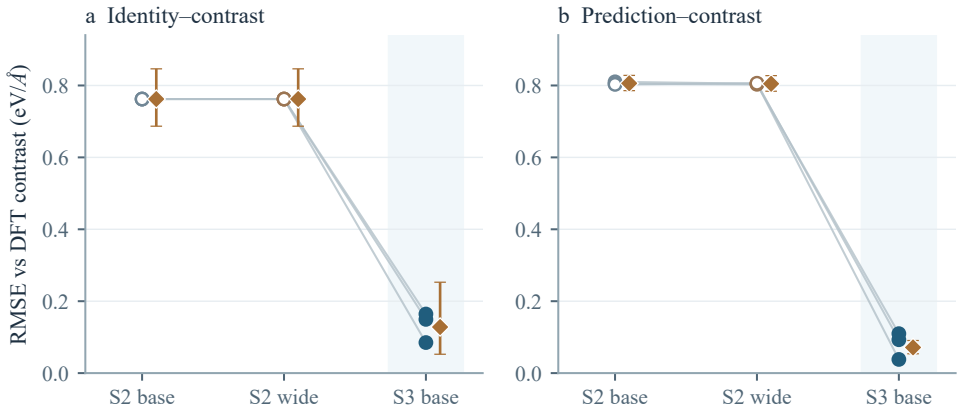


Figure 4: **SchNet recovers the depth-over-width response advantage.** Lines connect matched runs; diamonds show fitted-ensemble estimates with 95% structure intervals. S3 is compared with base and width-matched S2 controls.

Table 3: **Complete SchNet-style extension.** Identity–contrast RMSE compares the model-only fixed-geometry Al-to-Mg sensitivity with the held-out DFT doped-minus-undoped group contrast; it is not error against a paired DFT target. All RMSE values are in eV/Å, and max identity is the largest fixed-geometry Al-to-Mg effect.

| Variant | seed | parameters | force RMSE | ID–contrast RMSE | predictive RMSE | max identity |
|---|---|---|---|---|---|---|
| S2-base | 20260901 | 10,145 | 0.0665 | 0.7622 | 0.8102 | 0.00e+00 |
| S2-base | 20260902 | 10,145 | 0.0655 | 0.7622 | 0.8049 | 0.00e+00 |
| S2-base | 20260903 | 10,145 | 0.0653 | 0.7622 | 0.8030 | 0.00e+00 |
| S2-wide | 20260901 | 14,664 | 0.0648 | 0.7622 | 0.8050 | 0.00e+00 |
| S2-wide | 20260902 | 14,664 | 0.0649 | 0.7622 | 0.8032 | 0.00e+00 |
| S2-wide | 20260903 | 14,664 | 0.0643 | 0.7622 | 0.8063 | 0.00e+00 |
| S3-base | 20260901 | 14,881 | 0.0307 | 0.0850 | 0.0380 | 9.49e-01 |
| S3-base | 20260902 | 14,881 | 0.0305 | 0.1644 | 0.1097 | 1.12e+00 |
| S3-base | 20260903 | 14,881 | 0.0291 | 0.1494 | 0.0928 | 1.07e+00 |

### 4.6 Ag–Pd separates structural support from learned fidelity

In a second chemistry, label-blind graph analysis places every selected Ag–Pd pair outside D1 support at 3.0 Å. Float64 evaluation bounds the largest aggregate-label response by $2.37 \times 10^{-12}$ eV/Å$^2$, as predicted by the exact exclusion theorem. For a predicted force-constant block $\widehat{J}$ and DFT block $J$, define $\text{closure} = 1 - \text{RMSE}(\widehat{J}, J)/\,\text{RMSE}(0, J)$; negative closure is worse than predicting zero. With aggregate labels, the permitted D1@5 model lowers force RMSE to 13.6 meV/Å yet has closure $-1134.9 \pm 302.6\%$, so aggregate accuracy does not establish response fidelity.

Structure-disjoint harmonic supervision produces positive mean direction and block-ranking signals in selected permitted cells, but all mixed-stage closures remain negative. The cutoff–depth interaction is positive in 2/3 runs, while its structure-bootstrap interval $[-38.7, 486.8]$ includes zero. Under harmonic-only training, the permitted D2@3 response converges to numerical zero in all three seeds. A post hoc 8.0 Å direct edge raises the response-norm ratio to 0.851, but closure is $-30.0 \pm 33.3\%$ and one seed has negative cosine. Appendix C gives the complete stage and seed results. Reach permits a response; training must still recover its direction and scale.

Table 4: **Ag–Pd separates structural status, direction, and magnitude.** Direction entries are signed cosine/block-norm correlation; magnitude entries are response-norm ratio/closure. The 3.0 Å structural null is prospective; supervision changes are exploratory and the direct-edge row is post hoc. Values are three-seed means; every seed is reported in the Supplement.

| Stage | evidence role | structural status | direction/block alignment | magnitude/closure |
|---|---|---|---|---|
| Aggregate E/F | prospective | D1@3 $< 2.4 \times 10^{-12}$ | D2@3: 0.01/-0.40; D1@5: -0.19/0.67 | D2@3: 9.94/-904.1%; D1@5: 12.12/-1134.9% |
| Mixed harmonic | exploratory | D1@3 $< 2.4 \times 10^{-12}$ | D2@3: 0.18/0.46; D1@5: 0.17/0.75 | D2@3: 1.88/-102.7%; D1@5: 4.18/-312.7% |
| Harmonic-only | exploratory | D1@3 $< 2.4 \times 10^{-12}$ | D2@3: -0.03/-0.00; D1@5: 0.15/0.67 | D2@3: 0.00/0.0%; D1@5: 0.01/0.2% |
| Direct edge, D1@8 | post hoc | one-hop permitted | 0.05/0.61 | 0.85/-30.0% |

## 5 Discussion

**A negative reach margin identifies structural impossibility.** For the finite-reach local-energy class, an attribute–force or coordinate–force cross term vanishes beyond $2L$ graph hops when the periodic graph remains fixed. Source-level MACE analysis and the 3.0 Å Ag–Pd test verify these conditions; float64 re-evaluation observes the predicted numerical zero in a second chemistry.

The $Au_2$/MgO interventions connect this theorem to predictive behavior. At 3.5 Å, depth crosses the boundary, improves the held-out DFT contrast, and activates aligned Al-to-Mg sensitivity, whereas a width control matched within 0.091% in parameter count does not. A larger D2 cutoff closes 57.0% of the predictive gap (interval $[55.2, 58.8]\%$), although it also changes graph density and radial features. An independent continuous-filter model reproduces the exclusion and depth–width ordering in every seed, showing that restored support improves response without determining held-out accuracy.

**Where the result applies.** The theorem excludes global, reciprocal-space, charge-equilibration, and other nonlocal channels, and its coordinate corollary requires a stable periodic graph. Training seeds measure optimization variation; structure bootstraps condition on the fitted ensemble, with 1,080 Ag–Pd entries nested in 120 blocks from three equilibria. In this scope, a negative margin rejects a local architecture before training; a nonnegative margin calls for held-out tests of response direction and scale. This two-stage comparison remains meaningful across message functions and prevents aggregate error from concealing an excluded or poorly learned interaction.

**A response-specific model choice.** The comparison begins by fixing a physical source, target, and response before training, then computing the reach margin on label-blind geometries. A negative margin at a geometry rules out that local response there, regardless of its aggregate force RMSE. For a nonnegative margin, the held-out response must be compared with the zero-response baseline in direction, scale, and error, while ordinary force error is reported separately. Coordinate probes additionally require stability of the complete periodic edge/image set along the displacement path. Structure-bootstrap intervals and training-seed variation answer different uncertainty questions and should remain separate. This ordering distinguishes a need for graph access from a failure to learn an available interaction without selecting the architecture on its test response.

