

Carried but Conditionally Used: A Frozen Held-Out Audit of Whether Chirality-Aware Molecular Fingerprints Predict Measured Same-Assay Enantioselectivity

Dylan Ashraf

dylanashraf56014@gmail.com

American High School

Research Article

Keywords: chirality, enantioselectivity, Morgan fingerprints, molecular representations, QSAR, scaffold extrapolation, ChEMBL, reproducibility, stereochemistry

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10530515/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Chirality strongly influences drug potency, yet many widely used two-dimensional molecular representations omit stereochemistry or encode it only through local stereochemical features, and it is often unclear whether a model that *carries* stereochemical information also *uses* it to reproduce measured biology. Using same-assay enantiomer pairs from ChEMBL 37 and a prospectively specified, cryptographically frozen held-out analysis protocol, we separated two questions: whether radius-3 Morgan fingerprints represent enantiomers distinctly (carry), and whether an exactly antisymmetric model predicts measured potency differences between enantiomers under scaffold extrapolation (use). Carry was near-complete at radius 3; in the audited targets, the elevated radius-2 collision rate was radius-limited rather than hash-induced, being identical across 1,024–8,192 bits and unhashed identifiers. For use, chirality-aware Morgan fingerprints supported significant scaffold-extrapolation prediction of measured enantiomer potency differences for **five of eleven held-out targets** (Benjamini–Hochberg recomputed within the held-out family); **eight of fourteen** targets were positive across the complete map when three development targets were included. **Performance varied substantially by target, while ridge and random-forest models showed comparable skill, suggesting that much of the accessible stereochemical signal may be additive in signed-fingerprint space (formal model equivalence is not claimed).** A signal-to-variability association observed across the complete target map did not reach significance in the held-out cohort. Achiral controls and several biological targets returned null or negative skill, showing the procedure did not automatically manufacture predictability. **Scientific Contribution.** This work separates whether chirality-aware molecular fingerprints *carry* stereochemistry from whether an antisymmetric model *uses* it to predict experimentally measured, same-assay enantiomer potency differences, target by target, under scaffold-extrapolation evaluation and a prospectively specified, cryptographically frozen development/held-out separation. Carry is near-complete at radius 3 while predictive use is real but strongly target-dependent – positive for five of eleven held-out targets and null or negative for others – and the elevated low-radius collision rate is shown to be a neighborhood-radius effect rather than a hashing artifact. The contribution is a claim-disciplined audit that reframes "chirality-aware" around measured out-of-sample use rather than representational capacity, reporting balanced positive, null, and negative findings.

1. Introduction

Enantiomers—non-superimposable mirror-image molecules—frequently differ by orders of magnitude in target affinity, and stereochemistry is a first-order concern in medicinal chemistry. Many widely used two-dimensional molecular representations omit stereochemistry or encode it only through local stereochemical features: extended-connectivity fingerprints [1] without the chirality flag assign identical vectors to two enantiomers, and conventional message-passing networks treat them identically, motivating chirality-aware architectures [7, 8]. Prior work has established the *phenomenon*—that many representations collide enantiomers and that a large fraction of stereoisomer pairs differ in bioactivity [2, 3]—and has produced chirality-aware descriptors and models [3, 7, 8]. What remains under-examined is a distinct, measurable question: given real, same-assay measured potency differences between enantiomers, does a representation that technically encodes chirality allow a model to *predict* those differences out-of-sample, and does this hold uniformly or only for particular targets?

We do not introduce a new modelability index—dataset modelability has been formalized by MODI and RMODI [4, 5]—nor a new chirality-aware representation. Our contribution is an audit that (i) separates *carry* (a

deterministic representation property) from *use* (out-of-sample prediction of measured Δ potency by an antisymmetric model related to pairwise-difference learning [6]), (ii) uses only same-assay measured pairs so that within-pair comparisons are internally controlled, (iii) resolves results at the level of individual targets, and (iv) is executed under an analysis protocol frozen and hashed before any held-out modeling, with development targets separated from the confirmatory correction family. All decisions, thresholds, and hyperparameters were fixed in PROTOCOL_FROZEN.json (sha256 d758fc5e...) prior to held-out evaluation.

