AUTOMATED CIRCUIT DISCOVERY FOR EQUIVARIANT MOLECULAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural network force fields predict molecular energies and forces with near-quantum accuracy and increasingly run as foundation models that users never retrain. Whether such a model can be trusted beyond its test set depends on what it computes: which geometric features it responds to, where in the network the response is carried, and whether the same computation recurs from molecule to molecule. Mechanistic interpretability answers questions of this kind for language models by finding *circuits*, small sets of internal components that together implement one behaviour, and validating them with interventions. Its tools assume discrete inputs, hidden states that can be copied between examples and evidence from single inputs, whereas molecular geometry is continuous, equivariant hidden states are tied to a coordinate frame, and a mechanism is established only when it recurs across molecules. We present an automatic circuit-discovery pipeline for GNN force fields built around these conditions. Typed chemical triangles define families of small, valid perturbations; an exact Möbius decomposition over all 2^6 switch combinations of six internal units assigns every part of the response to a set of units; components whose effect recurs and survives dose and content controls are localised by a greedy search over units typed by their role relative to the triangle, so a circuit found on one set of molecules is defined on every molecule of its family; and each circuit is confirmed on molecules the search never saw by sufficiency, necessity and a random-subgraph contrast under one multiplicity correction. Run as a single command on MACE, ORB and NEQUIP, the pipeline returns 44 confirmed circuits whose interventions read as chemistry: angle circuits separate sp^3 from sp^2 centres and bond-stretch circuits behave as harmonic springs.

1 INTRODUCTION

Simulating how a molecule moves, reacts or relaxes requires the force on every atom for millions of successive arrangements (Frenkel & Smit, 2002; Allen & Tildesley, 2017). Quantum-mechanical calculations give accurate forces far too slowly for this (Kohn & Sham, 1965; Car & Parrinello, 1985), so practitioners use a *force field*, a function from atomic species and coordinates to a total energy and per-atom forces (Behler & Parrinello, 2007; Bartók et al., 2010; Unke et al., 2021). Graph neural networks (GNNs) have become the dominant way to learn such functions from quantum reference data (Gilmer et al., 2017; Duval et al., 2023): a GNN force field treats a molecule as a graph whose nodes are atoms and whose edges connect atoms within a cutoff, message passing lets each atom summarise its local environment, and a readout sums per-atom energies into a molecular energy whose gradient gives the forces (Schütt et al., 2017; Batzner et al., 2022; Batatia et al., 2022). These models reach near-quantum accuracy at a small fraction of the cost and now drive simulation and screening (Chanussot et al., 2021; Deng et al., 2023; Merchant et al., 2023).

Test accuracy is a weak measure of how far such a model can be trusted. Force fields run in long simulations where small errors compound (Stocker et al., 2022; Fu et al., 2023), on chemistry outside their training data (Vita & Schwalbe-Koda, 2023), and increasingly as foundation models that users never retrain (Batatia et al., 2023). To audit extrapolation, diagnose failures and compare architectures by what they compute, a practitioner needs to know which geometric features the model responds to, where in the network that response is carried, and whether the same computation is used wherever the same chemistry occurs.

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.

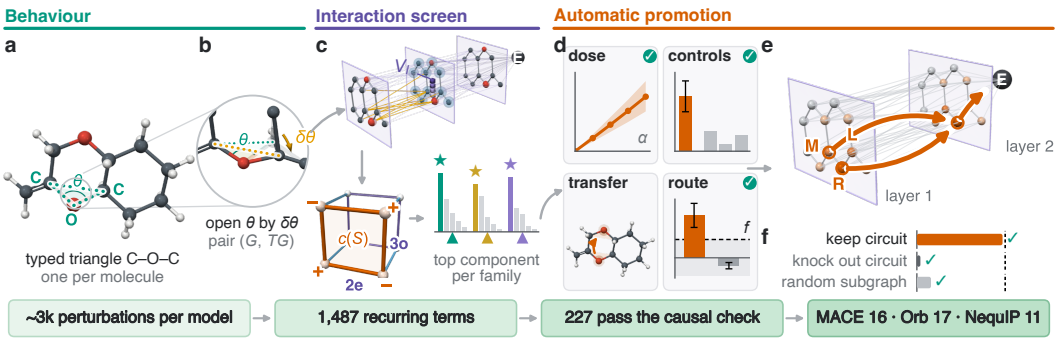

Figure 1: **Automatic circuit discovery for a GNN force field.** (a,b) A typed triangle defines a behaviour and its perturbed pair. (c) The exact decomposition over six units scores every unit set and the strongest recurring component per family is kept. (d) A causal check with doses and controls. (e) Localisation over role-typed units. (f) Confirmation on held-out molecules.

Mechanistic interpretability asks exactly these questions (Olah et al., 2020; Elhage et al., 2021; Bereska & Gavves, 2024). Its unit of explanation is the *circuit*, a small set of internal components that jointly implement one identifiable behaviour (Cammarata et al., 2020; Olsson et al., 2022; Wang et al., 2023), and its evidential tool is the *intervention*: the analyst changes the value a component takes during a forward pass and checks whether the output moves as the proposed explanation predicts (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024); a circuit that survives such interventions and their controls is causally validated (Goldowsky-Dill et al., 2023; Geiger et al., 2025). Automatic circuit discovery turns this into a search that prunes a computational graph while the behaviour is retained (Conmy et al., 2023; Syed et al., 2024; Bhaskar et al., 2024; Marks et al., 2025), and a circuit is judged by its faithfulness, whether it alone reproduces the behaviour, and its completeness, whether removing it destroys the behaviour (Wang et al., 2023; Hanna et al., 2024; Miller et al., 2024).

Why this does not carry over directly. Three properties of equivariant molecular models break the assumptions behind these tools. The input is a continuous three-dimensional geometry, so there is no natural set of corrupted inputs to patch from, and an implausible deformation moves the network for the wrong reason (Zhang & Nanda, 2024); choosing the perturbation is part of the problem. The hidden states are tensors whose values depend on the coordinate frame (Thomas et al., 2018; Geiger & Smidt, 2022), so an activation must be written in a way that respects the symmetry (Hansen et al., 2024). And the computational graph changes with the molecule: a component found on one molecule has no obvious counterpart on the next, so a finding on one input is an anecdote, whereas the same local chemistry recurs across thousands of molecules, which are the right unit of evidence.

This paper. We present a fully automatic pipeline that takes a trained GNN force field and a population of molecules and returns confirmed circuits (Figure 1). Every molecule is decomposed into its typed local *triangles*, the smallest unit that records both distances and an angle; the triangles partition the chemistry into families, and each family defines one small, valid perturbation, such as opening a C–O–C angle by 2° or stretching a C=C bond by 0.02 Å. At one internal boundary, six units are switched between their clean and perturbed values in all 2^6 combinations, and Möbius inversion assigns an exact coefficient to every unit set; the sets whose coefficient recurs across held-out molecules are the recurring terms, and each family's strongest ones pass through a causal check with graded doses and wrong-content controls. Every component that passes is expanded into a graph of internal units typed by their *role* relative to the triangle, and a greedy search finds the smallest set of them that reproduces the component's effect. Because the units are typed by role, the circuit is defined on every molecule of the family, and it is confirmed on molecules the search never saw: keeping only the circuit retains the component's effect, resetting only the circuit destroys it, and the circuit beats random subgraphs of the same size, under one Bonferroni correction over every confirmation test. A circuit explains the effect of a *component*: in a message-passing force field every atom within a few bonds contributes to every energy change, so the object that can have a small circuit is the part of the response carried by a named set of units at one boundary (Section 3).

Run as a single command on three force fields of different design, MACE (Batatia et al., 2022), ORB (Neumann et al., 2024) and NEQUIP (Batzner et al., 2022), the pipeline reduces about $3{,}000$ perturbations per model to $1{,}487$ recurring terms, 227 components that pass the causal check and 44 confirmed circuits. The circuits are small relative to the units that carry any effect, they outperform the same number of units nearest to the triangle, and their interventions read as chemistry: angle circuits separate sp^3 from sp^2 centres, with the MACE circuits within a degree of the ideal angles, and stretch circuits behave as harmonic springs.

Related work. Equivariant architectures (Thomas et al., 2018; Geiger & Smidt, 2022; Batzner et al., 2022; Batatia et al., 2022; Neumann et al., 2024) fix how a channel transforms and say nothing about what a trained channel encodes or whether the output uses it; graph explanation methods (Ying et al., 2019; Wu et al., 2023) score inputs and cannot say which internal route transmits a geometric change. Single-component patching (Vig et al., 2020; Zhang & Nanda, 2024) misses interactions between components, and automated circuit discovery (Conmy et al., 2023; Syed et al., 2024; Bhaskar et al., 2024) operates on one fixed computational graph. Our decomposition adapts any-order interaction indices (Harsanyi, 1963; Grabisch, 2016; Fumagalli et al., 2023) to interventions on an internal interface, and our search adapts ACDC to a graph typed by chemical role so that one circuit is defined across molecules (Appendix H).

Contributions. (i) A formulation of circuit discovery for equivariant molecular networks: chemically labelled behaviours, exact interaction coefficients on a declared interface, role-typed unit graphs and a ladder from recurring term to confirmed circuit with the molecule as the unit of replication (Section 2). (ii) An automatic pipeline that implements the ladder end to end as one command (Section 3). (iii) A three-model study of MACE, ORB and NEQUIP under one evidence standard: 44 confirmed circuits, ablations against locality baselines and across every design choice, and case studies in which the circuits reproduce textbook geometry (Section 4).

2 PROBLEM FORMULATION

2.1 GNN FORCE FIELDS AND BEHAVIOURS

A molecule is a geometric graph $G = (V, \mathbf{Z}, \mathbf{R})$ with N atoms, species $\mathbf{Z} \in \mathbb{N}^N$ and coordinates $\mathbf{R} \in \mathbb{R}^{N\times 3}$; atoms i and j are joined when they lie within a cutoff r_c, and $\mathcal{N}(i)$ denotes the neighbours of i. A GNN force field f_θ computes per-atom states $h_i^{(\ell)}$, $\ell = 1, \dots, L$, by message passing and reads out an energy whose gradient with respect to the coordinates gives the forces:

$$\begin{aligned} h_i^{(\ell)} &= \mathrm{Update}^{(\ell)}\Big(h_i^{(\ell-1)}, \textstyle\sum_{j\in\mathcal{N}(i)} \mathrm{Msg}^{(\ell)}\big(h_i^{(\ell-1)}, h_j^{(\ell-1)}, \mathbf{r}_i - \mathbf{r}_j\big)\Big), \\ E_\theta(G) &= \textstyle\sum_{i\in V} E_{\theta,i}\big(h_i^{(L)}\big), \qquad \mathbf{F}_{\theta,i} = -\nabla_{\mathbf{r}_i} E_\theta(G). \end{aligned} \tag{1}$$

Equivariant architectures build each state from irreducible representation (irrep) blocks that transform under rotation as scalars, vectors or higher-order tensors, labelled by rotation order and parity ($0e$ a scalar, $1o$ a vector, $2e$ and $3o$ higher orders), so that E_θ is invariant and $\mathbf{F}_\theta$ rotates with the molecule. Every behaviour is judged through a declared scalar *estimand* $y_B(G) = \phi_B(f_\theta(G))$; for the force fields studied here ϕ_B is the model's total energy.

Definition 1 (Chemical family and behaviour)**.** A *chemical family* $\mathcal{F}$ is a typed local motif: for a bonded pair, the two species and the bond type (C=O); for a three-atom triangle, the centre atom i bonded to j and k, its three species and two bond types (H–C–O). Each *occurrence* in a molecule carries the distances r_{ij}, r_{ik}, r_{jk} and the angle θ_{jik}, which are not part of the label. A *perturbation specification* T for $\mathcal{F}$ names the changed coordinate, the moving atom, the step size η and the quantities held fixed (open the H–C–O angle by 2° with both bond lengths fixed); it is valid when it preserves connectivity, chirality and minimum interatomic distances, range-constrained when the changed coordinate stays within the range observed for $\mathcal{F}$ in the population $\mathcal{D}$, and it commutes with rigid motions, $T(gG) = gT(G)$. The pair $B = (\mathcal{F}, T)$ is a *behaviour* of f_θ on $\mathcal{D}$ when a valid, range-constrained occurrence exists in at least a declared number of molecules $\mathcal{D}_\mathcal{F} \subseteq \mathcal{D}$ and the response $\Delta y_B(G) = y_B(TG) - y_B(G)$ exceeds a threshold δ_0 on a declared fraction of them.

Running example. In every molecule that contains an H–C–O angle we open one occurrence by 2° with both bond lengths fixed, record the change in predicted energy, and ask which internal

components carry it and whether the same components do so in the other molecules. The centre atom of the perturbed triangle is the *receiver* R, the moving neighbour M and the fixed neighbour L.

2.2 Interfaces, coefficients and circuits

Definition 2 (Interface, intervention, connected coefficient)**.** An *interface* $V_I = \{c_1, \dots, c_m\}$ is a declared finite set of nodes of the computation graph of f_θ at one boundary, the *units*: a unit may be an irrep block of one atom's state, a set of edge messages, or a grouped route such as "all other atoms". The *activation* of unit c on G is its value $a_c(G)$, and its natural difference $\Delta a_c = a_c(TG) - a_c(G)$ is how that value changes when the model processes the perturbed molecule. An *intervention* on a subset $A \subseteq V_I$ at strength $\alpha \in [0, 1]$ evaluates the model on G with a_c replaced by $a_c(G) + \alpha\, \Delta a_c$ for $c \in A$ and recomputes every downstream node; when the written value comes from a molecule in a different orientation it is transported through the unit's group action. $v(A)$ denotes y_B under this intervention at $\alpha = 1$. The interface is *admissible* when $v(\emptyset) = y_B(G)$ and $v(V_I) = y_B(TG)$ within tolerance: intervening on nothing returns the clean output and intervening on every unit reproduces the perturbed output, so the interface intercepts every route from the perturbation to the estimand. The *connected coefficient* of $S \subseteq V_I$ is its Möbius inverse (Harsanyi, 1963; Grabisch, 2016),

$$c(S) = \sum_{A \subseteq S} (-1)^{|S|-|A|}\, v(A), \qquad \textstyle\sum_{\emptyset \neq S \subseteq V_I} c(S) = \Delta y_B, \tag{2}$$

the part of Δy_B that requires switching all of S jointly. A singleton coefficient is an ordinary one-unit patch; a pair coefficient is the part of the response that needs both units. The labelled set $(\mathcal{F}, S)$ is a *recurring term* when $|c_G(S)| > \delta$ on more than half of the family's held-out molecules under multiple-testing control.

In the example, an admissible interface could be the irrep blocks of the message arriving at the carbon, one grouped unit for all other atoms and one for geometry reused downstream.

Definition 3 (Component, role-typed unit graph, circuit)**.** Fix a behaviour B and an admissible interface V_I. A *component* C is the joint patch of a recurring term's units; its *effect* on a molecule is $e_C = v(S) - v(\emptyset)$, the change in y_B when only C takes its perturbed value. C is a *candidate* when, on molecules not used to choose it, the aligned share of Δy_B it reproduces, $\rho_C = \sum_i \operatorname{sgn}(\Delta y_{B,i})\, e_{C,i} \big/ \sum_i |\Delta y_{B,i}|$, is positive with a coherent dose response, and wrong content of the same size written at the same units reproduces less. The *unit graph* U of C is a finite set of internal objects that together carry e_C: objects upstream of the boundary that build C's perturbed value, the pieces of C itself, and objects downstream that carry its change to the output, each typed by the *roles* of the atoms it touches relative to the perturbed triangle (receiver R, moving arm M, fixed arm L, their neighbours nM and nR, atoms two bonds away $B2$, and the rest X) and by its layer, so that U is the same list on every molecule of the family. For $S \subseteq U$, $v(S)$ is y_B with the objects in S passing their perturbed values and all others their clean values; $v(U) = e_C$ and $v(\emptyset) = 0$. The *sufficiency* of S is $v(S)/e_C$, the behaviour retained when only S is kept, and its *necessity* is $1 - v(U \setminus S)/e_C$, the behaviour destroyed when only S is reset. Here v measures the change in y_B, so $v(\emptyset) = 0$. $S^* \subseteq U$ is a *confirmed circuit* for C when, on molecules never used to find it, its sufficiency lies within $(0.5, 1.5)$ and its necessity exceeds 0.5 with confidence, its sufficiency exceeds that of random subgraphs of the same size, and it uses at most half of the units that carry any effect.

Problem 1 (Automatic circuit discovery in a GNN force field)**.** Given f_θ, a population $\mathcal{D}$ split into discovery and held-out molecules, a family vocabulary with perturbation specifications, an estimand ϕ_B and an admissible interface V_I: (i) enumerate the behaviours these declarations induce; (ii) for each, compute every $c(S)$ and return the recurring terms; (iii) decide which of each family's strongest terms are candidates; (iv) for each candidate, find a circuit S^* in its role-typed unit graph on discovery molecules and confirm or reject it on held-out molecules, by one procedure fixed before any outcome is seen, with a ledger of everything tested.

3 Method: Automatic Circuit Discovery

The pipeline takes a trained force field, a population of molecules and a declared interface, and returns confirmed circuits without any per-candidate human choice (Algorithm 1, detailed in Appendix B).[1]

[1]Code, frozen designs and evidence tables: `https://anonymous.4open.science/r/enterpretability-00E3`.

Algorithm 1 Automatic circuit discovery for one model (per family $\mathcal{F}$).

Input: force field f_θ; molecules $\mathcal{D}$ split into discovery $\mathcal{D}_\mathrm{d}$ and held-out $\mathcal{D}_\mathrm{h}$; interface V_I ($|V_I| = 6$); thresholds $\delta, \tau = 0.8$; tests per candidate $= 3$.

1: **Perturb:** for each $G \in \mathcal{D}$ choose one occurrence of $\mathcal{F}$ by hash and build TG ($+2^\circ$ angle or $+0.02$ Å stretch, range-constrained, valid).
2: **Decompose:** evaluate $v(A)$ for all $A \subseteq V_I$; compute $c(S)$ by Equation 2; keep the terms with $|c(S)| > \delta$ on more than half of $\mathcal{D}_\mathrm{h}$ under Benjamini–Yekutieli control (*recurring terms*).
3: **Screen:** score each recurring S by $\sum_i |c_i(S)| / \sum_i |\Delta y_{B,i}|$ on $\mathcal{D}_\mathrm{d}$; keep the top-1, the top-1 among receiver-side units and the top-3 (*components*).
4: **Causal check** on $\mathcal{D}_\mathrm{d}$: write C's change at doses $\alpha \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ and wrong-content controls; pass iff $\rho_C > 0$ with a coherent dose curve and every live control class is beaten (*candidates*).
5: **Localise** on $\mathcal{D}_\mathrm{d}$: build the role-typed unit graph U of C; rank units by leave-one-out patches; take the shortest prefix with $F(S) \in [\tau, 2-\tau]$; prune ACDC-style; draw 20 stratum-matched random subgraphs (*circuit* S^*).
6: **Confirm** on $\mathcal{D}_\mathrm{h}$: sufficiency $v(S^*)/e_C$, necessity $1 - v(U \setminus S^*)/e_C$ and the random contrast, with Bonferroni-adjusted molecule-bootstrap intervals over all $m = 3N_5$ tests; **return** S^* as *confirmed* iff all three pass and $|S^*| \leq \frac{1}{2}|\mathrm{live}(U)|$.