**Reproducibility statement.** Appendices A–C give the proofs, assumptions, splits, optimization settings, seed-level results, and statistical units. The anonymous supplement provides configurations and evaluated endpoints for all 93 independent fits.

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

## REFERENCES


Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022. doi: 10.52202/068431-0830.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5.

Andreas Burger, Luca Thiede, Nikolaj Rønne, Varinia Bernales, Nandita Vijaykumar, Tejs Vegge, Arghya Bhowmik, and Alan Aspuru-Guzik. Shoot from the HIP: Hessian interatomic potentials without derivatives, 2025. URL https://arxiv.org/abs/2509.21624.

Bingqing Cheng. Latent ewald summation for machine learning of long-range interactions. *npj Computational Materials*, 11:80, 2025. doi: 10.1038/s41524-025-01577-7. URL https://doi.org/10.1038/s41524-025-01577-7.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Liò, and Michael M. Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7865–7885. PMLR, 2023. URL https://proceedings.mlr.press/v202/di-giovanni23a.html.

Shahaf E. Finder, Ron Shapira Weber, Moshe Eliasof, Oren Freifeld, and Eran Treister. Improving the effective receptive field of message-passing neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17203–17220. PMLR, 2025. URL https://proceedings.mlr.press/v267/finder25a.html.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17875–17893. PMLR, 2025. URL https://proceedings.mlr.press/v267/fu25h.html.

Marius Herbold and Jörg Behler. A hessian-based assessment of atomic forces for training machine learning interatomic potentials. *The Journal of Chemical Physics*, 156(11):114106, 2022. doi: 10.1063/5.0082952. URL https://doi.org/10.1063/5.0082952.

Tsz Wai Ko, Jonas A. Finkler, Stefan Goedecker, and Jörg Behler. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nature Communications*, 12:398, 2021. doi: 10.1038/s41467-020-20427-2.

Teddy Koker, Abhijeet Gangan, Mit Kotak, Jaime Marian, and Tess Smidt. PFT: Phonon fine-tuning for machine learned interatomic potentials. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.07742. arXiv:2601.07742.

Arthur Kosmala, Johannes Gasteiger, Nicholas Gao, and Stephan Günnemann. Ewald-based long-range message passing for molecular graphs. In *Proceedings of the 40th International Conference*

*on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17544–17563. PMLR, 2023. URL https://proceedings.mlr.press/v202/kosmala23a.html.

Pingbing Ming and Han Wang. Why multi-layer message passing works: Completeness theory for graph neural network interatomic potentials, 2026. URL https://arxiv.org/abs/2609.00528.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, 2023. doi: 10.1038/s41467-023-36329-y.

Conrad W. Rosenbrock, Konstantin Gubaev, Alexander V. Shapeev, Livia B. Pártay, Noam Bernstein, Gábor Csányi, and Gus L. W. Hart. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *npj Computational Materials*, 7:24, 2021. doi: 10.1038/s41524-020-00477-2. URL https://doi.org/10.1038/s41524-020-00477-2.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 991–1001. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7UmjRGzp-A.

Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Sauceda, and Klaus-Robert Müller. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12:7273, 2021. doi: 10.1038/s41467-021-27504-0.

# Graph Reach Sets the Remote-Response Limit of Local Interatomic Potentials

## Supplementary Material

## A Graph-Reach Theory and Implementation

Let $G_r = (V, E)$ be the undirected periodic neighbor graph returned by the fixed neighbor builder. For $v \in V$, $B_L(v)$ denotes its closed $L$-hop ball. The configured MACE models have a local-energy decomposition

$$E_\theta(x, z) = \sum_{v \in V} \varepsilon_{v,\theta}\big(x_{B_L(v)}, z_{B_L(v)}\big),$$

where each interaction block adds at most one graph hop and the product and readout blocks are nodewise. The implementation checks below describe how these assumptions are realized.

**Proposition 3** (Attribute-response reach)**.** *Fix $x$ and $G_r$. If every local energy depends only on its $L$-hop ball, then, for any atomic attribute $z_s$,*

$$d_{G_r}(s, t) > 2L \quad \Longrightarrow \quad \frac{\partial F_t}{\partial z_s} = 0.$$

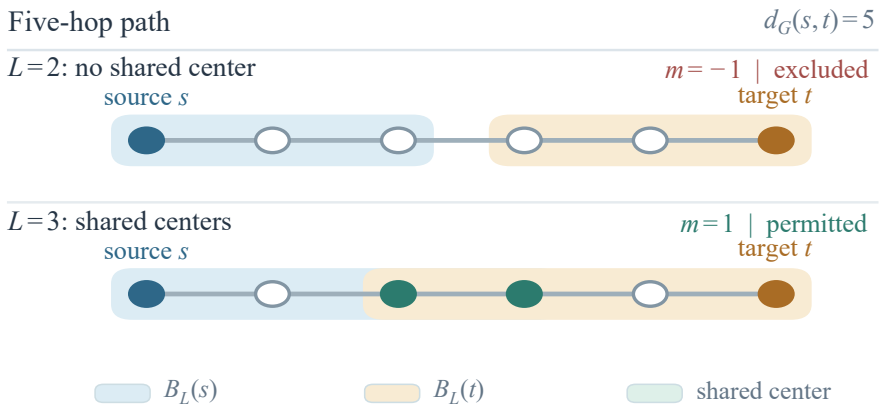


Figure 5: **What the reach solution looks like on a five-hop path.** The shaded source and target neighborhoods contain no common local-energy center at $L = 2$; at $L = 3$ they overlap. This is a schematic of the support proof, not an observed atomic geometry. Permission at $L = 3$ does not imply a nonzero or accurate learned response.

*Proof.* Because $F_t = -\partial E_\theta / \partial r_t$,

$$F_t = -\sum_{v:\, t \in B_L(v)} \frac{\partial \varepsilon_{v,\theta}}{\partial r_t}.$$

A summand can depend on $z_s$ only if $s \in B_L(v)$. Joint membership $s, t \in B_L(v)$ implies $d_{G_r}(s,t) \le 2L$ by the triangle inequality. No summand depends on $z_s$ when the stated strict inequality holds. □

**Corollary 4** (Coordinate force-response reach)**.** *Suppose $G_r(x')$ and its periodic image assignments are constant on an open neighborhood $U$ of $x$, each $\varepsilon_{v,\theta}$ is twice continuously differentiable on $U$, and the local-energy support above holds. Then*

$$d_{G_r}(s,t) > 2L \quad \Longrightarrow \quad \frac{\partial F_t}{\partial r_s} = -\frac{\partial^2 E_\theta}{\partial r_t\, \partial r_s} = 0.$$

*If the complete path $x + \tau e_{s,j}$, $|\tau| \le \delta$, lies in $U$, then the centered finite difference*

$$\frac{F_t(x + \delta e_{s,j}) - F_t(x - \delta e_{s,j})}{2\delta} = 0$$

*in exact arithmetic.*

*Proof.* Use the displayed decomposition of $F_t$. A mixed coordinate derivative can be nonzero only if one local term contains both $r_t$ and $r_s$, which again requires $s, t \in B_L(v)$ and hence $d_{G_r}(s,t) \le 2L$. Under the path condition, the derivative is identically zero at every point between the two finite-difference endpoints, so both endpoint forces have the same relevant component. □

Both results are one-way support statements. A nonnegative reach margin $2L - d_G(s,t)$ removes this obstruction but neither guarantees a nonzero response nor its correct direction or magnitude.