2. Methods

Data source and inclusion. Bioactivity records were drawn from ChEMBL 37 [9, 10]. Frozen inclusion criteria: standard type IC50, standard relation "=", units nM, a defined pChEMBL value, data_validity_comment null, potential_duplicate excluded, assay type "B" (binding), assay confidence_score = 9, and a single-protein target. These criteria isolate a clean, internally comparable slice at the cost of breadth. The overall workflow, including the prospectively frozen development/held-out separation, is summarized in Fig. 1.

Target eligibility (frozen, data-only). A target entered the analysis only if it satisfied all of: ≥ 80 unique structural enantiomer pairs; ≥ 40 Bemis–Murcko scaffold/series groups [11]; ≥ 5 source documents; and grouped-fold support sufficient for 5-fold scaffold cross-validation (≥ 50 scaffold groups, so each fold contained $\geq \sim 10$ scaffolds and $\geq \sim 15$ pairs). No criterion referenced measured $|\Delta y|$, model performance, or any outcome; target selection was therefore not effect-size enriched. Fourteen targets qualified (confirmatory_manifest.json, which records immutable target identifiers; its bytes are covered by the detached checksum confirmatory_manifest.json.sha256); the complete 14-target characterization is reported in Table 1. Target nomenclature: figures and tables use compact gene-symbol labels; the four targets with common synonyms are ERK2 (MAPK1), ROR- γ (RORC), KCNH2 (hERG/Kv11.1), and PI3K δ (PIK3CD), with full identifiers (ChEMBL target ID, gene symbol, UniProt accession) tabulated in target_metadata_14.csv.

Table 1
— Complete 14-target characterization. Per-target results across all fourteen targets (development targets included).

Source: FINAL_RESULTS.json (family_14).

target	development	n	scaffolds	S_delta	spearman	p	q(q14)	positive	Rt
BACE1	yes	112	59	0.241	0.53	0.011	0.022	YES	4.0
HDAC6	no	83	53	0.237	0.32	0.0015	0.0035	YES	0.67
JAK1	no	207	119	0.172	0.43	0.0005	0.0017	YES	1.33
JAK2	no	199	117	0.16	0.4	0.0005	0.0017	YES	1.07
ERK2	no	102	64	0.149	0.38	0.0005	0.0017	YES	3.18
BTK	yes	266	114	0.142	0.48	0.001	0.0028	YES	1.34
ROR-γ	no	201	119	0.087	0.28	0.0005	0.0017	YES	1.67
EGFR	no	82	64	0.038	0.17	0.032	0.0498	YES	0.82
TYK2	no	95	51	0.021	0.22	0.063	0.0801		1.07
IRAK4	no	172	82	0.015	0.22	0.063	0.0801		1.02
hERG	no	129	95	-0.007	0.09	0.0305	0.0498		0.47
PI3Kδ	no	98	57	-0.026	0.32	0.3198	0.3444		0.71
JAK3	yes	139	78	-0.039	0.09	0.1194	0.1393		0.42
SYK	no	82	52	-0.041	0.13	0.4698	0.4698		0.73

Enantiomer pairing and label construction. Structures were reduced to their largest organic fragment (salt/solvent removal) and re-paired on the parent, using RDKit 2026.03.3 [12]. True enantiomers were identified via the InChI /m-flag construction (InChI software 1.07.3) [13] (identical connectivity and relative-configuration layers, opposite absolute configuration), retaining the /t layer so that diastereomers and double-bond stereoisomers are excluded, and excluding isotopically labeled records. Both members were required to be fully specified single stereoisomers (RDKit FindPotentialStereo, no unspecified elements). Labels were built in this order: (1) collapse repeated measurements of the same compound within a single assay to their median pChEMBL; (2) within each shared assay, compute the enantiomer difference $\Delta y = \text{pChEMBL(A)} - \text{pChEMBL(B)}$; (3) aggregate the assay-level Δy values for the same target–pair by their median; (4) retain one Δy label per target–pair. This yielded **1,967 structural pairs across 14 targets** (cohort_pairs.csv, sha256 de0f79af...).