Its first half finds *what* recurs: typed triangles turn the population into labelled geometric tests, the exact decomposition assigns each response to unit sets, and a causal check keeps the components that carry their share. Its second half finds *how*: a greedy search over units typed by chemical role returns the smallest set that reproduces the component's effect on discovery molecules, and that set is confirmed on molecules the search never saw. Design, code and molecule splits are hashed before any outcome is seen, so every number is a function of the model and the data.

Perturbation families and exact decomposition (stages 1–2). Every molecule is decomposed into typed triangles (j, i, k) and bonded pairs; grouping by label gives the chemical families, each with one perturbation specification whose admissible range is frozen on discovery molecules (Definition 1), and one occurrence per molecule chosen by a fixed hash. The interface declares, before any coefficient is computed, six complete units at one internal boundary: the receiver's irrep blocks (MACE, NEQUIP) or its node and edge states (ORB), closed under the symmetry group, plus grouped units for all other atoms and for geometry reused downstream, which together intercept every route from the perturbation to the energy (Appendix A). Evaluating $v(A)$ for all 64 subsets and applying Equation 2 gives every connected coefficient exactly, with the sum to Δy_B as a closure check; the 64 forward passes per perturbation buy the interactions that one-unit patching cannot express, 56–63% of the recurring terms (Section 4.1). For each family and unit set an exact binomial test asks whether $|c(S)| > \delta$ on more than half of the family's held-out molecules, with one Benjamini–Yekutieli correction over all labels of a model (Benjamini & Yekutieli, 2001). Per family, the recurring terms are ranked by the share of the response they carry on discovery molecules, and three screens are kept: the top-1, the top-1 among terms built only from receiver-side units, and the top-3. The joint patch of a term's units is its *component* C.

Causal check (stage 3). On a fresh cohort of discovery molecules the component's natural change is written at doses $\alpha \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ with full propagation, which gives its aligned share ρ_C and its dose curve. Content controls write wrong content of the same size at the same units: a Haar-random orthogonal mixing of the channel axis, which commutes with rotation and preserves the norm (four seeds); a random direction orthogonal to the natural change, rescaled to its norm (two seeds); a channel permutation (two seeds); and, for geometry units, a displacement of the moving atom by the same length in a random or orthogonal direction. A control class is *live* when its median unsigned effect exceeds ten times the numerical noise. The component passes when the 95% molecule-bootstrap lower bound of ρ_C is positive, the population dose curve keeps its sign and grows, and for every live class the contrast between the component and the class's mean effect over seeds has a positive lower bound; fewer than 20 admitted molecules leaves the outcome unavailable.

Localisation over role-typed units (stage 4). The component's computation is expanded into a graph U of typed units in three regions: *upstream* objects that build C's perturbed value (messages and edges by layer), the pieces of C itself (channel bands of each irrep block, rows by role, typed geometric edges) and *downstream* objects that carry its change to the output (messages, self-updates and readouts by layer). Each unit is typed by the roles of the atoms it touches relative to the perturbed

triangle (R, M, L, nM, nR, $B2$, X) and by its layer, so the list is identical on every molecule of the family and a circuit found on discovery molecules is defined, without change, on held-out ones. Patching is ACDC-style object patching (Conmy et al., 2023): for an active set S, objects in S pass their perturbed values and all others their clean values, whole irreps at a time; $v(U) = e_C$ and $v(\emptyset) = 0$ hold exactly. The search scores a set by its faithfulness $F(S) = \sum_i \mathrm{sgn}(e_i)\, v_i(S) / \sum_i |e_i|$ over admitted discovery molecules, with $F(U) = 1$. Units are *live* when they act on at least one molecule and their aggregate effect exceeds 10^{-3} of $\sum_i |e_i|$. When $|U| \leq 200$ the units are ranked by real patches, the drop in F when the unit alone is removed plus F of the unit alone within its region; larger graphs are ranked by attribution, gradient times the hybrid-minus-clean difference. The circuit is the shortest prefix of live units whose F lies in the band $[\tau, 2 - \tau]$ with $\tau = 0.8$, pruned least-important-first in chunks of $|S|/8$ units, halving a rejected chunk and keeping a removal only if F stays in the band. The band is two-sided because a prefix that drops paths which cancel one another can overshoot the effect several-fold. Twenty random subgraphs with the same unit count in each region are drawn from the live units for the next stage.

Confirmation (stage 5). The circuit S^*, its unit graph and the random draws are hashed before any held-out molecule is loaded; the held-out molecules of a family were never evaluated by any earlier stage. On each admitted molecule the pipeline evaluates $v(U)$, $v(\emptyset)$, $v(S^*)$, $v(U \setminus S^*)$ and the twenty random subgraphs with their complements. *Sufficiency*, $\sum_i \mathrm{sgn}(e_i) v_i(S^*) / \sum_i |e_i|$, keeps only the circuit and asks how much of the behaviour is retained; *necessity*, $1 - \sum_i \mathrm{sgn}(e_i) v_i(U \setminus S^*) / \sum_i |e_i|$, resets only the circuit and asks how much is destroyed; the *random contrast* is sufficiency minus the mean sufficiency of the random subgraphs. All three are ratios of sums with molecule-bootstrap intervals, adjusted by Bonferroni over $m = 3N_5$ tests (Dunn, 1961), where N_5 counts the candidates of every model that reached this stage. A circuit is *confirmed* when its sufficiency interval lies inside $(0.5, 1.5)$, its necessity lower bound exceeds 0.5, its random-contrast lower bound exceeds 0, and it is *localised*, using at most half of the live units; every other outcome is labelled by the tests it passed and kept in the ledger.

What a circuit explains. A circuit accounts for the effect e_C of its component, and the pipeline reports beside it the share of the full response that the component carries. This is the right target in a message-passing force field: every atom within the receptive field contributes to every energy change, so the whole response has no small circuit, whereas the part carried by a named set of units at one boundary does. The decomposition tells which components recur, the causal check that each carries its share, and the circuit how that share is produced and transmitted.

4 RESULTS

We run the pipeline on three force fields of different design, used as released: MACE-OFF23 medium (Kovács et al., 2025), the conservative ORB checkpoint trained on OMol (Neumann et al., 2024; Levine et al., 2025), whose latents are unconstrained scalars in a fixed frame, and the organic-molecule NEQUIP checkpoint (Batzner et al., 2022). The population is 400 molecules from the MACE-OFF23 test set (Eastman et al., 2023); 38 typed families have enough support, giving 6,439 perturbations per model (ORB, whose neighbour graph excludes some, evaluates 5,877 in 37 families). The causal check and the search use a discovery cohort of 30 molecules per family; confirmation uses a reserved set of up to 50 molecules per family that no earlier stage evaluated, the *strict* reading. Nine families had fewer than 20 reserved molecules; a *topped-up* reading completes them to 40 with molecules seen only in earlier discovery cohorts (Appendix C) and is reported as a secondary column.

4.1 WHAT THE PIPELINE FINDS

Table 1 and Figure 2 give the funnel. From about 3,000 perturbations per model the decomposition returns 549, 508 and 430 recurring terms for MACE, ORB and NEQUIP; the screen keeps 115, 111 and 123 distinct components; 68, 86 and 73 pass the causal check; and of the 51, 64 and 53 graded on reserved molecules, 16, 17 and 11 are confirmed circuits, 44 in all, with five, five and four more in the topped-up reading. Every confirmed circuit passes all three tests under Bonferroni correction over the $m = 813$ tests of this stage, and an independent recomputation from the raw records with separate code reproduced every number and label (Appendix C); Appendix D lists every circuit.

Model	Discovery perturbations	Recurring terms	Screened components	Pass causal check	Strict reading				Confirmed circuits	
					Graded	Sufficient	Necessary	Beats random	Strict	Topped-up
MACE	3,191	549	115	68	51	19	37	37	16	21
ORB	2,911	508	111	86	64	18	54	51	17	22
NEQUIP	3,191	430	123	73	53	12	42	30	11	15
All	9,293	1,487	349	227	168	49	133	118	44	58

Table 1: **The funnel per model.** Strict reading; the topped-up column completes the held-out sets of nine short families (Appendix C).

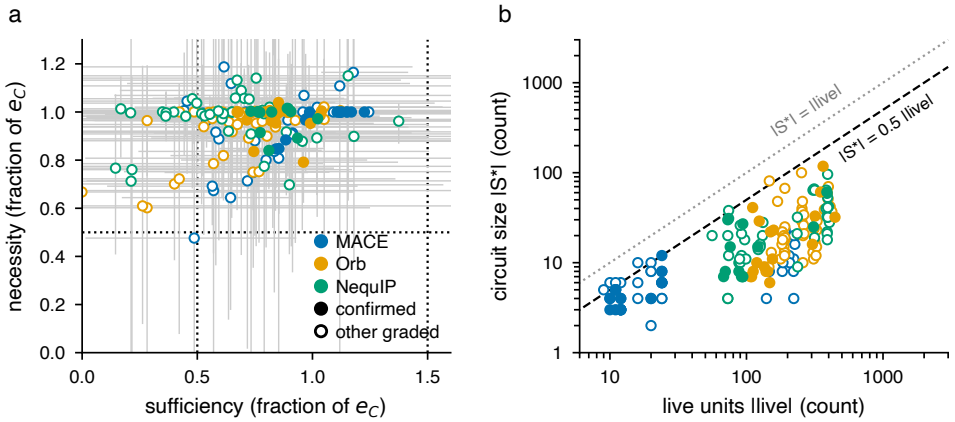

Figure 2: **The confirmation tests and the circuit sizes.** (a) Sufficiency and necessity of every graded candidate with adjusted intervals. (b) Circuit size against live units; the dashed line is the localisation rule. Filled markers are confirmed circuits.

Sufficiency is the binding test (Table 1, Figure 2a), and almost every sufficient candidate is also necessary and beats random. Among the confirmed circuits the median necessity is 1.00 in every model, and random subgraphs of the same size reproduce a median 0.39 of the effect in MACE, whose unit graphs are small, and none of it in ORB and NEQUIP.

The circuits are small (Figure 2b): 3–12 of 10–24 live units in MACE, 6–118 of 107–445 in ORB and 7–61 of 68–388 in NEQUIP, and their components carry a median 22%, 18% and 24% of the full energy change (2–73%; Appendix D). They cover 13, 14 and 10 families, angles and bond stretches alike, and their composition repeats across families: the MACE angle circuits hold the first-layer message from the moving atom M into the receiver R, one band of R's $3o$ block and R's own readout; the MACE stretch circuits hold the R–M edge geometry and the layer-1 messages sent along it; the ORB circuits pair a band of the receiver's node state and its input to the energy head with the messages from M into R in earlier blocks.

The decomposition behind the first stage, whose interaction terms fix which units belong together in a component, is reported in Appendix E.

4.2 ABLATIONS

Circuits against the nearest units. A natural alternative to a search is locality: take the same number of live units closest to the target triangle. We build this baseline for every confirmed circuit with a frozen rule, ordering units by the largest graph distance of any atom they touch from the triangle ($R, M, L = 0$; $nM, nR = 1$; $B2 = 2$; $X = 3$), and put it through the same tests on the same held-out molecules (Algorithm 6). The found circuits win (Figure 3): on sufficiency they beat the nearest units in 53 of 58 circuits, with median sufficiency 0.86 against 0.00, and a stricter baseline that matches each circuit's number of upstream, inside and downstream units still loses in 54 of 58. The nearest units carry the behaviour and fail to reproduce it on their own: knocking them out removes it (median necessity 0.97; 43 of 58 baselines pass the necessity criterion), keeping only them retains almost none (4 of 58 pass sufficiency), and 3 of the 58 baselines would meet the confirmation criteria. Nearest-first ordering fills the budget with the receiver's own channel bands and readout and

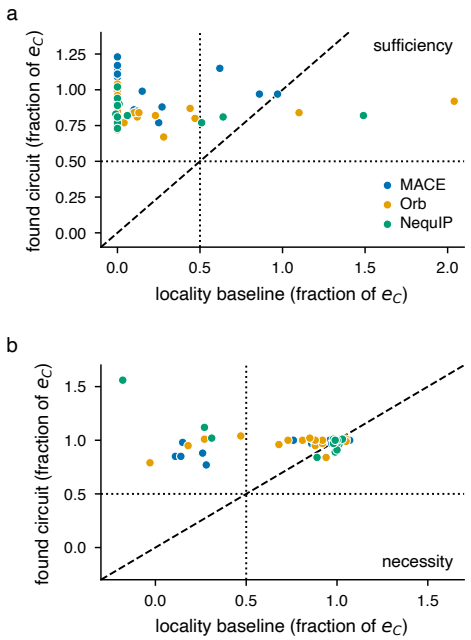

Figure 3: **Found circuits against the same number of nearest units.** One point per confirmed circuit; dashed line equality, dotted lines the pass criteria.

Setting	MACE	ORB	NEQUIP
Multiplicity			
Bonferroni, $m = 813$ (frozen)	**16 / 21**	**17 / 22**	**11 / 15**
Benjamini–Yekutieli, 5% FDR	25 / 33	25 / 35	13 / 22
Stage-3 control reading			
Mean control (frozen)	**16 / 21**	**17 / 22**	**11 / 15**
Worst seed	6 / 8	17 / 21	4 / 8
Per seed	14 / 19	17 / 22	10 / 14
Localisation ratio $\lvert S^*\rvert/\lvert\text{live}\rvert \leq$			
0.25	6 / 7	15 / 19	9 / 13
0.50 (frozen)	**16 / 21**	**17 / 22**	**11 / 15**
1 (none)	19 / 25	18 / 23	12 / 16
Stage-2 screen			
Top-1	1 / 2	2 / 3	6 / 8
Top-1 site-local	10 / 12	7 / 9	8 / 11
Top-3	16 / 21	17 / 22	11 / 15
Union (frozen)	**16 / 21**	**17 / 22**	**11 / 15**
Faithfulness band τ			
0.7	17 / 23	—	—
0.8 (frozen)	**16 / 21**	—	—
0.9	18 / 24	—	—

Table 2: **Sensitivity.** Confirmed circuits (strict / topped-up) when one setting is changed and the rest stay frozen (bold).

misses the incoming message from the moving atom: locality says where the behaviour lives, the circuit which path carries it.

Hyperparameter sensitivity. Table 2 changes one setting at a time and recounts the confirmed circuits; the frozen settings reproduce the published counts exactly. The multiplicity rule moves the counts most: Benjamini–Yekutieli control in place of Bonferroni gives 25, 25 and 13 strict circuits. The control reading is the second lever and is model-dependent: under the worst seed over control draws, ORB keeps all 17 of its circuits while MACE and NEQUIP keep 6 and 4, because their control writes have larger gain. The localisation rule binds mainly for MACE, whose graphs have 10–24 live units, and almost every circuit comes from the top-3 screen. Rerunning localisation and confirmation for MACE at $\tau = 0.7$ and 0.9 gives 17 and 18 strict circuits against 16; at 0.9 the circuits grow and 12 of the 16 frozen circuits remain, so the count is stable while the identity of the circuits shifts with the band (Appendix F).

4.3 CASE STUDIES: WHAT THE CIRCUITS COMPUTE

The confirmation tests say that a circuit carries its component's effect; the case studies ask what it computes. For the most localised circuits of MACE and ORB we sweep the perturbation through negative and positive doses, knock each unit out, transfer the circuit to other occurrences of its family, and read the implied geometry off its first-order response (Figure 4).

Stretch circuits are springs. The MACE C=C circuit, four units of twenty (the R–M edge geometry and the layer-1 messages along it), reproduces the component at every dose from -0.04 to $+0.04$ Å while its knockout removes the response; the component carries 73% of the full energy change. Its response is linear in the displacement, with a zero-force length of 1.234 Å against 1.331 Å for the full model and a curvature of $15.3\,\text{eV}/\text{Å}^2$ against 67.9: a short, soft two-body spring that the rest of the model stiffens and shifts. The ORB C=C circuit, nine units of 135, places its zero force at 1.325 Å against 1.327 for the full model and 1.33–1.34 in textbooks. Full-model stiffness follows bond order, as Badger's rule relates it to bond length (Badger, 1934), and the circuits carry 3–23% of it.

Angle circuits read out hybridisation. The three-unit MACE H–C–C circuit has its zero force at 108.8° on sp^3 carbons ($n = 25$) and 120.4° on sp^2 carbons ($n = 10$), against the ideal 109.5° and

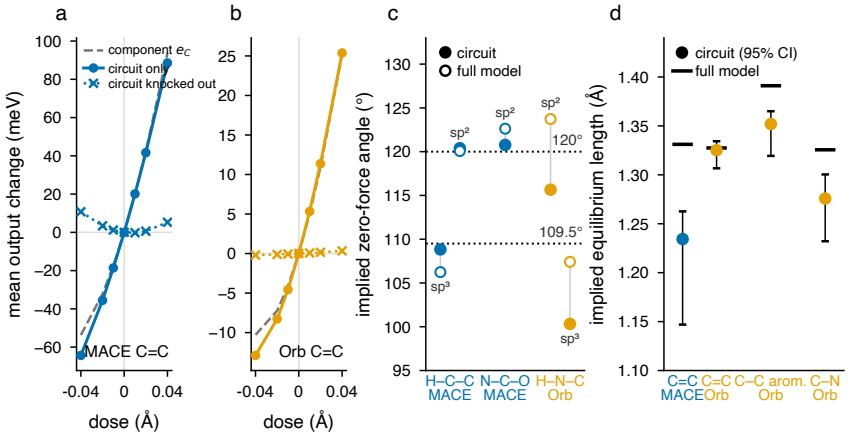

Figure 4: **Interventions on confirmed circuits.** (a,b) Energy change against dose for the MACE and ORB C=C stretch circuits: component, circuit alone, circuit knocked out. (c) Zero-force angle implied by the angle circuits, by hybridisation of the centre atom. (d) Zero-force bond length implied by the stretch circuits against the full model.

120° (Pauling, 1931; Gillespie & Nyholm, 1957), and the N–C–O circuit, all sp^2, at 120.8° [119.7, 121.8]: three units encode the centre atom's hybridisation. The ORB H–N–C circuit separates the two classes by the same 15°, both angles about 8° below the full model's.

Written at other occurrences of the same family in the same molecules, the circuits keep sufficiency between 0.86 and 1.29 and necessity at 0.90 or above: the role typing that defines a circuit across molecules also defines it across sites. Appendix G gives every case in full, with component-level interventions.

5 DISCUSSION

Three lessons carry beyond the three models. First, the chemistry can supply the structure that circuit discovery on language models takes from the architecture: typed triangles give a finite vocabulary of behaviours with molecule-level replication built in, and role typing gives a unit graph that is the same on every molecule of a family. Second, necessity is easy to satisfy and locality alone satisfies it; sufficiency separates a circuit from a neighbourhood, so both must be reported (Wang et al., 2023; Hanna et al., 2024). Third, the count of confirmed circuits depends on the multiplicity rule and the control reading more than on the search itself, so those choices belong in the frozen design and in the report.

Limitations. A circuit explains the effect of a component at one boundary; the share of the full response that the component carries ranges from 2% to 73% across the case studies and is reported with every circuit. The interventions establish causal structure inside the model (Pearl, 2009; Geiger et al., 2021); whether that structure corresponds to a term of the quantum-mechanical energy is a separate question, on which the case studies offer first evidence.