Figure 5 makes the support argument concrete for a five-hop source–target path, the nearest-dopant distance observed in the $Au_2$/MgO topology analysis. The drawing depicts a graph witness rather than atom coordinates or a fitted response.

**Architecture and graph-stability conditions.** The experiments use MACE 0.3.17 at source commit `163ef2a7e297b9e3ec442a943ce3104e184c5ccf`. The interaction blocks gather sender features and scatter them to receivers once per block; residual and product operations are nodewise; readouts produce atomic energies that are summed; and the configurations contain no reciprocal or global channel. Numerical support tests independently recover the same one-hop growth.

For Ag–Pd we additionally compare the complete directed edge and periodic image set along every prescribed $\pm 0.01$ Å displacement. At 3.0 Å, all 720 selected source/axis/sign checks preserve the complete graph, with a minimum cutoff margin of 0.06596 Å. This is the setting to which Corollary 4 is applied. At 5.0 and 8.0 Å, the unweighted 1–3-hop balls and selected-pair hop classes remain unchanged, but 245/720 and 49/720 checks, respectively, change at least one periodic edge image. We therefore treat those settings as empirical graph-permitted or direct-edge comparisons and do not transfer the fixed-graph zero theorem to them.

## B $Au_2$/MgO Experiment and Complete Results

**Dataset and prespecified split.** The experiment uses the $Au_2$/MgO structures released with Ko et al. Ko et al. (2021). The 5,000 structures are balanced across doped versus undoped and wetting versus non-wetting groups. A fixed seed, 20260812, assigns 1,000/125/125 structures from each group to training, validation, and test sets, respectively. The split was fixed before any trained-model endpoint was evaluated.

**Graph reach before training.** The neighbor graph uses the minimum-image 3.5 Å cutoff applied identically to all variants. Graph distances are computed from atomic numbers, positions, and periodic cells only; energies and forces are not read. For all 2,500 doped structures, the sorted pair of Au-to-nearest-Al graph distances is $(5, 5)$ or $(5, 6)$, with 1,250 structures in each class. A random-weight implementation test changes the remote Al identity to Mg at fixed geometry and obtains exactly zero D2 Au-force model identity sensitivity in float64, while exposing a nonzero D3 computational path. This test checks the implementation of Proposition 1; it is not evidence about a trained model.

**Correction of the excluded pretrained comparator.** Exact deserialization of the two legacy MACE-MP weight files previously described as "small" and "medium" shows that both have a 6.0 Å cutoff and two interaction blocks; they contain 3,847,696 and 4,688,656 parameters, respectively. They therefore differ in representation size, not nominal reach, and no outcome from this pair is used as reach evidence.

**Models and optimization.** All models use MACE with eight radial basis functions, five cutoff basis functions, maximum angular degree three, correlation order three, an Adam optimizer, learning rate 0.01, weight decay $5 \times 10^{-7}$, EMA decay 0.99, and energy/force loss weights 1/100. D2-base uses 32 channels, two interaction blocks, and 102,320 parameters. D2-wide uses 46 channels, two blocks, and 165,222 parameters. D3-base uses 32 channels, three blocks, and 165,072 parameters. Each variant receives at most 200,000 optimizer updates at batch size four, early-stopping patience 30, and seeds 20260812–20260814. Training used NVIDIA A40 GPUs under Python 3.10.16, PyTorch 2.6.0+cu124, CUDA 12.4, and upstream MACE 0.3.17 at commit `163ef2a7e297b9e3ec442a943ce3104e184c5ccf`.

**Signed-force estimand.** Let $(a^*, o^*)$ be the globally closest Au–O pair under the periodic minimum-image convention, and let $\hat{d}(x) = (r_{o^*} - r_{a^*})/\|r_{o^*} - r_{a^*}\|$ point from that Au atom toward that oxygen. The per-structure response is the net Au force projected onto this single direction,

$$S(x) = \left( \sum_{a \in \mathrm{Au}} F_a(x) \right) \cdot \hat{d}(x).$$

The primary contrast subtracts the undoped group mean from the doped group mean within the four non-wetting globally closest Au–O-distance bins $[2.05, 2.15)$, $[2.15, 2.25)$, $[2.25, 2.35)$, and $[2.35, 2.45)$ Å. Their doped/undoped counts are 21/9, 15/22, 19/15, and 22/25, respectively, so the primary contrast uses 148 test structures (77 doped and 71 undoped). The co-primary model measure

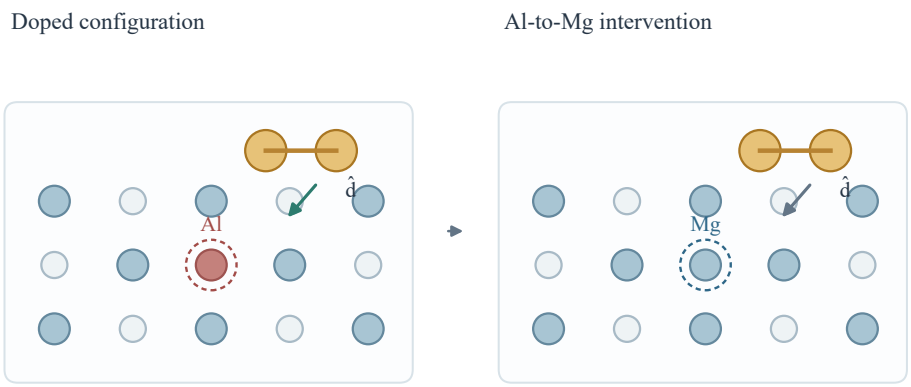


Figure 6: **The paired model intervention in a representative surface schematic.** The Au dimer, lattice, cell, and projection direction stay fixed while the Al source identity is replaced by Mg. Gold, blue, pale, and red identify Au, Mg, O, and Al, respectively. Positions are illustrative; the quantitative result is the held-out curve in Fig. 7.

is paired within each of those 77 doped test structures: all Al identities are replaced by Mg while positions, cell, and graph remain fixed. Because the Au and O coordinates are unchanged, the same $\hat{d}(x)$ is used, and the resulting $S(x) - S(x_{\mathrm{Al}\to\mathrm{Mg}})$ is recorded. Because DFT was not evaluated for this fixed-geometry replacement, the quantity is a model identity sensitivity rather than a paired physical response.

Figure 6 visualizes this intervention. It shows the identity change and the shared projection direction, without implying that the diagram reproduces any particular structure in the dataset.

**Uncertainty and reporting.** Reference and predictive contrasts use 10,000 stratified bootstrap replicates with seed 20260812. The resampling unit is the held-out structure in eight fixed bin-by-doping strata (four bins crossed with two doping groups); the paired model measure uses the four doped-bin strata. Shared doped draws link its identity–contrast alignment RMSE to the DFT group contrast. The plotted ensemble averages predictions across the three training runs before endpoint calculation. The structure-level interval estimates held-out sampling variation conditional on those fitted models and does not include training-population uncertainty; the run-specific endpoints show optimization variation separately. Energy comparisons across doped and undoped compositions are omitted because their atomic compositions differ.

**Evaluation criterion.** We regard depth as consequential when D3-base reduces held-out contrast error relative to both D2-base and D2-wide, while its model-only identity sensitivity has the same sign as the held-out DFT group contrast. The comparison also asks whether parameter-matched width can reproduce the gain and whether the effect extends beyond a short-range force-error control. These questions are evaluated on the cutoffs, depths, and widths fixed for the experiment.

**Ensemble curves and uncertainty.** The complete nine-fit matrix appears in main-text Table 1; the Appendix adds bin-level curves and uncertainty calculations. Table 5 provides the exact displayed values underlying Figs. 3 and 7. D3 tracks the DFT group-contrast curve across all four distance bins. Under the paired Al→Mg model intervention, its identity sensitivity also has the same sign as the held-out DFT group contrast in every bin. By contrast, the largest absolute individual D2 response across all doped test structures and seeds is only $1.15 \times 10^{-5}$ eV/Å for D2-base and $1.91 \times 10^{-5}$ eV/Å for D2-wide.

