ATROPOBENCH: A BENCHMARK FOR GEOMETRIC CHIRALITY IN MOLECULAR MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Restricted rotation around a bond can give molecules left- and right-handed configurations, a form of axial chirality. Recognising their handedness and predicting resistance to interconversion are distinct tasks: reflection reverses handedness but preserves the rotational energy barrier. We introduce AtropoBench, a diagnostic benchmark with nearly one million exact mirror pairs and over $21{,}000$ computational rotational-barrier annotations from ligand and catalyst candidates. Its protocols separately test whether models can recognise handedness, generate novel structures with requested handedness, and predict rotational barriers on unfamiliar chemistry. Experiments with eleven generators and eight stability-prediction configurations reveal that high generation quality can coexist with negligible graph novelty, while similar classification accuracy can conceal substantially different errors on readily rotating molecules. Controlled interventions further distinguish limitations in reflection sensitivity, bond localisation, and uncertainty calibration. AtropoBench supports systematic model comparison while identifying specific capabilities that require improvement.

1 INTRODUCTION

Chiral molecules have mirror-image forms that cannot be superimposed, like left and right hands, and can interact differently with biological targets (LaPlante et al., 2011). The familiar atom-centered form arises from the arrangement of groups around a single atom. Axial chirality instead arises from their arrangement around an axis (Figure 1). When interconversion is sufficiently slow, these configurations can be isolated as separate compounds, called *atropisomers* (Basilaia et al., 2022). Modeling them requires answering two questions: which handed configuration is present, and how readily can it interconvert? These concern distinct properties: reflection reverses handedness but preserves interatomic distances and the rotational energy barrier.

Geometric quality alone cannot establish these capabilities. A distance-only representation cannot distinguish mirror images, while a generator can produce plausible structures without controlling their handedness or generating new molecular graphs. Existing work provides chirality-aware representations and evaluations of conformation, generation, or stability in complementary settings (Chen et al., 2025; Shi et al., 2026b; Wang et al., 2026b; Xu et al., 2026). What remains missing is a common axial-chirality evaluation that tests these capabilities separately and reveals limitations concealed by overall scores.

We introduce **AtropoBench**, a diagnostic benchmark drawn from a library of ligand and catalyst candidates built by systematic chemical modification of published, chemist-verified cores. It provides $985{,}896$ exact mirror pairs for controlled handedness tests and a separately selected set of $21{,}268$ computational rotational-barrier annotations for stability prediction. The annotations describe scan-derived electronic barriers rather than experimental lifetimes, retaining intervals and bounds when calculations do not determine an exact value (Section 3).

The evaluation separates three capabilities. **Handedness recognition** tests distinguish identifying the relevant bond from assigning its handedness. **Generation** tests distinguish producing valid, novel axially chiral structures from following a requested handedness. **Stability prediction** tests assess both accuracy and predictive uncertainty on unfamiliar chemical classes, with readily rotating

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.

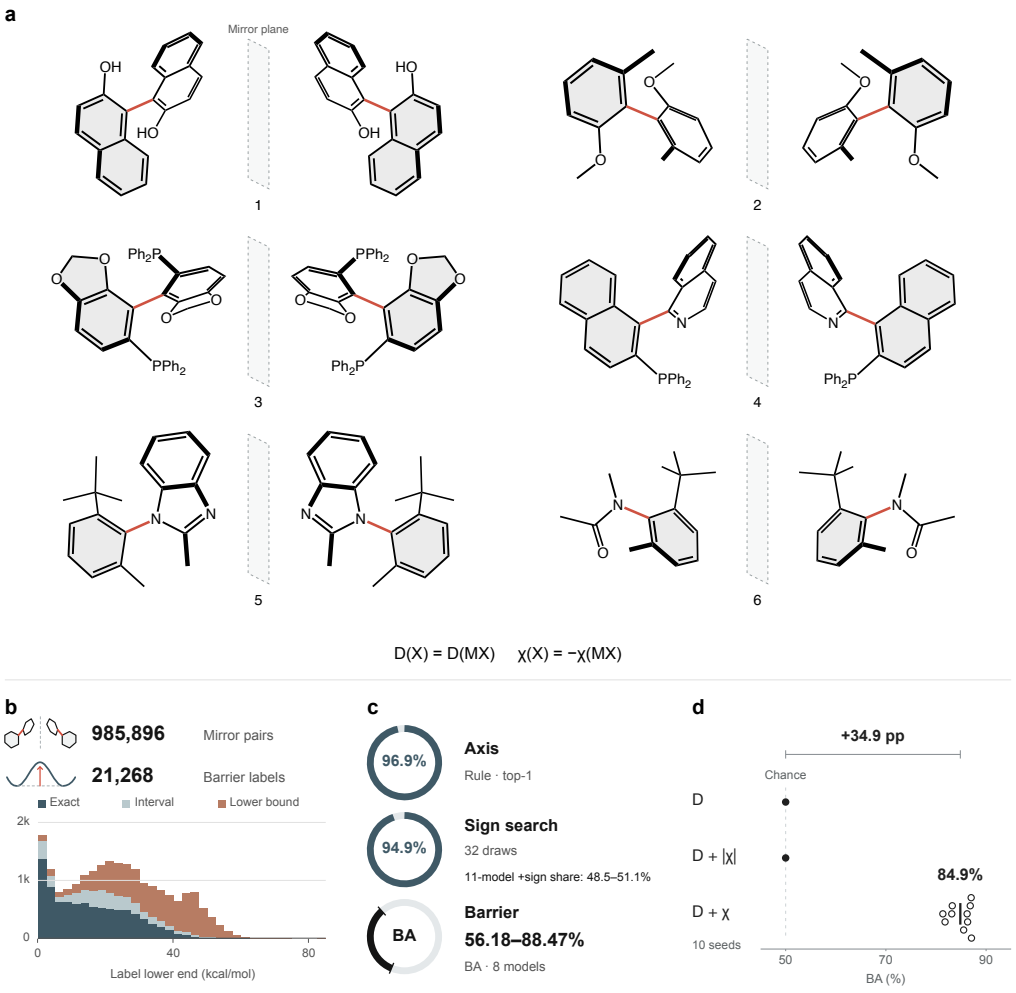


Figure 1: **AtropoBench overview. (a)** Axial-chiral molecules and mirrors; highlighted bonds mark the axes. **(b)** Mirror pairs and separately selected barrier labels; the histogram plots label lower endpoints, which equal the label value for exact observations. **(c)** Original-axis and sign-search diagnostics, plus the barrier balanced-accuracy range across the eight deep learning configurations in Table 1. **(d)** A signed feature lifts a reflection-blind network above chance.

molecules serving as controls against false high-barrier predictions. This organization makes clear what each score measures, rather than treating axial-chirality modeling as a single task.

We compare eleven three-dimensional generators and eight stability-prediction configurations. The results demonstrate why these distinctions matter: high generation quality can coexist with almost no graph novelty, and predictors with similar classification accuracy can make markedly different errors on readily rotating molecules. Strong validation performance also does not reliably indicate transfer to unseen bond classes. These findings expose limitations that overall quality or accuracy alone would miss.

Finally, we use controlled interventions to probe which model components limit performance. We separately modify reflection-sensitive features, the bond-localization model, and uncertainty calibration, then examine which capabilities each change improves. For example, reflection-sensitive features improve handedness discrimination, whereas recalibration improves probabilistic barrier scores without changing predicted barriers (Section 6). These comparisons connect benchmark results to specific model limitations, making AtropoBench useful for diagnosis and improvement, not only ranking.

2 Related Work

Chirality-aware representations. Handedness can be encoded through atom features (Pattanaik et al., 2020), ordered-neighbourhood message passing (Gaiński et al., 2023), chiral determinant kernels (Shi et al., 2026b) and pseudoscalars from an atomistic foundation model (Wedig et al., 2026), or computed directly by rule (Shi et al., 2026a; van Staalduinen & Bannwarth, 2026). These approaches address the fact that interatomic distances alone cannot distinguish mirror images (Pattanaik et al., 2020; Dumitrescu et al., 2025).

Benchmarks and evaluation protocols. Existing benchmarks differ in what they score. Geometry-centric resources (Axelrod & Gómez-Bombarelli, 2022; Xu et al., 2021), generative evaluation suites (Baillif et al., 2024) and torsion-profile collections (Rai et al., 2022; Jing et al., 2022) primarily assess structural or energetic quantities, many of which are unchanged by reflection. Related resources and methods address complementary aspects: 3DCS probes conformational sensitivity (Wang et al., 2026b), ACSD-GAT predicts atropisomeric barriers (Xu et al., 2026), and AtropDiff targets generation (Chen et al., 2025). In a related setting, parity diagnostics in crystal-tensor prediction show how representation choices constrain the targets a model can express (Polat et al., 2026).

3 Dataset Construction

3.1 The pool, the two selections, and the axis correction

AtropoBench was built by systematically modifying the peripheral substituents of 127 published, chemist-verified axial-chiral cores while keeping each core and its stereogenic axis fixed, yielding 5,219,960 candidate molecules. The library spans $1{,}278$ design scaffolds and 857 chemical families. It consists of ligands and catalysts, with a median of 49 heavy atoms.

Two datasets are selected from this library and released as fixed, leakage-audited tables (Table 11, Appendix I). The mirror-pair dataset contains each selected molecule and its exact reflection. The prediction dataset combines axially chiral molecules with 10,478 matched non-axial controls; excluding $2{,}470$ molecules with radical defects leaves $101{,}328$ usable molecules out of 103,798. Neither dataset is a subset of the other. For SEGPHOS-type scaffolds, a graph-based axis detector corrects annotations that originally pointed to an in-ring bond. Anchors for the handedness descriptor are selected using Cahn–Ingold–Prelog ranks, independently of the three-dimensional handedness being evaluated (Appendix D).

The public release represents the candidate library by a deterministic stratified sample of $500{,}000$ records covering $1{,}278$ family$\times$scaffold strata, enabling the reported composition audit. The standardization protocol specifies the element set {H, B, C, N, O, F, P, S, Cl, Br, I}; in a $100{,}000$-molecule audit sample, 1.66% fall outside this set (Appendix C). Figure 2 summarises dataset construction and barrier-label censoring.

3.2 Handedness definition and validation

We assign axial handedness from molecular geometry using a scalar triple product. For axis atoms i and j, let $\mathbf{u} = \mathbf{r}_j - \mathbf{r}_i$, with anchors a and c selected according to CIP priority at the respective ends of the axis. We define

$$s \;=\; \det\big[\,\mathbf{u},\; \mathbf{r}_a - \mathbf{r}_i,\; \mathbf{r}_c - \mathbf{r}_j\,\big]. \tag{1}$$

This quantity is invariant under rigid translations and rotations but changes sign under reflection. For non-degenerate geometries ($s \neq 0$), its sign therefore assigns opposite handedness labels to a molecule and its mirror image.

Handedness labels are computed from the corresponding molecular coordinates to ensure consistency between each geometry and its annotation. We verify the sign-reversal relation under reflection on all $986{,}285$ mirror pairs in the validation set, with the derivation and validation details provided in Appendix E.

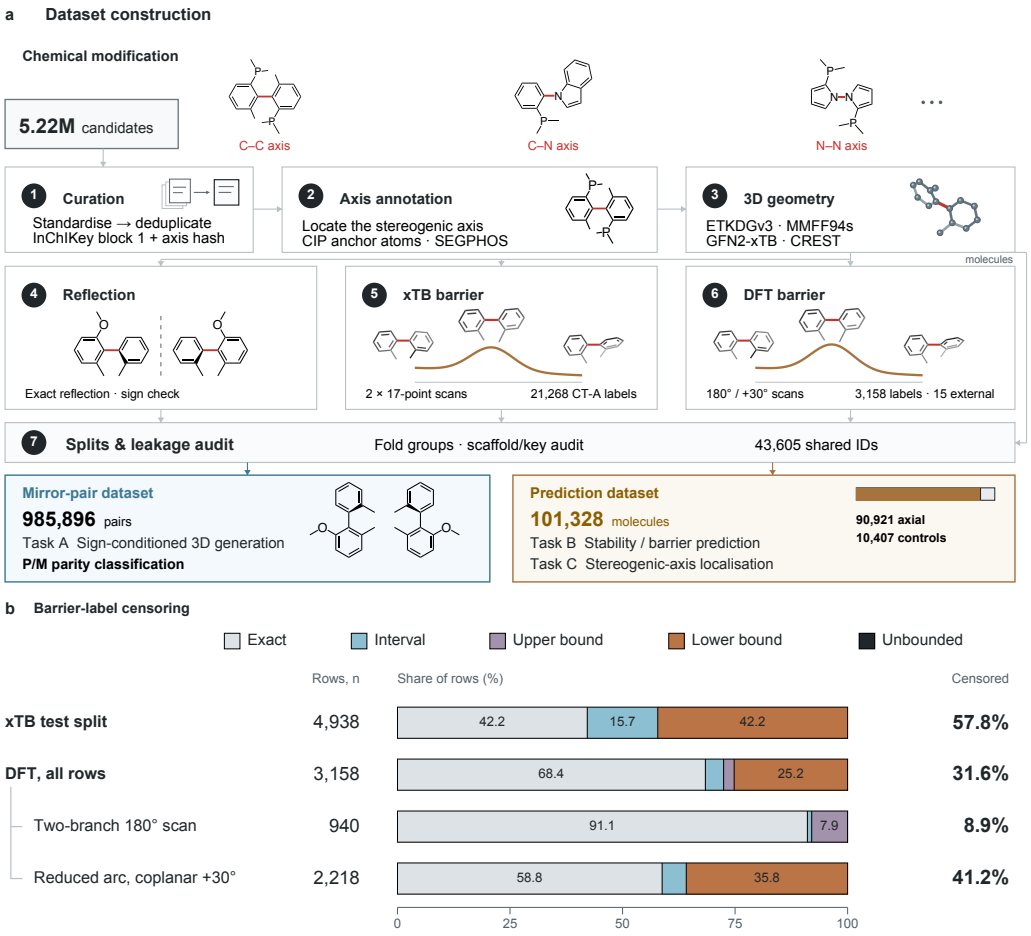


Figure 2: **Dataset construction and barrier-label censoring.** **(a)** Chemical modification produces a candidate library that is curated and annotated; geometry generation precedes separate reflection and barrier-scan branches and a leakage audit. The mirror-pair and prediction selections share 43,605 molecular identifiers; neither contains the other. Box sizes do not encode counts. The barrier-crest structures are illustrative torsional conformers rather than optimised transition states. **(b)** Shares of exact, interval, upper-bound, lower-bound and unbounded barrier labels in the xTB test split and DFT set; every non-exact label is censored. The bracketed DFT subsets (940 full-arc and 2,218 reduced-arc rows) use different scan protocols at the same PBE0-D3BJ/def2-SVP level of theory (Section 3.3, Appendix G).

3.3 The two barrier label sets and their censoring

Both label sets derive barriers from one-dimensional constrained relaxed scans of the axis dihedral, with different levels of theory and scan extents (Table 11). The semiempirical set uses a single GFN2-xTB protocol, with 21,662 molecules attempted and a 99.69% completion rate. The resulting labels comprise exact values (41.4%), intervals (15.1%) and lower bounds (43.5%). The axis-type test split contains 57.8% interval or lower-bound labels; evaluation therefore uses censored log-likelihood as the primary metric and MAE on exact labels as a complementary measure (Appendix F). The quantum-chemical set (3,158 labels) is computed at PBE0-D3BJ/def2-SVP in the gas phase. It comprises 940 full-180° two-branch scans and 2,218 reduced-arc scans extending 30° beyond the coplanar geometry. All 665 labels in the frozen full-arc core are reproduced from the raw scans with a maximum deviation of 0.0. Binary classification uses thresholds of 20 and 30 kcal mol^{-1} for the semiempirical and quantum-chemical sets, respectively: positive at or above the threshold and negative below it. Balanced accuracy is reported on class-determinate samples to compare models within each task.

Fidelity tag F-C. We use gas-phase electronic rotational barriers, $\Delta E^{\ddagger}$, obtained from relaxed scans along the axial dihedral as prediction targets. Each label carries the F-C fidelity tag. Additional calculations assess the effects of thermochemical contributions and transition-state refinement on barrier values and stability classifications. GFN2 biased-Hessian RRHO on 651 DFT geometries puts $\Delta G^{\ddagger}(298\,\mathrm{K}) - \Delta E^{\ddagger}$ at $+0.19 \pm 0.68\,\mathrm{kcal\,mol^{-1}}$, uncorrelated with the target, flipping the three-class label on 4.6% of rows; the same calculations show zero of those 651 scan maxima are first-order saddles. On a 72-molecule stratified sample of the semiempirical set, sampling rule fixed before any run, 0 of 72 maxima are stationary points and a saddle search reaches the axis-rotation transition state on 15, where the offset has median -0.21 and mean $+0.75\,\mathrm{kcal\,mol^{-1}}$ across -4.5 to $+7.8$: small, of either sign, and unmeasurable on the majority (Appendix H).

3.4 Data splits and evaluation settings

Scaffold splits hold out axis-local molecular frameworks, family splits test transfer from common to rare chemical families, and axis-type splits hold out bond classes defining the chiral axis. Molecules are grouped by InChIKey connectivity to keep conformers and mirror partners in the same partition.

Semiempirical barrier benchmark. The xTB barrier benchmark uses the CT-A axis-type split: train on C–C/C–O and test on C–N/N–N, with 14,701 training, 1,629 validation and 4,938 test records (Table 8). Audits find no shared scaffolds, connectivity keys, enantiomer groups or uids between training and test. A separate subset of exact labels within 17–23 $\mathrm{kcal\,mol^{-1}}$ probes classification near the $20\,\mathrm{kcal\,mol^{-1}}$ threshold (Appendix J).

DFT evaluation. The DFT axis-type split uses 2,090 C–N/N–N training records and 554 C–C/C–O test records. Calculations use the same level of theory, but all test records have full 180° scans and 1,713 training records (82.0%) have reduced-range scans. We also evaluate models with training restricted to the 377 full-range records while keeping the test set fixed. The full split thus evaluates concurrent shifts in axis chemistry and scan protocol; the restricted comparison also changes the training sample. Protocol definitions, label composition and comparisons are detailed in Appendix J.

4 Tasks and Metrics

Three tasks evaluate sign-conditioned 3D generation, rotational-barrier and configurational-stability prediction, and axis localisation. Each uses task-specific metrics, with explicit rules for handling invalid generated structures and censored barrier labels.

Task A, sign-conditioned 3D generation (§N). An unsigned graph specifies a bond but not its handedness, so the task is *conditional* on a target sign t_i, scored by target-sign accuracy with invalid outputs counted as failures, so a model cannot gain accuracy by making hard cases invalid. A target-sign metric admits a shortcut on an axis-only subset: reflecting any structure whose measured sign is wrong scores 100% while preserving every distance. The primary ranking therefore runs on a rebuilt *central-chirality confounder subset*, gated on configuration rather than constitution, where that shortcut gains 0.000 against $+0.448$ under the released gate (Appendix N). Stereochemical scores also require a novelty measure (§A.4); the reference evaluator is verified against the benchmark's own labels (Appendix M).