6 CONCLUSION

We formulated circuit discovery for GNN force fields and built a pipeline that carries it out automatically, from typed triangles and an exact decomposition through a causal check and a greedy search over role-typed units to confirmation on held-out molecules. On MACE, ORB and NEQUIP it returns 44 confirmed circuits that outperform the nearest units and compute recognisable chemistry; several boundaries per model and cross-architecture agreement of circuits are the next steps.

REPRODUCIBILITY STATEMENT

Every stage of the pipeline is specified in Appendix B with all constants (Table 5); the models, molecule panels, interfaces and statistical settings are in Appendix A; the freezing, molecule splits

and multiplicity family are in Appendix C; and every confirmed circuit with its statistics is in Appendix D. The pipeline runs as one command per model and family, and the code, frozen design files, per-molecule records and evidence tables are available at `https://anonymous.4open.science/r/enterpretability-00E3`. All models are used as publicly released, and the molecules come from a public test set.

ETHICS STATEMENT

This work analyses the internal computation of publicly released molecular force fields on public molecular data. It involves no human subjects, no personal data and no release of new datasets, and we see no direct path from its findings to harmful applications.

AI USE STATEMENT

Large language models were used to assist with editing the text, formatting tables and figures, and writing plotting scripts. All LLM-assisted text was reviewed against the underlying results, and every plotting script reads only the released evidence files and prints the numbers it plots, which were checked against the result tables. All methods, experiments, results, and claims were designed, run, and checked by the authors, who take full responsibility for the content, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9525–9536, 2018.

Michael P. Allen and Dominic J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2 edition, 2017.

Richard M. Badger. A relation between internuclear distances and bond force constants. *The Journal of Chemical Physics*, 2(3):128–131, 1934. doi: 10.1063/1.1749433.

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010. doi: 10.1103/PhysRevLett.104.136403.

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, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, William J. Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of E(3)-equivariant atom-centred interatomic potentials. *Nature Machine Intelligence*, 7(1):56–67, 2025. doi: 10.1038/s42256-024-00956-x.

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.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety – a review. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=ePUVetPKu6`.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://openreview.net/forum?id=8oSY3rA9jY`.

Klara Bonneau, Jonas Lederer, Clark Templeton, David Rosenberger, Lorenzo Giambagli, Klaus-Robert Müller, and Cecilia Clementi. Peering inside the black box by learning the relevance of many-body functions in neural network potentials. *Nature Communications*, 16:9898, 2025. doi: 10.1038/s41467-025-65863-0.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 5(6), 2020. doi: 10.23915/distill.00024.003.

Roberto Car and Michele Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55(22):2471–2474, 1985. doi: 10.1103/PhysRevLett.55.2471.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood,

Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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(9):1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019. doi: 10.1103/PhysRevB.99.014104.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293): 52–64, 1961. doi: 10.1080/01621459.1961.10482090.

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael M. Bronstein. A hitchhiker's guide to geometric GNNs for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10:11, 2023. doi: 10.1038/s41597-022-01882-6.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2 edition, 2002.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11515–11551, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Ronald J. Gillespie and Ronald S. Nyholm. Inorganic stereochemistry. *Quarterly Reviews, Chemical Society*, 11: 339–380, 1957. doi: 10.1039/QR9571100339.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Michel Grabisch. *Set Functions, Games and Capacities in Decision Making*, volume 46 of *Theory and Decision Library C*. Springer, 2016. doi: 10.1007/978-3-319-30690-2.

Thomas A. Halgren. Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5–6):490–519, 1996. doi: 10.1002/(SICI)1096-987X(199604)17:5/6⟨490::AID-JCC1⟩3.0.CO;2-P.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2403.17806.

Andreas Abildtrup Hansen, Anna Calissano, and Aasa Feragen. Interpreting equivariant representations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17538–17553, 2024.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025. doi: 10.1021/jacs.4c07099.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01235-7. Preprint: arXiv:2512.05349.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Pierre Mahé, Liva Ralaivola, Véronique Stoven, and Jean-Philippe Vert. The pharmacophore kernel for virtual screening with support vector machines. *Journal of Chemical Information and Modeling*, 46(5):2003–2014, 2006. doi: 10.1021/ci060138m.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2407.08734.

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.

Maximilian Muschalik, Fabian Fumagalli, Paolo Frazzetto, Janine Strotherm, Luca Hermes, Alessandro Sperduti, Eyke Hüllermeier, and Barbara Hammer. Exact computation of any-order shapley interactions for graph neural networks. In *International Conference on Learning Representations*, 2025.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), 2020. doi: 10.23915/distill.00024.001.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Linus Pauling. The nature of the chemical bond. application of results obtained from the quantum mechanics and from a theory of paramagnetic susceptibility to the structure of molecules. *Journal of the American Chemical Society*, 53(4):1367–1400, 1931. doi: 10.1021/ja01355a027.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Sayan Ranu and Ambuj K. Singh. Novel method for pharmacophore analysis by examining the joint pharmacophore space. *Journal of Chemical Information and Modeling*, 51(5):1106–1121, 2011. doi: 10.1021/ci100503y.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: Atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, 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, 2017.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388, 2021.

Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Günnemann, and Johannes T. Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Machine Learning: Science and Technology*, 3(4):045010, 2022. doi: 10.1088/2632-2153/ac9955.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268, 2020.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16): 10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Joshua A. Vita and Daniel Schwalbe-Koda. Data efficiency and extrapolation trends in neural network interatomic potentials. *Machine Learning: Science and Technology*, 4(3):035031, 2023. doi: 10.1088/2632-2153/acf115.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14:2585, 2023. doi: 10.1038/s41467-023-38192-3.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

APPENDIX OVERVIEW

The appendices hold what the main text relies on but does not need in full. They follow the order of the paper: the experimental setup, then the algorithms and the confirmation protocol behind Algorithm 1, the complete confirmation results, the decomposition that produces the candidates, the ablations, the case studies and the extended related work. Throughout, energies stay in each model's native convention, the molecule is the unit of replication, and every count is either a deterministic decision or an estimate with a molecule-bootstrap interval.

A Experimental setup . . . 14
B Algorithms . . . 16
C Confirmation protocol and disclosures . . . 20
D Full confirmation results . . . 21
E The decomposition behind the circuits . . . 23
E.1 Numerical reliability of the smallest threshold . . . 24
E.2 Random initialisation and a classical comparator . . . 24
E.3 Matched-depth re-test . . . 25
F Ablation details . . . 26
F.1 Circuits against the nearest units . . . 26
F.2 Hyperparameter sensitivity . . . 28
F.3 Choices in the decomposition stage . . . 29
F.4 Grain of the interface . . . 30
F.5 Perturbation step size . . . 31
F.6 Cost of the exact decomposition . . . 32
G Case studies . . . 33
G.1 Interventions on confirmed circuits . . . 33
G.2 Component-level interventions . . . 33
H Extended related work . . . 37

- **Appendix A, Experimental setup.** The three checkpoints and their accuracy, the behaviour vocabulary, the declared interface of every model and the statistical settings.
- **Appendix B, Algorithms.** Each stage of the pipeline written out with every constant, the locality baseline of the ablation, the role-typed unit graph of every model and the implementation details.
- **Appendix C, Confirmation protocol and disclosures.** Freezing, molecule splits, the strict and topped-up readings, the multiplicity family, the control readings and the independent recompute.
- **Appendix D, Full confirmation results.** Every count of the confirmation stage and every confirmed circuit with its statistics.
- **Appendix E, The decomposition behind the circuits.** What the exact decomposition returns on each model, which units interact, the complete ledger, numerical reliability, the random-initialisation and classical comparators, and the matched-depth re-test.
- **Appendix F, Ablation details.** The locality comparison per circuit, the full sensitivity sweep and the τ rerun, the choices in the decomposition stage, the grain of the interface, the step size and the cost of the decomposition.
- **Appendix G, Case studies.** The seven circuit case studies in full and the component-level interventions.
- **Appendix H, Extended related work.**

A Experimental setup

Models and estimands. The three checkpoints are used as released, with no fine-tuning and no fitted energy offset: MACE-OFF23 medium (Kovács et al., 2025), trained on SPICE (Eastman et al., 2023); the conservative Orb v3 checkpoint trained on OMol (Neumann et al., 2024; Rhodes et al., 2025; Levine et al., 2025); and the released organic-molecule NequIP checkpoint (Batzner et al., 2022). The estimand ϕ_B is the total energy for MACE, the total energy at the stored orientation for Orb, and the learned per-atom energy sum before the constant atomic reference energies for NequIP. Table 3 gives their accuracy on the 200 held-out molecules under one protocol. The force errors lie within a factor of 2.4 of one another; the raw energy errors of Orb and NequIP are on their own reference conventions and are reported as such. MACE and Orb run in double precision throughout; the NequIP decomposition games ran in single precision, for which Appendix E reports a precision screen, and its circuit stages run the released weights in double-precision arithmetic.

Behaviour vocabulary. The population is 400 molecules from the OFF23 test set, split 200/200 into discovery and held-out identities by hash before any model output is seen. Triangles are formed about every centre together with bonded pairs; family labels are species plus bond order, and neighbour roles are canonicalised by species and geometry (Algorithm 2). Enumeration records 42,978 occurrences in 126 families. A family is eligible when at least 20 discovery and 20 held-out identities contain an admissible occurrence; 38 qualify (23 angular, 15 bond-stretch) and 198 discovery and 200 held-out

Model	Energy MAE (meV/atom)	Force MAE (eV/Å)
MACE	0.82 [0.72, 0.94]	0.0197 [0.0185, 0.0209]
ORB	516.3 [501.2, 533.1]†	0.0360 [0.0352, 0.0368]
NEQUIP	515.9 [500.8, 532.7]†	0.0473 [0.0459, 0.0490]

Table 3: **Accuracy on the held-out panel.** 200 molecules, OFF23 reference labels, raw native outputs, 95% molecule-bootstrap intervals. †Raw total energy on the model's own reference convention with no fitted offset. Forces are reference-free.

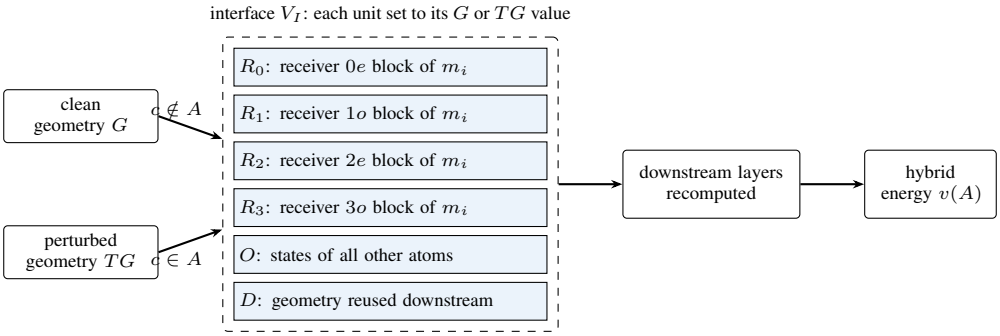

Figure 5: **The MACE interface.** Units in A take their perturbed value and the rest their clean value; the downstream layers are recomputed and the energy read out.

identities contribute. The angular probe rotates one neighbour about the centre by $+2^\circ$ with both bond lengths held; the radial probe stretches one bond by $+0.02$ Å. Admissible ranges are the 1st–99th percentiles of the discovery molecules. ORB builds its own neighbour graph with a 6 Å radius; the 524 perturbations that change that graph, and the one family (C–S–O) that then falls below the support minimum, are removed before any ORB output is seen, leaving 5,877 perturbations in 37 families.

The interfaces. Each model declares six units at one internal boundary (Table 4). Every switch set that contains a receiver unit is receiver-associated; sets built only from the other-atoms and geometry units (three per model) measure bypass routes.

Model	Boundary	Six units
MACE	incoming message to the receiver, before the first product block	receiver $0e$, $1o$, $2e$, $3o$ blocks; all other atoms; geometry reused downstream
ORB	input to the last of five message-passing blocks	receiver node; edges into the receiver; edges out of it; other nodes; other edges; geometry reused downstream
NEQUIP	output of the third of five interaction layers	receiver $0e$, $0o$, $\ell{=}1$, $\ell{=}2$ blocks; all other atoms; geometry reused downstream

Table 4: **The declared interface of each model.** One boundary and six units per model; the last two units of each row are the bypass units.

MACE. Per-atom states are irreps of orders $0e, 1o, 2e, 3o$; each interaction block multiplies several incoming irreps, and the number of factors is the correlation order. Let m_i be the receiver's aggregated incoming message before the first product block. The interface is $V_I = \{R_0, R_1, R_2, R_3, O, D\}$ (Figure 5): the complete $0e, 1o, 2e, 3o$ blocks of m_i, each switched with all channels and components together; O, the natural changes at all other atoms, as one grouped unit; and D, the geometry the second interaction reads again. O and D are included because a response attributed to the receiver's incoming blocks may travel through the neighbouring atoms' own updated states or through geometry the network reads a second time. Endpoint reconstruction holds because the four irreps, the other atoms and the downstream geometry exhaust the second interaction's dependence on the perturbation; the measured residual is at most 1.7×10^{-13} eV over all games. The first sender embedding is species-only, so an angular probe leaves the receiver $0e$ message unchanged and every coefficient with $R_0 \in S$ vanishes on angular games.

ORB. The latent states are unrestricted scalars in a fixed frame, so each unit is a complete 256-channel group; the geometry unit holds the final-block cutoffs and the graph passed to the repulsion branch. The endpoint tolerance is 10^{-8} eV. NEQUIP. The $\ell{=}1$ unit holds the $1o$ and $1e$ blocks together and

the ℓ=2 unit the $2e$ and $2o$ blocks; the other-atoms unit holds the complete layer states of every other atom.

Statistical settings. Coefficients are absolute changes in eV under the fixed step size. The thresholds are $\delta \in \{10\,\mu\mathrm{eV}, 0.1\,\mathrm{meV}, 1\,\mathrm{meV}\}$ with $10\,\mu$eV primary: it sits well above the reconstruction error of MACE and ORB and measures internal sensitivity. The prevalence boundary is $\pi_0 = \frac{1}{2}$, so "recurring" means the effect is present in a majority of the family's held-out molecules. All labels of a model, every family, switch set and threshold, enter one Benjamini–Yekutieli correction at 5% FDR (Benjamini & Yekutieli, 2001); nonrecurrence bounds are 95% simultaneous Bonferroni-adjusted exact binomial upper bounds. The molecule is the unit of replication throughout, and every estimate carries a molecule-bootstrap interval.

Compute. MACE and NEQUIP run on CPU and ORB on one GPU. Per candidate, the causal check costs a median 16, 2 and 47 CPU-seconds per molecule, localisation 83 s, 245 s and 44 min of wall time on eight workers, and confirmation 48 s, 33 s and 6 min, for MACE, ORB and NEQUIP. Code, configuration files and all evidence tables will be released with the paper.

B ALGORITHMS

This appendix writes out each stage of Algorithm 1 as it is implemented, with every constant. Throughout, molecules are indexed by i, x_i is the clean geometry and x_i' the perturbed one, $\Delta_i = y(x_i') - y(x_i)$, and e_i is the component's effect at full dose. Every interval is a molecule bootstrap of a ratio of sums: rows are resampled with replacement, the statistic $\sum_i \mathrm{num}_i / \sum_i \mathrm{den}_i$ is recomputed per draw, and a percentile interval is taken; seeds are deterministic hashes of the candidate and statistic name. Table 5 collects the settings.

Algorithm 2 Stage 1: typed-triangle families, the 2^6-way game and recurrence.

Input: molecules $\mathcal{D}$ with bond graphs; model f_θ; six units $V_I = (u_0, \ldots, u_5)$ at one boundary; threshold $\delta = 10\,\mu$eV; null prevalence $\pi_0 = \frac{1}{2}$; FDR level 0.05.

1: **Occurrences.** For every bonded pair $\{i, j\}$, order the atoms by (species, distance to centroid, index): the first is the receiver R, the second the moving atom M; family label Sym_R–Sym_M:bond type. For every atom i and unordered pair $\{j, k\}$ of its bonded neighbours, order (j, k) by (species, bond length, index): $R = i$, $L = j$, $M = k$; label Sym_L–Sym_R–Sym_M.

2: **Perturbation.** Radial: move M along the bond so that $r_{RM} \leftarrow r_{RM} + 0.02$ Å. Angular: rotate $\mathbf{r}_M - \mathbf{r}_R$ about the axis $(\mathbf{r}_L - \mathbf{r}_R) \times (\mathbf{r}_M - \mathbf{r}_R)$ by $+2^\circ$, leaving both bond lengths unchanged.

3: **Admission.** An occurrence is admissible when the angle lies in $(5^\circ, 175^\circ)$, the perturbed bond length (radial) or the two radii and angle (angular) lie inside the family's 1st–99th percentile window on discovery molecules, no interatomic distance falls below $0.65(r_a + r_b)$ in covalent radii, and the covalent graph, the 5 Å neighbour graph and, for ORB, the 6 Å directed pair set are unchanged.

4: **Selection.** Among admissible occurrences of a family in a molecule keep the one minimising $\mathrm{sha256}(\text{seed} \mid \text{identity} \mid \text{atoms})$.

5: **Corners.** For each mask $A \subseteq \{0, \ldots, 5\}$ build the hybrid state in which unit u_b takes its value on x' iff $b \in A$ and its value on x otherwise, re-run every layer downstream of the boundary with the model's own modules, and record $v(A) = y(\text{hybrid}) - y(x)$. Gate: $|v(\emptyset)| \le 10^{-8}$ eV and $|v(V_I) - \Delta| \le 10^{-8}$ eV.

6: **Inversion.** $c(S) \leftarrow \sum_{A \subseteq S} (-1)^{|S \setminus A|} v(A)$ for all 63 non-empty S, by the in-place transform **for** $b = 0..5$, **for** every mask with bit b set: $c[\text{mask}] \leftarrow c[\text{mask}] - c[\text{mask} \oplus 2^b]$. Check $\left|\sum_{S \neq \emptyset} c(S) - \Delta\right| \le$ tolerance (10^{-8}, 10^{-6}, 10^{-4} eV for MACE, ORB, NEQUIP) on every game.

7: **Recurrence.** For each family and S, on the held-out split with $n \ge 20$ molecules: $k \leftarrow \#\{i : |c_i(S)| > \delta\}$, $p \leftarrow \Pr[\mathrm{Bin}(n, \frac{1}{2}) \ge k]$. Apply Benjamini–Yekutieli over all families, sets and the three thresholds $\{10^{-5}, 10^{-4}, 10^{-3}\}$ eV of a model jointly: $q_{(j)} = \min_{l \ge j} p_{(l)} M H_M / l$. The term recurs iff $q \le 0.05$ at the primary threshold.

Output: the recurring terms $(\mathcal{F}, S)$ of the model with their coefficients $c_i(S)$ and Δ_i on discovery games.

Algorithm 3 Stages 2–3: screen and causal check.

Input: recurring terms of one family with discovery-game coefficients; a discovery cohort of 30 molecules per family, scaffold-split from the held-out molecules; doses $(\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1)$; control seeds random 4, orthogonal 2, permutation 2 (latent units) and random 4, orthogonal 2 (geometry units); $n_{\min} = 20$; $B = 10{,}000$ draws.