The ensemble primary error is 0.0095 [0.0068, 0.0161] eV/Å for D3-base, compared with 0.8002 [0.7833, 0.8176] and 0.7985 [0.7817, 0.8159] for the two D2 controls. The corresponding control-minus-D3 differences are 0.7907 [0.7726, 0.8055] and 0.7891 [0.7712, 0.8038] eV/Å; all 10,000 bootstrap draws are positive for each comparison. For the identity–contrast alignment endpoint,

Table 5: Four-bin signed-force values reported in Figs. 3 and 7 (eV/Å). DFT and predictive columns are doped-minus-undoped group contrasts; the final column is the model-only fixed-geometry Al→Mg identity sensitivity of D3-base. No paired DFT counterfactual is available. The D2 identity sensitivities are numerical zero and are bounded separately in the text.

| Au–O bin (Å) | DFT group | D2-base | D2-wide | D3-base | D3 identity |
|---|---|---|---|---|---|
| $[2.05, 2.15)$ | -0.731 | 0.120 | 0.119 | -0.727 | -0.849 |
| $[2.15, 2.25)$ | -0.843 | -0.012 | -0.013 | -0.837 | -0.835 |
| $[2.25, 2.35)$ | -0.746 | 0.045 | 0.041 | -0.729 | -0.769 |
| $[2.35, 2.45)$ | -0.723 | -0.001 | -0.001 | -0.721 | -0.719 |

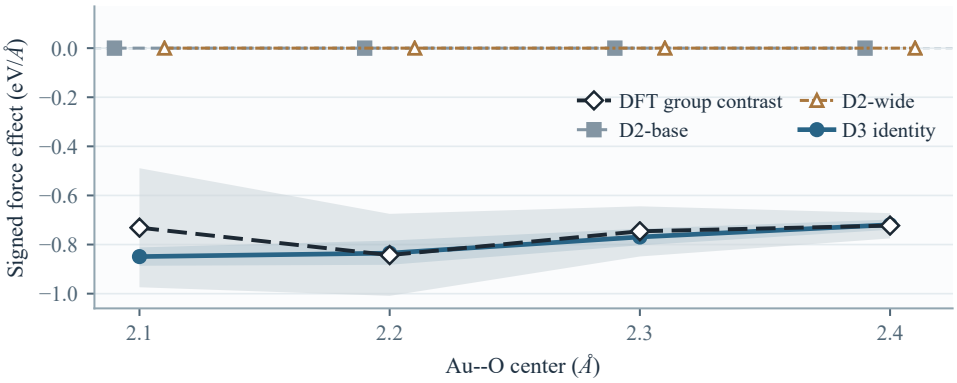


Figure 7: **D3 develops aligned identity sensitivity.** Fixed-geometry Al-to-Mg curves are compared with the held-out DFT contrast across four bins. Bands are 95% stratified structure-bootstrap intervals.

the D2-minus-D3 differences are 0.7021 [0.5152, 0.7858] and 0.7021 [0.5152, 0.7858] eV/Å, again positive in every draw.

**Secondary controls and final decision.** Across seeds, mean per-Au force-component MAE is 0.0944, 0.0940, and 0.0105 eV/Å for D2-base, D2-wide, and D3-base. Mean within-group relative-energy RMSE is 0.694, 0.827, and 0.445 meV/atom. These secondary measures favor D3 but were not assigned a post-hoc threshold. D3 satisfies the decision rule because it beats both D2 controls on the predictive endpoint and its model identity sensitivity has the same sign as the held-out DFT group contrast in all four bins.

**Prospective cutoff–depth intervention.** After fixing the original 3.5 Å depth comparison, we declared a second matrix before reading its trained-model outcomes. It uses the same 4,000/500/500 structure split, cutoffs 3.5 and 4.5 Å, D2-base, D2-wide, and D3-base, and three new training runs (20260822–20260824) in each of the six cells. In the 4.5 Å graph, the two nearest-dopant hop counts per structure are $(3, 3)$ for 1,249 doped structures and $(3, 4)$ for 1,251; thus each Au atom's nearest Al dopant is within the $2L$ force cross-term boundary for D2. This comparison focuses on nearest-dopant targets: 1,250 of the 15,000 specific Al–Au pairs remain five hops apart. All optimization settings remain unchanged. This matrix contains 18 independently trained models and evaluates the same four predictive and model identity-sensitivity bins used in the primary experiment.

For each run, the sensitivity interaction subtracts the cutoff-induced change in D3 model identity-sensitivity magnitude from the corresponding change in D2 magnitude. The predictive interaction analogously subtracts the D3 reduction in four-bin contrast RMSE from the D2 reduction. A positive interaction therefore means that enlarging the cutoff benefits the previously excluded D2 model more than the already permitted D3 model. The prespecified sensitivity interactions are 0.414, 0.457, 0.472, and the predictive interactions are 0.415, 0.465, 0.459; every value is positive.

The point estimate for D2 predictive-gap closure is 57.0%, with a shared-draw structure-bootstrap 95% interval of $[55.2, 58.8]\%$. This result exceeds the chosen 50% lower-limit requirement but not the separate 75% point threshold. We therefore interpret the cutoff intervention as a consistent but partial remedy. It crosses the verified support boundary and activates the group-aligned model identity-sensitivity sign, but it does not erase the remaining D2-to-D3 accuracy gap.

Table 6: **Prespecified cutoff–depth criteria and outcomes.** The intervention satisfies nine of ten criteria. Predictive-gap closure is 57.0%, above the 50% lower-bound criterion and below the 75% point target.

| Criterion | Observed evidence |
|---|---|
| 3.5 Å D2 model identity sensitivity is numerical zero | Maximum individual sensitivity below $10^{-4}$ eV/Å |
| 3.5 Å D3 identity sensitivity aligns in sign with the held-out DFT group contrast | Aligned sign in 4/4 bins for all three runs |
| 4.5 Å D2 has nonzero, sign-aligned model identity sensitivity | D2-base and D2-wide satisfy all bins and runs |
| Identity-sensitivity cutoff–depth interaction is positive | $0.414, 0.457, 0.472$ |
| Predictive cutoff–depth interaction is positive | $0.415, 0.465, 0.459$ |
| Width does not rescue excluded 3.5 Å reach | Both D2 variants remain at numerical zero |
| 4.5 Å D2-base is within 5% of D2-wide | Noninferiority holds in 3/3 runs |
| D2 closes at least 75% of the original predictive gap | Point estimate $57.0\%$ |
| Gap-closure bootstrap lower limit exceeds 50% | 95% interval $[55.2, 58.8]\%$ |
| All trained endpoints are complete | All 18 trained models contribute to the matrix |

**Remote-identity-free force-shape control.** The missing Al response of the two-block models does not imply a collapsed force map. Without refitting, we use the 250 undoped test structures from each run and center reference and predicted net-Au forces separately within the two geometry strata. Both D2 controls retain detailed variation: across the six geometry-by-seed cells, Pearson correlation is at least 0.9971 for D2-base or 0.9973 for D2-wide. Their centered RMSEs are 0.0665 $\pm$ 0.0016 and 0.0725 $\pm$ 0.0064 eV/Å, while D3-base further reduces this descriptive error to 0.0327 $\pm$ 0.0038 eV/Å. Thus, D2 learns structured variation without acquiring the remote Al-to-Au identity term. This descriptive analysis remains separate from the prespecified decision rule.