Task B, stability and rotational barrier (Appendix N). The target is the *electronic* $\Delta E^{\ddagger}$ of Section 3.3, distinct from a $\Delta G^{\ddagger}$ or a half-life. The primary metric on both tiers is censored log-likelihood per row (Eq. (4), Appendix N), which incorporates exact, interval and bound observations. Within each tier, MAE on exact-label rows and balanced accuracy on class-determinate rows at the specified threshold provide complementary measures of point prediction and classification performance (§A.2).

Task C, axis localisation (Appendix N). Top-1 localisation accuracy is the fraction of molecules for which the highest-ranked predicted axis matches the annotated axis. Section 6 further examines the effect of localisation methods through controlled comparisons. Handedness (P/M) is evaluated with the axis *supplied* (§A.5).

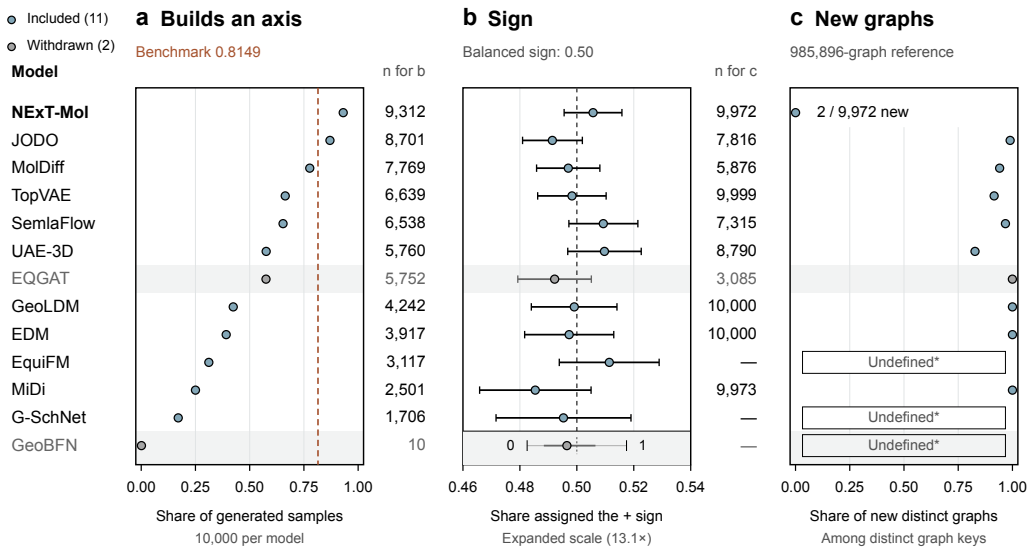


Figure 3: **Axis formation, sign balance and novelty across thirteen generators.** 10,000 samples each. Panels **(a,b)** use one coordinates-only axis detector; panel **(c)** reports graph novelty after post-hoc valence repair where bond-order graphs are available (*Undefined** means no usable bond-order graph). Blue: eleven included models; grey: withdrawn EQGAT and GeoBFN. Panel **(b)** uses 0.46–0.54 with null 0.50; GeoBFN's ten axes use a 0–1 inset. Exact values and ranked models: Table 6.

5 Benchmark Results

5.1 Molecular generation

Axis formation and handedness. We trained thirteen generators from scratch on 985,896 mirror pairs and evaluated 10,000 samples per model with the same coordinates-only detector (Figure 3). Eleven models passed the sample consistency checks; the exclusions of EQGAT and GeoBFN are documented in Appendix K. Only JODO (0.8701) and NExT-Mol (0.9312) exceed the reference library's fraction of molecules with detectable stereogenic axes (0.8149). Among detected axes, positive-sign fractions range from 0.4854 to 0.5114, and all eleven Wilson intervals include 0.50. These unconditional generators thus produce both handedness signs at approximately equal rates. Detector coverage and error rates are given in Table 6.

Novelty and sign-conditioned generation. NExT-Mol leads the reported quality metrics. In the post-hoc repaired-graph novelty diagnostic, 9,998 of its 10,000 samples match training-set molecular graphs, compared with at most 1,521 for any other model. Its positive-sign fraction is 0.5057. The quality scores therefore describe mostly structures already represented in the training set. To evaluate whether generation can follow a requested handedness, we separately compare three baselines: target-sign accuracy is 0.4843 for unconditional generation, 0.9490 for rejection sampling with thirty-two draws, and 0.8030 for an axial-dihedral rule (Appendix N).

5.2 Rotational-barrier and stability prediction

We evaluate eight deep learning configurations and graph-based baselines on the semiempirical axis-type split, with C–C/C–O axes used for training and C–N/N–N axes held out for testing. Every model receives the target axis, and each deep learning configuration uses five random seeds. Table 1 reports barrier prediction, stability classification at 20 kcal/mol, and errors on free-rotor controls under this common protocol. The complete configuration grid is given in Appendix L.

Barrier prediction. Censored log-likelihood evaluates the probability assigned to both exact values and censored observations (Eq. (4)). Four of the eight deep learning configurations outperform the validation-selected constant baseline on this primary metric. GROVER obtains the highest score, improving on the constant by +0.57 nats per row.

Table 1: **Barrier prediction, binary classification and free-rotor controls.** Eight deep learning configurations (five seeds each), three non-neural baseline recipes and one constant comparator (all *italic*), with the target axis supplied: threshold 20 kcal/mol, 4,187 class-determinate test records, 2,082 exact labels and 599 class-determinate controls from 647 free rotors. All model, feature, checkpoint, learning-rate and scale selection uses validation data. The primary metric is censored log L/row (Eq. 4; higher is better). The validation-selected constant predicts positive for every molecule; positive-class prevalence among the free-rotor controls is 0. Bars span 0–1 and 0–100%, with reference ticks at chance and zero; pale-only bars denote zero values.

Recipe	Censored log L/row	Balanced accuracy	MAE, exact (kcal/mol)	Free rotors predicted positive
GROVER	−1.9801	0.8847	4.43	7.18%
Uni-Mol v2	−2.0642	0.8805	4.73	3.17%
ChiDeK	−2.1584	0.8638	6.32	1.50%
RDKit + axis-local descriptors	−2.5760	0.8266	5.70	0.50%
second-shell count, one feature	−2.2298	0.7995	9.81	4.34%
ChIRo	−2.4979	0.7949	4.72	53.09%
3D-Infomax/PNA	−2.5940	0.7879	5.06	0.00%
SE(3)-Transformer	−2.7121	0.7810	5.31	3.17%
EquiformerV2-small	−4.7092	0.5730	6.36	1.67%
MACE	−4.2215	0.5618	5.84	2.50%
ECFP4 fingerprint	−3.6468	0.5243	6.32	3.17%
best constant	−2.5503	0.5000	19.53	100.00%

Table 2: **Discrimination, barrier ranking and decision-threshold transfer on held-out chemistry.** Models are trained on C–C and C–O axis types and evaluated on the held-out N–N axis type: 2,128 class-determinate records, 1,003 exact labels and three seeds. *Decision recovered* is the fraction of the test-fitted oracle threshold's balanced-accuracy gain recovered by fitting a threshold on a separate C–N axis type. This column pools each architecture's seeds and capacity settings (Table 14).

Model	ROC area	Spearman	MAE, exact (kcal/mol)	Decision recovered
D-MPNN (2-D graph)	0.9411	0.7510	4.07	86.9%
PaiNN (3-D)	0.7589	0.4216	6.27	56.1%
SchNet (3-D)	0.6561	0.1289	7.03	2.6%
Constant	0.5000	—	14.05	—

Stability classification. GROVER also achieves the highest balanced accuracy, 0.8847. A ridge-regression baseline whose sole input is the number of atoms in the second shell around the target axis reaches 0.7995, higher than five of the eight deep learning configurations. This result shows that a simple measure of the axis environment already provides substantial information for stability classification. A random forest with two hundred and fifty-eight descriptors reaches 0.8266, compared with 0.5243 for the ECFP4 baseline. Performance by axis type and the effect of validation-based model selection are reported in Appendix L.

Free-rotor controls. We next examine whether a model incorrectly predicts the positive class for molecules that can rotate freely. ChIRo and 3D-Infomax/PNA have similar balanced accuracy: their difference is 0.0071, with a paired scaffold-bootstrap interval of $[-0.0051, +0.0192]$. On the same 599 class-determinate free-rotor controls, however, ChIRo makes 318 false-positive predictions and 3D-Infomax/PNA makes none (Table 1). The control set therefore identifies a difference in false-positive predictions that is not apparent from their overall classification accuracy.

Generalisation across axis types. Across the eight configurations, balanced accuracy ranges from 0.9382 to 0.9604 on validation data and from 0.5618 to 0.8847 on the held-out axis types. Validation performance is therefore insufficient to characterise transfer to new axis chemistry. Appendix L reports the class-specific errors, validation–test rank comparisons and a separate earlier nine-configuration experiment.

5.3 CALIBRATION ON HELD-OUT CHEMISTRY

To examine how calibration affects transfer to new axis types, we conduct a separate experiment with D-MPNN, PaiNN and SchNet. These models are trained on C–C/C–O data using the recorded barrier bounds as point targets. C–N data are then used to fit the decision threshold or the uncertainty scale, and N–N data are reserved for testing (Table 2).

Decision-threshold calibration. The benefit depends on how well predictions order the test molecules. D-MPNN achieves a ROC area of 0.9411; transferring the C–N-fitted threshold recovers 86.9% of the balanced-accuracy improvement attainable with a threshold fitted directly on the test set. SchNet has a ROC area of 0.6561 and recovers 2.6%. Thus, threshold calibration is more effective in this comparison when the model already discriminates positive from negative samples.

Uncertainty-scale calibration. Across the seventeen runs covering these three architectures, random seeds and capacity settings, fitting the scale on C–N data improves test censored log-likelihood in every run. The median improvement recovers 88.6% of the gain from fitting the scale directly on the test set, and family-clustered intervals exclude zero in fourteen runs. Recalibration leaves the point predictions and their ordering unchanged; it improves the probability assigned to the observed barrier intervals. All seventeen runs remain below the best constant baseline. These results concern calibration of the observed-bound predictors; Table 1 evaluates models trained with censored-interval supervision.

6 HANDEDNESS, AXIS LOCALISATION AND UNCERTAINTY CALIBRATION

To improve prediction on axial-chiral molecules, we introduce reflection-sensitive geometric features, improve axis localisation and recalibrate predictive uncertainty. Each modification addresses a specific requirement of the prediction task. We describe the modifications below and evaluate their effects through controlled comparisons (Table 3).

Table 3: **Effects of feature, localiser and calibration changes on three evaluation targets.** Each cell gives the change in its row's metric; bold entries identify the change designed for that target. Intervals use family×seed bootstrap with 20,000 draws. Three zero effects follow analytically; unchanged axis selection is also verified in 592,560 comparisons over 60 runs. For handedness, balanced accuracy increases $0.5000 \rightarrow 0.8486$ and mirror consistency $0.0000 \rightarrow 1.0000$; seed variability is reported instead of an interval (Appendix O). Eight cells are measured or analytically established; the remaining cell was not evaluated.

Evaluation target	Signed feature	Improved localiser	Refitted error scale
Handedness discrimination *balanced accuracy*	**+0.3486** sd 0.0223, 10 seeds	0, exact by theorem, measured	0, exact by theorem
Axis localisation *top-1 localisation*	−0.0071 [−0.0253, +0.0100]	**+0.1434** [+0.0775, +0.2178]	0, exact by theorem, measured
Uncertainty on held-out chemistry *score per molecule, nats*	−0.2367 [−0.5549, +0.0287]	*not measured*	**+1.3841** [+0.6733, +2.1359]

Reflection-sensitive geometric features. Distinguishing mirror images requires features that respond to reflection. We therefore augment the SchNet readout with an axis-local signed volume, $\chi = \det[\hat{\mathbf{u}}, \mathbf{v}_L, \mathbf{v}_R]$. With the target axis supplied, this modification increases handedness balanced accuracy from 0.5000 to 0.8486, averaged over ten seeds, and mirror consistency from 0.0000 to 1.0000 (Table 3). A parameter-matched control uses $|\chi|$, retaining the feature's magnitude while removing its sign. The control remains at balanced accuracy 0.5000 on every seed, supporting the contribution of the signed information to the improvement. A label-free reflection test confirms that the signed feature changes the model's response to reflection (Figure 5); the analytic CIP reference is reported separately in Table 22.

We also compare the signed and absolute-value features on stability classification and barrier ordering, whose targets are unchanged by reflection. Across the six feature × target comparisons, the signed feature improves handedness balanced accuracy by $+0.3486$, whereas its absolute value gives zero gain on handedness by symmetry. Relative to the parameter-matched control, the signed

feature's effects on stability and barrier ordering have intervals containing zero; the absolute-value feature improves stability classification (Figure 6, Appendix O). Additional tests using external ECD labels are reported in the same appendix.

Improved axis localisation. When the target axis is not supplied, the model must select it from candidate bonds. The baseline (`d2`) uses a GIN on the perceived bond graph without coordinates. We replace it with the O(3)-invariant SchNet representation (`inv3d`) used by the handedness head, giving the localiser access to the same three-dimensional geometry. On $4{,}938$ mirror pairs, mean top-1 accuracy increases from 0.7926 to 0.9360 over twenty seeds, a gain of $+0.1434$ $[+0.0775, +0.2178]$. Adding the signed pseudoscalar head to `inv3d` gives 0.9289; its difference from `inv3d` has an interval containing zero (Table 3). The three arms and their common training settings are detailed in Appendix K.4. A strictly increasing recalibration of candidate scores preserves their ordering and leaves axis selection unchanged.

Uncertainty-scale recalibration. To adapt predictive uncertainty to held-out chemistry, we refit the error scale while keeping the predicted barrier means fixed. In a twenty-one-seed paired comparison, this improves censored log-likelihood by $+1.3841$ nats per molecule (Table 3). Adding the signed feature has an effect whose interval includes zero on this metric. The recalibration gain reflects better agreement between the predicted distributions and the observed barrier intervals, with the barrier means and their ordering unchanged.

7 Conclusion and Limitations

AtropoBench jointly evaluates axial-chiral generation, handedness discrimination and configurational stability prediction. Its results show why generation quality, novelty, handedness and stability require separate assessment. Reflection-sensitive features improve handedness discrimination, geometric SchNet improves localisation over the GIN baseline, and uncertainty-scale recalibration improves censored log-likelihood on held-out chemistry without changing point predictions.

Three considerations define the scope. Barrier labels are gas-phase electronic rotational barriers $\Delta E^{\ddagger}$ from relaxed scans, with computational fidelity recorded per label (Section 3.3). The library covers known axial-chiral ligand and catalyst cores. Improvement effects are established for the evaluated tasks, models and splits; transfer to other architectures and chemical distributions requires further evaluation (Appendix O).

Within this scope, AtropoBench tests whether three-dimensional generators can form axially chiral structures and produce a specified handedness, alongside geometric quality and novelty. Broader chemical coverage, free-energy calculations and experimental measurements could support generators that jointly target handedness and configurational stability.

Reproducibility Statement

Under Apache-2.0 we release the $985{,}896$ exact mirror pairs, the $21{,}268$ rotational-barrier labels each carrying its censoring type, the 647 freely rotating controls, the chemistry-held-out splits, the pre-registrations and the issue register. The evaluation code will be made publicly available upon acceptance. The supplied-axis handedness results are distinct from the end-to-end diagnostic; composing a trained localiser in front of the sign head instead costs 0.0297 $[0.0138, 0.0531]$ balanced accuracy (Appendix K). We document the source files, data fields and computational code for the reported results. Appendix R specifies the release scope; the candidate pool and per-row predictions for several neural campaigns are not included in the release. Published intervals use the clustering scheme specified for each experiment; aggregate summaries do not suffice to regenerate them. Appendix R lists the frozen tables that make up the release, the SHA256 digest, byte size and row $\times$ column shape of every file, and the per-file `MANIFEST.tsv` of the two dataset packages built from those tables. The construction of the pool, the two selections and the corrected axis annotation is in Section 3.1 and Appendix D; the handedness convention and its reflection test, verified on all $986{,}285$ pairs, are Eq. (1) and Appendix E. Both label tiers, their levels of theory, their censoring and the fidelity tag every label carries are in Section 3.3. The splits, their leakage audit and the protocol confound the provenance column makes visible are in Appendix J. Every

metric is defined in Appendix N, including the censored log-likelihood of Eq. (4) and the two-level cluster bootstrap over chemical family and seed that every interval in Table 3 is drawn under. Each figure and body table is written by a producer that reads its values from a results file at build time and writes a provenance dump naming the file and key behind each one. The predictions this paper reports as pre-registered were registered in `preregistration/` before the runs that test them, and the registrations record the falsifying values as well as the predicted ones.

ETHICS STATEMENT

This work introduces a computational dataset of small molecules. It involves no human subjects, no personal or identifiable data, no crowdsourcing and no user study, and it raises no privacy concern. The molecules are derived by chemical modification of published, chemist-verified axial-chiral cores rather than drawn from any proprietary corpus, and the release carries an Apache-2.0 licence and a datasheet.

Two boundaries are worth stating because a reader could otherwise act on the labels directly. First, every barrier label here is a computed gas-phase electronic $\Delta E^{\ddagger}$ read off a relaxed scan, not a measured free energy and not a located transition state; Section 3.3 ships a fidelity tag on the label itself for exactly this reason, and these values should not be used as configurational stability estimates for a real synthesis or formulation decision without experimental validation. Second, the library is a controlled stress test of complex stereochemistry rather than a sample of medicinal chemistry, so a model's standing on this benchmark is not evidence about its behaviour on a drug-discovery pipeline. Beyond those, we see no foreseeable harm from the release: the chemistry it covers is ligand and catalyst space based on core structures already published in the literature.

AI USE STATEMENT

In this work, we used generative AI tools for writing and editing the manuscript, for implementing and debugging experiment, analysis and figure code, for literature search, and for adversarial review passes over the manuscript's own numbers and claims. We did not use generative AI tools to produce any result, measurement or number reported here: every value is computed by the analysis code used for this study and is read from a results file at build time by the script that typesets it. Generative AI was not used to generate research hypotheses that are presented as findings, and it is not an author.

We have reviewed all AI-assisted work. Every number in the paper is checked against the results file it is read from, and each figure and table is produced by a script that refuses to emit its output when an assertion over that file fails. Bibliographic entries were checked against the cited work rather than accepted as generated; one entry left with a placeholder author list was corrected that way. AI-assisted code was verified using tests and assertions during development, and the pre-registered predictions were committed before the runs that test them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Keir Adams, Lagnajit Pattanaik, and Connor W. Coley. Learning 3D representations of molecular chirality with invariance to bond rotations. In *International Conference on Learning Representations (ICLR)*, 2022.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4.

Benoît Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using GenBench3D: Ligand conformation quality matters. *arXiv preprint arXiv:2407.04424*, 2024.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176.