1: **Score.** For each recurring S: $\mathrm{score}(S) = \sum_i |c_i(S)| / \sum_i |\Delta_i|$ and $\mathrm{signed}(S) = \sum_i \mathrm{sgn}(\Delta_i) c_i(S) / \sum_i |\Delta_i|$ over discovery games; rank by $(-\mathrm{score}, -\mathrm{signed}, \mathrm{mask})$.

2: **Screen.** Keep rank 1 (*top-1*), ranks 1–3 (*top-3*), and the best-ranked S whose units are all receiver-side blocks, geometry excluded (*top-1 site-local*). Each distinct S is run once; its component C is the joint patch of its units.

3: **Arms.** For each cohort molecule: clean and perturbed forwards give y_0, y_1, Δ and the numerical noise. For each dose α write $a_c(x) + \alpha\, \Delta a_c$ into every unit of C and record $e_i(\alpha) = y - y_0$ and whether every written block norm lies inside the envelope of the clean and perturbed states (support); $e_i \leftarrow e_i(1)$. For each control class and seed k write wrong content of matched norm (Haar-random channel mixing; a random direction orthogonal to Δa_c rescaled to its norm; a channel derangement; for geometry, a displacement of M by $|\Delta \mathbf{r}_M|$ in a random or orthogonal direction, redrawn if any graph changes) and record $e_{i,k}$.

4: **Admission.** Keep molecules with all gates passed, $|\Delta_i| \geq \max(10^{-4}\,\mathrm{eV},\ 100 \cdot \mathrm{noise}_i)$ and every dose arm in support. If fewer than $n_{\min}$ remain: *unavailable*.

5: **Gate 1.** $\rho_C = \sum_i \mathrm{sgn}(\Delta_i) e_i / \sum_i |\Delta_i|$ with bootstrap interval; $\rho(\alpha)$ likewise per dose. Pass iff the lower bound of ρ_C exceeds 0, the four $\rho(\alpha)$ share one sign and are non-decreasing in magnitude, and more than half of the molecules have a per-molecule dose curve of one sign that is non-decreasing within $\max(10 \cdot \mathrm{noise}_i, 10^{-9})$.

6: **Gate 2.** A control class is *live* when its median $|e_{i,k}|$ exceeds $10 \cdot \max(\mathrm{median}_i\, \mathrm{noise}_i, 10^{-9})$. For every live class compute the contrast $\sum_i [\mathrm{sgn}(\Delta_i) e_i - \mathrm{mean}_k\, \mathrm{sgn}(\Delta_i) e_{i,k}] / \sum_i |\Delta_i|$ on the admitted molecules; pass iff every lower bound exceeds 0 and every class is complete on the admitted set. (Worst-seed and per-seed contrasts are recorded as sensitivity readings.)

Output: *candidate* iff Gate 1 and Gate 2 pass.

Algorithm 4 Stage 4: greedy localisation over role-typed units.

Input: candidate C with its admitted discovery molecules and effects e_i; $\tau = 0.8$; real-patch cap 200 units; prefix grid dense to 64 then $\times 1.25$; prune chunks $|S|/8$; live threshold $10^{-3} \sum_i |e_i|$; $K = 20$ random subgraphs; at least 10 molecules.

1: **Roles.** With $\mathrm{tri} = \{R, M\} \cup \{L\}$: atom a is R, M or L if it is that atom; nM if bonded to M; nR if bonded to R or L; $B2$ if at graph distance 2 from tri; else X. Units are typed by role rather than by atom index so that the graph is the same on every molecule of the family and a circuit found on discovery molecules is defined on every held-out molecule.

2: **Unit graph.** U = upstream objects that build $C(x')$ (messages and node updates by layer and sender>receiver roles), the pieces of C (dyadic channel bands of each block in a discovery-fitted PCA basis, by role; typed geometric edges), and downstream objects (messages, self-updates and readouts by layer and role). $v(S)$: upstream pass on x' with objects in S hybrid and others clean; write only C's slots, projected onto the in-units in S (whole irreps); downstream pass on x with objects in S hybrid and others clean. Check $|v_i(U) - e_i|$ and $|v_i(\emptyset)| \leq 10^{-6} \max(|e_i|, |\Delta_i|) + 10 \cdot \mathrm{noise}_i + 10^{-8}$ per molecule; drop failures; require ≥ 10 molecules.

3: **Objective.** $F(S) \leftarrow \sum_i \mathrm{sgn}(e_i) v_i(S) / \sum_i |e_i|$, so $F(U) = 1$.

4: **Rank.** If $|U| \leq 200$: for each u, $\mathrm{key}_u = F(e - v(U \setminus \{u\})) + F(v(\{u\} \cup \text{other two regions}))$, $\mathrm{mag}_u = \sum_i |e_i - v_i(U \setminus \{u\})| + |v_i(\{u\} \cup \cdots)|$. Else use attribution $A_{i,u} = (\text{hybrid} - \text{clean}) \cdot \nabla y$: $\mathrm{key}_u = F(A_{\cdot,u})$, $\mathrm{mag}_u = \sum_i |A_{i,u}|$.

5: **Live.** u is live iff it binds an object on some molecule and $\mathrm{mag}_u > 10^{-3} \sum_i |e_i|$. Order live units by key descending.

6: **Prefix.** Grid $k = 1, \ldots, \min(L, 64)$, then $k \leftarrow \min(L, \max(k{+}1, \lceil 1.25k \rceil))$. $S_0 \leftarrow$ the first prefix with $\tau \leq F \leq 2 - \tau$, refined densely between the previous grid point and the hit; if none, $S_0 \leftarrow U$ (recorded as a fallback). The band is two-sided because a set that overshoots by dropping a cancelling path does not reproduce the behaviour either.

7: **Prune.** Queue the units of S_0 least important first in chunks of $\lceil |S_0|/8 \rceil$. Pop a chunk; if $F(S \setminus \text{chunk})$ stays in the band, remove it; else if the chunk has more than one unit, split it in two and re-queue both halves at the front; else keep the unit. $S^* \leftarrow S$.

8: **Random subgraphs.** For $k = 1, \ldots, K$, with seed $\mathrm{sha256}(\text{candidate} \mid \text{random} \mid k)$, draw R_k from the live units of each stratum with the same number of units as S^* has there.

Output: S^*, its unit graph U, ranking, prefix curve, prune log and the $K = 20$ random subgraphs $R_1, \ldots, R_{20}$, hashed before any held-out molecule is loaded.

Stage	Setting	Value
1	Perturbation	$+2^\circ$ per angle, $+0.02$ Å per bond
1	Units	six per model (fixed bit order)
2	Screen score, tie-break	$\sum \lvert c(S)\rvert / \sum \lvert\Delta\rvert$; signed share; mask
2	Screen variants	top-1, top-1 site-local, top-3 (union)
3	Doses	0.25, 0.5, 0.75, 1.0
3	Control classes (seeds)	random (4), orthogonal (2), permutation (2), geometric random (4), geometric orthogonal (2)
3	Live control class	median $\lvert e_{\mathrm{ctrl}}\rvert > 10\times$ noise
3	Admission	prepare gates; $\lvert\Delta\rvert \geq \max(\text{floor}, 100 \times \text{noise})$; every dose in support
3	Minimum admitted molecules	20
3	Bootstrap	10,000 identity draws, 95% percentile
4	Faithfulness band τ	0.8, i.e. $[0.8, 1.2]$
4	Real-patch ranking cap	200 units (attribution above)
4	Prefix grid	$1, \ldots, 64$, then $\times 1.25$; dense in the last gap
4	Prune chunks	$\lvert S\rvert/8$, halved on rejection
4	Live unit	bound on ≥ 1 molecule and aggregate $\lvert\text{effect}\rvert > 10^{-3} \sum \lvert e\rvert$
4	Random subgraphs	20, stratum-matched
4	Minimum discovery molecules	10
5	Sufficiency	adjusted interval inside $(0.5, 1.5)$
5	Necessity	adjusted lower bound > 0.5
5	Beats random	adjusted lower bound of $\mathrm{Suff} - \mathrm{Suff}_{\mathrm{rand}} > 0$
5	Localised	$\lvert S^*\rvert/\lvert\text{live}\rvert \leq 0.5$ (a size rule, not a test)
5	Multiplicity	Bonferroni, $m = 3N_5 = 813$; two-sided $1 - 0.05/m$ intervals

Table 5: **Frozen hyperparameters of the discovery pipeline, by stage.** Settings not listed here are given in the implementation-details paragraph of this appendix.

Algorithm 5 Stage 5: confirmation on held-out molecules.

Input: frozen S^*, U and random subgraphs $R_1, \ldots, R_{20}$; the family's reserved molecules, never evaluated before; N_5 = candidates reaching this stage over all models; $\alpha = 0.05$; $n_{\min} = 20$.

1: **Verify.** Check the circuit hash, the code digest and that no held-out identity is a discovery identity.
2: **Arms.** Per molecule: $v(U)$, $v(\emptyset)$, $v(S^*)$, $v(U \setminus S^*)$, $v(R_k)$ and $v(U \setminus R_k)$ for all k, and the four dose arms of C (giving e_i).
3: **Admission.** As in stage 3, plus $\lvert v_i(U) - e_i\rvert$ and $\lvert v_i(\emptyset)\rvert$ within tolerance and $e_i \neq 0$; fewer than $n_{\min}$: *unavailable*, still counted in N_5.
4: **Statistics.** With $w_i = \mathrm{sgn}(e_i)$, $a_i = \lvert e_i\rvert$, $s_i = v_i(S^*)$, $c_i = v_i(U \setminus S^*)$, $\bar{r}_i = \mathrm{mean}_k v_i(R_k)$:
$\mathrm{Suff} = \sum_i w_i s_i / \sum_i a_i$, $\mathrm{Nec} = \sum_i (a_i - w_i c_i) / \sum_i a_i$,
$\mathrm{Suff}_{\mathrm{rand}} = \sum_i w_i \bar{r}_i / \sum_i a_i$, $\mathrm{Contrast} = \sum_i w_i (s_i - \bar{r}_i) / \sum_i a_i$.
5: **Intervals.** $m \leftarrow 3N_5$, three tests for every candidate that reaches this stage, so that the family is fixed before any outcome is known; two-sided level $1 - \alpha/m$ with $\max(10^4, \lceil 20 \cdot 2m/\alpha \rceil)$ bootstrap draws (here $m = 813$, 650,400 draws, so that the adjusted tail holds at least 20 draws); unadjusted 95% intervals are printed beside them.
6: **Tests.** Sufficiency: adjusted interval inside $(0.5, 1.5)$; necessity: adjusted lower bound > 0.5; beats random: adjusted lower bound of Contrast > 0. Localised: $\lvert S^*\rvert/\lvert\text{live}(U)\rvert \leq 0.5$ (a size rule, not a test).

Output: `confirmed_circuit` iff all three tests pass and localised; `confirmed_not_localised` iff all three pass; otherwise `partial`: ⟨tests passed⟩ or `not_confirmed`. Also recorded: the empirical rank of Suff among the 20 random draws, the share of Δ carried by C and by S^*, and per-molecule medians.

Algorithm 6 Locality baseline for the ablation of Section 4.2.

Input: a confirmed circuit S^* with its unit graph U and live mask; its frozen admitted held-out molecules and random draws; role distances $R, M, L \mapsto 0$, $nM, nR \mapsto 1$, $B2 \mapsto 2$, $X \mapsto 3$; role ranks $R < M < L < nM < nR < B2 < X$; $D = 20$ draws; Bonferroni family $m_{\mathrm{loc}} = 122$.

1: **Distance key.** For each unit u, roles(u) = the roles named in its type; $\mathrm{dist}(u) = \max_\rho \mathrm{dist}(\rho)$; key$(u) = (\mathrm{dist}(u), \sum_\rho \mathrm{dist}(\rho),$ sorted ranks$)$.

2: **Nearest units.** $k \leftarrow |S^*|$. Group live units into tiers of equal key in ascending order; take whole tiers while they fit in k; fill the remainder from the boundary tier uniformly without replacement with seed $\mathrm{sha256}(\text{candidate} \,|\, \text{loc} \,|\, d)$, giving $\mathcal{L}_d$, $d = 1, \ldots, D$.

3: **Stratum-matched variant.** Repeat step 2 inside each stratum of S^* with that stratum's unit count, union over strata: $\mathcal{L}^s_d$.

4: **Arms.** Evaluate $v(\mathcal{L}_d)$, $v(U \setminus \mathcal{L}_d)$ and likewise for $\mathcal{L}^s_d$ on the frozen admitted molecules through the unchanged adapters; check that $v(S^*)$, $v(U \setminus S^*)$ and $v(U)$ reproduce the frozen records.

5: **Statistics.** With w_i, a_i, s_i, c_i as in Algorithm 5, $\bar{l}_i = \mathrm{mean}_d v_i(\mathcal{L}_d)$ and $\bar{l}^c_i = \mathrm{mean}_d v_i(U \setminus \mathcal{L}_d)$:
$\mathrm{Suff}_{\mathcal{L}} = \sum_i w_i \bar{l}_i / \sum_i a_i$, $\mathrm{Nec}_{\mathcal{L}} = \sum_i (a_i - w_i \bar{l}^c_i) / \sum_i a_i$,
$\Delta\mathrm{Suff} = \sum_i w_i (s_i - \bar{l}_i) / \sum_i a_i$, $\Delta\mathrm{Nec} = \sum_i w_i (\bar{l}^c_i - c_i) / \sum_i a_i$.

6: **Grading.** Grade $\mathcal{L}$ with the criteria of Algorithm 5 at $m = 813$ (against the circuit's own random draws). Paired differences get two-sided $1 - 0.05/122$ intervals: *found better* if the lower bound exceeds 0, *locality better* if the upper bound is below 0, otherwise not distinguishable; *identical* if all D draws equal S^*.

Output: per circuit: the grade of $\mathcal{L}$ and of $\mathcal{L}^s$, and the paired differences $\Delta\mathrm{Suff}$ and $\Delta\mathrm{Nec}$ with their adjusted intervals.

Per-model unit graphs. MACE *(medium, two interactions):* the boundary is the input of the first product block, the receiver's $0e$, $1o$, $2e$, $3o$ blocks of the aggregated message; upstream units are the layer-0 messages into the component's rows by sender>receiver role; in-units are dyadic channel bands $\{0\}, \{1\}, \{2,3\}, \{4..7\}, \ldots$ of each block in a PCA basis fitted on the discovery molecules, and, for the geometry unit, the layer-1 edge attributes and radial weights of every edge touching M; downstream units are the receiver's readout and skip connection and the layer-1 messages it sends. Unit graphs have 16–24 units. ORB *(v3 conservative, five blocks):* the boundary is the input of block 5; upstream units are encoder edge latents, block 1–4 messages with their gates and block 1–4 node updates in the component's cone; in-units are nine dyadic bands of the 256-channel latent per slot and role, the block-5 cutoff weights and the pair-repulsion terms; downstream units are the block-5 send and receive terms, node update and the node's input to the mean-pooled head. Unit graphs have 231–593 units. NEQUIP *(five interaction layers):* the boundary is the output of layer 2; upstream units are the messages and self-connections of layers 0–2 into the component's rows; in-units are dyadic bands of the $0e$, $0o$, $1o$, $1e$, $2e$, $2o$ blocks and the perturbed edge embeddings of layers 3–4 on M-incident edges; downstream units are the messages, self-connections and readouts of layers 3–4. Unit graphs have 76–445 units. Every adapter verifies $v(U) = e_C$, $v(\emptyset) = 0$, determinism, rotation invariance where the model is exactly equivariant, and that units outside the downstream region alone carry nothing for latent components.

Implementation details. Each model runs in its released framework: MACE in PyTorch in double precision on CPU, ORB in PyTorch in double precision on one GPU, and NEQUIP in JAX on CPU. One command,

```
python -m autopromote run --model M --family F --stages all,
```

runs stages 1–5 for one model and one perturbation family; it is a thin driver over the frozen stage code and adds no statistic or threshold of its own. Every finished stage writes its own record, and the localisation stage writes one circuit file per candidate (unit graph, ranking, prefix curve, prune log, S^* and the 20 random subgraphs), which is hashed into a per-model ledger before any confirmation job starts; at the end the command writes a result card with the stage-by-stage counts, the chosen component, the circuit, its sufficiency and necessity and the runtime per stage. Every design file and the code snapshot are hashed before the stage they govern runs, and the confirmation stage refuses to start unless the design hashes verify, the running code's digest equals the reviewed one, every circuit hash matches the ledger, the checkpoint hash equals the localisation run's and no held-out identity is a discovery identity. Bootstrap seeds are deterministic hashes of the candidate and statistic name. Settings that Table 5 does not list: the stage-1 thresholds $\delta \in \{10\,\mu\text{eV}, 0.1\,\text{meV}, 1\,\text{meV}\}$ with $10\,\mu$eV primary, the prevalence boundary $\pi_0 = \frac{1}{2}$ and the Benjamini–Yekutieli level of 5%; the stage-3 admission floor of 10^{-4} eV; the stage-5 bootstrap size $\max(10^4, \lceil 20 \cdot 2m/0.05 \rceil) = 650{,}400$ draws at $m = 813$; and the contract tolerance $10^{-6} \max(|e_i|, |\Delta_i|) + 10 \cdot \text{noise}_i + 10^{-8}$ eV of stages 4–5. Per-candidate costs are in the compute paragraph of Appendix A; end to end, one family with three components takes 33 minutes for MACE on four CPU workers and 19 minutes for ORB on three workers and one GPU, and one family with its top-1 component takes about three hours for NEQUIP on two CPU workers; the screen takes under a second in every run.

C CONFIRMATION PROTOCOL AND DISCLOSURES

Frozen design. The designs of stages 1–3 and of stages 4–5 were written and hashed before any outcome of the stage they govern existed; the threshold $\tau = 0.8$, the 0.5 criteria and the number of random subgraphs were set a priori on exposed development molecules and a toy model, and no per-candidate choice exists anywhere in the pipeline. The code snapshot that ran each stage is hashed, every circuit file is hashed before the first held-out molecule is loaded, and the confirmation stage refuses to run unless the design hashes verify, the code digest matches the reviewed one, the held-out cohort is the family's live reserved set and no held-out identity appears among the discovery identities. Freezing is what makes the confirmation counts interpretable: with every threshold fixed in advance, a confirmed circuit cannot be the product of a choice made after seeing the held-out molecules.

Molecules. Stage 1 uses the 400-molecule panel of Appendix A, split 200/200 by hash into discovery and held-out identities; recurrence is decided on the held-out half. Stages 3–4 use one cohort of 30 discovery molecules per family drawn from the MACE-OFF23 test set and shared

by all models; the cohort's molecules are grouped by Murcko scaffold, and the held-out sets are scaffold-disjoint from them, so that a circuit cannot be confirmed by close analogues of the molecules it was found on. Stage 5 uses a reserved set of up to 50 fresh molecules per family (1,107 in all), allocated before stage 3 ran and never evaluated by any earlier stage of any model. Two families received none and seven more fewer than 20; their candidates are *unavailable* in the strict reading and still count in m.

Strict and topped-up readings. The strict reading confirms only on the reserved set. The topped-up reading exists because nine families have too few reserved molecules to grade at all: it completes their held-out sets to 40 molecules with 296 molecules that appeared only in discovery halves of earlier development cohorts, never in any held-out half, circuit selection or stage-1 panel, and whose scaffolds are absent from the family's discovery set. These molecules are independent of the selection of every candidate and circuit; they are not fresh in the strict sense, so the topped-up counts are the secondary column. The 60 affected candidates use their frozen circuits unchanged, and every other candidate is identical in both readings.