Table 7: **Post-hoc undoped force-shape analysis.** Centered RMSE is mean $\pm$ seed standard deviation; correlation and slope ranges span both geometry strata and all three seeds. This comparison is descriptive and was not prespecified.

| Variant | Centered RMSE (eV/Å) | Pearson $r$ | OLS slope |
|---|---|---|---|
| D2-base | $0.0665 \pm 0.0016$ | 0.9971–0.9997 | 0.988–1.088 |
| D2-wide | $0.0725 \pm 0.0064$ | 0.9973–0.9997 | 0.987–1.089 |
| D3-base | $0.0327 \pm 0.0038$ | 0.9978–0.9999 | 0.993–1.001 |

## C Further Experiments and Reproducibility

**Ag–Pd structural-null and path-use experiments.**

**Stage-level response details.** With aggregate energy–force labels, D2@3 has closure $-904.1 \pm 818.9\%$ and signed cosine $0.009$; D1@5 has closure $-1134.9 \pm 302.6\%$ and cosine $-0.191$. In the mixed harmonic stage, D2@3 reaches signed cosine $0.178$ and block-norm correlation $0.463$; D1@5 reaches $0.170$ and $0.753$, while the parameter-matched D1-wide@5 reaches $0.229$ and $0.806$. Their mean response-norm ratios are $1.876$, $4.175$, and $4.087$, respectively; closures remain negative ($-102.7\%$, $-312.7\%$, and $-297.9\%$). The cutoff–depth interaction is positive after seed pooling ($153.2$ closure points), but its structure-bootstrap interval in the main text includes zero; with only three equilibrium clusters this is exploratory evidence of partial alignment.

Harmonic-only D1@5 reaches $0.2 \pm 0.2\%$ closure, cosine $0.147$, and block-norm correlation $0.671$, yet its response-norm ratio is only $0.012$. For the D1@8 direct-edge extension, seed cosines are 0.209, 0.227, and $-0.288$, with closures $-19.1\%$, $-3.5\%$, and $-67.4\%$. Direct connectivity can restore response scale without resolving direction or total block error.

Table 8: **Post hoc Ag–Pd direct-edge extension.** The D1-base harmonic-only model moves from two-hop support at 5 Å to a direct edge at 8 Å.

| Cutoff (Å) | run | closure (%) | cosine | norm ratio | block $r$ | force RMSE (meV/Å) |
|---|---|---|---|---|---|---|
| 5.0 | 20260831 | 0.18 | 0.1160 | 0.0168 | 0.6758 | 7.66 |
| 5.0 | 20260832 | 0.38 | 0.3310 | 0.0116 | 0.6309 | 7.09 |
| 5.0 | 20260833 | -0.01 | -0.0065 | 0.0067 | 0.7067 | 7.49 |
| 8.0 | 20260831 | -19.05 | 0.2093 | 0.8884 | 0.6328 | 7.27 |
| 8.0 | 20260832 | -3.50 | 0.2274 | 0.5781 | 0.6941 | 7.68 |
| 8.0 | 20260833 | -67.40 | -0.2876 | 1.0854 | 0.5009 | 7.21 |

We use the open Ag–Pd alloy data and phonon calculations released by Rosenbrock et al. (2021). The aggregate source file contains 1,680 configurations. Of these, 982 carry both DFT energy and force labels and 698 carry neither; no frame carries only one of the two required labels. We exclude the unlabeled frames without imputation and partition the labeled configurations by a UUID-based split with seed 20260824, yielding 796/97/89 train/validation/test structures. Every UUID is unique. A separate geometry comparison over species, positions, cells, and periodicity finds no exact match, at eight-decimal precision, between any of the 20 examined phonon equilibria and any source frame.

The reach comparison was selected without reading a force constant or model prediction. We first computed minimum-image neighbor graphs for the first 20 published phonon supercells over cutoffs from 2.9 to 8.0 Å. These 30–32 atom cells have graph diameter at most three in the meaningful cutoff range, so they do not admit a valid D2/D3 comparison. They do admit a complete D1/D2 boundary on the fixed 6–8 Å metric shell. At a 3.0 Å cutoff, all 120 unordered pairs in this shell are exactly three graph hops apart: D1 excludes every force cross term, whereas the general bound no longer excludes D2. At 5.0 Å, the same pairs are exactly two hops apart, so the bound no longer excludes D1. The latter is the smallest tested cutoff with this D1 reach status in both the 6–7 and 7–8 Å bins. The prescribed $\pm 0.01$ Å displacements preserve this selected-pair hop classification in all 1,440 cutoff/source/axis/sign checks: every low-cutoff pair remains at three hops and every high-cutoff pair at two. This statement concerns hop class, not complete-graph stability. The latter holds in all 720 checks at 3.0 Å but fails in 245 of 720 checks at 5.0 Å, as detailed in the graph-stability checks of Appendix A.

Each stage contains the same 18-cell architectural matrix: two cutoffs, D1-base/D1-wide/D2-base, and three seeds. D1-base and D2-base use 32 channels; label-blind parameter matching selects 74 channels for D1-wide, leaving a 0.49% parameter-count gap to D2-base. The first stage trains on the 796/97/89 aggregate energy–force split with seeds 20260825–20260827 and a 200-epoch ceiling. The second stage adds structure-disjoint harmonic frames to the aggregate split, giving 2,032/277/269 train/validation/test frames, and uses seeds 20260828–20260830 with a 100-epoch ceiling. The third stage is a force specialist trained only on 1,236/180/180 harmonic frames with seeds 20260831–20260833, a learning rate of 0.003, and a 200-epoch ceiling. Across all stages, structures 0–19 form an untouched response set; structures 20–54, 55–59, and 60–64 are assigned to training, validation, and auxiliary test partitions, respectively. No response structure enters optimization or early stopping.

Within a stage, cutoff, interaction count, and the declared width control are the only architectural changes. The original aggregate-label matrix uses the same weighted energy–force objective in every cell. The mixed stage changes only the training distribution, while the specialist stage changes the distribution and uses a force-only objective in every cell. We report the stages separately rather than selecting one retrospectively. Because the training configuration uses float32, the structurally excluded cells of the aggregate-label matrix also receive an independent float64 derivative control; the mixed and specialist response evaluations are performed directly in float64.

For each canonical pair $s < t$, we displace source atom $s$ by $\delta = 0.01$ Å along each Cartesian direction and estimate $J^{\mathrm{coord}}_{t\leftarrow s,j}(x;\delta)$. The DFT target is $J^{\mathrm{DFT}}_{t\leftarrow s,j} = -\Phi^{(:,j)}_{t,s}$ from the archived phonopy force-constant tensor. The primary analysis pools all 1,080 entries from the 120 three-by-three blocks. These pairs occur in three of the 20 examined supercells; the remaining cells are retained for response-set force evaluation but do not contain a 6–8 Å pair under the minimum-image convention. We do not select a component or high-signal subset. The analysis reports signed cosine, zero-baseline-normalized RMSE closure, response-norm ratio, and block-norm correlation; 10,000 bootstrap draws resample the three contributing equilibrium structures and retain every pair within a sampled structure. The same three conditions organize each stage: structural zero from both D1 models below the boundary; positive DFT alignment and improvement over the zero baseline for D2 at 3.0 Å and D1 at 5.0 Å in at least two seeds; and a positive cutoff–depth interaction in at least two seeds and after seed pooling. The auxiliary held-out force endpoint tests transfer to five additional phonon structures. Ordinary held-out errors provide complementary context, while the remote-response measurements carry the mechanistic claim.