Mariami Basilaia, Matthew H. Chen, Jim Secka, and Jeffrey L. Gustafson. Atropisomerism in the pharmaceutically relevant realm. *Accounts of Chemical Research*, 55(20):2904–2919, 2022. doi: 10.1021/acs.accounts.2c00500.

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 (NeurIPS)*, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2024.

Letian Chen, Xi Wang, Gufeng Yu, Caihua Shan, and Yang Yang. Atropdiff: Data-scarce atropisomer generation via multi-task pretrained classifier-guided diffusion. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexandru Dumitrescu, Dani Korpela, Markus Heinonen, Yogesh Verma, Valerii Iakovlev, Vikas Garg, and Harri Lähdesmäki. E(3)-equivariant models cannot learn chirality: Field-based molecular generation. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (NeurIPS)*, 2020.

Piotr Gaiński, Michał Koziarski, Jacek Tabor, and Marek Śmieja. ChiENN: Embracing molecular chirality with graph neural networks. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, Lecture Notes in Computer Science, pp. 36–52. Springer, 2023. doi: 10.1007/978-3-031-43418-1_3.

Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with metadynamics simulations based on tight-binding quantum chemical calculations. *Journal of Chemical Theory and Computation*, 15(5):2847–2862, 2019. doi: 10.1021/acs.jctc.9b00143.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning (ICML)*, 2022.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2D & 3D diffusion models for complete molecule generation. *arXiv preprint arXiv:2305.12347*, 2023.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – efficient 3D molecular generation with latent attention and equivariant flow matching. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 258, pp. 3772–3780, 2025.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-1489.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi S. Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Steven R. LaPlante, Lee D. Fader, Keith R. Fandrick, Daniel R. Fandrick, Oliver Hucke, Ray Kemper, Stephen P. F. Miller, and Paul J. Edwards. Assessing atropisomer axial chirality in drug discovery and development. *Journal of Medicinal Chemistry*, 54(20):7005–7022, 2011. doi: 10.1021/jm200584g.

Tuan Le, Frank Noé, and Djork-Arné Clevert. Representation learning on biomolecular structures using equivariant graph attention. In *Learning on Graphs Conference (LoG)*, PMLR 198, pp. 30:1–30:17, 2022.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof T. Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3D molecule generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NExT-Mol: 3D diffusion meets 1D language modeling for 3D molecule generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Yanchen Luo, Zhiyuan Liu, Yi Zhao, Sihang Li, Hengxing Cai, Kenji Kawaguchi, Tat-Seng Chua, Yang Zhang, and Xiang Wang. Towards unified and lossless latent space for 3D molecular latent diffusion modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lagnajit Pattanaik, Octavian-Eugen Ganea, Ian Coley, Klavs F. Jensen, William H. Green, and Connor W. Coley. Message passing networks for molecules with tetrahedral chirality. In *Machine Learning for Molecules Workshop, NeurIPS*, 2020.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. MolDiff: Addressing the atom-bond inconsistency problem in 3D molecule diffusion generation. In *International Conference on Machine Learning (ICML)*, PMLR 202, pp. 27611–27629, 2023.

Can Polat, Mustafa Kurban, Erchin Serpedin, and Hasan Kurban. A single design choice determines whether machine learning models of materials make physically impossible predictions. *arXiv preprint arXiv:2608.18714*, 2026.

Brajesh K. Rai, Vishnu Sresht, Qingyi Yang, Ray Unwalla, Meihua Tu, Alan M. Mathiowetz, and Gregory A. Bakken. TorsionNet: A deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *Journal of Chemical Information and Modeling*, 62(4):785–800, 2022. doi: 10.1021/acs.jcim.1c01346.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*, PMLR 139, 2021.

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 (NeurIPS)*, 2017.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning (ICML)*, PMLR 139, pp. 9377–9388, 2021.

Runhan Shi, Chi Zhang, Gufeng Yu, Xiaohong Huo, and Yang Yang. A unifying geometric framework for computational representation of stereoisomers based on mixed product. *Cell Reports Physical Science*, 7(1):103065, 2026a. doi: 10.1016/j.xcrp.2025.103065.

Runhan Shi, Zhicheng Zhang, Letian Chen, Gufeng Yu, and Yang Yang. Learning molecular chirality via chiral determinant kernels. In *International Conference on Learning Representations (ICLR)*, 2026b.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3D molecule generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3D molecules with bayesian flow networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D infomax improves GNNs for molecular property prediction. In *International Conference on Machine Learning (ICML)*, PMLR 162, pp. 20479–20502, 2022.

Nils van Staalduinen and Christoph Bannwarth. Automated classification and representation of sp^2–sp^2 atropisomers in the MolBar identifier. *Journal of Chemical Information and Modeling*, 66(17):10935–10955, 2026. doi: 10.1021/acs.jcim.6c00688.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 14170 of *Lecture Notes in Computer Science*, pp. 560–576. Springer, 2023. doi: 10.1007/978-3-031-43415-0_33.

Xi Wang, Jiahan Li, Yuxuan Xia, Yingcheng Wu, Shaoyi Zheng, and Shengjie Wang. Smoothing dark areas in molecular latent diffusion. *arXiv preprint arXiv:2606.13955*, 2026a.

Xi Wang, Yang Zhang, Yingjia Zhang, Yejia Cai, and Shenji Wan. 3dcs: Datasets and benchmark for evaluating conformational sensitivity in molecular representations. In *International Conference on Learning Representations*, volume 2026, pp. 60906–60930, 2026b.

Steffen Wedig, Felix Burton, Rokas Elijošius, Christoph Schran, and Lars L. Schaaf. Rem3Di: Learning smooth, chiral 3D molecular descriptors from atomistic foundation models. *arXiv preprint arXiv:2607.19977*, 2026.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Haisong Xu, Tingting Du, Jiali Lin, Feiying You, Yingbo Shao, Qi Yang, and Xin Li. Physicochemically informed axial chirality descriptors enable accurate prediction of atropisomeric stability. *Angewandte Chemie International Edition*, 65(5):e21349, 2026. doi: 10.1002/anie.202521349.

Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *International Conference on Machine Learning (ICML)*, PMLR 202, pp. 38592–38610, 2023.

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3D: A benchmark for predicting 3D geometries from molecular graphs. *arXiv preprint arXiv:2110.01717*, 2021.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

SUPPLEMENTARY INFORMATION

The following appendices provide extended methods, controls, and results.

A WHAT CURRENT MODELS AND METRICS MISS

Appendix K tabulates the full baseline matrix behind this section. Two conventions hold throughout: the primary barrier metric is censored log-likelihood per row, σ from the held-out validation split (Section N), and every accuracy is labelled *per run* or *ensemble* (Table 4), since neither is uniformly larger. A third is forced by measurement: on the trained parity-odd arm re-running one seed on one card varies more than changing the seed does, standard deviation 0.0280 over three repeats of one seed against 0.0223 over ten different seeds, so we quote no interval there below ten seeds (Appendix O).

A.1 EQUIVARIANCE IS NOT CHIRALITY

Table 4: **Layer (i): mirror consistency and absolute handedness are different axes.** MC is the fraction of pairs assigned *opposite* signs, a relative property of a pair; balanced accuracy is absolute. The trained-model and random-anchor rows use the same $4{,}938$ mirror pairs (109 families, effective 66.92) with the axis supplied as an oracle; the untrained analytic probe instead tests 256 mirror pairs (512 structures). Intervals are family-clustered 95% bootstraps on accuracy. *Per run* scores seeds separately and averages, *ensemble* scores the pooled predictor once: NequIP is 0.6365 ± 0.1164 per seed against 0.8161 ensembled. MACE uses a frozen MACE-MPA-0-medium backbone with a learned parity-odd head.

representation / architecture (seeds)	MC ens. ↑	MC per run ↑	bal. acc. ↑	family 95% CI
analytic probe: reflection-invariant features, no training				
SchNet / EGNN / PaiNN, distance-only	—	—	0.5000	all 512/512 tied
trained parity-odd heads, axis supplied				
NequIP (13)	**1.000000**	**1.000000**	**0.8161**	[0.7658, 0.8639]
EquiformerV2 (5)	0.7284	0.4993	0.7462	[0.6945, 0.7957]
MACE, frozen backbone + head (5)	0.9763	—	0.5681	[0.5235, 0.6144]
SE(3)-Transformer (5)	0.2683	0.3744	0.5785	[0.5478, 0.6116]
random admissible anchor (200 draws)	—	—	0.4996	[0.4864, 0.5158]

A representation built only from interatomic distances assigns a molecule and its mirror image identical features: on 256 double-precision mirror pairs SchNet (Schütt et al., 2017), EGNN (Satorras et al., 2021) and PaiNN (Schütt et al., 2021) all return verdict *even* at maximum relative deviation $\leq 1.5 \times 10^{-14}$, so the distance-only row of Table 4 is an identity, not an estimate. The claim is narrow: a strictly reflection-invariant representation cannot express axial handedness (Dumitrescu et al., 2025; Pattanaik et al., 2020), not that equivariant models cannot learn chirality. The trained rows pull the two axes apart. NequIP enforces parity antisymmetry architecturally and holds mirror consistency at exactly 1.000000 on all 13 seeds and both evaluation splits, yet its per-seed accuracy reaches down to the random-admissible-anchor floor, 0.4996 ± 0.0073: exact parity equivariance and near-chance accuracy coexist in the same network. MACE makes the point from the other side, mirror consistency 0.9763 at balanced accuracy 0.5681, and a *learned* symmetry does not survive the shift at all: per run EquiformerV2's mirror consistency is chance and SE(3)-Transformer's is well below it, so ensembling, not the architecture, supplies most of whatever antisymmetry EquiformerV2 shows, and it moves SE(3)-Transformer further below chance. The CIP-anchor exclusion applies to every P/M number here (Appendix O).

A.2 LIKELIHOOD REVERSES THE OBSERVED-BOUND LEADERBOARD; ITS NEURAL RUNS DO NOT BEAT A CONSTANT

A metric that reads a bound as a value cannot see an error on the censored 58% of the test half, where the reported target *is* the bound (Section 3.3). Table 13 shows the consequence: the two ECFP4 models hold MAE ranks 4 and 5 of 23 and likelihood ranks 22 and 23, because on the $2{,}083$

lower-bound rows they place confident mass *below* bounds the label says the barrier exceeds (-5.61 against a constant's -0.73) while tying the constant on exact rows. An oracle σ refit on test, a diagnostic and never ranked, lifts the better of them only to -2.7507, still below every constant, and over the 23 predictors the two metrics correlate at Spearman -0.1982. The effect is specific to heavy censoring: on the small, largely exact DFT tier it does not reproduce once the predictive scale is fitted out of fold (Appendix G).

Under observed-bound supervision, 0 of 17 neural runs beat the best constant (Table 14). D-MPNN loses 0.80 nats/row *per run* while holding the tier's best MAE on exact labels and balanced accuracy 0.695 against a constant's 0.500: a benchmark scoring only MAE or accuracy would call this tier solved. The mechanism is uniform out-of-distribution mis-calibration: across all 17 runs, three architectures and a $23.6\times$ parameter range, the validation-fitted σ lies in 5.13–5.62 where the test half wants 12.6–21.1, the whole deficit falls on the censored rows, and under an oracle σ only 1 of 9 three-seed runs recovers. This finding concerns observed-bound training; the censored-supervision Barrier-20 campaign in Appendix L has four recipes above its own constant. Capacity does not close the gap, and within a family it widens it: in an eight-run capacity $\times$ depth sweep the $207{,}169$-parameter D-MPNN beats SchNet's $4{,}887{,}041$ parameters on all three metrics, 0.6925 against 0.6662 (Table 14), and all eight stay below the best constant.

A.3 Clustering at the chemical family removes most orderings

For the audited neural runs in this chemically modified candidate library, scaffold clustering yields narrower intervals than family clustering: the $3{,}332$ xTB test scaffolds collapse to 109 families, inflating the median width $4.46\times$ (Appendix O). The 2D-versus-3D Kendall-τ contrast on the barrier tier becomes unresolved under family clustering. Handedness rankings are also split-dependent: EquiformerV2 $-$ NequIP is $+0.0734$ on extrapolation and -0.0700 on test. The family-level tests resolve the test contrast but not the extrapolation contrast (Appendix K); they do not support a ranking that transfers across splits.

A.4 A stereochemical metric alone selects a lookup table

Table 5: **Generation read from the shipped bond tables: thirteen 3D generative models,** $10{,}000$ **samples each; seven rows are displayed.** |*cand.* conditions axis incidence on the samples carrying a candidate axis at all; novelty is against the $985{,}896$-molecule generative training set. ‡ EQGAT is excluded from the body ranking because of an export defect (Appendix K).

model	validity ↑	connected ↑	axis incid. ↑	|cand. ↑	novel vs. train ↑
NExT-Mol	1.0000	1.0000	0.9788	0.9996	0.0002
JODO	0.6605	0.6375	0.6252	0.9950	0.9874
MolDiff	0.5638	0.5248	0.5151	0.9913	0.9384
TopVAE	0.5294	0.5225	0.4768	0.9974	0.8407
SemlaFlow	0.5166	0.4514	0.4630	0.9867	0.9561
UAE-3D	1.0000	0.9806	0.4567	0.9969	0.8274
EQGAT‡	0.1569	0.1218	0.1238	0.9806	0.9994

NExT-Mol attains the maximum on validity, connectedness, InChI round-trip, geometry validity *and* axis incidence, and is last on novelty by a factor of $4{,}000$: $9{,}998$ of its $10{,}000$ samples are exact canonical-graph matches to a training molecule and *not one* lies outside the candidate pool, so its axis incidence is no modelling result. Adding a stereochemical metric to a generative leaderboard is not sufficient; without a novelty column a lookup table wins outright. Over the seven displayed rows the P-fraction spans 0.4988–0.4999, so this detector agrees with the coordinates-only reading of the same files (Table 6) that no model prefers a handedness. Six of the thirteen are excluded from this ranking by their export, not ranked badly: their shipped bond tables carry no bond order this detector can build a candidate axis from, and re-perceiving one from coordinates is no neutral remedy, depressing axis incidence $17\times$ to $34\times$ on exactly the three of them whose files do record bond orders. The thirteen are NExT-Mol (Liu et al., 2025), JODO (Huang et al., 2023), MolDiff (Peng et al., 2023), TopVAE (Wang et al., 2026a), SemlaFlow (Irwin et al., 2025), UAE-3D (Luo et al., 2025), EQGAT-diff (Le et al., 2024; 2022), GeoLDM (Xu et al., 2023), EDM (Hoogeboom et al., 2022), EquiFM (Song et al., 2023), MiDi (Vignac et al., 2023), G-SchNet (Gebauer et al., 2019)

Table 6: **Eleven generative models build a stereogenic axis at very different rates, and none of them produces a preferred handedness.** 10,000 samples each under one coordinates-only detector; the last row is the benchmark's own molecules under it. *Sign* is read over the eleven models with more than a hundred axes. *New graphs* is graph novelty relative to each model's training set after post-hoc valence repair and is available for the nine files that record bond orders (— elsewhere). The biaryl detector has zero recall by construction on B–C, C–O and C–S axes, 0.8115 bond-level recall on the positive xTB benchmark molecules, and unconditional false-positive rate 0.0879 on separate negative ETKDG controls.

Model	Builds an axis	Samples with an axis (n)	Sign (share $+$)	New graphs
NExT-Mol	0.9312	9,312	0.5057	0.0002
JODO	0.8701	8,701	0.4914	0.9894
MolDiff	0.7769	7,769	0.4970	0.9409
TopVAE	0.6639	6,639	0.4983	0.9156
SemlaFlow	0.6538	6,538	0.5093	0.9677
UAE-3D	0.5760	5,760	0.5097	0.8274
GeoLDM	0.4242	4,242	0.4991	1.0000
EDM	0.3917	3,917	0.4973	1.0000
EquiFM	0.3117	3,117	0.5114	—
MiDi	0.2501	2,501	0.4854	1.0000
G-SchNet	0.1706	1,706	0.4953	—
Benchmark molecules	0.8149	—	—	—

and GeoBFN (Song et al., 2024). Every one was trained from scratch on the released parity layer, the $985{,}896$ mirror pairs of Appendix E and not the smaller Task-A table, so the comparison is between architectures on one corpus rather than between checkpoints trained on different ones, and the novelty column of Table 6 is novelty against the corpus each model saw.

A.5 End to end, without the oracle axis

The handedness-classification results in Table 4 use the axis *supplied*. In a separate end-to-end composition experiment, pairing a localiser with a trained sign head on the same $4{,}938$ pairs obeys the measured accuracy transfer function, $\mathrm{acc} = p \cdot \mathrm{acc} \mid \mathrm{correct} + (1-p) \cdot \mathrm{acc} \mid \mathrm{wrong}$, whose conditionals are constant across localisers spanning top-1 0.126–0.927 and reproduce every measured value to ≤ 0.0016: all of the loss is localisation error propagating and none of it is the sign model degrading. At the published top-1 of 0.8631, the composite's accuracy is 0.7890 against 0.8457 with the oracle axis. The failure is not uniform: on the C–S/B–C probe, where the oracle-axis figure is *highest* (0.8784), end-to-end accuracy is 0.5887 $[0.5568, 0.6155]$, so the best-looking configuration collapses hardest once the axis must be found. The composition costs -0.0297 $[-0.0531, -0.0138]$ balanced accuracy at top-1 0.9270 and -0.0738 $[-0.1022, -0.0484]$ at top-1 0.8339, the first being the price Section 7 quotes for the oracle-axis assumption every P/M number in this paper makes. End-to-end mirror consistency is exactly 1.000000, bought by construction rather than earned: the localiser reads a canonical stereo-free graph, blind to the conformer the head reads.

B The candidate pool: taxonomy and composition

Each record in the 5,219,960-candidate pool ($\approx 5.2M$; 127 chemist-verified axial-chiral cores) carries a design *scaffold* ($1{,}278$ of them) and a family label (857 of them) grouping the chemistry of the stereogenic axis — biaryl and heterobiaryl C–C bisphosphines, C–N anilides and N-aryl azoles, diaryl ether and thioether C–O/C–S axes, N–N and rarer heteroatom axes, spiro and helicene-like scaffolds, ligand-specific scaffolds (CPA, phosphoramidite, SDP, SPINOL), and matched non-axial controls.

The pool is withheld because releasing its axis annotation and chemical-modification rules would turn a reproducibility artefact into a pretraining set. The released $500{,}000$-row sample uses proportional Hamilton allocation over the $1{,}278$ family$\times$scaffold strata, with a floor of one row per stratum, selected by a seeded `blake2b` rank so that no RNG state is involved.

Table 7: AtropoBench versus existing molecular benchmarks. ✓: first-class labeled/evaluated; –: absent; ∼: partial/implicit only. "Chirality type" names what is *evaluated*, not merely present. The DFT column reflects a relaxed scan at fidelity F-C, with no transition state located (Section 3.3); the half-life column is ∼ for the same reason — what ships is a stability class at a barrier threshold, and no half-life and no free-energy barrier are computed anywhere in this release (Appendix R). † the pool is the candidate library the two released selections are drawn from and is not itself shipped (Appendix B).