Multiplicity. $m = 3N_5 = 813$ counts three tests for each of the 271 candidates that reached stage 5 across all models graded under one ledger. The family is every test whose outcome could be reported as a confirmation, which is the set a reader needs corrected for; Bonferroni is used rather than a false-discovery-rate procedure because each confirmed circuit is an individual claim, not a member of a set whose error rate is the quantity of interest. The ledger includes candidates from a further architecture outside the scope of this paper, which makes the correction conservative for the three models reported.

Control readings. The causal check reports the contrast against the mean of each control class as the primary reading, with the worst-seed and per-seed contrasts beside it. The mean over seeds estimates the expected response to wrong content of the same norm, which is the quantity the control is meant to bound; the worst-seed reading depends on how many seeds are drawn and tightens with every additional seed. Appendix F.2 gives the counts under all three readings.

Independent recompute. A reviewer with no access to the pipeline's code recomputed every stage-5 statistic and label from the raw per-molecule records and the frozen circuits: 30,135 floating-point values and all 542 candidate-by-reading labels matched bitwise, with $N_5 = 271$ and $m = 813$; the locality ablation and the sensitivity recount were reproduced the same way. The one-command interface, the per-stage records and the hashing of designs, code and circuits are described in the implementation details of Appendix B.

D FULL CONFIRMATION RESULTS

Table 6 gives every count of the confirmation stage per model and reading, and Table 7 lists every circuit confirmed in at least one reading with its size and test statistics. Two readings of the table are useful. For the small MACE circuits in receiver-block graphs of 10–12 live units, random subgraphs of the same size also carry much of the behaviour (random sufficiency 0.25–0.70), so these circuits pass on the contrast bound and on necessity, and their empirical rank among the 21 subgraphs is descriptive. The ORB, NEQUIP and larger MACE circuits are the best of the 21 with random sufficiency near zero. Several necessity points exceed one: resetting the circuit then overshoots the behaviour, and the criterion is the adjusted lower bound above 0.5. Almost every confirmed circuit comes from the top-3 screen; the top-1 screen alone confirms 1, 2 and 6 circuits and the site-local screen 10, 7 and 8 for MACE, ORB and NEQUIP.

Reading	Passes	Graded	Unavail.	Localised	Suff.	Nec.	Beats random	All three	Circuits	Conf., not localised	Partial	Not conf.
MACE												
strict	68	51	17	47	19	37	37	19	16	3	24	8
topped-up	68	65	3	60	25	46	48	25	21	4	30	10
ORB												
strict	86	64	22	63	18	54	51	18	17	1	39	7
topped-up	86	81	5	80	23	71	67	23	22	1	51	7
NEQUIP												
strict	73	53	20	52	12	42	30	12	11	1	34	7
topped-up	73	71	2	70	16	59	41	16	15	1	47	8

Table 6: **Confirmation outcomes per model and reading.** All screen variants pooled; a component counts once. Passes: components that passed the stage-3 causal check; unavailable: fewer than 20 admitted confirmation molecules; suff. / nec.: sufficient / necessary; circuits: confirmed circuits.

Table 7: **Every circuit confirmed in at least one reading.** 58 circuits: MACE 21, ORB 22, NEQUIP 15. Statistics are those of the reading the circuit qualified in; intervals and lower bounds are Bonferroni-adjusted at $m = 813$.

Family	Screen	Reading	$\lvert S^* \rvert$ / live / all	n	Sufficiency	Necessity	Random suff.	Contrast lower bd.	Share of Δ
MACE (21 circuits)									
C–C–Cl angle	1L, 3	strict	3 / 11 / 17	31	1.12 [0.83, 1.40]	1.00 (1.00)	0.56	0.42	0.11
C–C–F angle	1L, 3	top-up	3 / 10 / 17	40	1.09 [0.93, 1.21]	1.00 (1.00)	0.44	0.56	0.17
C–C–O angle	3	strict	12 / 24 / 24	42	0.97 [0.76, 1.24]	0.97 (0.75)	0.34	0.45	0.39
C–C–O angle	1L, 3	strict	3 / 12 / 17	41	0.99 [0.72, 1.17]	1.00 (1.00)	0.25	0.53	0.07
C–N–C angle	3	strict	6 / 24 / 24	38	0.99 [0.79, 1.19]	0.98 (0.79)	0.02	0.76	0.43
C–O–C angle	3	strict	6 / 24 / 24	32	0.88 [0.70, 1.06]	0.88 (0.70)	0.06	0.64	0.52
C–O–C angle	1L, 3	strict	5 / 11 / 17	32	1.15 [1.08, 1.22]	1.00 (1.00)	0.70	0.42	0.17
C–S–C angle	3	top-up	4 / 24 / 24	40	0.86 [0.73, 0.98]	0.85 (0.73)	−0.00	0.73	0.51
H–C–C angle	1L, 3	strict	3 / 12 / 17	45	1.23 [1.08, 1.37]	1.00 (1.00)	0.43	0.70	0.02
H–N–C angle	3	strict	8 / 24 / 24	38	0.84 [0.68, 1.02]	0.85 (0.68)	0.12	0.56	0.35
H–O–C angle	3	top-up	6 / 20 / 24	40	0.77 [0.65, 0.87]	0.77 (0.66)	0.11	0.55	0.43
N–C–N angle	3	strict	8 / 24 / 24	33	0.85 [0.60, 1.05]	0.85 (0.58)	0.11	0.50	0.41
N–C–O angle	1L, 3	strict	3 / 12 / 17	35	1.15 [0.90, 1.45]	1.00 (1.00)	0.46	0.54	0.17
O–C–O angle	1L, 3	strict	4 / 12 / 17	23	1.13 [0.92, 1.35]	1.00 (1.00)	0.52	0.46	0.17
C=C stretch	1L, 3	strict	5 / 11 / 16	36	1.11 [0.99, 1.29]	1.00 (1.00)	0.70	0.33	0.15
C=C stretch	1, 3	strict	4 / 20 / 20	36	0.97 [0.86, 1.11]	0.97 (0.86)	0.07	0.79	0.73
C–Cl stretch	1L, 3	strict	3 / 10 / 16	32	0.90 [0.81, 1.04]	1.00 (1.00)	0.16	0.63	0.28
C–F stretch	1, 1L, 3	top-up	5 / 10 / 16	40	1.15 [1.11, 1.20]	1.00 (1.00)	0.64	0.48	0.24
C–F stretch	3	top-up	3 / 10 / 16	40	0.86 [0.73, 1.05]	1.00 (1.00)	0.16	0.57	0.22
C–O stretch	1L, 3	strict	5 / 11 / 16	31	1.17 [1.07, 1.28]	1.00 (1.00)	0.45	0.62	0.18
H–C stretch	1L, 3	strict	4 / 10 / 16	45	1.11 [0.95, 1.33]	1.00 (1.00)	0.56	0.45	0.08
ORB (22 circuits)									
C–C–Cl angle	3	strict	118 / 364 / 593	20	0.67 [0.51, 0.84]	1.00 (1.00)	0.06	0.44	0.23
C–C–F angle	1, 3	top-up	30 / 250 / 383	38	0.75 [0.56, 1.00]	1.01 (0.93)	0.00	0.56	0.56
C–N–C angle	3	strict	32 / 445 / 593	28	0.85 [0.59, 1.10]	1.04 (0.91)	0.00	0.59	0.36
H–C–H angle	3	strict	61 / 349 / 593	39	0.82 [0.57, 1.11]	0.96 (0.85)	0.01	0.57	0.31
H–C–N angle	1, 3	strict	33 / 318 / 383	38	0.99 [0.69, 1.27]	0.95 (0.87)	0.00	0.69	0.59
H–N–C angle	1, 3	strict	16 / 304 / 383	34	0.96 [0.57, 1.37]	0.79 (0.62)	0.00	0.56	0.58
H–O–C angle	1L, 3	top-up	16 / 144 / 231	34	0.77 [0.65, 0.90]	1.00 (0.99)	0.00	0.65	0.10
C–C (arom.) stretch	1L, 3	strict	8 / 143 / 231	39	0.84 [0.70, 1.00]	0.95 (0.92)	0.00	0.71	0.18
C=C stretch	1L, 3	strict	9 / 135 / 231	35	1.03 [0.89, 1.20]	0.99 (0.97)	0.00	0.88	0.20
C=C stretch	3	strict	22 / 150 / 257	36	0.81 [0.72, 0.91]	1.01 (0.99)	0.00	0.72	0.16
C–Cl stretch	1L, 3	strict	7 / 109 / 231	30	0.72 [0.51, 0.96]	0.97 (0.94)	−0.00	0.51	0.25
C–Cl stretch	3	strict	29 / 124 / 257	28	0.80 [0.66, 0.98]	1.00 (1.00)	0.01	0.65	0.17
C–F stretch	1L, 3	top-up	34 / 102 / 231	40	0.92 [0.83, 1.02]	1.00 (1.00)	0.00	0.83	0.22
C–F stretch	3	top-up	20 / 121 / 257	38	0.84 [0.69, 1.00]	1.01 (1.00)	0.00	0.69	0.14
C=N stretch	3	strict	11 / 155 / 257	33	0.75 [0.61, 0.84]	0.84 (0.77)	0.00	0.61	0.16
C–N stretch	1L, 3	strict	6 / 148 / 231	38	0.85 [0.60, 1.12]	1.00 (0.96)	−0.00	0.60	0.08
C=O stretch	1L, 3	strict	41 / 111 / 231	44	0.84 [0.67, 1.00]	1.00 (1.00)	0.06	0.60	0.18
C=O stretch	3	strict	10 / 119 / 257	43	0.85 [0.68, 0.99]	1.00 (0.99)	0.00	0.69	0.16
C–O stretch	3	strict	23 / 158 / 257	28	0.84 [0.62, 1.13]	0.99 (0.93)	−0.00	0.63	0.13
H–C stretch	1L, 3	strict	7 / 107 / 231	44	0.89 [0.63, 1.19]	1.00 (0.99)	0.00	0.63	0.22
H–N stretch	1L, 3	strict	8 / 112 / 231	33	1.04 [0.77, 1.32]	1.00 (0.99)	0.00	0.77	0.18
H–O stretch	3	top-up	58 / 248 / 593	31	0.87 [0.58, 1.20]	1.02 (0.90)	0.01	0.56	0.17
NEQUIP (15 circuits)									
C–N–C angle	1, 3	strict	61 / 388 / 445	26	0.89 [0.55, 1.26]	1.02 (0.96)	0.00	0.54	0.57
C–O–C angle	1L, 3	strict	15 / 121 / 151	32	0.91 [0.79, 1.03]	1.00 (0.98)	−0.00	0.79	0.08

continued on next page

Table 7, continued

Family	Screen	Reading	$\lvert S^* \rvert$ / live / all	n	Sufficiency	Necessity	Random suff.	Contrast lower bd.	Share of Δ
C–S–C angle	1, 3	top-up	45 / 386 / 445	32	0.90 [0.53, 1.39]	1.12 (1.01)	0.00	0.52	0.57
C–S–O angle	1L, 3	top-up	14 / 94 / 146	40	0.82 [0.54, 1.03]	0.96 (0.94)	0.00	0.55	0.13
H–O–C angle	1L, 3	top-up	16 / 96 / 151	34	0.78 [0.71, 0.84]	0.97 (0.95)	0.00	0.71	0.06
C–C (arom.) stretch	1L, 3	strict	7 / 94 / 115	40	0.94 [0.76, 1.17]	0.89 (0.85)	−0.00	0.76	0.24
C=C stretch	1L, 3	strict	7 / 92 / 115	36	0.77 [0.56, 0.98]	0.91 (0.88)	−0.00	0.56	0.24
C=C stretch	1, 3	strict	30 / 73 / 76	36	0.81 [0.59, 1.04]	0.84 (0.68)	0.17	0.45	0.35
C–C stretch	1, 3	strict	25 / 310 / 354	46	0.83 [0.55, 1.16]	1.56 (1.24)	0.01	0.55	0.10
C–Cl stretch	1, 1L, 3	strict	7 / 68 / 115	32	0.89 [0.75, 1.15]	1.01 (0.99)	−0.00	0.75	0.41
C–F stretch	1, 1L, 3	top-up	9 / 60 / 115	40	0.81 [0.72, 0.93]	1.01 (1.01)	−0.00	0.72	0.69
C–N stretch	1L, 3	strict	27 / 94 / 115	39	0.82 [0.68, 1.01]	1.00 (1.00)	0.04	0.64	0.11
C=O stretch	1L, 3	strict	15 / 76 / 115	43	0.77 [0.63, 0.91]	1.00 (1.00)	0.00	0.63	0.19
C–O stretch	1, 1L, 3	strict	8 / 88 / 115	31	1.02 [0.85, 1.35]	0.97 (0.93)	−0.00	0.85	0.32
H–N stretch	1, 1L, 3	strict	8 / 70 / 115	34	0.73 [0.56, 0.86]	1.00 (1.00)	0.00	0.56	0.24

Screen 1, 1L, 3: top-1, top-1 site-local, top-3 screen; top-up: confirmed in the topped-up reading only. Units: circuit $\lvert S^* \rvert$ / live units / all typed units. Sufficiency: point [adjusted interval]; necessity: point (adjusted lower bound); random: mean sufficiency of 20 stratum-matched random subgraphs; contrast: adjusted lower bound of sufficiency minus random sufficiency; share of Δ: fraction of the full energy change carried by the component on the confirmation molecules.

E THE DECOMPOSITION BEHIND THE CIRCUITS

The exact decomposition makes the funnel's first stage possible, and its own results explain why the components look as they do (Table 8). No term of order four or higher recurs in any model, and only two order-three terms do, both in MACE. Between 56% and 63% of the recurring terms are interactions between units, which one-unit patching cannot express, yet they carry only 2–8% of the coefficient mass: rebuilding the perturbed energy from single-unit terms alone leaves a mean error of 1.1, 0.5 and 4.0 meV. The response to a small geometric change is thus largely additive with a small interaction component that recurs across molecules, and in MACE that component is learned: randomly initialised copies respond more than a hundred times more weakly and put a smaller share of their response into pairs in 35 to 38 of the 38 families (Section E.2). The interaction terms fix which units belong together in a component; their small mass is why a few carriers can dominate the resulting circuits. The rest of this appendix gives how to read the recurring terms, which units interact in each model, the complete verdict ledger, the numerical reliability of the smallest threshold, the random-initialisation and classical comparators, and a matched-depth re-test.

Reading the recurring terms. Each recurring term is a (family, switch-set) label whose connected coefficient exceeds the threshold in a majority of that family's held-out molecules under joint false-discovery control; the same switch set in two families counts twice. The interaction share, the fraction of recurring terms of order two or higher, is 63% for MACE, 56% for ORB and 56% for NEQUIP. Orders are bookkeeping over stored coefficients. Raising the threshold removes the pairs first (Figure 6b, Table 10): in MACE every single block that recurs at 10 μeV still recurs at ten times it, most pairs drop out, and at 1 meV no interaction survives while 92 single blocks do.

Which units interact. In every model the most widespread pair couples a bypass unit to the downstream geometry or the other atoms; receiver-internal pairs are where the models differ. MACE mixes irrep orders: $1o \times 2e$ recurs in 34 of 38 families, $2e \times 3o$ in 32 and $1o \times 3o$ in 25, and within the H–C–O family each of these pairs recurs in 71–75 of 80 molecules. Its only recurring order-three terms join a receiver block to both bypass units, in the angular C–O–C and C–S–C families. ORB pairs the receiver node with its incoming edges (28 families) and its incoming with its outgoing edges (26). NEQUIP's receiver-internal pairs, $0e \times \ell{=}2$ and $0e \times \ell{=}1$, reach at most 9 families, while its receiver-block–geometry interactions recur in 38 families for $\ell{=}1$ and 37 for $\ell{=}2$. The bypass-only terms, 114 in MACE and NEQUIP and 175 in ORB, are the grouped routes through the rest of the molecule and all recur, which is why the screen's site-local variant exists.

The complete ledger. Table 9 splits every primary-threshold label of every model into recurrent, bounded-nonrecurrent and inconclusive. In every model 74–80% of the receiver-associated labels are bounded away from recurrence, and no claim of order four or higher recurs in any model: of the 2,486 such claims, all are bounded-nonrecurrent except eight inconclusive. These are population- and threshold-dependent bounds.

Model	Perturbations	Families	Recurring terms	Receiver / bypass	Order 1 / 2 / ≥ 3	Interaction mass	Additive-only error
MACE	6,439	38	549	435 / 114	205 / 342 / 2	3.3%	1.08 meV
ORB	5,877	37	508	333 / 175	222 / 286 / 0	1.9%	0.47 meV
NEQUIP	6,439	38	430	316 / 114	190 / 240 / 0	7.5%	4.03 meV

Table 8: **The exact decomposition on each model.** Recurring terms by receiver/bypass and by order; interaction mass is the share of coefficient mass at order ≥ 2; the last column is the error of an additive-only rebuild.

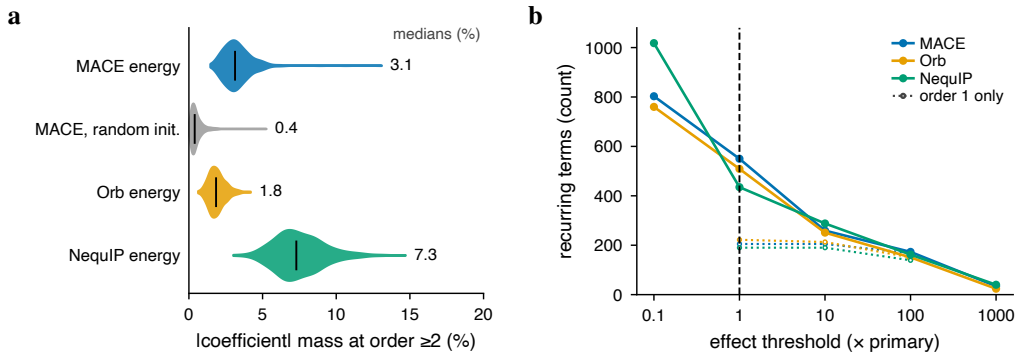

Figure 6: **Interactions are small and disappear first as the threshold rises.** (a) Share of coefficient mass at order two and higher over the 398 molecules, with three randomly initialised copies of MACE in grey; bars mark medians. (b) Recurring terms as the threshold is scaled from 0.1 to 1000 δ_1; dotted: single-unit terms at the registered thresholds.

	Receiver-associated claims				Bypass claims				Order ≥ 4	
Model	n	Recur.	Bounded	Incon.	n	Recur.	Bounded	Incon.	n	Bounded
MACE	2,280	435	1,698	147	114	114	0	0	836	836
ORB	2,072	333	1,643	96	259	175	53	31	814	814
NEQUIP	2,280	316	1,827	137	114	114	0	0	836	828

Table 9: **Every label gets a verdict.** Primary-threshold claims split into recurrent, bounded-nonrecurrent (95% simultaneous exact binomial upper bound on prevalence below one half) and inconclusive; the last two columns count claims of order four or higher and how many are bounded.