The primary matrices test the exact support boundary, where D1 is no longer excluded by the reach bound and the response at 5.0 Å is mediated through a shared energy centre. After the first primary specialist cells were inspected, we added one explicitly post hoc reach-margin extension: D1-base at 8.0 Å for the same three specialist seeds. The original label-blind graph analysis already showed that all 120 pairs become direct neighbours at this cutoff. The extension reuses the unchanged harmonic-only split, loss, stopping rule, response set, and float64 evaluation; no scale or component is fitted after training. Its decision rule requires positive closure and signed cosine in at least two seeds and an improvement in mean closure over the paired D1@5 runs. We report this extension separately from the prospective 18-cell matrices regardless of outcome.

**Precision control at fixed fitted weights.** The native float32 finite difference leaves a residual of order $10^{-3}$ eV/Å$^2$ in structurally excluded cells because it subtracts nearly equal forces. Evaluating the same six fitted models in float64 reduces every forbidden response below $2.37 \times 10^{-12}$ eV/Å$^2$, with a maximum response-norm ratio of $8.73 \times 10^{-11}$. Table 9 reports all paired evaluations. This comparison recovers the structural zero predicted by Proposition 1; the aggregate-label outcomes retain their native-precision values.

Table 9: **Float64 evaluation resolves the structurally forbidden response.** Using the same fitted weights, the D1@3 maximum derivative falls from roughly $10^{-3}$ to $10^{-12}$ eV/Å$^2$.

| Variant | Run | native float32 | float64 | reduction |
|---|---|---|---|---|
| D1-base | 20260825 | $1.25 \times 10^{-3}$ | $2.37 \times 10^{-12}$ | $5.27 \times 10^{8}$ |
| D1-base | 20260826 | $1.04 \times 10^{-3}$ | $2.27 \times 10^{-12}$ | $4.56 \times 10^{8}$ |
| D1-base | 20260827 | $1.17 \times 10^{-3}$ | $2.01 \times 10^{-12}$ | $5.81 \times 10^{8}$ |
| D1-wide | 20260825 | $1.33 \times 10^{-3}$ | $2.04 \times 10^{-12}$ | $6.50 \times 10^{8}$ |
| D1-wide | 20260826 | $9.26 \times 10^{-4}$ | $1.93 \times 10^{-12}$ | $4.81 \times 10^{8}$ |
| D1-wide | 20260827 | $1.17 \times 10^{-3}$ | $1.89 \times 10^{-12}$ | $6.19 \times 10^{8}$ |

**Independent continuous-filter architecture.** The architecture-level extension uses a SchNet-style scalar continuous-filter network (Schütt et al., 2017) implemented independently of MACE's equivariant tensor products. Each interaction forms messages only along the fixed neighbor edges, aggregates them at the receiver, and applies a nodewise residual update. The energy is a sum of nodewise atomic readouts. Consequently, an $L$-interaction instance satisfies the local-energy decomposition in Appendix A; no global state, reciprocal operation, charge equilibration, or graph rewiring enlarges its support. A random-weight float64 test obtains an exact zero model identity sensitivity beyond $2L$ and a nonzero computational path once the boundary is crossed.

The matrix was specified before any extension test prediction. S2-base uses two interactions and 32 channels; S2-wide uses two interactions and 39 channels; S3-base uses three interactions and 32 channels. Their parameter counts are 10,145, 14,664, and 14,881, leaving a 1.46% gap between the width and depth controls. Each cell uses seeds 20260901–20260903, the original 3.5 Å graph, and the same $Au_2$/MgO train/validation/test structures. The optimizer is AdamW with learning rate 0.001 and weight decay $5 \times 10^{-7}$; energy and force weights are 1 and 100; batch size is eight; and

validation loss alone selects the saved model within the 40-epoch ceiling. No response or test label participates in optimization or early stopping.

All nine training summaries were complete before the one-shot test evaluation. Table 3 reports every run. Both two-block cells are exactly invariant to Al-to-Mg replacement in float64, including the parameter-matched width control. Every S3 run has the same sign as the held-out DFT group contrast in all four bins and improves both prespecified endpoints relative to its matched S2 runs. The identity–contrast alignment advantage over either S2 control has a shared-draw structure-bootstrap interval $[0.444, 0.787]$ eV/Å. The bootstrap resamples held-out structures within the fixed distance and doping strata after averaging the three fitted models; run-specific rows display optimization variability separately.

**Complete primary matrices.** Tables 10 and 11 expose every seed and architectural cell from the aggregate-label and mixed-harmonic stages. No row is filtered by response sign, magnitude, ordinary force error, or response classification. The harmonic-only specialist matrix follows the two prospective stage aggregates; the post hoc direct-edge extension is reported above.

Table 10: **Complete Ag–Pd aggregate-label reach matrix.** Panel **a** reports response structure; panel **b** reports error endpoints. Closure is relative to the zero-response baseline. Force RMSE uses 89 held-out configurations; all RMSE values are in meV/Å.

**a** Response structure

| Cutoff (Å) | Variant | Run | Closure (%) | Cosine | Norm ratio | Block $r$ |
|---|---|---|---|---|---|---|
| 3.0 | D1-base | 20260825 | 0.03 | 0.0282 | 0.0427 | -0.3359 |
| 3.0 | D1-base | 20260826 | -0.03 | 0.0129 | 0.0401 | -0.3738 |
| 3.0 | D1-base | 20260827 | 0.03 | 0.0269 | 0.0403 | -0.4150 |
| 3.0 | D1-wide | 20260825 | -0.03 | 0.0108 | 0.0357 | -0.3234 |
| 3.0 | D1-wide | 20260826 | -0.03 | 0.0082 | 0.0324 | -0.3039 |
| 3.0 | D1-wide | 20260827 | 0.10 | 0.0446 | 0.0388 | -0.4095 |
| 3.0 | D2-base | 20260825 | -841.57 | -0.0937 | 9.2692 | -0.4131 |
| 3.0 | D2-base | 20260826 | -1752.41 | 0.0436 | 18.5408 | -0.3531 |
| 3.0 | D2-base | 20260827 | -118.21 | 0.0768 | 2.0178 | -0.4325 |
| 5.0 | D1-base | 20260825 | -1427.71 | -0.2826 | 14.9644 | 0.7662 |
| 5.0 | D1-base | 20260826 | -823.36 | -0.1630 | 9.0178 | 0.6160 |
| 5.0 | D1-base | 20260827 | -1153.72 | -0.1279 | 12.3700 | 0.6347 |
| 5.0 | D1-wide | 20260825 | -1979.66 | 0.0648 | 20.8374 | 0.4088 |
| 5.0 | D1-wide | 20260826 | -1076.33 | -0.3063 | 11.4184 | 0.7405 |
| 5.0 | D1-wide | 20260827 | -1678.01 | -0.1580 | 17.5946 | 0.6809 |
| 5.0 | D2-base | 20260825 | -883.03 | 0.0264 | 9.8057 | 0.3117 |
| 5.0 | D2-base | 20260826 | -1577.81 | -0.0309 | 16.7174 | -0.1032 |
| 5.0 | D2-base | 20260827 | -693.73 | -0.1224 | 7.7526 | 0.0225 |