		Chirality		Stability			Tasks		
Benchmark	Size (# mols)	Evaluated type	Axial chir.	Barrier label	Half-life / stab.	DFT hard	Gen. task	Repr. task	Scaffold OOD
MoleculeNet (Wu et al., 2018)	∼700K	none	–	–	–	–	–	✓	∼
GEOM (Axelrod & Gómez-Bombarelli, 2022)	450K / 37M conf.	central	–	–	–	∼	✓	✓	–
Molecule3D (Xu et al., 2021)	∼4M	central	–	–	–	✓	✓	✓	∼
TorsionNet (Rai et al., 2022)	∼7K profiles	none	–	∼	–	✓	–	✓	–
GenBench3D (Baillif et al., 2024)	suite	none	–	–	–	–	✓	–	–
ChiDeK (Shi et al., 2026b)	task sets	central+axial	✓	–	–	–	–	✓	–
AtropoBench (ours)	5.22M pool†; 103,798+52,800	**axial (geom.)**	✓	✓	∼	✓	✓	✓	✓

Two properties of the taxonomy matter for how a result on this benchmark should be read: candidate-pool size is not chemical diversity (5.22 million molecules collapse to 1,278 design scaffolds, and the fifteen most populous families carry 44.6% of the pool), and families are the correlated unit, which is why the intervals in Table 3 are family-clustered (Appendix O) and why 109 effective families cannot resolve models one to two points apart. Where per-row predictions were not retained the interval is scaffold-clustered and therefore understated; those runs are named in Appendix R.

Where this sits. Table 7 places AtropoBench against the closest prior work. ChiDeK (Shi et al., 2026b) encodes central and axial chirality for R/S classification, enantiomer ranking and ECD/OR prediction, but evaluates neither chirality-preserving 3D generation, configurational stability, nor barrier labels under a scaffold-OOD split; AtropDiff (Chen et al., 2025) generates atropisomers, and ACSD-GAT (Xu et al., 2026) predicts rotational barriers from axial descriptors as point predictions. AtropoBench instead declares a censoring type per barrier label and scores it with a proper scoring rule rather than MAE (Section N); no prior benchmark pairs reflection-sensitive axial handedness with per-row censored barrier labels (Section 3.3).

C Standardization protocol

Every molecule passes a fixed standardization before any 3D work, each step writing an explicit QC flag rather than silently dropping the sample. (1) Parse (RDKit `2026.03.3`; on failure set `parse_ok=false`, retain); 0 of 20,000 sampled SMILES fail (100%). (2) Fragment (largest organic fragment kept, original retained). (3) Charge (recorded explicitly; the pool is essentially neutral — 3 charged records per 100,000, all -1). (4) Tautomer (canonical tautomer generated, input preserved; barriers never mixed across tautomers). (5) Stereochemistry (central retained as drawn; axial chirality via the axis annotation and signed handedness descriptor, Appendix E). (6) Provenance (scaffold and family stored per record, so leakage is groupable). (7) Deduplication (InChIKey first block plus axis-local hash; atom-map numbers $1, 2$ are ignored by InChI, so annotation never perturbs deduplication).

The element whitelist is {H, B, C, N, O, F, P, S, Cl, Br, I}; 1.66% of a 100,000-molecule subsample falls outside it, almost entirely silicon (1,658 of 1,660 offending molecules, the remaining 2 iron metallocenes), so the sila-spiro and metallocene families can be excluded wholesale. Chemical review is automated: a stratified vision pass per chemical-modification recipe, not an independent expert audit of the 5,219,960 structures.

D Axis annotation, the SEGPHOS correction, and an invariance that does not hold

Each stereogenic axis is stored as the ordered pair (i, j) of axis atoms plus four anchor substituents (a, b) on i and (c, d) on j, using atom-map numbers 1 and 2 (which survive SMILES/SDF round-trips, are ignored by InChI, and are re-stamped by position after an xTB optimization). Anchors are chosen by a CIP-like priority — primary CIP rank, canonical rank as tie-break.

The SEGPHOS-type axis correction. SEGPHOS-type scaffolds stored the map-$1/2$ bond on an *in-ring* aromatic ipso–ortho bond rather than on the true rotatable biaryl inter-ring bond. Affected records are remapped to the true biaryl bond by a graph axis detector (candidate-axis enumeration plus a hindrance test), with axis atoms, anchors, handedness sign, signed dihedral and axis-local hash all regenerated. The count is population-dependent: $150{,}096$ records in the $1{,}000{,}000$-molecule Task-A selection and 11,481 in the frozen AtropoBench-Gen table.

Anchors must be pinned independently of the chirality measured. A descriptor that depends on RDKit's implicit atom ordering is not a chemical quantity; the failure is invisible to a reflection test, since any non-degenerate triple product has the right parity whatever three atoms build it. Twelve `RenumberAtoms` permutations of one biphenyl, through a descriptor whose anchors come from the 3D geometry, return four distinct values including both signs, and a molecule and its mirror receive the *same* sign ($\chi(P) + \chi(M) = +1.998983$, where the formula requires 0). The released convention avoids this by pinning both anchors by Cahn–Ingold–Prelog rank, at which point the algebra of Eq. (3) is exactly parity-odd: the fault is always the anchor rule, never the pseudoscalar. The original generation-side evaluator is superseded by the repaired path measured in Appendix M; every reported Task-A sign metric uses that repaired evaluator (Appendix N).

What *is* verified, on every pair rather than a sample, is the reflection property, a different guarantee (Appendix E). Neither a permutation-invariance test nor a reflection test is a test of *chemical* axis correctness: a deterministic detector that always selects the same wrong bond passes both. Physical axis correctness requires a stratified chemical audit against a known-compound axis-gold subset, which remains future work.

E HANDEDNESS CONVENTION AND THE REFLECTION-PARITY CORRECTION

The convention is a deterministic, reflection-sensitive handedness sign, with axis atoms i (map 1), j (map 2), $\mathbf{u} = \mathbf{r}_j - \mathbf{r}_i$, and a, c the CIP-selected near and far anchors.

Parity of a triple product. Anchor differences and $\mathbf{u}$ transform by any rigid motion $\mathbf{Q} \in O(3)$, so

$$\det[\mathbf{Q}\mathbf{v}_1, \mathbf{Q}\mathbf{v}_2, \mathbf{Q}\mathbf{v}_3] = \det(\mathbf{Q})\, \det[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]\,: \tag{2}$$

a triple product is rotation-invariant and reflection-odd — exactly the parity a handedness label needs.

The corrected descriptor. AtropoBench uses the single cross-end pseudoscalar

$$s \;=\; \mathbf{u} \cdot \big[(\mathbf{r}_a - \mathbf{r}_i) \times (\mathbf{r}_c - \mathbf{r}_j)\big] \;=\; \det[\,\mathbf{u},\; \mathbf{r}_a - \mathbf{r}_i,\; \mathbf{r}_c - \mathbf{r}_j\,], \tag{3}$$

and $\operatorname{sign}(s)$ is a reflection-odd handedness class; R_a/S_a is derived independently by the standard CIP test, not assumed to map globally onto one (P, R_a) pairing.

Correcting the design-plan formula (a reflection-even bug). The originally proposed descriptor multiplied two per-end pseudoscalars, one at each axis terminus, and set the label to their sign. By Eq. (2) a product of two triple products transforms as $\det(\mathbf{Q})^2 = 1$: it is a true scalar, reflection-invariant, and useless as a handedness label. AtropoBench uses the cross-end form of Eq. (3) instead, and releases the two per-end pseudoscalars only as reflection-invariant relative-frame features.

Validation on the parity layer, on every pair rather than a sample. The released parity layer holds $985{,}896$ mirror pairs, each one molecule stored as a GFN2-xTB optimised conformer plus its exact reflection. Over the $986{,}285$ parsed pairs, 0 are quarantined, $\det(\mathbf{Q}) = -1$ on 100%, improper-superposition RMSD $\leq 4.02 \times 10^{-14}$ Å against a minimum proper-rotation RMSD of 0.504 Å with zero superimposable pairs, and the negated axis dihedral's maximum deviation is $0.0°$; the sign of Eq. (3) flips under reflection on 100% (100% of pairs give $\det = -1$).

The shipped handedness column, and which one to use. One correction matters for anyone training on this release: the original handedness column in the prediction universe table was computed on a different ETKDG conformer from the geometry shipped beside it, and because ETKDG handedness

is a coin flip across seeds the two agree at chance — 0.5043 over $22{,}233$ molecules, and 0.5038 away from the near-planar degeneracy, ruling out a coplanarity artefact. A handedness label is meaningful only relative to its coordinates; for an atropisomer both wells are real molecules, so there is no conformer-independent absolute sign. Use the ground-state handedness column ($22{,}233$ rows, full split coverage), recomputed on the shipped ground-state geometry and validated $600/600$ on each of the three pseudoscalar identities: sign flip under reflection, magnitude preservation, and rigid-motion invariance.

F THE GFN2-XTB BARRIER TIER: PROTOCOL, COMPLETION, CENSORING AND SPLITS

The scalable barrier tier runs *one* protocol on every molecule, `xtb_gfn2_scan_fixedbudget_17pt`: an ETKDGv3 conformer with MMFF94s cleanup, a GFN2-xTB (Bannwarth et al., 2019) ground-state optimisation, then *two* constrained relaxed scans of the axis dihedral to the nearest coplanar angle plus a 30° overshoot, 17 points per branch (CREST (Grimme, 2019) sampling upstream for high-flexibility molecules). Post-scan QC checks structural and chemical validity; failed scans are retained with an explicit failure taxonomy (`embed_fail`, `opt_fail`, `bond_break`, `axis_lost`, `chirality_ambiguous`, `profile_discontinuous`, `multi_path`, `ts_not_bracketed`, `ring_passing`), never dropped.

Completion, and the one chemistry that fails. Of $21{,}662$ molecules submitted, $21{,}594$ completed (99.69%); of the 68 that did not, 67 are hypervalent centres, chiefly pentafluorosulfanyl (`ANC generation failed`), and one fails conformer embedding. The loss is negligible throughout (99.5% retained even in the worst size class).

The label is a minimum over two directions. A molecule rotates whichever way is easier, so the label is $y = \min(\max_{\text{arc}} e_+,\ \max_{\text{arc}} e_-)$ and the censoring is decided by the branch attaining that minimum: exact where that branch turned over inside its arc, interval-censored where it ran out but the other turned over, lower-bounded where neither did (Section 3.3 gives the three shares). An earlier build's over-conservative rule flagged a row censored whenever *either* branch ran to its last frame; correcting it moved $3{,}868$ exact labels to $8{,}940$ without changing any label value.

Splits. `CT-A` is the axis-type split: train on C–C and C–O, test on C–N and N–N (Table 8). Scaffold, InChIKey-14, enantiomer-group and uid overlaps are all zero by measurement; training also excludes anything sharing a scaffold with the frozen DFT-tier test set, leaving 326 completed molecules (304 C–C, 22 C–O) shipped with `ok=True` but belonging to no released fold.

Table 8: The `CT-A` axis-type split of the xTB tier (train C–C/C–O, test C–N/N–N). "eff. groups": inverse-Simpson effective scaffold count. "determinate" counts rows whose censoring interval lies entirely on one side of the 20 kcal mol^{-1} threshold; only those rows are scored for classification.

split	n	scaffolds	eff. groups	exact	interval	bound	determinate	pos. rate
train	14,701	10,264	9,167	6,124	2,208	6,369	13,325	0.667
val	1,629	1,147	1,023	632	244	753	1,447	0.670
test	4,938	3,332	2,992	2,082	773	2,083	4,187	0.556

Why 20, and the ceiling that travels with it. Measured on the 654 paired GFN2-xTB/DFT molecules, scored where censoring determines a class, the threshold trades coverage against agreement: 15 kcal mol^{-1} gives 97.4% coverage at 0.964; 20 gives 94.8% at 0.927; 30 gives 82.0% at 0.778 and drives the test-side positive rate to 0.224. For orientation only, a free-energy barrier of 20 kcal mol^{-1} would correspond to 25.5 seconds of symmetric racemisation half-life at 298.15 K (Appendix N); this conversion is not applied to the electronic scan labels. Two ceilings travel with any score here: a model reproducing the shipped label *perfectly* agrees with the DFT label on only 92.7% of the 654 molecules carrying both, and the shipped geometries reproduce the campaign's own axis dihedral to within 1° for 94.6% of molecules.

The scan profiles are an oracle. The per-point scan energies determine the label exactly: recomputing $\min(\max e_+, \max e_-)$ from them reproduces the shipped label on $21{,}268/21{,}268$ split rows at maximum absolute deviation 0.000000, so they are a target, not a feature, and their exclusion is re-tested on every build.

G DFT PROTOCOL, CENSORING, AND THE TWO PROVENANCE LAYERS

Level of theory, protocol, and fidelity. The released barrier labels were generated without transition-state optimisation, frequency calculations, IRC, thermal corrections or solvation (fidelity F-C, §3.3). The level is PBE0-D3BJ / def2-SVP, gas phase, (99,590) grid, density fitting against `def2-universal-jkfit`, `scf_conv` $= 10^{-10}$, `GAU` convergence, geomeTRIC-driven optimisation. The read-out is a constrained relaxed scan of the axis dihedral, every other degree of freedom relaxed at each step: $\Delta E^{\ddagger} = \max_{\text{arc}} E(\theta) - E(\text{GS})$ (`barrier_kind` = `electronic_dE_cross`), run on two engines.

A double-counted dispersion term, corrected where it can be. Cross-validating the $1{,}517$ branches computed by both engines found a double-counted $D3(BJ)$ dispersion term on the Gaussian side, affecting 198 of $3{,}158$ rows (6.3%): 124 reduced-arc rows, corrected exactly per branch before the minimum is taken, and 74 full-arc rows that cannot be corrected from anything retained, whose error is bounded (not corrected) by an 870-molecule proxy at mean $|\Delta| = 1.93$, median 1.49, maximum $10.23\,\text{kcal mol}^{-1}$.

Two provenance layers, and the censoring rule. The DFT campaign has two layers, recorded per row, both one-dimensional relaxed dihedral scans at the same level of theory, with different arc extents and scan-step schedules. `dft_profiles_2branch180` takes the full $180°$ in two branches at 15–30° steps; `arc24_2026_08` stops at coplanar plus $30°$ at a nominal $24°$ step (measured 20.0–32.8°); censoring follows the release's four-case convention (Eq. (4)), and Table 9 gives the counts by layer. The frozen 665-molecule DFT core is truncated to $[2.06,\ 44.88]\,\text{kcal mol}^{-1}$, so a likelihood over it must renormalise on that window.

Table 9: The two DFT provenance layers of the shipped merged table ($3{,}158$ rows, build v2), with different arc extents and scan-step schedules at the same level of theory. "unb." is the one row unbounded on both branches. The counts use build-v2 censoring after the dispersion correction described above.

layer	arc	step	n	exact	low. bd.	upp./int./unb.	exact rate
full arc	full 180°, two branches	15–30°	940	856	**0**	74 / 9 / 1	**0.9106**
reduced arc	coplanar +30°	20.0–32.8°	2,218	1,304	**795**	0 / 119 / 0	**0.5879**

No F-B labels ship. An F-B ("TS-verified") label requires one imaginary frequency on the axis-rotation mode, connecting the two minima, IRC-confirmed; no released label meets this full protocol, so zero rows ship at F-B. `dft_ts_dG_kcal` and `dft_ts_class` are placeholders (0 non-null); `dft_ts_verified` is 103,798 non-null and uniformly `False`, not null, so `notna()` will not catch it. Drop all three on load.

The predictive scale is fitted out of fold. Fitting σ in-sample is large enough to invert a leaderboard — `desc_rf/naive` moves from rank 16 of 23 to rank 1 on this choice alone — so every log-likelihood used for ranking on this tier and the xTB tier is calibrated out of sample. Test-fitted oracle scales are diagnostic exceptions and are never ranked.

Cross-fidelity against DFT and against experiment. Every quantity in Table 10 is measured from a released file. No calibration factor is applied to any released label; scan integrity is reported per molecule instead, because the discrepancy tracks barrier magnitude, not scan continuity (affine correction slope 0.594 $[0.309,\ 0.880]$, excluding 1, on a 63-molecule external check).

Table 10: Measured cross-fidelity of the released barrier protocols. Row 1 is xTB vs. DFT on the paired core; rows 2–3 are our protocol and the originating lab's own DFT protocol against the same 288 external barriers, a solution-phase $\Delta G^{\ddagger}$.

comparison	n	bias	MAE	Pearson r	Spearman ρ	Kendall τ
xTB vs. DFT (paired core)	654	$+0.7890$	5.9467	0.8129	0.7582	0.5746
ours vs. experiment (external)	288	—	8.4735	—	0.493	—
their DFT vs. experiment	288	—	1.6401	—	0.946	—

H How far a shipped barrier label is from a located saddle

Section 3 states no label comes from a located saddle; this section measures how wrong that makes it, on the GFN2 tier's 21,268 xTB-tier molecules only. It does not transfer to the 665-molecule PBE0-D3BJ core: the mechanism applies there only in form, and its DFT magnitude is unmeasured.

The sampling rule, fixed before any number existed. A 72-cell design over censoring, axis side, scan-integrity, heavy-atom count and barrier band draws one molecule per cell under `RandomState(20260906)`, oversampling flagged cells to 50% against 28.2% of the frame because the label is most suspect there. Nothing is filtered for size or tractability, failures are reported as failures, and four textbook atropisomers are positive controls drawn with the sample.

The saddle searches, and the gap. Mode-following P-RFO from the label geometry reaches the *axis-rotation* saddle on 15 of 72 under the repository's four acceptance criteria. On these, $\Delta E^{\ddagger}(\text{saddle}) - \Delta E^{\ddagger}(\text{label})$ is median -0.21, mean $+0.75$, sd 3.36, range -4.52 to $+7.84$ against the scan's own zero ($6/15$ positive), and median $+1.23$, mean $+3.60$ against the minimum the saddle actually connects: the two differ because 37 of 72 relax below the campaign's ETKDG ground state, drifting the label's zero too. Search success is predicted almost entirely by an imaginary mode already present in the label geometry: 12 of 26 such molecules yield the axis saddle against 3 of 46 without one (46% against 6.5%), and none of the 72 has both an imaginary mode and a converged gradient.

What this does not license. Not a correction applied to the labels: the sign splits $6/9$ and the spread on 15 molecules is $3.4\,\text{kcal mol}^{-1}$ against a mean of 0.75; the gap is small only on the minority where a saddle is reachable, unmeasured on the rest, and a failed search is evidence, not proof, that no saddle exists.