Native endpoint	Unit	δ_1	Recurrent at $10\delta_1$	Recurrent at $100\delta_1$
MACE energy	eV	10^{-5}	256	173
ORB energy	eV	10^{-5}	251	151
NEQUIP energy	eV	10^{-5}	288	160

Table 10: **Recurrence at the second and third thresholds.** Receiver and bypass terms together on the games and families of Table 8, under joint Benjamini–Yekutieli control at 5% (7,182 claims for MACE and NEQUIP, 6,993 for ORB); counts are decisions, not effect sizes.

E.1 NUMERICAL RELIABILITY OF THE SMALLEST THRESHOLD

Exact reconstruction of a stored game tests its algebra. MACE and ORB run in double precision and reconstruct their endpoints to 1.7×10^{-13} and 3.6×10^{-12} eV, far below $\delta_1 = 10\,\mu$eV. The NEQUIP games ran in single precision, so we subtract an empirical error margin from every coefficient and re-run the recurrence test: the margin is the empirical corner discrepancy, 1.05×10^{-5} eV, scaled by $2^{|S|}$. All 190 recurrent single units survive and 115 of the 240 recurrent pairs (Figure 7), which is why the NEQUIP pair count is an upper bound. The screen is an empirical sensitivity check and leaves the registered labels unchanged.

E.2 RANDOM INITIALISATION AND A CLASSICAL COMPARATOR

Random initialisation. Following the model-randomisation test of Adebayo et al. (2018), for each of three seeds a fresh network with the MACE-OFF23 medium architecture is built with the constructor's default initialisers; the fixed buffers (atomic reference energies, scale and shift) are copied from the checkpoint, so only learned weights differ. Each copy is evaluated on the same 6,439

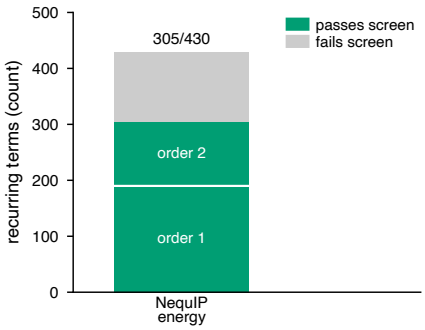

Figure 7: **Numerical reliability.** Recurrent NEQUIP claims by order that still recur after the empirical precision margin is subtracted (coloured) and those that do not (grey).

games with the same interface and its own joint correction. The only terms that recur are single units, the other-atoms unit and the receiver scalar block, 45, 48 and 50 claims against 205 single units, 342 pairs and two triples in the trained model (Figure 8a). The random copies respond far more weakly (median $|\Delta y_B|$ of 0.09–0.24 meV against 30 meV), so the share of coefficient mass at each order, which is scale-free, is the fair comparison: pairing each held-out game of the trained model with the same game of a random copy, the trained model's order-two share is higher in 35, 38 and 38 of the 38 families with molecule-bootstrap intervals excluding zero (Figure 8b).

A classical comparator. As a structural reference we decompose the MMFF94s angle-bend energy (Halgren, 1996), whose angle term is by construction a function of two neighbours. The players are the receiver's bonded neighbours, and a subset's value is the change in the angle terms whose two neighbours are both present, so every coefficient of order other than two vanishes. Of the 3,722 angular games, 3,714 admit MMFF94s parameters and an exact match between its chemical graph and the panel graph, covering 23 families. The selected triangle's order-two coefficient recurs in all 23 families at each of the three thresholds (Figure 8c): the population test recovers an interaction where one exists by construction.

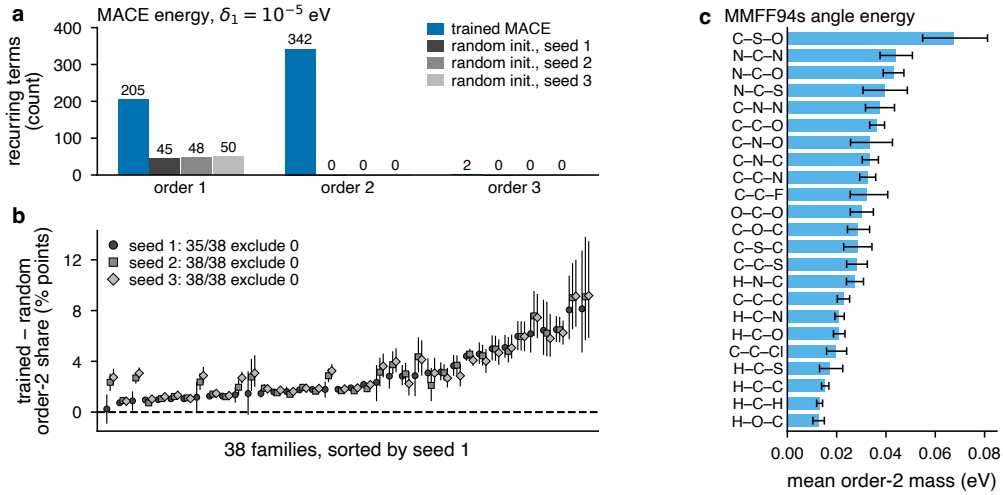

Figure 8: **Learned against structural interactions.** (a) Recurring MACE terms by order for the trained model and three randomly initialised copies. (b) Per-family paired difference, trained minus random, in the order-two share of coefficient mass, with 95% intervals. (c) Order-two mass of the MMFF94s angle comparator by family.

E.3 MATCHED-DEPTH RE-TEST

The declared interfaces sit at relative depth 0.5 (MACE), 0.8 (ORB) and 0.6 (NEQUIP). To compare the models at the same depth, each is re-tested on the 5,877 games all three evaluate at two boundaries chosen by one rule, the completed stage nearest to relative depths $\frac{1}{2}$ and 1: before the first product block of MACE, after two of five blocks of ORB and NEQUIP, and for every model the last boundary

before the output head (Figure 9, Table 11). These are the nearest realisable boundaries, not identical interfaces. Each cell keeps six units of the same kinds, and depth changes what they mean: at the last boundary of MACE each receiver unit carries its incoming message with the residual skip and the sixth unit is the direct first-layer readout. Some units cannot affect the output at a boundary, the edge units of ORB at its node-only energy head and the absent $0o$ block of NEQUIP mid-network; claims containing them are excluded from the 2,331 tested per cell, which is why fewer claims are eligible in those cells. Counts are family×switch-set claims recurring at the primary threshold under joint Benjamini–Yekutieli control at 5% per model and boundary; receiver units are the receiver's blocks in MACE and NEQUIP and its node and incoming and outgoing edges in ORB; a receiver-internal pair is a recurring order-two claim on two receiver units; distinct sets ignore family; and the NEQUIP counts at both depths carry no precision screen. Mid-network, the models agree: every one couples the receiver to its context (187, 218 and 174 recurring receiver–context pairs) and pairs the receiver's own units (106, 94 and 39). At the last boundary they part: MACE keeps 121 pairs among its receiver's irrep blocks, which its next product block multiplies within the atom, whereas in ORB and NEQUIP only single units recur, and ORB's drop is partly structural since its energy head reads nodes only. Comparable force accuracy thus comes with different internal organisation, and the circuits of Section 4.1 reflect it: MACE circuits are built from a receiver band and its readout, while the ORB and NEQUIP circuits reach back into earlier blocks.

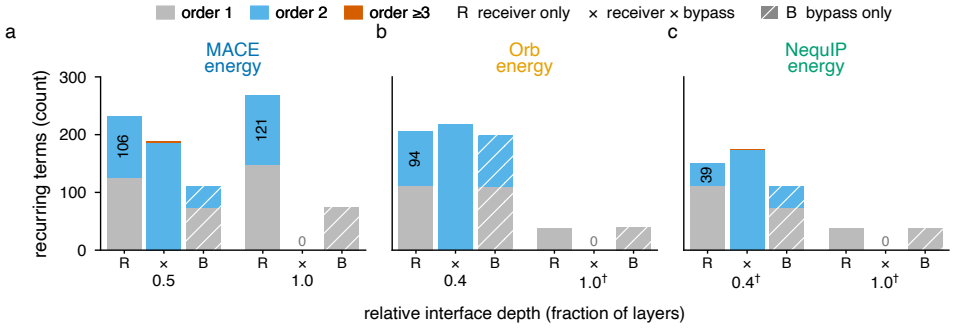

Figure 9: **Same inputs, matched depths, different organisation.** Recurring terms on the 5,877 games all three models evaluate, re-tested at a mid-network boundary and at the last boundary before the output head, split into receiver-only (R), receiver×bypass (×) and bypass-only (B) terms and stacked by order; numbers are pairs among the receiver's own units.

Endpoint	Relative depth	Primary claims Eligible/tested	Primary claims Recurrent	Order 1 / 2 / ≥3	Unit type R / R×B / B	Receiver-internal pairs	Distinct sets
MACE energy	0.50 (1/2)	2,331/2,331	532	200 / 330 / 2	232 / 189 / 111	106	23
	1.00 (2/2)	2,331/2,331	343	222 / 121 / 0	269 / 0 / 74	121	12
ORB energy	0.40 (2/5)	2,331/2,331	622	222 / 400 / 0	205 / 218 / 199	94	20
	1.00 (5/5)	259/2,331	76	76 / 0 / 0	37 / 0 / 39	0	3
NEQUIP energy	0.40 (2/5)	1,147/2,331	436	185 / 250 / 1	150 / 175 / 111	39	16
	1.00 (5/5)	1,147/2,331	74	74 / 0 / 0	37 / 0 / 37	0	2

Table 11: **Recurrence at two matched depths.** Primary-threshold claims recurring on the 5,877 shared games at the boundary nearest relative depth $\frac{1}{2}$ and at the last boundary before the output head, split by order and by unit type (receiver-only R, mixed R×B, bypass-only B).

F ABLATION DETAILS

F.1 CIRCUITS AGAINST THE NEAREST UNITS

Algorithm 6 defines the baseline; Table 12 summarises the comparison per model and Figure 10 gives every paired difference with its Bonferroni-adjusted interval. The baselines were evaluated through the unchanged adapters on the same frozen admitted held-out molecules as the confirmation stage, where $v(S^*)$, $v(U \setminus S^*)$ and e_C reproduce the frozen records (bit-identical for MACE and ORB; within 1.2×10^{-14} eV for NEQUIP). The Bonferroni family for the paired differences is $2 \times 61 = 122$, two differences for every circuit that could have been confirmed, fixed before the baselines ran and therefore slightly conservative for the 58 circuits analysed.

		Found beats baseline (suff.)		Locality passes	Median, found vs locality			
Model	Circuits	Locality	Stratum-matched	suff. / nec. / all	Sufficiency	Necessity	Non-local unit	Identical to baseline
MACE	21	19	19	3 / 15 / 2	0.99 vs 0.00	1.00 vs 0.92	1	1
ORB	22	20	22	1 / 16 / 1	0.84 vs 0.00	1.00 vs 0.92	7	0
NEQUIP	15	14	13	0 / 12 / 0	0.82 vs 0.00	1.00 vs 0.99	6	0
All	58	53	54	4 / 43 / 3	0.86 vs 0.00	1.00 vs 0.97	14	1

Table 12: **Found circuits against the same-size locality baseline.** Nearest live units, evaluated on the same confirmation molecules, for the 58 circuits confirmed in either reading. "Non-local unit": the circuit contains a unit touching an atom two or more bonds from the target; "identical": every baseline draw equals the circuit.

The baseline wins the sufficiency comparison for three circuits, and two of these are overshoots: the ORB C–F and NEQUIP C–N baselines keep 2.04 and 1.49 times e_C and fail the sufficiency band. On the strict reading alone (44 circuits) the found circuit wins on sufficiency in 40. Three circuits are the cases where locality is the whole story: the MACE C=C stretch circuit in the geometry graph is identical to its baseline in all 20 draws (it is the R–M edge geometry and the layer-1 messages along it), and the MACE C–C–O and ORB C=O baselines pass all three criteria, the latter with a sufficiency of 1.10 against the circuit's 0.84. In 12 of the 49 circuits with an upstream unit no locality draw contains any upstream unit; in the rest the draws contain upstream units and miss the $M \to R$ message path. Thirty-two of the 58 circuits contain a unit off the triangle, and 14 a unit two or more bonds away, mostly the larger circuits of ORB and NEQUIP.

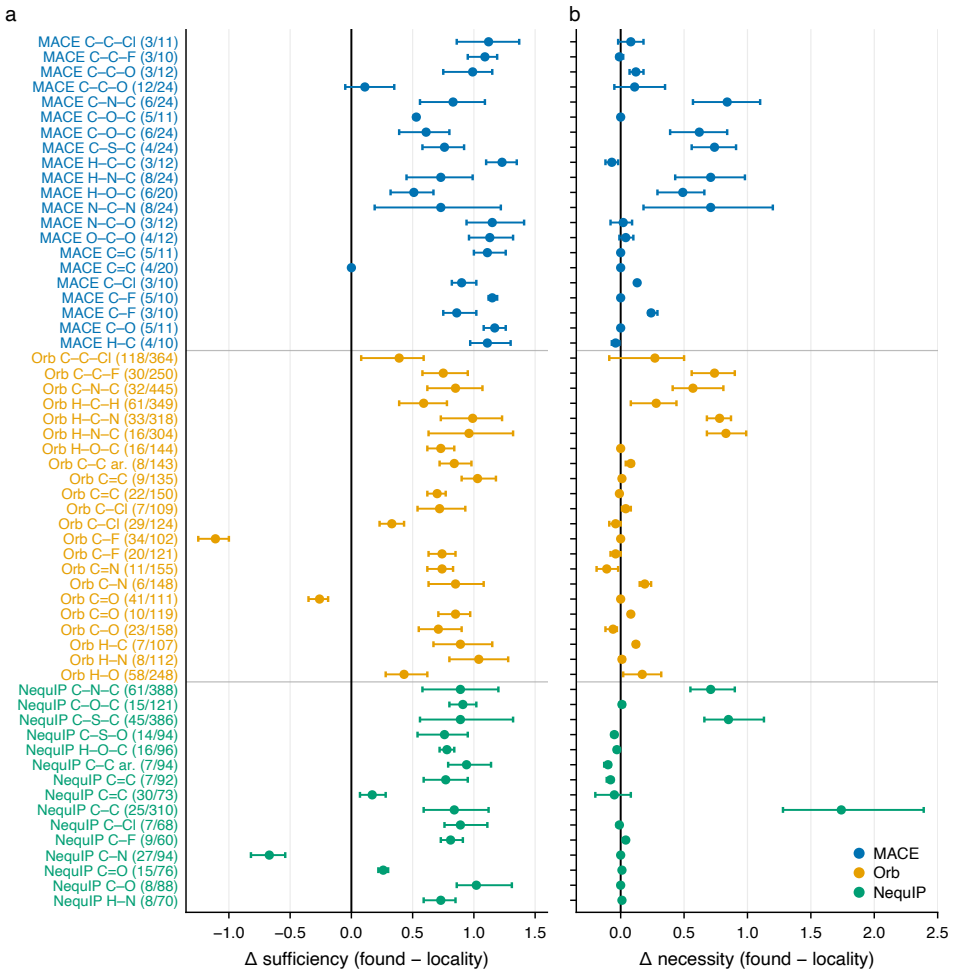

Figure 10: **Found circuit minus locality baseline.** Paired differences in sufficiency (left) and necessity (right) for every confirmed circuit, with Bonferroni-122 intervals.

Why these comparisons. Circuit discovery on equivariant force fields has no established alternative to compare against: earlier circuit methods are defined on a fixed computational graph with a discrete corrupted input, and porting each of them to a graph that changes with the molecule would require the same role typing, perturbation families and patching semantics that this pipeline introduces, at which point the comparison would be between variants of our own search. The two comparisons reported here are the ones that test the claims the paper makes. The nearest-units baseline tests whether the search finds anything beyond proximity, the one hypothesis a chemist would raise first; it is built with the same unit vocabulary and tested with the same statistics, so the comparison is exact. The stratum-matched random subgraphs test whether the specific choice of units matters given their number and their distribution over regions. Their sufficiency is low in the large ORB and NEQUIP graphs because a random set rarely contains a complete upstream–inside–downstream path, which is precisely what makes a found circuit informative there; in the small MACE graphs random sets often do contain such a path, and the contrast is correspondingly harder to pass. The ranking used inside the search covers the natural gradient-based alternative: for graphs above 200 units the units are ordered by attribution (gradient times hybrid-minus-clean), so an attribution top-k set is the unpruned prefix that the band test and the pruning then refine, and the sensitivity analysis of the band (τ) reports how much that refinement changes the result.

Scope of the evaluation. The pipeline is the first of its kind for this class of models, and its evaluation was designed to test the claims made rather than to exhaust the space of variants. Each model is analysed at one internal boundary chosen a priori, with a matched-depth re-test of the decomposition at two further boundaries (Appendix E.3); the estimand is the energy, from which the forces follow by differentiation; recurrence is decided at one declared step size, whose non-linearity is characterised in Section F.5; the τ rerun covers the model whose localisation rule binds most, MACE; and the confirmation statistics use percentile bootstrap intervals of ratios of sums, adjusted by Bonferroni, the most conservative of the corrections compared in Table 2. Running the same stages at several boundaries, on force components and at several step sizes are natural extensions that the one-command interface makes routine; the results reported here are the frozen first pass, with every setting and its sensitivity disclosed.

F.2 HYPERPARAMETER SENSITIVITY

Figure 11 plots the full sweep behind Table 2. Every row except the τ rerun is a recount of the frozen stage-3 cards, circuits and stage-5 records with one setting changed. For the false-discovery-rate rows, each test's p-value is read off its frozen adjusted bootstrap ($B = 650{,}400$ draws, two-sided, with a $+1$ correction): $p = \min(1, 2(1 + \#\{T^* \leq 0.5\})/(B+1))$ for necessity, $p = \min(1, 2(1 + \#\{T^* \leq 0\})/(B+1))$ for the random contrast, and the intersection–union form over both ends of the band for sufficiency; the unavailable candidates keep their three tests in m with $p = 1$, and Bonferroni on these p-values reproduces the interval-based counts exactly. Under Benjamini–Yekutieli at 5% the rejection threshold is $p \leq 0.0030$ (strict) and 397 of the 813 tests are rejected, against 336 under Bonferroni. The stage-3 readings are exact recounts: a candidate is confirmed under a reading iff it passes stage 3 under that reading and is confirmed in the frozen stage 5, because the worst-seed and per-seed passes are subsets of the mean-control passes in every model.

The τ rerun. τ is the one setting that changes the circuit itself, so it cannot be recounted from frozen records: all 68 MACE candidates were run through stages 4 and 5 at $\tau \in \{0.7, 0.9\}$ on the same discovery and held-out molecules and graded with the frozen criteria at $m = 813$ (at $\tau = 0.8$ this reproduces every frozen circuit and label). At $\tau = 0.7$ the circuits shrink (median 4.5 units; 66 of 68 localised; 29 circuits identical to the frozen ones) and 17 / 23 are confirmed, with 15 of the 16 strict and 19 of the 21 topped-up frozen circuits kept. At $\tau = 0.9$ they grow (median 8 units; 47 localised; 13 identical) and 18 / 24 are confirmed: four strict and five topped-up frozen circuits still pass all three tests and lose localisation, and six strict and eight topped-up circuits are newly confirmed. The proxy in Figure 11f, which counts frozen circuits whose sufficiency point lies in the band, falls steeply at $\tau = 0.9$ because the stage-5 criterion is the adjusted interval inside $(0.5, 1.5)$ and the circuit itself changes with τ.