**b** Error endpoints

| Cutoff (Å) | Variant | Run | Max $\lvert J \rvert$ (eV/Å$^2$) | Force RMSE | Response-set RMSE |
|---|---|---|---|---|---|
| 3.0 | D1-base | 20260825 | 1.25e-03 | 42.95 | – |
| 3.0 | D1-base | 20260826 | 1.04e-03 | 42.94 | – |
| 3.0 | D1-base | 20260827 | 1.17e-03 | 43.13 | – |
| 3.0 | D1-wide | 20260825 | 1.33e-03 | 42.91 | – |
| 3.0 | D1-wide | 20260826 | 9.26e-04 | 43.25 | – |
| 3.0 | D1-wide | 20260827 | 1.17e-03 | 43.25 | – |
| 3.0 | D2-base | 20260825 | 2.36e-01 | 38.04 | – |
| 3.0 | D2-base | 20260826 | 3.63e-01 | 37.23 | – |
| 3.0 | D2-base | 20260827 | 3.63e-02 | 39.17 | – |
| 5.0 | D1-base | 20260825 | 6.75e-01 | 13.21 | – |
| 5.0 | D1-base | 20260826 | 4.05e-01 | 13.58 | – |
| 5.0 | D1-base | 20260827 | 5.49e-01 | 14.06 | – |
| 5.0 | D1-wide | 20260825 | 8.86e-01 | 13.44 | – |
| 5.0 | D1-wide | 20260826 | 4.91e-01 | 14.29 | – |
| 5.0 | D1-wide | 20260827 | 7.69e-01 | 13.68 | – |

**b** Error endpoints (continued)

| Cutoff (Å) | Variant | Run | Max $\lvert J\rvert$ (eV/Å$^2$) | Force RMSE | Response-set RMSE |
|---|---|---|---|---|---|
| 5.0 | D2-base | 20260825 | 2.68e-01 | 9.64 | – |
| 5.0 | D2-base | 20260826 | 4.23e-01 | 10.37 | – |
| 5.0 | D2-base | 20260827 | 2.34e-01 | 9.90 | – |

Across the aggregate-label runs, both D1 variants remain at the structural-zero scale below the boundary. At 5 Å, every D1 run develops a nonzero remote block and block-norm correlation rises to 0.409–0.766. The matrix therefore separates consistent support activation from direction and scale, which depend on task-aligned supervision.

Table 11: **Complete Ag–Pd harmonic-supervised reach matrix.** Panel **a** reports response structure; panel **b** reports error endpoints. Closure is relative to the zero-response baseline. Force RMSE uses 269 held-out configurations; all RMSE values are in meV/Å.

**a** Response structure

| Cutoff (Å) | Variant | Run | Closure (%) | Cosine | Norm ratio | Block $r$ |
|---|---|---|---|---|---|---|
| 3.0 | D1-base | 20260828 | -0.00 | -0.0008 | 0.0000 | -0.5952 |
| 3.0 | D1-base | 20260829 | 0.00 | 0.0533 | 0.0000 | -0.4989 |
| 3.0 | D1-base | 20260830 | -0.00 | -0.0136 | 0.0000 | -0.4200 |
| 3.0 | D1-wide | 20260828 | -0.00 | -0.0360 | 0.0000 | -0.5641 |
| 3.0 | D1-wide | 20260829 | 0.00 | 0.0424 | 0.0000 | -0.5724 |
| 3.0 | D1-wide | 20260830 | -0.00 | -0.0325 | 0.0000 | -0.4510 |
| 3.0 | D2-base | 20260828 | -206.91 | 0.1001 | 3.0035 | 0.0136 |
| 3.0 | D2-base | 20260829 | -87.12 | 0.2009 | 1.7952 | 0.6768 |
| 3.0 | D2-base | 20260830 | -14.08 | 0.2337 | 0.8304 | 0.6981 |
| 5.0 | D1-base | 20260828 | -304.61 | 0.1787 | 4.1033 | 0.8291 |
| 5.0 | D1-base | 20260829 | -387.08 | 0.2262 | 4.9985 | 0.5926 |
| 5.0 | D1-base | 20260830 | -246.39 | 0.1057 | 3.4238 | 0.8359 |
| 5.0 | D1-wide | 20260828 | -338.24 | 0.2423 | 4.5160 | 0.8377 |
| 5.0 | D1-wide | 20260829 | -287.82 | 0.2238 | 3.9776 | 0.8024 |
| 5.0 | D1-wide | 20260830 | -267.61 | 0.2217 | 3.7661 | 0.7783 |
| 5.0 | D2-base | 20260828 | -707.39 | -0.0205 | 7.9913 | 0.2841 |
| 5.0 | D2-base | 20260829 | -438.04 | -0.1461 | 5.1426 | 0.1418 |
| 5.0 | D2-base | 20260830 | -528.41 | -0.1055 | 6.0994 | 0.7663 |

**b** Error endpoints

| Cutoff (Å) | Variant | Run | Max $\lvert J\rvert$ (eV/Å$^2$) | Force RMSE | Response-set RMSE |
|---|---|---|---|---|---|
| 3.0 | D1-base | 20260828 | 1.76e-12 | 25.97 | 40.33 |
| 3.0 | D1-base | 20260829 | 2.18e-12 | 27.80 | 41.43 |
| 3.0 | D1-base | 20260830 | 1.84e-12 | 26.99 | 40.66 |
| 3.0 | D1-wide | 20260828 | 1.67e-12 | 25.16 | 39.45 |
| 3.0 | D1-wide | 20260829 | 1.76e-12 | 26.67 | 42.87 |
| 3.0 | D1-wide | 20260830 | 2.01e-12 | 26.76 | 41.77 |
| 3.0 | D2-base | 20260828 | 6.00e-02 | 24.29 | 26.19 |
| 3.0 | D2-base | 20260829 | 8.24e-02 | 21.89 | 23.09 |
| 3.0 | D2-base | 20260830 | 4.49e-02 | 22.14 | 25.17 |
| 5.0 | D1-base | 20260828 | 2.52e-01 | 21.96 | 33.58 |
| 5.0 | D1-base | 20260829 | 1.84e-01 | 20.06 | 38.57 |
| 5.0 | D1-base | 20260830 | 1.60e-01 | 19.15 | 46.53 |
| 5.0 | D1-wide | 20260828 | 2.31e-01 | 16.65 | 42.11 |
| 5.0 | D1-wide | 20260829 | 1.62e-01 | 15.26 | 36.56 |
| 5.0 | D1-wide | 20260830 | 1.65e-01 | 22.60 | 36.92 |
| 5.0 | D2-base | 20260828 | 2.73e-01 | 16.20 | 20.99 |
| 5.0 | D2-base | 20260829 | 1.26e-01 | 21.61 | 17.71 |
| 5.0 | D2-base | 20260830 | 2.47e-01 | 14.31 | 20.07 |

Harmonic supervision sharpens the permitted response. D2@3 has positive cosine in all three runs (0.100–0.234), as does D1@5 (0.106–0.226), while D1@3 remains below $2.18 \times 10^{-12}$ eV/Å$^2$. The corresponding D1@5 block-norm correlations span 0.593–0.836.

Table 12: **Complete Ag–Pd harmonic-only force-specialist reach matrix.** Panel **a** reports response structure; panel **b** reports error endpoints. Closure is relative to the zero-response baseline. Force RMSE uses 180 held-out configurations; all RMSE values are in meV/Å.