A reproducibility defect found on the way. Regenerating the campaign's deleted scan geometries from unmodified code is itself a measurement, and it does not pass cleanly: the median |re-run − shipped| is $0.0037\,\text{kcal mol}^{-1}$, but 16 of 72 labels move by more than $0.5\,\text{kcal mol}^{-1}$ and one by 19.46, and 11 of the 72 settle on a different optimised ground-state dihedral, moving by more than a degree and up to 147° — the same instability, read on the geometry rather than on the label. The shipped GFN2 label is deterministic given a conformer and not otherwise, the ETKDG-to-GFN2 ground state being unstable across RDKit versions.

I DATASET STATISTICS (EXTENDED)

Table 11: **The frozen AtropoBench release**: one candidate pool, two overlapping selections drawn from it (43,605 shared uids; neither contains the other), and two barrier tiers on the same identifier. **No row in either label set carries a located transition state.**

Pool and selections		**Barrier tiers (frozen)**	
Statistic	Value	Statistic	Value
Pool / valid / cores	5,219,960 / 100% / 127	xTB completed / in splits	21,594 / 21,268
Design scaffolds / families	1,278 / 857	xTB exact / int. / bound	8,940 / 3,269 / 9,385
Barrier-eligible axes	4,233,488 (81.1%)	xTB test / determinate	4,938 / 4,187
Parity pairs released / built	985,896 / 986,285 verified	DFT core / label reproduces	665 / 665, $\max\|\delta\|$=0.0
Universe shipped / usable	103,798 / 101,328	DFT exact / upper bound	649 / 16
Universe axial / controls	93,320 / 10,478	DFT determinate at 30	664, pos. rate 0.5377
Gen mol / records / pairs	52,800 / 105,600 / 3,212	TS-verified rows (F-B)	**0** (labels are F-C)

Table 11 gives the frozen release counts; Figure 2 (§3) gives the censoring composition of the two barrier tiers. Beyond the median heavy-atom count already given in §3.1, the molecules also have mean 51.3 heavy atoms, median molecular weight 703, median 9 rotatable bonds and 7 rings. We run no GEOM/QM9 control, so we record the shift and attribute no measured result to it; the 2D-versus-3D comparison on the barrier tier does not resolve at the family level (Appendix O).

The released tables, in counts. AtropoBench-Gen holds 52,800 molecules and 105,600 records over 427 families, 7,808 Bemis–Murcko scaffolds and 4,242 radius-3 axis-local motifs, with 11,481 SEGPHOS-corrected axes; its scaffold-OOD-moderate split is 97,666/1,510/6,424, its family-OOD test half 25,590 records, and its enantiomer-consistency track 3,212 P/M pairs. Of the 52,800 molecules, 293 carry a seed geometry in the shipped table and the remaining 52,507 are produced by the geometry-generation harness. AtropoBench-Pred holds 103,798 rows — 93,320 axial positives against 10,478 matched non-axial controls, 93,320 of them handedness-labelled — over 561 families and 17,309 scaffolds, of which 101,328 are usable after excluding the 2,470 radical-defect uids the repository excludes from everything (90,921 positives, 10,407 controls, 549 families, 16,864 scaffolds). Its released test halves are 20,783 rows (random), 9,094 (scaffold-OOD moderate), 294 (scaffold-OOD strict) and 33,713 (family-OOD), and 19,999 rows were the xTB tier's build target. Performance evaluations apply the 2,470-uid exclusion; inventory counts explicitly distinguish shipped rows from usable rows. One shipped artefact does not yet obey it: `data/v2/dft_split_s4r/` retains three excluded uids in its test halves.

J SPLITS, LEAKAGE, AND MEASUREMENT-PROTOCOL EFFECTS

The DFT split conformance fix. The shipped split carried 557 rows; three radical-defect uids that our own release rule 6 excludes from everything were removed — a conformance fix, not a refresh. Re-running every baseline changes no ranking and no conclusion, at maximum $|\Delta| = 0.1257$ nats on a single in-sample figure already retracted.

Grouping key, as implemented. The group key (§3.4) is the InChIKey connectivity block's 14-character first block, so tautomers and conformers of one constitution are co-assigned along with enantiomer pairs (3,212 P/M pairs on the generative side); where an axis-local scaffold or a multi-axis parent spans several connectivity blocks, those blocks are co-assigned too.

Split families.

- *Random* — a sanity upper bound only; 20,783 test rows on the universe.
- *Scaffold shift* — grouped on axis-local scaffolds; test↔train maximum Tanimoto capped at strict (< 0.4, 294 rows, a tiny diagnostic cell since the pool shares many modified core structures) and moderate (< 0.6, 9,094 rows).
- *Family shift* — train on common axis chemistries, test on rare ones; 33,713 test rows on the universe.

- *Axis type* — on each barrier tier separately (Table 8 for the xTB tier; §3.4 and the rest of this section for the DFT tier).
- *Near threshold* — the exact-label rows within $\pm 3\,\mathrm{kcal\,mol^{-1}}$ of 20, i.e. the band 17–23; this label-selected subset evaluates prediction performance near the classification boundary.
- *Label source* — screen on the xTB tier, evaluate on the DFT reference. Both report an electronic $\Delta E^{\ddagger}$, so the quantity is fixed; the level of theory and, on the shipped DFT split, the scan-arc convention differ.

Leakage, measured and scoped. The zero-overlap measurement is stated in §3.4; on the xTB tier, training additionally excludes anything sharing a scaffold with the frozen DFT-tier test set. This overlap check is narrower than "no leakage": a user who can reconstruct a test label from elsewhere in the release has no held-out set whatever the split contains, so we audit reconstructability directly, and it is that audit, not the overlap check, that bounds the exposure. Searching 58 frozen test splits against $1{,}758$ released tables over six join keys and their deterministic reductions finds $\mathbf{2{,}430}$ *complete* recovery routes — every row of a split at $\max|\delta| = 0.0$ — across 47 splits and 295 distinct files. The HuggingFace release alone suffices: no clone is needed, and the `t4_near_threshold_*` configs alone cover $4{,}510$ of the $4{,}938$ CT-A test molecules (91.3%). `data/dft_profiles/` is tracked in git in full, though its README calls it an oracle. And the routes are not all copied columns either: `dist_to_threshold` is exactly $|y - 20|$, so y follows from it and `cls_20`, recovering $1{,}264$ of $1{,}264$ near-threshold test rows with the label column deleted. A glob-based audit alone would miss this: it exits non-zero, but its pool covers only the top-level release tables, omitting the localisation and parity tracks, the cheap tier and every dropped-test split. The split *construction* is itself clean: 0 enantiomer-partner, 0 duplicate-InChIKey and 0 uid or scaffold overlap across every partition, with the best single-feature threshold attack reaching only 0.5944.

The sealed release, and the limit of what sealing can buy. Because every published test target is recoverable by a `merge`, we ship a *sealed* variant beside the release: identical file set and identical schemas, with the target columns withheld on held-out rows and nothing else changed, plus a SHA-256 commitment to the withheld values so the answer key can be published later without being trusted now. Three independent implementations verify it (a recoverability audit, ten named attacks, a key-broadcast closure scan) rather than the one that built it, then attack it with the answer key in hand: 0 of 191 attacks succeed against $44{,}091$ hits on the unsealed tree.

Three further limits apply: sibling splits drawn from one pool cannot all be held out at once; P1 and P2 keep the label and the split marker in one table, so their evaluations are transductive; and no packaging makes a deterministic label unreconstructable, since re-running GFN2-xTB reproduces the $4{,}938$ CT-A test labels for roughly 398 core-hours. The seal moves the price of the answer key from a `pandas` merge to 398 core-hours, and that is the strongest true claim available — not that the labels are secret. Second, `barrier_high_kcal` — the *other* branch's barrier, when the label is the minimum of two — was live in the public release from 2026-08-27 to 2026-09-02, over which window a single threshold (26.37) fitted on the training half scores test balanced accuracy 0.8798 against the best structure-only model's 0.7332. It was removed at source on 2026-09-02, and the C–S extrapolation probe, which predated the fix, was repaired on 2026-09-04; any result produced against the affected files between these dates is invalid. Third, one shipped column is a near-copy of a task label rather than a split exposure. Near-symmetric anchors raise the handedness ambiguity score and are excluded as abstentions — $9{,}719$ of $103{,}798$ universe rows (9.4%) carry the flag — and the flag is never set on a molecule that has an axis (0 of $90{,}921$ usable axial rows, $9{,}719$ of $10{,}407$ usable non-axial rows flagged): on these $101{,}328$ usable rows, the rule "predict axial iff not flagged" has sensitivity 1.0000, specificity 0.9339 and balanced accuracy 0.9669, calculated from the stated counts. The column ships in the `universe` config and in no task split, but the master table is where a user joins features from, so any axis-existence baseline that consumes the ambiguity flag is reading its own label and must exclude it.

The DFT axis-type split (S4R) confounds axis type with measurement protocol. Beyond the headline shift given in §3.4, the shipped training half is also far more censored than the test half: a training lower-bound fraction of 0.283 against the test half's 0.000. Performance on this split reflects concurrent changes in chemistry, scan protocol and censoring. The reduced-arc layer contains

Table 12: Decomposing the DFT axis-type split's train–test shift by matching the measurement protocol. Barriers are medians over exact-label rows, in kcal mol^{-1}. Counts and medians are under the corrected build-v2 censoring of Table 9, not the over-conservative rule the split files ship with, so the exact counts here add against that table.

training half	n (rows)	n exact	median barrier	test minus train
as shipped, mixed protocol	2,090	1,380	19.93	+16.18
protocol-matched (`2branch180` only)	377	329	27.98	+8.13
test half (C–C/C–O, 100% `2branch180`)	554	526	36.11	—

1,049 C–N, 1,166 N–N and three C–C molecules (2,218 total), while the frozen test half is entirely full-arc. Of the 2,090 training rows, 1,713 (82.0%) are reduced-arc. This establishes the confound in the frozen split without assuming that future campaign composition is fixed. The analysis below separates the read-out-window offset from the exact-row selection effect.

DFT scan protocols. The DFT data were calculated at the same level of theory using two dihedral scan protocols. The full 180° scans yield 940 records, of which 856 (91.1%) have exact labels. The reduced-range scans yield 2,218 records, including 1,304 exact labels (58.8%) and 795 lower bounds. All lower-bound records in this campaign come from the reduced-range scans. Each record retains its scan protocol and censoring type.

The protocol confound, decomposed. The layer names mislead about the direction of the label-arc offset. `two_branch_180` *scans* 180° but *reads* the label only as far as each branch's first coplanar crossing, while `reduced_arc_coplanar_plus30` reads 30° past coplanar, so the reduced arc's read-out window is a strict *superset* of the other's and its maximum can only rise.

The two protocols differ in how far beyond the coplanar geometry they read the energy maximum; the wider read-out window increases the reported barrier by a median +9.9 kcal mol^{-1}.

The train–test shift does not decompose into protocol and chemistry. Restricting the training half to the test half's layer (Table 12) moves the apparent shift in median exact barrier from +16.18 to +8.13 kcal mol^{-1} — naively an even protocol/chemistry split, but that reading is wrong. The statistic is a median over exact rows and the shipped training half is 28% right-censored, so its exact subset is a low-barrier selection: its median-over-exact is 19.93 against a Kaplan–Meier median over *all* its rows of 24.55. The control half carries no right-censoring at all, and there the two agree to 1.35 kcal mol^{-1} (27.98 and 26.63). Restricting to the full-arc layer therefore removed a *censoring-selection* bias along with the protocol; on a Kaplan–Meier estimator over every row the movement decomposes as 80% censoring rule and 20% protocol.

Matching the protocol does not break the confound either — it exchanges it for a campaign-and-molecule-selection one: the 377 full-arc rows the match leaves are 18% of the shipped training half, whose other 82% is reduced-arc, and they come from a separate campaign whose molecules are smaller (median 39 against 45 heavy atoms). The single comparison free of both — one campaign, one grid, one read-out, 331 against 529 molecules — resolves a location difference of +11.3 [+4.3, +18.3] kcal mol^{-1} and *no* threshold difference at either 20 or 30 kcal mol^{-1}. This corroborates the control split's +8.13 under an independent read-out, rather than licensing a new attribution.

K THE BASELINE MATRIX THAT WAS RUN

Reporting contract. For every run we release the architecture, optimizer, schedule, batch size, early-stop criterion, parameter count, input featurisation (including whether a chirality tag or target-sign channel is provided), seed, split fingerprint and training-target convention, in the per-run JSON beside the result. The convention is stated rather than assumed because it moves Kendall τ by up to 0.11 depending on model family: the seventeen neural runs in Table 14 use `observed_bound`, treating the recorded bound as a point target. The eight Barrier-20 recipes instead use censored interval supervision (Appendix L).

K.1 THE xTB BARRIER TIER: STRUCTURE-ONLY PREDICTORS

Table 13 is the tier's leaderboard under its primary metric, Eq. (4), with one global σ per predictor fitted by maximum likelihood on the $1{,}629$ held-out validation rows and never on test. Three learned predictors clear the best constant, the best of them by $+0.0769$ nats/row (95% CI $[+0.0633, +0.0901]$, 0 of $2{,}000$ scaffold-bootstrap draws favouring the constant); the tier's honest intervals are family-clustered and roughly $4.5\times$ wider (Appendix O). The censoring-aware constant *loses* to the naive one on the full test set because the censored training rows sit much higher than the censored test rows (training lower-bound median 38.01 against test 24.78); it wins where it should, on the $2{,}083$ lower-bound rows, where it is the best predictor in the table (-0.4614 against the naive median's -0.7286), and pays for it on the exact rows.

Table 13: The structure-only leaderboard on the xTB tier's axis-type split (train C–C/C–O, test C–N/N–N; $4{,}938$ test rows, 57.84% censored). Primary metric is the censored log-likelihood per row at a validation-fitted σ; higher is better. "MAE_{ex}" is over the $2{,}082$ exact-label rows only; constants are marked $*$ and sit at balanced accuracy 0.5000 by construction. Eleven near-duplicate rows of the full 23-row leaderboard are omitted, five constants and six fingerprint variants; the # column keeps the full table's ranks.

#	predictor	regime	logL/row	σ_{val}	MAE_{all}	MAE_{ex}	bal.acc@20	τ
1	descriptors + RF	naive	-2.4266	15.61	7.494	7.361	0.6845	0.2844
2	descriptors + ridge	naive	-2.4448	18.34	7.099	6.887	0.7332	0.3584
3	`desc_gbm`	naive	-2.4676	15.02	7.895	7.462	0.6650	0.3068
4	the axis-type shortcut split*	naive	-2.5037	21.21	9.389	11.578	0.5000	0.0000
5	`shortcut_heavy_decile`	naive	-2.5131	22.10	8.767	9.759	0.6115	0.1589
9	`const_train_median`*	naive	-2.5300	22.88	9.389	11.578	0.5000	0.0000
12	censored-MLE constant*	censored	-2.5438	22.76	15.130	19.005	0.5000	0.0000
13	`desc_gbm`	drop	-2.6239	21.83	8.622	7.336	0.5807	0.2349
14	descriptors + ridge	drop	-2.6289	24.55	7.751	6.704	0.6017	0.2958
15	descriptors + RF	drop	-2.6639	23.27	8.774	7.066	0.5632	0.2313
16	`ecfp4_knn5_tanimoto`	drop	-2.6962	19.67	10.156	7.998	0.5172	0.0482
20	ECFP4 + descriptors + GBM	drop	-3.0564	12.42	11.571	8.186	0.5158	0.2247

K.2 THE xTB BARRIER TIER: END-TO-END NEURAL RUNS

Table 14 gives all 17 runs under the calibration reading of Section A.2: not one beats a constant, and every run is 1.6–$2.3\times$ less accurate on the unseen axis type than its own validation split implies. Refitting σ on the test set is an *oracle* recalibration, a diagnostic that separates a location error from a width error and never used to rank; under it exactly one run (`dmpnn_s0`, -2.4508) clears the best constant, so for that run the failure is calibration alone. Within the 2D family capacity *hurts* monotonically, 207k $\rightarrow$ 783k $\rightarrow$ 3.04M taking balanced accuracy $0.6925 \rightarrow 0.5705 \rightarrow 0.5677$. D-MPNN's 0.80 nats/row deficit against the best constant is a per-run mean; its best single seed loses 0.55 nats/row at balanced accuracy 0.7015.

Table 14: All 17 end-to-end neural runs on the xTB tier, same split and same protocol as Table 13. The upper block is three architectures × three seeds (mean ± seed s.d., scored *per run* and averaged, never ensembled); the lower block is the eight-configuration capacity sweep at seed 0 only, named (hidden width, layers), so no interval attaches to its rows. "oracle" is the log-likelihood at a scale refitted on the test set, never a ranked score: that scale runs over 12.6–21.1 across all 17 runs, against a validation-fitted σ confined to 5.13–5.62.

run	params	logL/row	σ_{val}	oracle	bal.acc@20
D-MPNN, 2D (3 seeds)	—	-3.3031 ± 0.2176	5.38–5.56	$-2.45/-2.53$	0.6947 ± 0.0125
SchNet, 3D (3 seeds)	—	-3.7824 ± 0.1800	5.41–5.49	$-2.55/-2.62$	0.6619 ± 0.0192
PaiNN, 3D (3 seeds)	—	-4.5685 ± 0.6532	5.13–5.41	$-2.63/-2.78$	0.5861 ± 0.0302
D-MPNN (192, 3)	207,169	-3.4121	5.468	-2.5266	**0.6925**
D-MPNN (192, 5)	207,169	-3.6195	5.498	-2.5633	0.6731
D-MPNN (768, 3)	3,040,513	-4.0805	5.601	-2.6609	0.5677
D-MPNN (384, 3)	782,977	-4.2405	5.528	-2.6760	0.5705
SchNet (128, 6)	481,921	-4.2969	5.619	-2.6889	0.6305
SchNet (128, 4)	337,025	-4.3569	5.296	-2.6522	0.6601
SchNet (512, 4)	4,887,041	-4.7232	5.248	-2.6896	0.6662
SchNet (256, 4)	1,263,873	-5.1903	5.342	-2.7615	0.5953
best constant	—	$\mathbf{-2.5037}$	21.21	—	0.5000
best descriptor (RF)	—	$\mathbf{-2.4266}$	15.61	—	0.6845

K.3 THE DFT BARRIER TIER

Table 15 gives the tier's baselines on both halves of the axis-type split of Appendix J, the shipped PRIMARY split and the protocol-matched CONTROL, with the scale fitted out of fold (GroupKFold(5) on Bemis–Murcko scaffold); no in-sample figure is reported, for the reason given in Appendix G. Fitting one constant by censored maximum likelihood rather than by naive median moves it from 23.42 to 27.80 kcal mol^{-1} on the PRIMARY training half, a $+4.38$ correction, and that constant then *beats* the naive-median one by $+0.0755$ nats/row ($p = 0.0024$, scaffold-clustered bootstrap, 10,000 resamples over 377 clusters); under the over-conservative censoring rule that correction read $+15.71$. A learned model that beats the constant on one half and loses on the other is the signature of the protocol confound, since a model can beat a constant by learning the offset between two label conventions and is then scored against the other one. Scored one way throughout, the best learned model leads the best constant on PRIMARY by $+0.0468$ nats/row at $p = 0.20$, indistinguishable, and on CONTROL every learned model loses to a constant, with both comparisons reflecting the axis-type and measurement-protocol conditions of their respective splits.