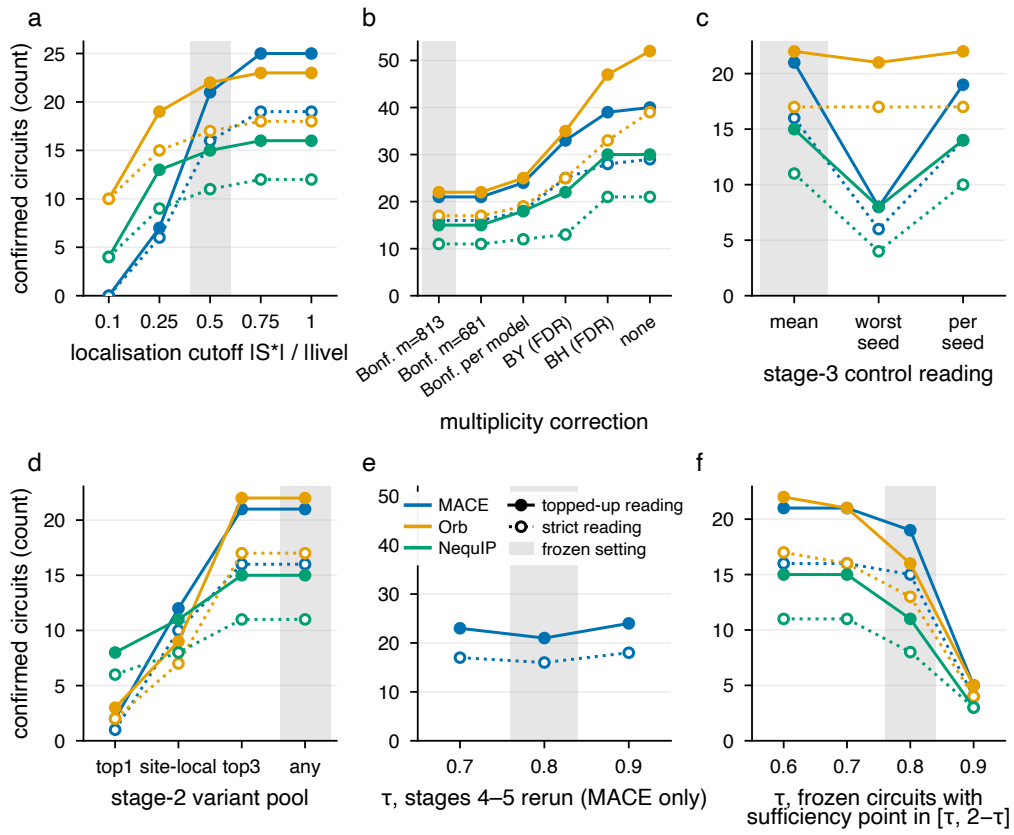

Figure 11: **Confirmed circuits as one setting is changed.** Solid: topped-up reading; dotted: strict; shaded: the frozen setting. (e) reruns localisation and confirmation for MACE; (f) counts the frozen circuits whose sufficiency point lies in $[\tau, 2-\tau]$.

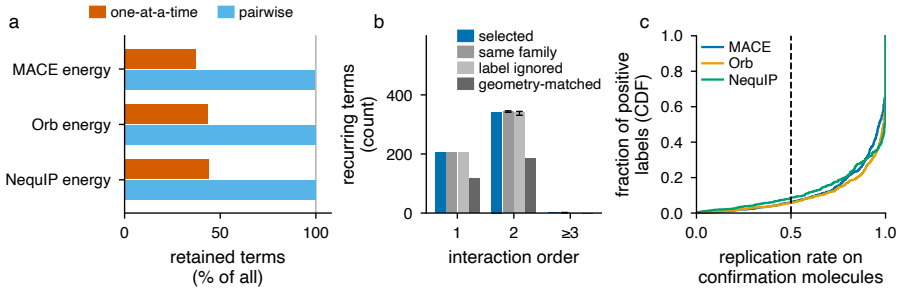

Figure 12: **What each simpler alternative in the decomposition stage loses.** (a) Recurring terms expressible by one-at-a-time and pairwise patching. (b) Recurring terms by order for the selected triangles and three substitutes (MACE, five seeds). (c) Held-out replication rate of single-molecule positives; dashed line $\pi_0 = \frac{1}{2}$.

F.3 Choices in the decomposition stage

All combinations against one at a time. Figure 12a gives, per model, the recurring terms that stay expressible under one-at-a-time and pairwise patching, and Figure 13 the error made when the perturbed energy is rebuilt from those terms alone. Retaining coefficients through orders two, three or four corresponds to 22, 42 or 57 of the 64 evaluations; at the primary threshold, orders three and four retain every recurring term.

Typed triangles against other triangles. Figure 12b gives the recurring terms by order for the selected typed triangles and three substitutes on MACE (same family, label ignored, geometry-matched), with the between-seed range over five seeds; Figure 14 gives the paired molecule-level

difference in positivity rate against the selected arm. Only the label-ignored substitute lowers recurrence in every seed: the chemical label is what makes the perturbation family a behaviour that recurs.

Many molecules against one. Table 13 gives, per model, what a single-molecule positive is worth: the share whose label recurs, is bounded away from recurrence or stays inconclusive on the held-out test, and the median and tenth percentile of the label's replication rate; Figure 12c shows the full distribution, whose left tail is what the population test removes.

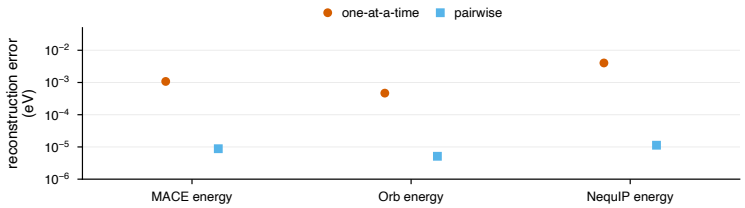

Figure 13: **Error of rebuilding the energy from fewer terms.** Mean molecule-level absolute error of the perturbed energy rebuilt from one-at-a-time or pairwise terms alone; 95% molecule-bootstrap intervals, 398 molecules per model.

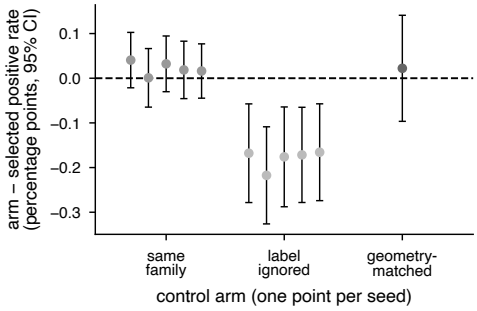

Figure 14: **Only ignoring the label changes the positivity rate.** Paired molecule-level difference in positivity rate between each substitute triangle and the selected typed triangle, one point per seed, 95% intervals; MACE.

Model	Single-molecule positives	In recurrent labels	In bounded labels	In inconclusive labels	Repl. median	Repl. p10
MACE	50,167	88.5%	2.4%	9.1%	0.97	0.63
ORB	42,679	88.4%	2.3%	9.3%	0.99	0.65
NEQUIP	40,437	85.2%	4.1%	10.8%	0.99	0.57

Table 13: **What a single-molecule finding is worth.** Every (discovery molecule, label) pair with $|c(S)| > \delta_1$ is one positive; columns give the share whose label is recurrent, bounded or inconclusive on the held-out test, and the median and tenth percentile of its replication rate.

Effect threshold and correction family. Table 10 gives the recurrence counts at the second and third thresholds. A threshold-only comparison must keep its multiplicity family fixed, so Figure 15a,b uses one five-tier family per model; a separate contrast varies one, three and five tiers at the unchanged primary threshold (Figure 15c) and moves the recurring set by under 2%.

F.4 GRAIN OF THE INTERFACE

The declared MACE interface has six units: four receiver irreps, all other atoms, and downstream geometry. We changed its grain in both directions (Table 14). In the coarser arms each accepted primary-tier recurring term is either still resolvable individually at the grouped interface or merged with others, and no term is re-tested; in the finer arms recurrence is re-tested with each row corrected in its own family over its hypotheses and three thresholds, and rows are never pooled. Counts are decisions and carry no resampling uncertainty.

Coarser. Grouping is an exact restriction of the stored games, with no new inference. Grouping the three nonscalar receiver irreps leaves 154 of the 549 MACE recurring terms resolvable and merges 395; the other two models' architecture-native groupings merge 50% and 73% of theirs.

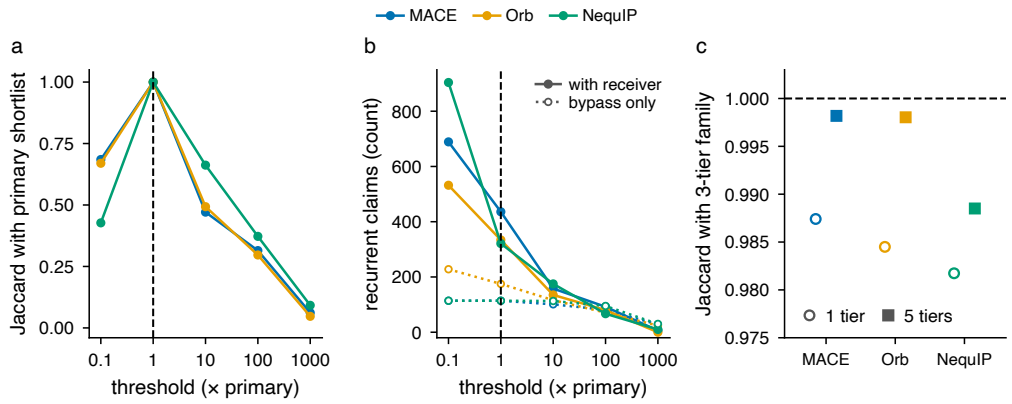

Figure 15: **Threshold and correction-family sensitivity.** (a) Overlap with the primary recurring set as the threshold varies within a fixed five-tier family. (b) Receiver-associated and bypass-only claims. (c) Changing the correction family at the primary threshold.

Finer: other atoms. Splitting other atoms into bonded neighbours and the remaining atoms (seven units, new evaluations) leaves every term without other atoms unchanged (molecule means agree within 4×10^{-14} eV); 395 of their 399 recurring terms recur again, the four losses arising only from the larger correction family. Of the 150 recurring terms that involve other atoms, 135 keep at least one recurring refinement; ten non-recurring terms gain one. The refinements divide evenly: 131 need only bonded neighbours, 120 only the remaining atoms, 32 both. The grouped unit bundles two routes of similar weight.

Finer: channels. Splitting each receiver irrep into two halves of its 128 channels (eight receiver-only units, everything else held clean) gives 493 recurring terms, none above order two. In 129 of 152 family–irrep cases both halves recur and in none only one, so an irrep does not conceal a channel-specific subset. The halves are coordinates of this checkpoint, and the arm switches only receiver channels in an otherwise clean molecule, so it measures a different quantity and its counts are not comparable with the other rows. Moving the boundary to the second product block is impossible: that block receives only scalar features.

Coarser: units grouped within the stored games (primary tier)

Endpoint	Grouping	Recurring	Resolvable	Merged
MACE energy	nonscalar receiver irreps together	549	154	395
NEQUIP energy	scalar vs. nonscalar features	430	114	316
ORB energy	incoming vs. outgoing receiver edges	508	255	253

Finer: MACE *units split, all 6,439 games re-evaluated*

Units	Hypotheses	Recurring at 10^{-5} / 10^{-4} / 10^{-3} eV		
Declared (six)	2,394	549	256	173
Other atoms split (seven)	4,826	678	322	200
Receiver channel halves (eight)	9,690	493	259	96

Table 14: **Changing the grain of the interface.** Top: primary-tier recurring terms that remain resolvable or are merged at a coarser interface. Bottom: recurrence re-tested at two finer MACE interfaces, each corrected in its own family.

F.5 PERTURBATION STEP SIZE

The screen perturbs every geometry by one declared step size η, $+2^\circ$ for angles and $+0.02$ Å for bonds. To test that choice, we re-evaluated all 64 switch combinations of every MACE game at four further step sizes per perturbation, half and twice the declared step with both signs, changing nothing else (22,244 new games). Admission was re-checked at every step size, so larger and negative steps admit fewer games (Table 15; the exclusions are 2,465 neighbour-graph changes and 1,047 endpoints outside the discovery support window). The games admissible at all five step sizes (2,070 angle and 1,751 bond games) give the paired curves of Figure 16; the table's last two rows are molecule means over these games, and the highest interaction order is counted with a fixed cut of 10^{-6} eV on the coefficient, so part of its rise reflects larger effects.

Two features are stable. Interactions carry a small share of the coefficient mass at every step size, 2.7–6.6% for angles and 0.8–3.4% for bonds (Figure 16b). Paired and per-step curves agree within 0.2 percentage points, so the changing admitted set does not drive them.

Two features change. The interaction share grows with the step size (angles 2.7, 4.7 and 6.5–6.6% at half, one and twice the declared step), consistent with interactions entering at second order. Nor is the response linear at the declared step size: doubling it from half to full multiplies the median full effect by 2.64 for angles and 2.31 for bonds, between linear (2) and quadratic (4) scaling (Figure 16c), and negating the half step reverses the sign of the full effect in only 68% and 81% of games (19% and 37% at twice the step). Recurrence was not re-tested at the new step sizes, so the recurring set remains a property of the declared step; at larger steps more interaction terms would be expected to cross a fixed threshold.

	Angle (°)					Bond length (Å)				
Step size η	-4	-1	$+1$	$+2$	$+4$	-0.04	-0.01	$+0.01$	$+0.02$	$+0.04$
Admitted games	2,511	3,357	3,701	3,722	3,029	2,039	2,535	2,712	2,717	2,360
Receiver-only share (%)	17.4	23.4	23.2	21.0	17.4	30.4	38.6	43.1	44.6	46.3
Highest order with mass	3.88	2.51	2.50	3.07	3.89	3.27	2.01	2.01	2.97	3.34

Table 15: **The MACE screen at five step sizes.** $+2°$ and $+0.02$ Å are the declared screen; admitted games pass the unchanged admission rule, and the last two rows are molecule means over the games admissible at all five step sizes.

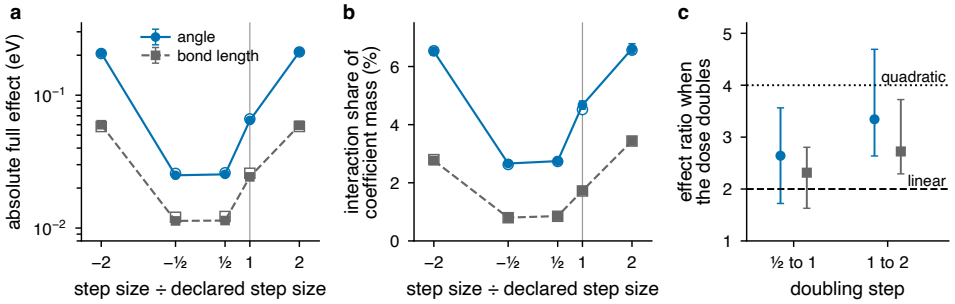

Figure 16: **MACE response as the step size varies.** (a) Absolute full effect and (b) interaction share of coefficient mass, molecule means with standard errors over the games admissible at all five step sizes (filled) and over each step size's own admitted games (open). (c) Median and interquartile range of the ratio of full effects when the step size doubles; 2 is linear and 4 quadratic.

F.6 Cost of the exact decomposition

Evaluating every switch combination costs 2^m hybrid evaluations per game. Figure 17 times the complete evaluation and inversion at interface sizes $m = 4, 6, 7, 8$ on twelve frozen games per model with three repeats each. Within every model the time scales close to 2^m; at $m = 8$ the median game takes 101 s for MACE and 19 s for NequIP on CPU and 0.92 s for Orb on one A100. Table 16 gives the six-unit production screens. Software stacks and devices differ across models, so these are within-model scaling curves.

Model	Timed games	Fitted games	Seconds per game (95% interval)
MACE	6,439	6,363	0.601 [0.578, 0.629]
Orb	5,877	5,877	0.238 [0.238, 0.238]

Table 16: **Time per six-unit game in the production screens.** Median over 398 molecular identities of each identity's median game time (64 switch combinations per game), with a 95% interval from 1,000 molecule-bootstrap resamples; NequIP recorded only aggregate session time.

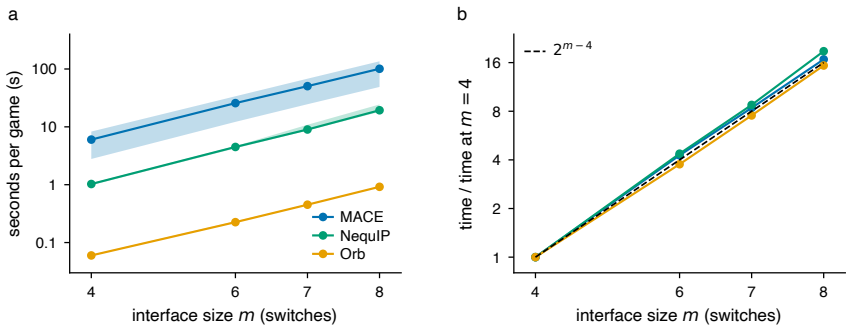

Figure 17: **Cost of the exact decomposition.** (a) Seconds per game for complete evaluation and inversion at interface sizes 4–8; twelve games per model, three repeats, min–max range shaded. (b) Time relative to size 4 against the 2^{m-4} reference.

G Case studies

G.1 Interventions on confirmed circuits

For the most localised confirmed circuits of MACE and Orb, one per family, we ran four interventions on the frozen held-out molecules: single-unit knockouts from the circuit and from the full model; dose sweeps from -0.04 to $+0.04$ Å and from $-4°$ to $+4°$ with the circuit alone active and with the circuit knocked out; transfer to every other occurrence of the family in the same molecules; and a read-out of the circuit's first-order response against the clean geometry across molecules, whose zero crossing is the implied equilibrium geometry. Statistics use 95% molecule-bootstrap intervals. Table 17 summarises the seven cases and Figures 18–19 give each one.

Dose response and knockout. At every dose the circuit alone reproduces 0.78–1.24 of the component's behaviour and knocking it out removes 0.79–1.00 of it; the smallest necessity along a sweep is 0.79 for the Orb H–N–C circuit and at least 0.94 for the others. The MACE circuits respond linearly across the sweep and the Orb circuits show the same curvature as their components. Dropping any one of the three units of the MACE angle circuits removes the whole effect (each carries the full 1.15 or 1.23 when dropped); the Orb circuits degrade unit by unit, with the receiver's node band and its input to the energy head the largest carriers and the upstream messages from M to R next.