Pairing validity. A stratified manual audit of 200 pairs, oversampling multi-stereocenter, macrocyclic, salt-stripped, and extreme/near-zero- Δy cases, found **0 pairing errors (95% Wilson CI [0%, 1.9%])**.

Measurement variability (approximate secondary reference). Repeated same-assay IC50 records were provenance-classified. Of 17,676 repeated (compound, assay) groups, **15.2% (2,688) were provenance-distinct, nonidentical repeated-measurement groups** (distinct source records and distinct values); **72.2% were non-identical but shared a source record** and are not independent. From the provenance-distinct groups, the robust

global single-measurement standard deviation was 0.318 log units (median), giving $\sigma(\Delta y) \approx 0.45$ as an **approximate secondary reference**; this value assumes measurement independence and a specific paired-error correlation and is used only for secondary analyses, never to filter the primary data. Per-target replication was too sparse for stratified error models (median 2 groups per target), so a single robust global value was used.

Representation. Primary: Morgan fingerprint, radius 3, 4,096 bits, binary, chirality enabled. Radius 2 and count fingerprints were prespecified sensitivity checks; the achiral radius-3 fingerprint was the exact collision-floor control.

Task and models. For each pair we formed the signed difference $\delta = \varphi(A) - \varphi(B)$. Pair orientation is not available at deployment, so orientation was randomized per pair (seed 7); because the model is exactly antisymmetric, real-metric values are invariant to this choice while the null is correctly centered. Predictions used the symmetrized estimator $\hat{\Delta}(\delta) = [h(\delta) - h(-\delta)]/2$, guaranteeing $\hat{\Delta}(B,A) = -\hat{\Delta}(A,B)$ exactly (verified: $\max |f(A,B) + f(B,A)| = 4 \times 10^{-16}$), with orientation augmentation kept within a fold. **Separate target-specific models were fitted; no labels were pooled across targets.** Base learners were a random forest (80 trees, minimum leaf 3) and ridge regression ($\alpha = 10$) on δ ; both hyperparameter sets were frozen a priori from the three-target development cohort and were not tuned on confirmatory metrics.

Splits. Because each target-pair contributes exactly one aggregated label, the modeling unit is the pair. Evaluation used scaffold-grouped 5-fold cross-validation (extrapolation to unseen Bemis-Murcko scaffolds) as primary, with connectivity-grouped interpolation as secondary. Both orientations of a pair (the augmentation copies) were assigned to the same fold; the source records underlying a pair's label had already been aggregated at label construction.

Statistics. The single primary statistic was the normalized zero- Δ skill score $S_{\Delta} = 1 - \text{MAE}(\hat{\Delta}, \Delta y) / \text{MAE}(0, \Delta y)$. Secondary metrics were Spearman($\hat{\Delta}, \Delta y$) and balanced directional accuracy (exact ties, $|\Delta y| \leq 10^{-9}$, excluded from the directional metric only and retained in S_{Δ} and Spearman). Target-level significance used a **scaffold-cluster-level sign-flip null** (2,000 repetitions per target, magnitude preserving) as primary and a **pair-level label-permutation null** (1,000 repetitions) as a robustness check, with full model refitting inside each repetition and empirical p-values. Benjamini-Hochberg false-discovery-rate (FDR) correction [14] was applied across targets. Because BH depends on the hypothesis family, the three development targets (BACE1, BTK, JAK3—used for pipeline construction, null calibration, and hypothesis formation) were excluded from the confirmatory correction family, and BH was recomputed within the eleven held-out targets.