Table 15: DFT-tier baselines under the corrected censoring rule, **one model and one scale protocol throughout**. PRIMARY is the axis-type split (2,090 training rows, 28.2% lower-bound censored; 554 test rows, 0%); CONTROL restricts training to the test half's own provenance layer (377 rows) against the same 554 molecules. Every model, *including both constants*, has σ refitted out of fold on GroupKFold(5) blocked by Bemis–Murcko scaffold, and every column in a row comes from that same model. No balanced-accuracy column is printed: this tier is classified at 30 kcal mol^{-1} because 20 is degenerate on it (Appendix N), and these runs retain no per-row predictions to rescore at 30.

model	split	logL/row	MAE_{ex}	Spearman
censored linear (tabular)	primary	$\mathbf{-4.1010}$	**11.60**	**0.392**
censored-MLE constant	primary	-4.1478	14.11	—
median constant	primary	-4.2233	16.00	—
GBDT (exact rows only)	primary	-4.2938	16.33	0.113
GBDT (bounds as points)	primary	-4.5904	15.83	0.185
median constant	control	$\mathbf{-4.3997}$	14.56	—
censored-MLE constant	control	-4.4750	15.28	—
censored linear (tabular)	control	-4.6305	**14.27**	0.268
GBDT (bounds as points)	control	-4.6872	16.22	0.269
GBDT (exact rows only)	control	-4.8845	16.66	**0.270**

One physics baseline, on the coverage it actually has. The xTB barrier is a legitimate feature for the DFT tier, unlike the xTB scan profiles for the xTB tier, which are an oracle. It is also thin: the legitimate xTB ground-state geometry set covers $1{,}172$ training and only 43 test molecules of the DFT split, and the DFT geometry set is excluded outright as an oracle. On those 43 the raw xTB barrier reaches MAE 5.27 kcal mol^{-1} and Spearman 0.687, far the best secondary metrics anywhere on this tier, at a log-likelihood of -4.2780 averaged over the 43 covered rows; the constant's reported likelihood is averaged over all 554 test rows.

K.4 Axial P/M classification and axis localisation

On the same $4{,}938$ pairs a gradient-boosted model over hand-built geometric features of the axis, trained on C–C axes alone, reaches balanced accuracy 0.9532 as a *single fit* against the 13-seed NequIP *ensemble*'s 0.8161, a paired margin of $+0.1371\,[+0.1250, +0.1494]$; that interval is scaffold-clustered, and applying the largest inflation factor measured on this split ($4.5\times$) would widen it to roughly $[+0.08, +0.19]$ as a sensitivity estimate. The weaker linear variant of the same feature set is *unresolved* at 0.8258 against the same ensemble, margin $+0.0097\,[-0.0053, +0.0253]$. Table 16 gives the four trained 3D architectures on the mirror-pair proxy, adding the extrapolation split to the reading of Section A.1. Two further comparisons characterise performance variation. First, the three lowest NequIP seeds (0.4806, 0.5016, 0.5170) straddle the random-admissible-anchor floor of 0.4996 ± 0.0073, $[0.4864, 0.5158]$ over 200 draws. Second, the relative performance depends on the split: EquiformerV2 − NequIP-13 is $+0.0734$ on the extrapolation split and *reverses* to -0.0700 on test, and an exact family-level sign-flip permutation test over all $2^{11} = 2{,}048$ assignments gives $p = 0.0996$ on the extrapolation split, against $[-0.1078, -0.0308]$ and exact $p = 0.00114$ on test. The internal supplied-axis P/M benchmarks use a $21{,}914$-pair proxy; they do not include the cost of finding the axis, and the release split has never been trained on. That split is the $964{,}623$ of the $985{,}896$ released pairs that carry an unambiguous axis of one of the four common types; it is not the proxy's complement, and the two counts do not add.

Table 16: Trained P/M architectures on the mirror-pair proxy, as seed ensembles, with family-clustered 95% percentile intervals. "MC" is mirror consistency, the fraction of pairs receiving opposite predicted signs. The test split carries $4{,}938$ pairs over 109 families (effective 66.92); the extrapolation split carries 647 pairs over 11 families (effective 7.51). MACE uses a frozen MACE-MPA-0-medium backbone with a learned parity-odd head.

architecture (seeds)	MC	bal.acc	family 95%	extrap.
NequIP (13)	**1.000000**	0.8161	[0.7658, 0.8639]	0.7403
EquiformerV2 (5)	0.7284	0.7462	[0.6945, 0.7957]	0.8138
SE(3)-Transformer (5)	0.2683	0.5785	[0.5478, 0.6116]	0.5232
MACE, frozen backbone + head (5)	0.9763	0.5681	[0.5235, 0.6144]	0.5216
random anchor (200 draws)	—	0.4996	[0.4864, 0.5158]	—

Axis localisation. The localisation track is 504 runs: four backbones (ChiENN (Gaiński et al., 2023), D-MPNN, GPS, GIN) $\times$ seven feature variants $\times$ three seeds $\times$ two candidate pools $\times$ three splits, with a worst-group column throughout. Table 17 gives the six cells. Its headline is a *split*-induced rather than metric-induced instability: the Kendall τ between the i.i.d. and axis-type-OOD leaderboards is $+0.138$, with 163 rank inversions across 28 models. On the i.i.d. and distractor splits every configuration sits between 0.93 and 0.99 and the ordering is essentially noise. A leaderboard built on the i.i.d. split would report a solved task.

Localiser representations and controlled comparison. The axis-localisation comparison in Table 3 evaluates three arms. The baseline, `d2`, is a GIN operating on the perceived molecular bond graph without coordinate inputs. The geometric arm, `inv3d`, uses the O(3)-invariant SchNet stack also used by the handedness head, so that axis localisation can use the same molecular geometry available to the downstream predictor. The parameter-matched signed-feature control, `inv3d_chi`, adds the parity-odd pseudoscalar head of Section 6 to `inv3d`.

All three arms use the same data split and common training code, with a training limit of 25 epochs, early-stopping patience 5, batch size 32, learning rate 10^{-3}, hidden dimension 160 and four layers;

the geometric cutoff is 5.0 Å. Each arm is evaluated with twenty seeds (100–119), disjoint from the three seeds used in the earlier comparison, giving 60 runs in total. The experimental arms, stopping rule and evaluation criteria were specified before these runs.

Across the $4{,}938$ mirror pairs, mean top-1 accuracy is 0.7926 for `d2`, 0.9360 for `inv3d` and 0.9289 for `inv3d_chi`. The geometric-arm contrast, `inv3d` − `d2`, is $+0.1434$ $[+0.0775, +0.2178]$. The signed-head contrast, `inv3d_chi` − `inv3d`, is -0.0071 $[-0.0253, +0.0100]$. Intervals use a two-level bootstrap over 109 chemical families and twenty seeds, with $20{,}000$ resamples. These comparisons quantify the gain from replacing the two-dimensional GIN localiser with the geometric SchNet localiser and the additional effect of the signed head.

Table 17: The 504-run axis-localisation matrix, as six cells of 28 configurations (4 backbones × 7 feature variants, 3 seeds each). "top-1" is the localisation top-1 accuracy and "worst" the worst-group value over the symmetry and CIP-depth strata; each column gives the range over the 28 configurations.

pool	split	configs	runs	top-1 range	worst-group range
clean	axis-type OOD	28	84	0.6598–**0.8631**	0.0130–0.6356
clean	distractor	28	84	0.9309–0.9826	0.8184–0.9604
clean	i.i.d.	28	84	0.9870–0.9937	0.9169–0.9589
matched	axis-type OOD	28	84	0.4886–**0.7690**	0.0101–0.5537
matched	distractor	28	84	0.9478–0.9801	0.8924–0.9437
matched	i.i.d.	28	84	0.9725–0.9888	0.8710–0.9583

K.5 The generative tier

Export defects affecting Table 5 and the omitted models. Six models omitted from Table 5 ship files whose bond tables admit no axis measurement, so a bond-table incidence of 0.0000 reflects the export rather than model performance: G-SchNet, EquiFM and GeoBFN carry no bonds at all (GeoBFN's $10{,}000$ records are $947{,}869$ chlorine atoms and 78 bromine, with no carbon), EDM and GeoLDM write every bond as single, and MiDi ships $993{,}484$ bonds of which 36 are aromatic — enough to pass a bond-count check and not enough to perceive an axis. ‡ Every one of EQGAT's $10{,}000$ records has coordinates uniformly ≈ 6.4–7% too long (nearest-neighbour median 1.1645 Å against 1.0877 for JODO), and a single scalar rescaling lifts its bond-length validity from 0.358 to 0.926, so we report its numbers as properties of this drop.

L Barrier-20 under one given-axis protocol, and the baselines beside it

Every recipe here answers the same question in the same way: given a molecule and one unordered target axis, predict a rotational barrier in kcal mol^{-1}, and assign the positive class at or above 20 kcal mol^{-1} and the negative class below it (Section N). The protocol fixes the target-axis interface, the censored interval supervision, the checkpoint and learning-rate selection rule—validation balanced accuracy, never a test or free-rotor label—and the seed budget. It does not equalise pretraining corpus, parameter count or input modality, so a row is a recipe and not an architecture. The eight recipes are MACE (Batatia et al., 2022; 2024), EquiformerV2 (Liao et al., 2024), the SE(3)-Transformer (Fuchs et al., 2020), 3D-Infomax (Stärk et al., 2022) on a PNA backbone (Corso et al., 2020), GROVER (Rong et al., 2020), Uni-Mol v2 (Ji et al., 2024), ChIRo (Adams et al., 2022) and ChiDeK (Shi et al., 2026b); the neural runs of Appendix K add D-MPNN (Yang et al., 2019), SchNet (Schütt et al., 2017) and PaiNN (Schütt et al., 2021). Every number in this section is computed *in this repository*, from the raw per-seed prediction files, against the frozen `CT-A` split of Appendix J plus the 647 free-rotor controls (624 carbon–sulfur, 23 boron–carbon).

Which constant is the bar. The reference in Table 1 and below is a constant selected the way every recipe here is, on the validation split and never on test: 34.25 kcal mol^{-1} at $\sigma = 22.92$, scoring -2.5503 per row. It classifies every molecule as positive, so its balanced accuracy is 0.5000 and it predicts positive for every free rotor. A constant whose value *and* scale are fitted on the test labels — an oracle no honest procedure reaches — gets to -2.4241, and 3 of these recipes and 2 of the baselines clear even that. Appendix A.2's leaderboard is a different campaign on the same split, with

Table 18: **The full Barrier-20 leaderboard under one given-axis protocol.** Upper block: eight deep recipes, five-seed ensembles. Lower block: twenty-one fingerprint and descriptor configurations fitted under the same protocol; the nine marked $*$ are deterministic and are run once. Classification on the 4,187 determinate test rows; MAE_{ex} over the 2,082 exact rows. *Val* is balanced accuracy on the validation split, whose axis chemistry the training split shares; *Test* is the held-out axis chemistry. *Rotors* counts determinate free-rotor false positives out of 599. *logL* is the PRIMARY metric of Section 4: the censored log-likelihood per row of Eq. (4) over all 4,938 test rows, at one global predictive σ fitted on the validation split and never on test; higher is better. The best validation-selected constant scores -2.5503 on the same rows. Baseline blocks: FP = ECFP4 fingerprint, AX = axis-local fingerprint, D = RDKit 2-D descriptors, G = axis-local geometric descriptors. Learners: RF = random forest, GBM = gradient boosting, 5-NN = Tanimoto nearest neighbours.

recipe	Val	Test	logL	AP	MAE_{ex}	fire	rotors
GROVER	0.9496	0.8847	−1.980	0.9568	4.43	0.5431	43
Uni-Mol v2	0.9604	0.8805	−2.064	0.9441	4.73	0.5816	19
ChiDeK	0.9382	0.8638	−2.158	0.9372	6.32	0.5990	9
ChIRo	0.9459	0.7949	−2.498	0.9444	4.72	0.3561	318
3D-Infomax/PNA	0.9490	0.7879	−2.594	0.9345	5.06	0.3843	0
SE(3)-Transformer	0.9584	0.7810	−2.712	0.9198	5.31	0.3745	19
EquiformerV2-small	0.9563	0.5730	−4.709	0.8602	6.36	0.0920	10
MACE	0.9584	0.5618	−4.222	0.7974	5.84	0.0860	15
D + G + ridge*	0.9062	0.8526	−2.635	0.8604	8.02	0.5806	7
D + G + RF	0.9500	0.8266	−2.576	0.9321	5.70	0.4387	3
FP + AX + D + G + RF	0.9490	0.8138	−3.079	0.9318	6.07	0.3986	0
second-shell count (1 feature)*	0.8734	0.7995	−2.230	0.8543	9.81	0.6771	26
G + ridge*	0.9056	0.7967	−2.175	0.9143	8.75	0.6826	37
G + RF	0.9469	0.7955	−2.982	0.8712	6.83	0.4149	32
FP + AX + D + G + GBM	0.9500	0.7525	−3.489	0.9107	6.91	0.3272	6
G + GBM	0.9376	0.7205	−2.966	0.8920	6.71	0.2959	3
D + G + GBM	0.9408	0.7087	−3.110	0.9126	6.97	0.2720	9
FP + AX + D + G + ridge*	0.9386	0.6917	−5.278	0.7797	10.79	0.3919	11
FP + AX + ridge*	0.9266	0.6003	−3.840	0.6655	10.33	0.5164	157
FP + ridge*	0.9229	0.5973	−3.528	0.6484	10.90	0.5474	263
FP + GBM	0.9328	0.5744	−4.310	0.7529	7.73	0.1220	21
D + ridge*	0.7840	0.5684	−2.855	0.6896	10.33	0.1784	307
D + RF	0.9093	0.5528	−3.153	0.6794	7.50	0.1034	20
FP + RF	0.9375	0.5243	−3.647	0.7332	6.32	0.0609	19
FP + 5-NN*	0.9036	0.5220	−3.064	0.5906	12.04	0.6074	322
D + GBM	0.9094	0.5218	−3.688	0.7093	7.74	0.0420	10
FP + AX + 5-NN*	0.9302	0.5101	−3.805	0.5898	13.10	0.6819	372
FP + AX + RF	0.9407	0.5025	−4.757	0.6973	7.18	0.0033	2
FP + AX + GBM	0.9349	0.4912	−5.468	0.6346	7.76	0.0134	4

its own supervision and its own best constant at -2.5037; every recipe that clears the oracle here clears that one too.

Descriptor, selection and control analyses. (i) Single-descriptor baseline. The second-shell count (1 feature) reaches 0.7995 on the held-out axis chemistry, and 5 of the 8 deep recipes score below it. The feature is ortho steric bulk read off the molecular graph, with no coordinates and no training beyond one fitted coefficient. Its exact-label MAE is $9.81\,\mathrm{kcal\,mol^{-1}}$ and it predicts positive on 0.6771 of held-out molecules against a prevalence of 0.5562. On C–N it reads 0.6027; On N–N it reads 0.9168; the pooled balanced accuracy combines groups with different class prevalences.

(ii) Validation and test rankings. Over the 21 fingerprint and descriptor configurations, all fitted under one protocol on one training split, validation and test balanced accuracy agree at Kendall $\tau_b = +0.1619$ ($p = 0.323$). Selecting on validation, as the protocol requires, returns FP + AX + D + G + GBM and costs 0.1001 of test balanced accuracy against the best configuration in the block. The deep recipes say the same thing with a narrower range: all eight read between 0.9382 and 0.9604 on validation and between 0.5618 and 0.8847 on the held-out axis chemistry.

(iii) Free-rotor errors by axis type. The free-rotor set splits into 23 B–C and 624 C–S, and the split is

Table 19: **The same eight recipes, split by the axis chemistry of the test half.** Balanced accuracy and positive-class recall, on each of the two held-out axis types. An aggregate score can be respectable while one half is at chance.

	C–N		N–N	
recipe	BA	recall	BA	recall
GROVER	0.8224	1158/1346	0.9322	902/983
Uni-Mol v2	0.8054	1260/1346	0.9207	881/983
ChiDeK	0.7822	1209/1346	0.9231	938/983
ChIRo	0.7724	792/1346	0.8199	647/983
3D-Infomax/PNA	0.7771	907/1346	0.7817	583/983
SE(3)-Transformer	0.7209	771/1346	0.8373	682/983
EquiformerV2-small	0.6215	365/1346	0.5000	0/983
MACE	0.5996	327/1346	0.5001	1/983

where the failures are: Among the 599 class-determinate controls, five recipes make false-positive predictions only on boron–carbon (Uni-Mol v2, ChiDeK, SE(3)-Transformer, EquiformerV2-small, MACE), two only on carbon–sulfur (GROVER, ChIRo), and 3D-Infomax/PNA makes none. The 23/624 counts describe the full 647-molecule set, not separate denominators for these class-determinate rates. Two recipes with the same aggregate rate can therefore be failing on opposite chemistries, and the aggregate cannot tell them apart.

Which orderings this benchmark can actually support. A paired scaffold bootstrap over the 2,971 Bemis–Murcko scaffolds of the determinate test rows resamples whole scaffolds, scores every recipe on the same resample, and takes the difference; 2000 draws, percentile interval. Of the 28 pairs among the deep recipes, 4 are left unresolved: ChIRo − SE(3)-Transformer, $+0.0139$ $[-0.0001, +0.0268]$; ChIRo − 3D-Infomax/PNA, $+0.0071$ $[-0.0051, +0.0192]$; 3D-Infomax/PNA − SE(3)-Transformer, $+0.0069$ $[-0.0055, +0.0183]$; GROVER − Uni-Mol v2, $+0.0041$ $[-0.0057, +0.0137]$; including the two highest scores in the table. The analysis is exploratory, conditional on the fixed five-seed ensembles, and carries no multiple-comparison correction; we report the orderings it does not support rather than ranking through them.

Reproduction. `experiments/barrier20/analyze_r2.py` scores the raw per-seed prediction CSVs under `results/barrier20_r2/` against `data/v2/t4/CT-A_test.parquet` and writes every number above to `results/barrier20_r2/barrier20_r2_metrics.json`.

L.1 Supplementary prediction comparisons

Model selection and axis-specific errors. Selecting a complete model configuration from the twenty-one-configuration grid on validation data gives test balanced accuracy 0.7525, which is 0.0470 below the selected single-descriptor baseline. EquiformerV2-small and MACE have validation scores 0.9563 and 0.9584 but together identify only one of the 983 positive-class N–N test molecules. The validation–test rank correlation is negative across the eight deep learning configurations. On the 599 determinate free-rotor controls, Uni-Mol v2 and SE(3)-Transformer each make 19 false-positive predictions, all on B–C axes; GROVER makes 43, all on C–S axes.