Hybridisation and bond order. The MACE H–C–C circuit's zero force lies at 108.8° for sp^3 centres ($n = 25$) and 120.4° for sp^2 centres ($n = 10$; the remaining 10 centres have no assigned hybridisation); the full model gives 106.2° and 120.1°. The N–C–O circuit, all sp^2, gives 120.8° against 122.6°. The Orb H–N–C circuit separates sp^3 (100.3°) from sp^2 (115.7°) centres by the same margin as the full model (107.4° and 123.7°), with both values shifted down by about 8°; a pooled zero crossing over mixed hybridisations is dominated by the majority class and is reported in Table 17 for completeness only, the per-class values being the meaningful read-out. For the stretch circuits the within-molecule curvature matches the across-molecule slope of the circuit's force against bond length (MACE C=C: 15.3 against 13.9 eV/Å^2), the signature of a transferable harmonic term, and the circuits carry 3–23% of the full-model stiffness (MACE C=C 0.23, Orb C=C 0.11, aromatic C–C 0.08, C–N 0.03). The Orb C–N circuit separates sp^2–sp^2 (1.310 Å), sp^2–sp^3 (1.276) and sp^3–sp^3 (1.415) bonds, in a different order from the full model (1.385, 1.580, 1.466); its equilibrium lengths sit below the model's throughout, as for the other stretch circuits, so what transfers is the class separation and the harmonic form.

Transfer. Written at the other occurrences of the family in the same molecules (9–130 sites per case), the circuits keep sufficiency between 0.86 and 1.29 and necessity at 0.90 or above, while the random subgraphs give 0.00–0.46.

G.2 Component-level interventions

Several components were also examined at the component level, without localisation, with the interventions of the causal check on dedicated fresh cohorts: dose response, content controls, transfer

Model	Family	$\lvert S^* \rvert$/live	n	Suff.	Nec.	Share of Δ	k circuit / full	q_0 circuit	q_0 full
MACE	C=C stretch	4/20	36	0.97 [0.91, 1.03]	0.97 [0.91, 1.03]	0.73	15.3 / 67.9 eV/Å^2	1.234 [1.147, 1.263] Å	1.331 Å
MACE	H–C–C angle	3/12	45	1.23 [1.16, 1.29]	1.00 [1.00, 1.00]	0.02	1.6 / 107.5 eV/rad^2	110.0 [106.9, 113.1]°	105.9°
MACE	N–C–O angle	3/12	35	1.15 [1.03, 1.29]	1.00 [1.00, 1.00]	0.17	2.6 / 15.3 eV/rad^2	120.8 [119.7, 121.8]°	122.6°
ORB	C–N stretch	6/148	38	0.85 [0.73, 0.98]	1.00 [0.98, 1.02]	0.08	1.8 / 53.3 eV/Å^2	1.276 [1.232, 1.300] Å	1.326 Å
ORB	H–N–C angle	16/304	34	0.96 [0.78, 1.15]	0.79 [0.72, 0.86]	0.58	76.4 / 119.1 eV/rad^2	—	127.4°
ORB	C–C (arom.) stretch	8/143	39	0.84 [0.78, 0.91]	0.95 [0.94, 0.97]	0.18	4.8 / 61.0 eV/Å^2	1.352 [1.319, 1.365] Å	1.391 Å
ORB	C=C stretch	9/135	35	1.03 [0.96, 1.11]	0.99 [0.98, 0.99]	0.20	7.3 / 67.0 eV/Å^2	1.325 [1.307, 1.334] Å	1.327 Å

Table 17: **Circuit case studies.** Sufficiency and necessity on the held-out molecules, share of the full response carried by the component, curvature k of the dose response and implied equilibrium geometry q_0 for the circuit alone and the full model; the ORB H–N–C zero crossing is undetermined (32.6° [32.6, 172.9]).

of the component's change to chemically equivalent and non-equivalent sites, and the share carried through a named downstream path. These studies use the same estimands and controls as stage 3 (Figure 20).

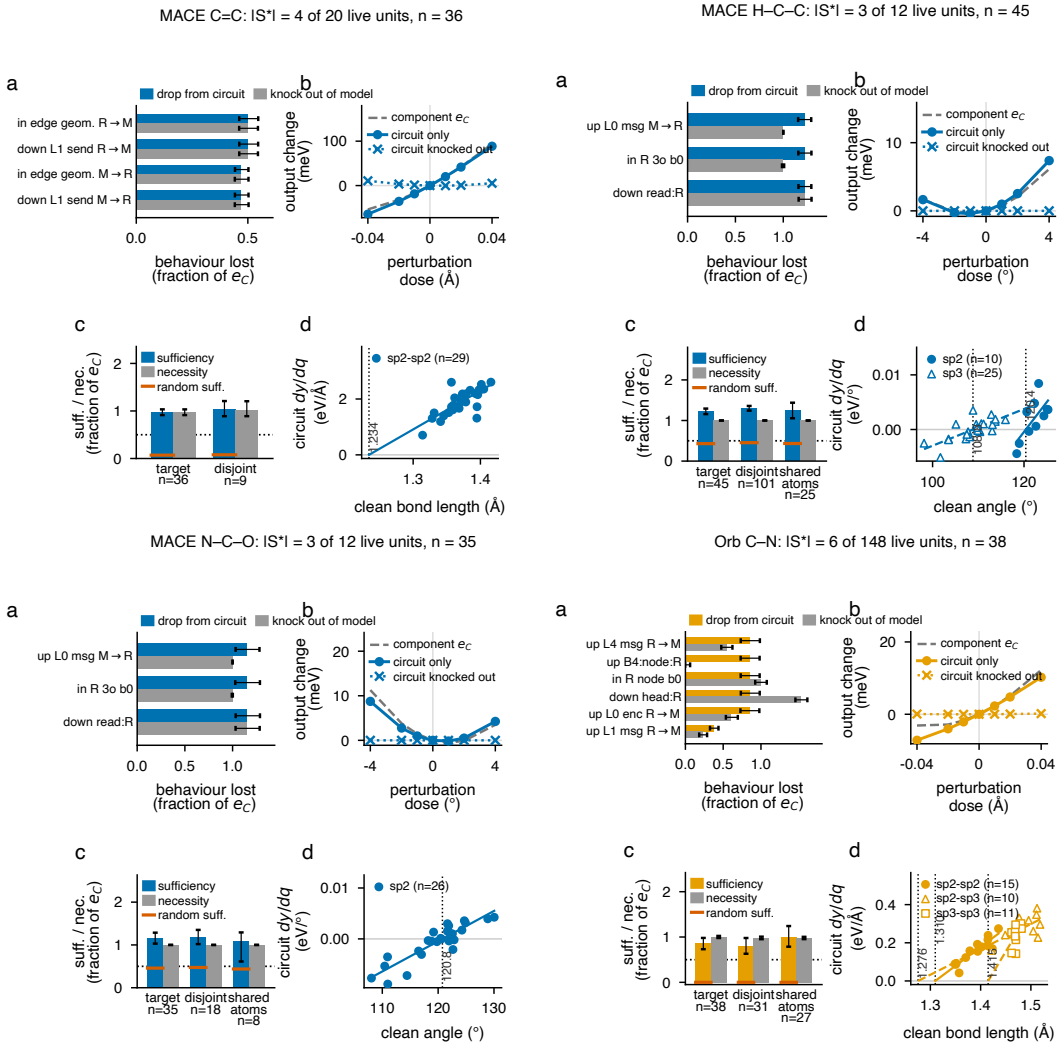

Figure 18: **Circuit case studies, MACE C=C, H–C–C and N–C–O and ORB C–N.** Per case: (a) behaviour lost when each unit is dropped from the circuit or knocked out of the full model; (b) dose response of the component, the circuit alone and the circuit knocked out; (c) sufficiency and necessity on the target and on other occurrences of the family, random sufficiency in red; (d) the circuit's first-order response against the clean geometry.

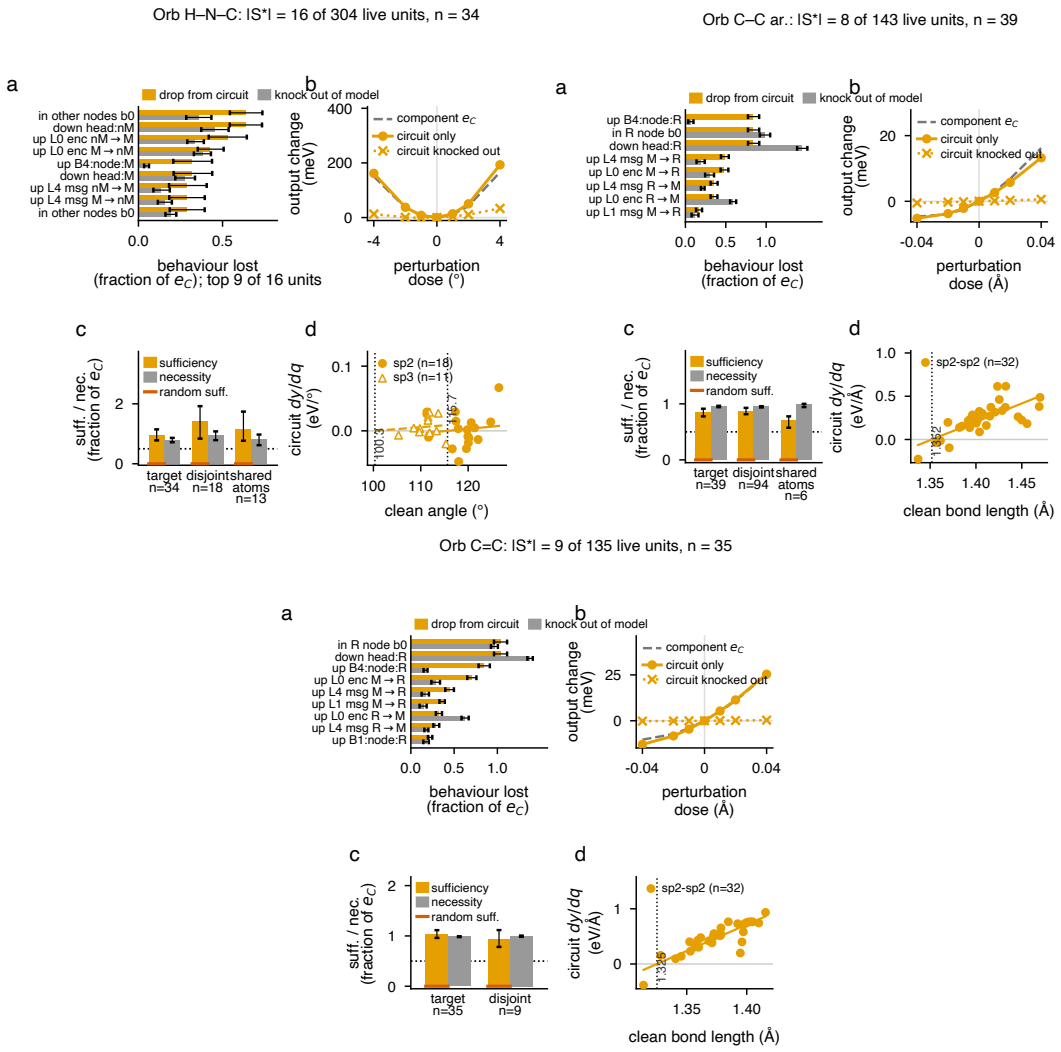

Figure 19: **Circuit case studies, ORB H–N–C, aromatic C–C and C=C.** Panels as in Figure 18.

MACE: the C–O–C receiver $2e$ block. On 91 molecules the $2e$ block of the ether oxygen's incoming message carries 0.195 [0.165, 0.230] of the energy change of a 2° C–O–C opening on the medium checkpoint and 0.135 [0.116, 0.157] on the large one, dose-coherently; random, orthogonal and permuted content of the same norm falls short by at least 0.135 and 0.105. Its own first-layer readout and self-connection carry 0.254 [0.214, 0.299] and 0.177 [0.149, 0.208], more than the whole block: the second-layer path carries -0.058 and -0.042, so a later layer cancels about a quarter of the block's direct effect. Rotated into the frame of another C–O–C oxygen by the model's own Wigner matrices, the change reproduces 0.967 [0.871, 1.060] and 0.929 [0.809, 1.043] of the effect (31 molecules); a second cohort of 120 molecules gives a block share of 0.182 and 0.124 and transfer to a matched oxygen of the same molecule of 0.985 and 0.991. Written at hydroxyl oxygens it reproduces 0.68 and 0.61.

MACE: the H–C–O angle minimum. On 24 molecules (12 discovery, 12 held-out) the H–C–O angle is set on a $0.1°$ grid from $103.7°$ to $122.6°$ by moving the hydrogen with both bond lengths fixed, and the complete vector block of the first-layer H→C message is either held at its observed-geometry value or allowed to follow the angle. Letting it follow shifts the energy minimum by $-0.364°$ [-0.500, -0.182] on the 11 held-out molecules with an interior minimum, with a coherent dose response in all 11, and carries 0.110 [0.067, 0.149] of the energy change of a 3° opening. Permuted and randomly redirected vector content shifts the minimum by $0.375°$ and $0.385°$ less (Figure 21); the shift is reproduced at an equivalent anchor (1.23 [0.85, 2.04], 7 molecules).

MACE: an H–O–C interaction term. The connected coefficient of the hydroxyl oxygen's $2e$ block with the downstream geometry is positive on 99–100% of 159 molecules and carries 0.27% [0.24, 0.30] and 0.62% [0.54, 0.71] of the energy change on the medium and large checkpoints; content controls fall short, and written jointly at another hydroxyl it reproduces 0.97 and 1.00 of itself. The term is resolved only in double precision.

ORB: the C=C edge. For a $+0.02$ Å C=C stretch on 59 molecules, the stretched bond's directed edge state after block 4 carries 0.112 [0.103, 0.122] of the energy change; content controls fall short by at least 0.111. The receiver's sender slot carries 0.085 [0.078, 0.093] and the rest of the edge's effect 0.026 [0.024, 0.029]. Transplanted to another non-aromatic C=C the change reproduces 1.11 [0.86, 1.37] of the effect, and to other bonds 0.35 [0.26, 0.44]: the edge's content is specific to its motif.

NEQUIP: the C–F receiver scalar block. For a 0.02 Å C–F stretch on 44 molecules, the receiver carbon's layer-2 scalar block carries 0.853 [0.680, 1.024] of the learned-energy change; orthogonal and permuted content falls short by at least 0.54. A route through the following layers carries 0.631 [0.484, 0.775] and the rest of the block's effect 0.222 [0.169, 0.277]. Transfer is 0.974 at another C–F carbon and 0.806 at other carbons: a generic carbon channel.

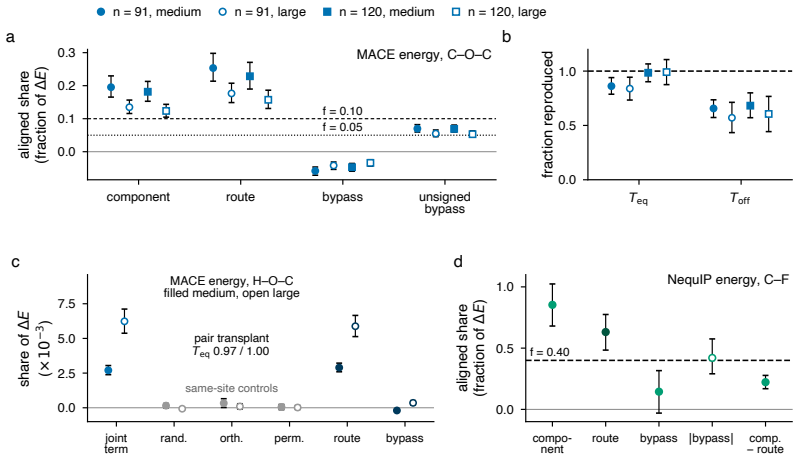

Figure 20: **Component-level interventions.** (a) MACE C–O–C block: component, route and bypass shares on two cohorts (circles 91, squares 120 molecules) and two checkpoints (filled medium, open large); (b) its transfer to equivalent and non-equivalent sites; (c) MACE H–O–C interaction term against same-site controls; (d) NEQUIP C–F block. Bars: 95% intervals.

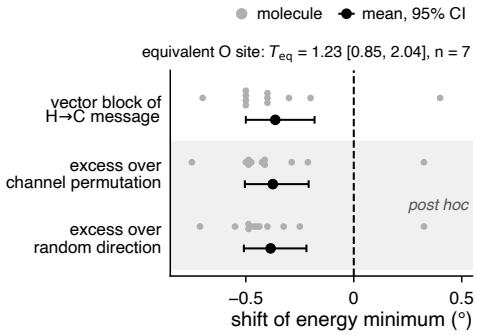

Figure 21: **H–C–O angle minimum on the medium checkpoint.** Shift of the energy minimum when the vector block of the H→C message follows the angle, and its excess over permuted and redirected content (shaded); grey: molecules, black: mean and 95% interval.

H EXTENDED RELATED WORK

GNN force fields. Message-passing and equivariant architectures (Schütt et al., 2017; Klicpera et al., 2020; Thomas et al., 2018; Fuchs et al., 2020; Schütt et al., 2021; Geiger & Smidt, 2022; Batzner et al., 2022; Musaelian et al., 2023; Drautz, 2019; Batatia et al., 2022) establish how to build accurate, symmetry-respecting potentials, and ORB (Neumann et al., 2024; Rhodes et al., 2025) reaches strong accuracy with non-equivariant latents in a fixed frame. This literature explains what the models can represent; equivariance fixes how a channel transforms under rotation and leaves open what a trained channel encodes and whether the output uses it, which is the question circuit discovery answers.

Graph explanation and triangle descriptors. GNNExplainer and chemistry-oriented masking (Ying et al., 2019; Wu et al., 2023) identify which input nodes, edges or substructures matter for a prediction. They score inputs and leave the internals untouched, so they cannot say which internal route transmits a geometric change. Three-point pharmacophore descriptors (Mahé et al., 2006; Ranu & Singh, 2011) encode atom-triple geometry for screening; we reuse the triangle only as the unit that makes the choice of perturbation enumerable and the role typing well defined.

Interpreting learned interatomic potentials. Relevance propagation through many-body terms (Bonneau et al., 2025), architectures designed so that each body order is readable by construction (Batatia et al., 2025), and comparisons of the latent geometry learned by different foundation potentials (Li & Walsh, 2026) each describe what a trained potential represents. They attribute a prediction to inputs or body orders, or compare representations; none intervenes on internal components to test which of them carry a specific geometric response, which is the question circuit discovery answers and which the interventions here are built for.

Mechanistic interpretability. Four levels of evidence should be distinguished. Attribution (Sundararajan et al., 2017; Lundberg & Lee, 2017) establishes correlation. Activation and path patching (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024; Goldowsky-Dill et al., 2023; Heimersheim & Nanda, 2024) show that a component can move the output but, applied one component at a time, miss interactions. Automated circuit discovery (Conmy et al., 2023; Syed et al., 2024; Bhaskar et al., 2024; Marks et al., 2025; Lieberum et al., 2023) assembles patched components into a subgraph that reproduces the behaviour on one fixed computational graph, and its output is judged by faithfulness and completeness (Wang et al., 2023; Hanna et al., 2024; Miller et al., 2024), the two tests our confirmation stage applies under the names sufficiency and necessity. Causal abstraction (Geiger et al., 2021; 2025; Pearl, 2001) supplies the notion of a circuit as an intervention-preserving map. Exact any-order interaction indices, from Harsanyi dividends and Möbius inversion (Harsanyi, 1963; Grabisch, 2016) to Shapley–Taylor and SHAP-IQ (Sundararajan et al., 2020; Fumagalli et al., 2023) and their graph form (Muschalik et al., 2025), decompose a prediction over input features; our connected coefficient applies the same inversion to interventions on an internal interface. No earlier framework jointly provides a perturbation vocabulary for continuous inputs, symmetry-respecting writes of equivariant states (Hansen et al., 2024), and molecule-level replication with a circuit defined across molecules.