**a** Response structure

| Cutoff (Å) | Variant | Run | Closure (%) | Cosine | Norm ratio | Block $r$ |
|---|---|---|---|---|---|---|
| 3.0 | D1-base | 20260831 | 0.00 | 0.0488 | 0.0000 | -0.2376 |
| 3.0 | D1-base | 20260832 | -0.00 | -0.0271 | 0.0000 | -0.2944 |
| 3.0 | D1-base | 20260833 | -0.00 | -0.0330 | 0.0000 | -0.1594 |
| 3.0 | D1-wide | 20260831 | -0.00 | -0.0496 | 0.0000 | -0.3876 |
| 3.0 | D1-wide | 20260832 | 0.00 | 0.0060 | 0.0000 | -0.2684 |
| 3.0 | D1-wide | 20260833 | -0.00 | -0.0062 | 0.0000 | -0.0420 |
| 3.0 | D2-base | 20260831 | -0.00 | -0.0378 | 0.0000 | -0.0199 |
| 3.0 | D2-base | 20260832 | -0.00 | -0.0070 | 0.0000 | 0.2359 |
| 3.0 | D2-base | 20260833 | -0.00 | -0.0403 | 0.0000 | -0.2270 |
| 5.0 | D1-base | 20260831 | 0.18 | 0.1160 | 0.0168 | 0.6758 |
| 5.0 | D1-base | 20260832 | 0.38 | 0.3310 | 0.0116 | 0.6309 |
| 5.0 | D1-base | 20260833 | -0.01 | -0.0065 | 0.0067 | 0.7067 |
| 5.0 | D1-wide | 20260831 | 1.31 | 0.2136 | 0.0737 | 0.7871 |
| 5.0 | D1-wide | 20260832 | -0.05 | -0.0355 | 0.0123 | 0.7832 |
| 5.0 | D1-wide | 20260833 | 0.12 | 0.0906 | 0.0141 | 0.7168 |
| 5.0 | D2-base | 20260831 | 0.01 | 0.0212 | 0.0042 | 0.3501 |
| 5.0 | D2-base | 20260832 | 0.86 | 0.2799 | 0.0323 | 0.7113 |
| 5.0 | D2-base | 20260833 | -0.07 | -0.0339 | 0.0173 | 0.4856 |

**b** Error endpoints

| Cutoff (Å) | Variant | Run | Max $\|J\|$ (eV/Å$^2$) | Force RMSE | Response-set RMSE |
|---|---|---|---|---|---|
| 3.0 | D1-base | 20260831 | 1.79e-13 | 7.95 | 12.06 |
| 3.0 | D1-base | 20260832 | 4.18e-13 | 9.07 | 20.62 |
| 3.0 | D1-base | 20260833 | 2.10e-13 | 8.39 | 14.44 |
| 3.0 | D1-wide | 20260831 | 4.82e-13 | 8.09 | 17.92 |
| 3.0 | D1-wide | 20260832 | 1.89e-13 | 8.81 | 13.65 |
| 3.0 | D1-wide | 20260833 | 2.19e-13 | 8.53 | 13.09 |
| 3.0 | D2-base | 20260831 | 1.58e-14 | 7.44 | 9.93 |
| 3.0 | D2-base | 20260832 | 1.05e-13 | 7.55 | 10.36 |
| 3.0 | D2-base | 20260833 | 2.09e-13 | 7.49 | 9.80 |
| 5.0 | D1-base | 20260831 | 8.15e-04 | 7.66 | 42.47 |
| 5.0 | D1-base | 20260832 | 4.40e-04 | 7.09 | 34.71 |
| 5.0 | D1-base | 20260833 | 2.57e-04 | 7.49 | 22.53 |
| 5.0 | D1-wide | 20260831 | 3.20e-03 | 8.70 | 74.65 |
| 5.0 | D1-wide | 20260832 | 5.35e-04 | 7.05 | 31.32 |
| 5.0 | D1-wide | 20260833 | 4.36e-04 | 7.31 | 31.08 |
| 5.0 | D2-base | 20260831 | 1.50e-04 | 7.14 | 31.28 |
| 5.0 | D2-base | 20260832 | 1.19e-03 | 6.63 | 41.95 |
| 5.0 | D2-base | 20260833 | 4.69e-04 | 6.89 | 33.80 |

Force-only specialization cleanly resolves structural status: every 3 Å cell is numerical zero, whereas 5 Å produces a response in every architecture. D1-base has positive alignment in two runs, block-norm correlation 0.631–0.707, and held-out force RMSE 7.09–7.66 meV/Å.

**Statistical units and reproducibility.** Table 13 separates closure and direction for the three independent Ag–Pd response clusters.

Table 13: **Ag–Pd equilibrium-level sensitivity.** The two panels separate magnitude recovery from direction for the three-run ensemble.

**a** Closure by equilibrium

| Cell | eq. 7 | eq. 16 | eq. 19 | pooled | LOO closure |
|---|---|---|---|---|---|
| Aggregate D2@3 | -75.5% | -763.1% | -1078.9% | -708.7% | $[-971.3, -410.3]$% |
| Mixed D2@3 | -7.8% | -75.6% | -88.1% | -54.7% | $[-83.5, -33.4]$% |
| Specialist D1@5 | 0.4% | -0.0% | -0.0% | 0.2% | $[-0.0, 0.3]$% |
| Direct D1@8 | -6.0% | -22.5% | -2.7% | -8.5% | $[-11.5, -4.5]$% |

**b** Signed cosine by equilibrium

| Cell | eq. 7 | eq. 16 | eq. 19 | pooled |
|---|---|---|---|---|
| Aggregate D2@3 | 0.00 | 0.24 | -0.10 | 0.00 |
| Mixed D2@3 | 0.54 | 0.12 | -0.08 | 0.20 |
| Specialist D1@5 | 0.49 | -0.00 | -0.10 | 0.23 |
| Direct D1@8 | 0.15 | -0.39 | -0.01 | 0.00 |

The $Au_2$/MgO structure bootstrap resamples held-out structures within eight fixed bin-by-doping strata (four Au–O bins crossed with two doping groups) after averaging the three fitted models; the paired model-intervention curve uses the four doped-bin strata. Its interval therefore describes structure variation conditional on that ensemble. The three seed-specific endpoints separately display optimization variation; bootstrap draws are not counted as additional training runs.

For Ag–Pd, Cartesian components remain nested within $3 \times 3$ response blocks, and blocks remain nested within three equilibrium structures. The descriptive cluster bootstrap and the leave-one-equilibrium-out calculation preserve this nesting. The table exposes the three raw equilibrium aggregates for representative prospective, exploratory, and post hoc cells. The anonymous supplement reports the complete per-run, per-equilibrium, and leave-one-equilibrium-out values.

Split-integrity tests verify 5,000 unique $Au_2$/MgO structure identifiers, the fixed 4,000/500/500 partition, and identical partitions across all architecture cells. The Ag–Pd response structures 0–19 are disjoint from every training, validation, and auxiliary-test structure, preventing response-set leakage.

Figure 8 summarizes the graph solution for the selected Ag–Pd 6–8 Å pairs. Across the tested cutoffs, their hop distance changes from three to two to one, giving D1 reach margins $-1$, 0, and 1. These graph classes establish access only; the response outcomes are reported in Tables 10–12.

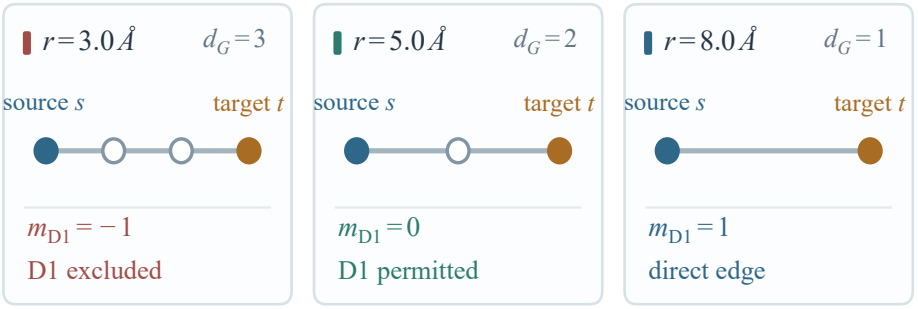


Figure 8: **Ag–Pd graph reach across the tested cutoffs.** The selected pairs are three, two, and one hops apart at 3, 5, and 8 Å, respectively. The 8 Å direct-edge extension was post hoc. Paths are graph schematics, not atomic coordinates; permission does not imply response fidelity.