Fine-tuned and frozen MACE. In a separate comparison, fine-tuned MACE-MPA-0 has balanced accuracy 0.4996 and ROC area 0.7903, while its frozen-backbone counterpart has 0.4999 and 0.5115, respectively. The similar thresholded classification accuracies therefore correspond to different discrimination performance.

Earlier nine-configuration comparison. Figure 4 reports an earlier experiment with nine configurations and the original full-control-set annotation. Its results are kept separate from the eight-configuration comparison in Table 1.

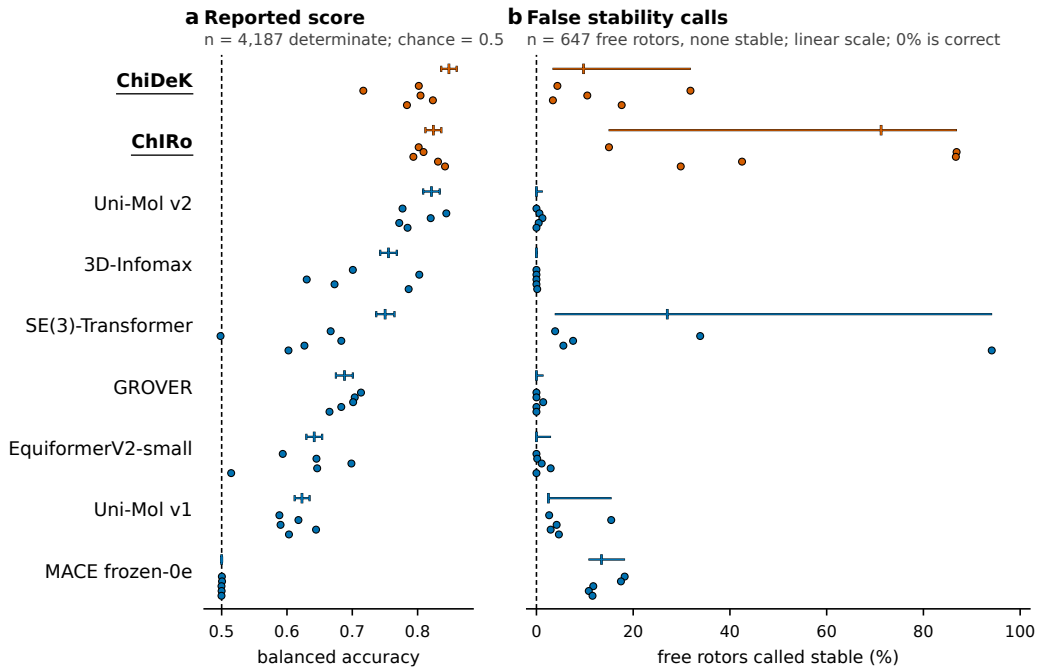


Figure 4: **Balanced accuracy and free-rotor errors in the earlier nine-configuration experiment.** Panel (a) uses 4,187 class-determinate test molecules at 20 kcal mol^{-1}; panel (b) retains the original annotation for 647 controls, where "called stable" denotes positive predictions under this binary criterion. ChiDeK and ChIRo are highlighted. Table 1 reports the separate eight-configuration experiment on 599 class-determinate controls.

M THE GENERATION-SIDE DESCRIPTOR REPAIR, MEASURED

Section N's reference `stereo()` descriptor was not parity-odd; the defect and the repair are quantified here on real axis-bearing molecules against the labels.

generation-side descriptor	agree	n	rate
as released (`eval_generated.py`)	1,914	2,883	0.6639
repaired (`eval_generated_fixed.py`)	2,852	2,860	**0.9972**

The released path is barely above a coin flip against the labels it reproduces. The eight residual disagreements are anchor tie-breaking, not a parity failure: with anchors pinned Eq. (3) is exactly parity-odd.

N TASK DEFINITIONS AND FULL METRIC DEFINITIONS

The three tasks of Section 4, with the diagnostic layer of *representation parity probes*, are each scored under a random split, a family macro-average and the leakage-controlled OOD splits (Section 3.4); Task B separately on its two label tiers, never pooled.

Task A (sign-conditioned 3D generation). Input $(G, \text{axis}, \text{target sign } t_i)$; reflection-tested label s_i (Eq. (3)); prediction $\hat{s}_i$. *Primary* metric: target-sign handedness accuracy $\mathrm{Acc}_{\mathrm{hand}} = \frac{1}{N}\sum_i \mathbb{1}[\hat{s}_i = t_i]$ with invalid outputs counted as failures; the validity-filtered variant is secondary. Also *Validity3D* (bond-length/angle, valence, clash and ring sanity, in the spirit of GenBench3D (Baillif et al., 2024)); *axis-torsion MAE* $\frac{1}{N}\sum_i |\mathrm{wrap}(\hat{\phi}_i - \phi_i)|$; *enantiomer consistency* $\mathrm{ECS} = \mathbb{1}[\hat{s}_A = -\hat{s}_B] \cdot \exp(-\mathrm{RMSD}_{\mathrm{mirror}}/\tau)$; the *raw-vs-relaxed gap* $\Delta_{\mathrm{relax}} = \mathrm{Acc}^{\text{post-relax}} - \mathrm{Acc}^{\mathrm{raw}}$; and graph and scaffold novelty, which are non-optional first-class columns rather than diagnostics, because retrieval satisfies every other metric on this list. For unconditioned generators (no sign channel) we report P/M coverage and balance instead.

One property bounds every Task-A number: on the axis-only subset the primary metric admits the reflection shortcut of Section 4, which is to emit, measure the sign, and reflect if wrong. The primary ranking is therefore taken on the 2,000-molecule central-chirality confounder subset, on which the shortcut is measured rather than assumed defeated; axis-only is secondary.

What defeats the shortcut, measured. A reflection inverts a genuine sp^3 centre, so the subset defeats the move exactly to the extent that its records specify a central configuration and its centres are genuinely stereogenic. The validity gate is therefore versioned: `is_valid_axial/1.0` tests constitution, as published, and reproduces every number reported under it; `is_valid_axial/2.0` tests configuration with the axial descriptor excluded. The rebuilt subset supplies central-configuration tags; the organometallic perception exceptions reported below limit which records are verified from the released SMILES, and that two-dimensional test agrees with the geometric one on 30,254 of 30,254 pool molecules.

Under the constitution gate the shortcut gains $+0.4913$ on the axis-only subset, $+0.4512$ on the confounder subset as first released and $+0.4480$ on the rebuilt one, each measured over 2,000 molecules. Under the configuration gate the same three gains are $+0.4855$, $+0.1685$ and 0.0000. Neither change works alone: the gate leaves the axis-only subset almost untouched, because those molecules carry no centre a reflection could violate, and it takes the subset as first released only part of the way. Only the two together reach zero, and there not one of the 1,973 reflected geometries the shortcut emits is valid at all.

A first baseline, and what it costs to be one. Under the configuration gate, with invalid outputs counted as failures, the unconditioned reference scores 0.4843. For uniformly random target signs independent of the output, the expected score is Validity3D/2. Rejection sampling against the reference evaluator scores $0.4395 \to 0.6650 \to 0.8405 \to 0.9177 \to 0.9390 \to 0.9490$ over $k = 1, 2, 4, 8, 16, 32$ draws, and a rule that sets the axial dihedral directly reaches only 0.8030, worse than eight draws, losing on validity rather than on sign. Seed spread is 0.0004 at $k = 32$, graph novelty is 1.0000 and scaffold novelty 0.9377, so none of this is retrieval, and turning compute into handedness is not learning it: at $k = 32$ the axis-torsion error is still $18.5°$.

The rebuilt subset's centres are specified and genuinely stereogenic for $\mathbf{1{,}978}$ of its 2,000 records, the 22 exceptions being organometallic records whose tagged centres RDKit misperceives, so every rung above is also computed on the 1,977 records containing no transition metal, the largest difference any rung shows being 0.011. Every Task-A sign metric in this paper is computed with the repaired reference evaluator, the released one having read only one record per file and carrying a `stereo()` descriptor that was not parity-odd as integrated (Appendix D); the repaired evaluator agrees with this benchmark's own labels on 0.9972 of 2,860 molecules (Appendix M).

Task B (stability and rotational barrier). Table 20 lists the subtasks. The primary metric is the censored log-likelihood per row, on *all* test rows, with one global predictive σ per model fitted on the held-out validation split; MAE, balanced accuracy and 3-way macro-F1 are secondary. Each model emits a point prediction μ_i, giving a predictive Gaussian $\mathcal{N}(\mu_i, \sigma^2)$ per row. With $z(x) = (x - \mu_i)/\sigma$, ϕ and Φ the standard normal density and CDF, and $[L_i, U_i]$ the shipped `y_lower`/`y_upper` pair, the per-row score is one of four cases keyed off the per-row censoring flag:

$$\ell_i = \begin{cases} \log \phi(z(y_i)) - \log \sigma, & \text{exact,} \\ \log[\Phi(z(U_i)) - \Phi(z(L_i))]\,, & \text{interval,} \\ \log[1 - \Phi(z(L_i))]\,, & \text{lower bound,} \\ \log \Phi(z(U_i))\,, & \text{upper bound,} \end{cases} \tag{4}$$

and the reported metric is $\frac{1}{N}\sum_i \ell_i$, higher is better. No shipped CT-A split has an upper-bound row, and every case's count is reported, zero included: the xTB test split is 4,938 rows, 2,082 exact, 773 interval, 2,083 lower bound and 0 upper bound, i.e. 57.84% censored. The tail direction is load-bearing: a lower-bounded row says the barrier is *at least* L_i and contributes the upper tail, and swapping the two tails moves the naive-median constant on the xTB test split by $+0.0947$ nats/row, from -2.5332 to -2.4385, more than the entire margin between the best learned model and the best constant. Handedness (B2) is *not* graph-determined, since an unsigned 2D graph does not encode axial P/M, so B2 requires a signed 3D conformer or an explicit axial stereotag and is excluded

Table 20: Task B subtasks and metrics. Axis existence (B1) is easy; difficulty concentrates in barrier regression (B5) and the near-threshold 3-way class (B4). B5's target is electronic and censored, never a $\Delta G^{\ddagger}$. No released barrier label is transition-state-verified and no half-life label is shipped (Appendix G). B6 defines a profile task; this paper reports no profile-prediction leaderboard.

Subtask	Output	Metrics
B1 Axis-existence detection	axial / non-axial	AUROC, AUPRC, MCC
B2 Handedness classification	P/M (sign)	accuracy, macro-F1, mirror consistency
B3 Binary barrier classification	positive / negative	censored logL, balanced acc., MCC
B4 3-way stability class	unstable / borderline / stable	macro-F1, ECE
B5 Barrier regression	$\Delta E^{\ddagger}$ (electronic, censored)	censored logL; MAE, RMSE, Spearman ρ
B6 Torsion-profile prediction	$E(\phi)$ curve	profile MAE, barrier MAE, minima error

from any graph-only track; the axis-existence, barrier and class subtasks remain well-posed from a 2D graph. The three-class task uses the categories unstable (< 20), borderline ($20 \leq y < 30$), and stable (≥ 30 kcal mol^{-1}), with the sweep of Appendix O. Binary labels are positive at or above the threshold and negative below it. The threshold is 20 kcal mol^{-1} for xTB and 30 kcal mol^{-1} for DFT. The xTB positive class therefore includes both the borderline and stable categories of the three-class task; the DFT positive class corresponds to the stable category. On the DFT tier, using 20 instead yields a positive fraction of 0.8242 over 660 determinate rows. Balanced accuracy compares models within each tier at its specified threshold.

A single kinetic convention is fixed for orientation, the Eyring rate $k = \frac{k_B T}{h} \exp(-\Delta G^{\ddagger}/RT)$ at 298.15 K with symmetric racemisation half-life $t_{1/2} = \ln 2/(2k)$, so that the 20 kcal mol^{-1} threshold can be read chemically. It is never applied to a shipped label: both tiers report an electronic $\Delta E^{\ddagger}$, and no physical half-life is computed anywhere in this release.

Task C (axis localisation). The input is a molecular structure; the prediction is an unordered atom pair defining the axis. Top-1 accuracy is the fraction whose predicted pair matches the annotated axis. Candidate-pool and split definitions are those of Table 17. Supplied-axis handedness scores do not include localisation error; Appendix A.5 reports the composed diagnostic separately.

A parameter-free rule using only the 2D structure reaches top-1 localisation accuracy 0.9689 $[0.9497, 0.9846]$ on $178{,}911$ held-out molecules in the axis-type split.

Representation parity probes (diagnostic). On the axis-only clean subset, so that a global reflection isolates the axis, freeze a representation and train a linear probe to read handedness from a geometry and its exact mirror. *Probe 1*: a frozen encoder embedding, the SchNet encoder being randomly initialised, not pretrained, since the claim is about the symmetry of the function class, which pretraining cannot change. *Probe 2*: a feature ablation over distances, distances+angles+|dihedral| (all reflection-invariant), each also augmented with an axis-local signed pseudoscalar. Strictly reflection-invariant features cannot separate mirror pairs (Pattanaik et al., 2020; Dumitrescu et al., 2025), so the zero point on the unsigned end is analytic rather than empirical and reported as a symmetry verdict (Section A.1). Appending a single parity-odd scalar moves the same backbones to verdict ODD, while the parameter-matched parity-*even* controls stay EVEN.

Parity metrics and the identity that relates them. For hard predictions $\hat{y} \in \{-1, +1\}$ on balanced mirror pairs, let $\mathrm{PairAcc} = \Pr[\hat{y}(P) = +1,\ \hat{y}(M) = -1]$ denote joint correctness and $\mathrm{MC} = \Pr[\hat{y}(P) = -\hat{y}(M)]$. Then per-example accuracy satisfies $\mathrm{Acc} = \mathrm{PairAcc} + \frac{1}{2}(1 - \mathrm{MC})$: same-sign pairs have exactly one correct member. This joint-correctness metric differs from the ranking statistic $\Pr[f(P) > f(M)]$, for which the identity need not hold. The reported metric audit satisfies the identity to $< 10^{-12}$ on 288 of 294 records; the six exceptions carry the documented zero-logit scoring defect. Tie handling must therefore be specified when reproducing the audit. For these hard decisions, $\mathrm{MC} = 1.0000$ is a consistency statement that forces $\mathrm{Acc} = \mathrm{PairAcc}$ while saying nothing about whether either is above chance (Table 16).

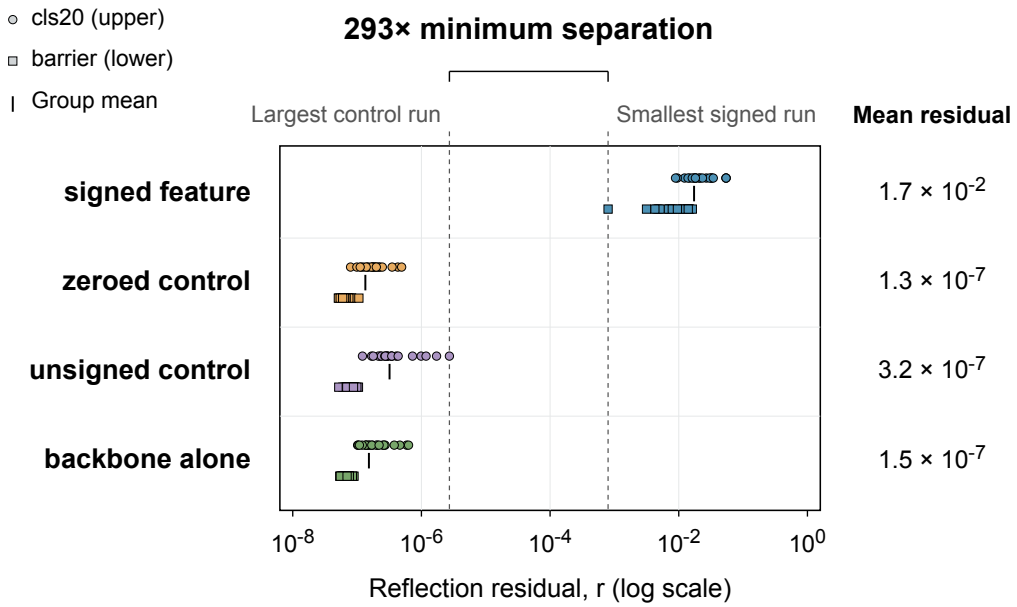


Figure 5: **Prediction changes under molecular reflection.** Each mark represents one seed–target run (21 seeds, two targets): `cls20` is stability classification at 20 kcal/mol and `barrier` is barrier prediction. For each run, $r = N^{-1} \sum_{i=1}^{N} \frac{|f(x_i)-f(Mx_i)|}{|f(x_i)|+|f(Mx_i)|}$ averages the relative prediction change over N held-out molecules, where M reflects the coordinates and f is the scalar model output. The horizontal axis is logarithmic; short vertical ticks mark group means. The signed-feature model responds to reflection, while the unsigned-feature control removes the sign. Reflection-invariant readouts have analytical residual $r = 0$; the small control residuals reflect floating-point precision. Task-level effects are reported in Table 3.

O ADDITIONAL ABLATIONS

Reflection-response controls. A separate reflection test compares predictions for a molecule and its mirror image without using handedness labels. In the twenty-one-seed comparison, the signed-feature model has mean reflection residuals about five orders of magnitude larger than the unmodified network, the zeroed-feature control and the absolute-value control (Figure 5). This confirms that the signed feature changes the model's response to reflection. Direct reconstruction from the CIP-defined target descriptor is reported as an analytic reference in Table 22.

Pre-specified comparisons. The expected effects and evaluation criteria for Table 3 were specified before the experiments. A prospective evaluation on Tox21, with predictions recorded before dataset access, is described in Appendix P.

Recalibration of reflection-invariant predictions. A reflection-invariant model assigns identical scores to a molecule and its mirror image. Applying the same recalibration function to these scores preserves their equality, so recalibration alone cannot improve handedness discrimination (Table 3). The trained two-dimensional graph model follows this prediction, with balanced accuracy 0.5000 and a seed-to-seed standard deviation of 0.0000.

Evaluation with external ECD labels. We also evaluate reflection consistency on 485 axially chiral mirror pairs from ChiDeK (Shi et al., 2026b), using the sign of a computed ECD band as the target. A projected prediction head $\frac{1}{2}[f(x) - f(Mx)]$ increases mirror consistency from 0.8955 to 1.0000. Mirror augmentation without a reflection-odd feature gives balanced accuracy 0.5000. This experiment extends the symmetry comparison to an independently defined target; mirror consistency measures opposite predictions for a pair, while balanced accuracy measures agreement with its labels.

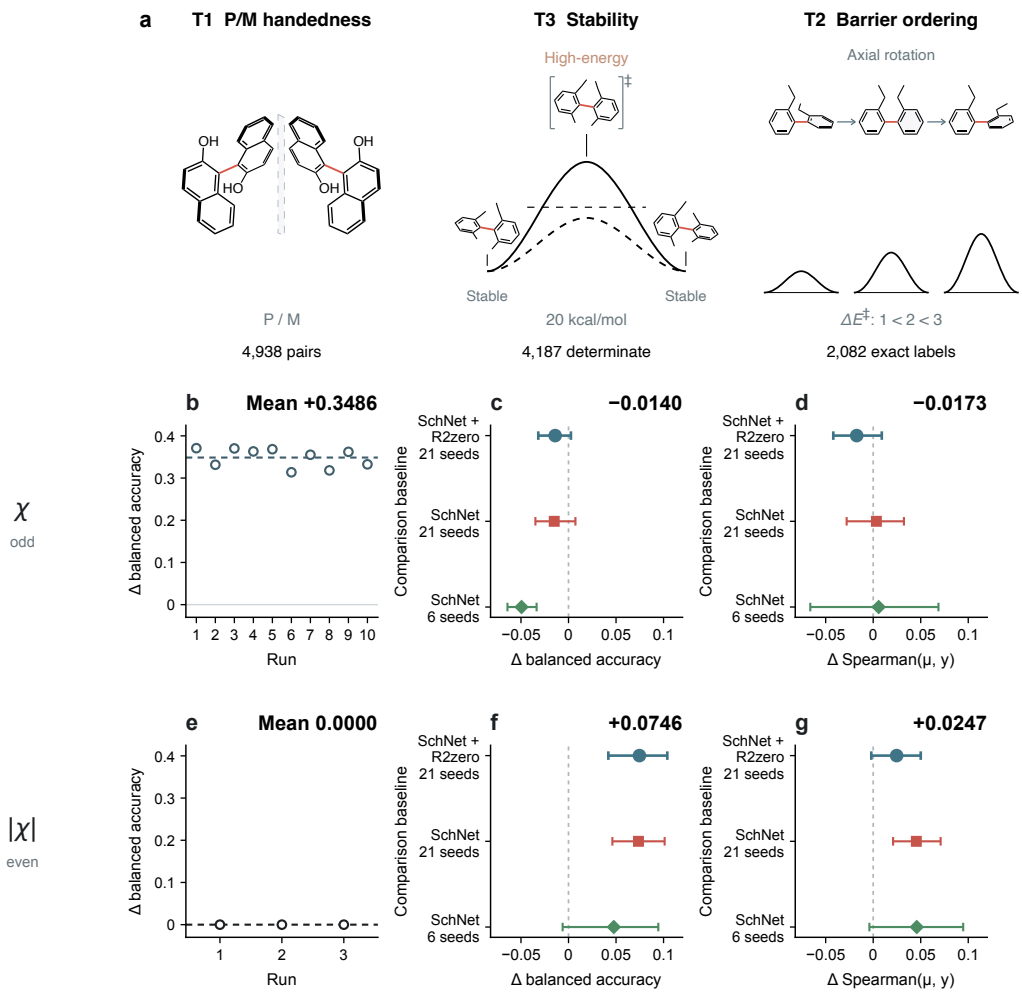


Figure 6: **The parity-odd feature helps the parity-odd target, and its parity-even twin helps a parity-even one.** The control $|\chi|$ is parameter-identical to the signed pseudoscalar χ, and all three targets sit on the same axis-type OOD split. The T3 stability task uses positive/negative labels at the 20 kcal/mol threshold. Both references are drawn in every parity-even cell because the plain base confounds the added feature with the added capacity, and on barrier ordering the two disagree. The likelihood metric on that target points the other way and is reported in Appendix O, not drawn here. The handedness column draws no interval because repeating one seed on one card spreads wider than ten different seeds do (Appendix O); its two parity-blind arms sit at exactly 0.5000 on all three of their seeds; the panel shows their change as $\Delta = 0.0000$. In greyscale, read the interval panels (c), (d), (f) and (g) against the dashed zero line.

Additional comparisons with the unmodified network. In the Table 3 campaign, removing the sign changes balanced accuracy by $+0.0600$ and -0.0395 on two parity-even targets; those intervention contrasts use the unmodified network. The cross-layer contrasts in Table 21 instead use the parameter-matched zeroed-head control; Figure 6 also shows the unmodified-network reference.

Run-to-run variation at a fixed seed. The trained parity-odd arm has substantial run-to-run variation even when the seed is held fixed. This variation must be considered alongside variation across seeds. Three consecutive runs of seed 0 on *one* card, same code and configuration, give balanced accuracies 0.8615, 0.8461 and 0.8072 — a spread of 0.0543, standard deviation 0.0280 — and the early-stopping epoch differs between them. Ten *different* seeds at the same configuration give sd 0.0223, range 0.8137–0.8708. In this small repeat check, the fixed-seed standard deviation exceeds that of the ten-seed sample, so we report no interval on that arm below ten seeds. It is not the harness: the parity-blind base at 0.5000 and the hard-masked nearest-anchor control at 0.5984 are

Table 21: **A repair helps exactly the layer whose symmetry it matches, and the parity-odd half survives on external labels no determinant defines.** Internal parity-even contrasts use a parameter-matched zeroed-head control; P/M gains use the parity-blind base. Where reported, 95% intervals resample Bemis–Murcko scaffolds and seeds. Bold marks intervals excluding zero and the P/M seed-mean gain, for which no interval is reported. Dashes denote unreported entries. Internal rows: one reflection-blind SchNet on the same axis-type split, scored per run; the denominators are 4,938 mirror pairs for P/M, 4,187 class-determinate rows for stability, and 2,082 exact-label rows for barrier ordering. External rows: ChiDeK's axially chiral ECD set (Shi et al., 2026b), 485 usable mirror pairs whose parity-odd label is the sign of a TD-DFT ECD band (CAM-B3LYP/6-311+G(d,p)) and whose parity-even label is its peak position, scaffold-split with zero leakage, every arm at 515,929 parameters. The axis is an oracle input throughout.

target (symmetry)	signed χ	$\lvert\chi\rvert$	head to head, $\lvert\chi\rvert-$signed
internal labels: P/M over 10 seeds, the parity-even rows over 21			
P/M (odd, bal. acc)	$+\mathbf{0.3486}$	0.0000 (level 0.5000), blind by theorem	$-\mathbf{0.3486}$
barrier class @20 (even)	-0.0140 $[-0.0318, +0.0025]$	$+\mathbf{0.0746}$ $[+0.0419, +0.1039]$	$+\mathbf{0.0887}$ $[+0.0629, +0.1156]$
barrier Spearman (even)	-0.0173 $[-0.0420, +0.0090]$	$+0.0247$ $[-0.0020, +0.0500]$	$+\mathbf{0.0420}$ $[+0.0109, +0.0741]$
censored logL (even)	—	-0.8740 $[-1.8304, +0.1094]$	$-\mathbf{1.0428}$ $[-2.0718, -0.0513]$
external ECD labels, 20 seeds; projected odd head 0.7331 beats signed χ by $+\mathbf{0.0370}$ $[+0.0042, +0.0688]$			
Cotton sign (odd)	0.6961: $+\mathbf{0.1961}$ $[+0.1258, +0.2700]$	0.5000, seed sd 0	—
peak position (even)	none of 9 contrasts excludes zero at 121 test pairs		

bit-identical on every seed, and a second, independently written trainer reproduces the effect at the same magnitude, three runs of one seed of the cross-layer 20 kcal mol^{-1} arm giving spread 0.0549, sd 0.0283.

Family-clustered versus scaffold-clustered intervals. Two clustering units are used in this paper and every caption names which. Table 3's intervals are clustered on chemical *family* and seed. Figure 6 and Table 21 use two-level bootstraps over Bemis–Murcko *scaffolds* and seeds where intervals are reported. In contrast, the Barrier-20 paired comparisons in Appendix L resample scaffolds conditional on fixed five-seed ensembles; they do not resample seeds. The handedness architecture tables use family-clustered intervals. These schemes quantify different sources of uncertainty, and the inflation factors below are measurements on specific runs, not universal conversion factors. A family-clustered interval is a bootstrap over *families*, not over Bemis–Murcko scaffolds. Molecules arise through chemical modification of a limited set of verified cores, so two scaffolds inside one family are not independent draws. Replacing a Bemis–Murcko scaffold bootstrap (n = 3,329) with a chemical-family bootstrap (n = 109) on 17 independent neural runs widens the 95% interval on balanced accuracy by a median factor of 4.46, over a range of 3.25–5.13; the point estimates do not move. The 2D-versus-3D difference on Kendall τ, paired between seed-averaged predictors on identical molecules, is $+0.1013$ with a scaffold-clustered interval of $[+0.0709, +0.1325]$ and a family-clustered interval of $[-0.0047, +0.1983]$ that contains zero. What survives at the family level is the separation of D-MPNN and SchNet from PaiNN on balanced accuracy ($+0.1186$ $[+0.0581, +0.1800]$ and $+0.1033$ $[+0.0582, +0.1522]$).

Analytic reference and learned-anchor models. Table 22 reports analytic target reconstruction alongside learned-anchor models and anchor-selection controls. The benchmark descriptor factorises as a stability mask, an ε-tensor contraction (the pseudoscalar of Eq. (3)) and an anchor selection, so a predictor handed the CIP-ranked anchor triple is handed a matrix whose determinant *is* the target: the zero-parameter analytic route scores balanced accuracy exactly 1.0000 on both held-out axis types, a definitional identity and not learning. The floor for the same task is a *random admissible* anchor, at 0.4996 ± 0.0073 on test over 200 draws. Recovering handedness therefore requires a parity-odd feature whose *anchor choice* is itself learned rather than supplied.

Signed-feature ablation, and the cross-layer transfer test. The parity-odd repair is transplanted from the parity-*odd* P/M target onto the parity-*even* targets of the same split, against a parameter-identical zeroed-head control, and which feature helps reverses. The four internal contrasts are the upper block of Table 21: the signed pseudoscalar carries P/M, where its magnitude is blind by theorem, and on the 20 kcal mol^{-1} classification target it is the magnitude that helps and the signed feature that does not. The mis-specification is visible directly as well: the signed arm's outputs differ between a molecule and its mirror by 1.2×10^{-2} relative, on a target where the two are identical by construction, against $\sim 10^{-7}$ on every parity-even arm. What twenty-one seeds resolve is the

Table 22: Analytic and learned-anchor handedness comparisons. Scored on the mirror-proxy test split (4,938 pairs; C–N 2,446, N–N 2,492) held out from a C–C/C–O training half. Model-row intervals use a 2,000-draw scaffold-cluster bootstrap; the random-anchor floor uses 200 draws. Row 1 is an analytic reference supplied with the target-defining descriptor. The learned row chooses its own anchors and is fitted on C–C only; the last row is the empirical floor, over 200 draws. § flat 2.05 Å bond cutoff, against $0.9532\ [0.9466, 0.9597]$ for the covalent-radius variant of the same fit.

predictor	status	bal.acc	95% CI	MC
`analytic_cip` (zero parameters)	analytic reference	1.0000	$[1.0000, 1.0000]$	1.0000
`hgb_cconly` (anchors learned)§	admissible	0.9544	$[0.9479, 0.9608]$	1.0000
`nearest_anchor` (control)	admissible	0.5984	$[0.5824, 0.6131]$	1.0000
random admissible anchor	floor	0.4996	$[0.4864, 0.5158]$	—

head-to-head, and it does so *in opposite directions on the two barrier metrics*: removing the sign helps on Spearman, which scores the location μ alone, and hurts on the censored log-likelihood, which also scores the predicted σ. Splitting that likelihood loss into the part a single global scale can recover (*calibration*) and the part no scale can (*location*) puts 95–101% of it in *calibration*: $|\chi|$ is better on every location metric and worse only because it predicts a narrower scale ($\Delta\bar{\sigma} = -0.335$ $[-0.559, -0.096]$) than its own residuals justify ($\Delta\,\mathrm{RMS}\,z = +0.373$ $[+0.067, +0.678]$), saturating the likelihood floor on 13.46% of rows against 6.11% ($\Delta = +0.0735$ $[+0.0427, +0.1048]$). The floor at $\log 10^{-12}$ truncates per-row penalties; aggregate log-score differences therefore depend on this convention. The reported coverage contrast at $\pm 1.96\sigma$ is unresolved ($\Delta = -0.0014$ $[-0.0420, +0.0394]$), so this coverage check does not resolve the scale-miscalibration effect. Every interval here resamples scaffolds and seeds together, 4,000 resamples, over twenty-one seeds; model selection is in-distribution, scoring is not. Figure 6 draws all six cells together.

Recipe portability, and two recipes that fail their own controls. The censoring-objective recipe was crossed over ten model×feature cells spanning four learners and four representations, three seeds, 14,665 training and 4,932 test rows, 2,000-draw scaffold bootstrap. Seven cells resolve and they resolve in *both* directions — two positive (MLP/PCA-128 $+0.0660$ $[+0.0466, +0.0849]$; gradient boosting/ECFP4 $+0.0232$ $[+0.0078, +0.0387]$) and five negative, as far as random forest/descriptors at -0.1099 $[-0.1301, -0.0900]$. Resolved cells that disagree in sign are not a weak effect but the absence of a portable one, and the pre-registered universality gate is decided against the recipe. A V-REx penalty over eight chemical-family environments gives $+0.0034$ $[+0.0005, +0.0064]$ on Kendall τ at full scale, *smaller* than the $+0.0055$ $[+0.0041, +0.0067]$ from environments assigned at random, so what it buys is generic regularisation.

Generative rank comparison: the stereo-aware ranking is largely the validity ranking. Over the seven rankable models of Table 5 the Kendall τ between the validity ranking and the axis-incidence ranking is $+0.5855$ ($p = 0.0683$, tau-b, a point estimate at $n = 7$), and against the *conditional* incidence $+0.6831$ ($p = 0.0334$).

P EXTERNAL VALIDITY: THE FULL THRESHOLD LADDER

A permuted-label null decomposes the in-block figure of Section 7: with no model, no features and no chemistry, only the twelve block shapes, it reaches 0.0410, 48% of the observed in-block lift. Against the deployable fit's observed $+0.0672$ the same null is 0.0059, under 9%, which is why the held-out number is the one to read (Table 23).

Table 23: **The understatement is real, and it survives a threshold the scoring never saw.** Lift in balanced accuracy over the conventional 0.5. *The registered fit sees the block it scores, so these counts are identities, not evidence.

	Tox21, prospective (36 pairs)			four datasets (12 pairs)	
threshold fitted on	lift	improved	range	lift	improved
the scored block itself (registered)	$+0.0977$	$36/36^*$	$[+0.0036, +0.1941]$	$+0.0756$	$12/12^*$
a held-out scaffold block	$+0.0672$	$31/36$	$[-0.0293, +0.1840]$	$+0.0220$	$8/12$
an in-distribution validation split	$+0.0627$	$31/36$	$[-0.0866, +0.1774]$	$+0.0152$	$8/12$
the better of the two held-out fits	$+0.0704$	$33/36$	$[-0.0293, +0.1840]$	$+0.0294$	$10/12$

Q DATA CARD

AtropoBench ships a Datasheets-style data card (`DATASHEET.md`) covering motivation, composition, collection and generation, preprocessing, label provenance, QC flags, splits, recommended and out-of-scope uses, limitations, licensing, maintenance and contact, each keyed to the appendix section that states it in full. Two of its conventions are not repeated elsewhere. Composition carries two counts per artefact — rows shipped and rows usable after the 2,470-uid radical-defect exclusion — because reporting only the shipped count is what made `data/` look inconsistent with the packaged build; and the out-of-scope list is explicit that this is axial chirality only, that ring-locked and spiro scaffolds are declined for the barrier task, that it is not a general drug-like atropisomer set, and that it is not a source of free-energy barriers or half-lives. The data contain no human-subject information.

Three data-card entries are load-bearing enough to repeat here. (i) The external 288-molecule comparison ships *two* prediction columns; use `ours2`, never `ours`, which is contaminated by a refinement bug — the difference is MAE 8.47 against 15.42 kcal mol^{-1} and Kendall τ $+0.348$ against $+0.025$. We shipped the contaminated column alongside the safe one and documented both rather than removing it, and an audit across the code, evaluation and results trees confirms that no reported number reads it. (ii) The 945-row DFT campaign summary is a *superset* of the 665-molecule labelled core, so 280 DFT-measured molecules carry no shipped label. (iii) the issue register, the claim ledger and the leakage audit are part of the artefact rather than commentary on it.

R REPRODUCIBILITY

What is released, and where. Under the repository's Apache-2.0 licence we release: the audit metadata and a deterministic 500,000-row stratified sample of the 5,219,960-candidate pool; the 103,798-row prediction universe with split membership, the handedness columns of Appendix E and the 2,470-uid defect list; the parity layer of 985,896 released mirror pairs (986,285 built and verified); the xTB barrier tier as 21,268 barrier labels, every row carrying its censoring type, with the per-point scan profiles behind them *withheld from every package* because $y = \min(\max(\text{branch}^+), \max(\text{branch}^-))$ recomputes the shipped label from them exactly, making them an oracle; the DFT tier with the per-row provenance layer column that makes the protocol confound of Appendix J visible at all; the generative samples and their scores; and the evaluation-code defect list. Two HuggingFace-standard packages are built from the same tables (`hf/` for the prediction universe and its six split configurations, `hf_dft/` for the DFT barriers), with the radical-defect exclusion applied at build time and a per-file `MANIFEST.tsv` of SHA256 digests.

Integrity. `MANIFEST.md` carries a SHA256 digest, byte size and row×column shape for every shipped file, and records which files the radical-defect rule has *not* yet been applied to. Two externally sourced tables are recorded as byte-identical extractions with their hashes, so that the DFT margins quoted in `KNOWN_ISSUES.md` are reproducible from a file in this repository rather than from a description of one. Versions are pinned (`rdkit 2026.03.3`, `xtb 6.7.1`, PBE0-D3BJ/def2-SVP with the grid, fitting basis and convergence settings of Appendix G) and seeds are fixed; every `data/v2/*/BUILD.json` records the builder script, its seed and its selection rule.

How a result in this paper is regenerated. The internal project organises each experiment as a directory under `experiments/` whose name matches the `results/` directory it writes; each

results directory contains the JSON read by its Markdown summary, and every figure in this paper reads its numbers from a file at plot time, with none transcribed.

What is not released, stated so that no reader assumes it. There is no hidden-test evaluation server: all $4{,}938$ xTB test labels are recoverable from sibling files in the same release at $\max|\delta| = 0.0$, so containment is an honour-system boundary (Appendix J). No released barrier label is transition-state-verified, and no released label is a free-energy barrier or a half-life; Appendix H reports a separate saddle-search audit. There is no re-exported generative batch, so the bond-table reading of the generative tier in Table 5 still leaves six of the thirteen models unranked; the body reads all thirteen of the same files from atomic positions instead and ranks the eleven that pass a check on their own contents (Table 6), where the one column still read from bond orders — novelty — stays empty for the three files that record none. The per-row predictions for the nine three-seed neural barrier runs are retrieved from a training box and are gitignored, and per-row predictions for the P/M architectures are not part of the reported release. Their published summaries nevertheless include family-clustered intervals (Tables 4 and 16). Aggregate summaries do not replace the per-row predictions needed to reproduce or re-cluster those intervals; release coverage must be stated per campaign rather than inferred from the interval label.