Forest-plot confidence intervals. Per-target S_{Δ} 95% confidence intervals (held-out forest plot) were computed by a scaffold-cluster bootstrap: scaffold clusters were resampled with replacement (the target's cluster set, matched count), the pairs belonging to the resampled clusters were pooled, S_{Δ} was recomputed on the fixed out-of-fold predictions, and the 2.5/97.5 percentiles over 2,000 resamples were reported.

Descriptive signal-to-variability index. A per-target ratio $R_t = \text{median}|\Delta y| / 0.45$ was specified in the frozen protocol; its association with skill across targets was evaluated descriptively.

Reproducibility. All artifacts and random seeds are hashed in FROZEN_ARTIFACTS.json (sha256 0bec1c7d...); null distributions (mean, quantiles, empirical p) are exactly regenerable from the recorded seeds (orientation 7, nulls 202) and cohort_pairs.csv.

3. Results

3.1 In the audited targets, radius-2 collisions were radius-limited, not hash-induced. Enantiomer "collisions" (identical feature vectors) under chirality-aware fingerprints were *identical* across 1,024, 2,048, 4,096, and 8,192 bits and under unhashed Morgan identifiers, excluding finite hashing. They depended on radius: the collision rate fell from 17.8% to 5.1% (BACE1), 3.5% to 0% (BTK), and 1.3% to 0% (JAK3) between radius 2 and 3 (Fig. 2). Among BACE1's radius-2 collisions, 15 of 21 resolved at radius 3 (radius-limited), 6 of 21 remained identical even with unhashed radius-3 identifiers (a genuine encoding limitation, ~ 5% of pairs), and none were hashing collisions. Carry is near-complete at radius 3; all collision percentages are configuration-specific and refer to the audited targets.

3.2 Controls behave as required. The symmetrized estimator was exactly antisymmetric (4×10^{-16}). The achiral representation produced $\delta \equiv 0$ for every enantiomer pair, forcing $\Delta \equiv 0$ and making S_{Δ} exactly equal to the zero- Δ baseline—an operational collision floor. Sign-flip nulls centered slightly below zero (mean - 0.041, range - 0.077 to - 0.017), as expected for a normalized skill score under scrambled labels; centering was uncorrelated with cluster imbalance ($\rho = -0.07$) or target size ($\rho = -0.26$), and empirical p-values account for it directly.

3.3 Held-out confirmation (primary). With the three development targets removed from the correction family and BH recomputed at $m = 11$, **five of eleven held-out targets showed FDR-significant positive S_{Δ} under scaffold extrapolation: JAK1, JAK2, ERK2 (MAPK1), ROR- γ (RORC), and HDAC6** ($q \leq 0.0033$; Table 2; Fig. 3). EGFR did not reach significance in the held-out family ($q_{\text{held-out}} = 0.0503$) and is treated as a boundary case (Section 3.4).

Table 2

— **Held-out confirmation (m = 11).** Benjamini–Hochberg recomputed within the eleven held-out targets; five positive.

Source: FINAL_RESULTS.json (family_11_heldout).

target	development	n	scaffolds	S_delta	spearman	p	q(q11)	positive	Rt
HDAC6	no	83	53	0.237	0.32	0.0015	0.0033	YES	0.67
JAK1	no	207	119	0.172	0.43	0.0005	0.0014	YES	1.33
JAK2	no	199	117	0.16	0.4	0.0005	0.0014	YES	1.07
ERK2	no	102	64	0.149	0.38	0.0005	0.0014	YES	3.18
ROR- γ	no	201	119	0.087	0.28	0.0005	0.0014	YES	1.67
EGFR	no	82	64	0.038	0.17	0.032	0.0503		0.82
TYK2	no	95	51	0.021	0.22	0.063	0.077		1.07
IRAK4	no	172	82	0.015	0.22	0.063	0.077		1.02
hERG	no	129	95	-0.007	0.09	0.0305	0.0503		0.47
PI3K δ	no	98	57	-0.026	0.32	0.3198	0.3518		0.71
SYK	no	82	52	-0.041	0.13	0.4698	0.4698		0.73

3.4 Complete-map characterization and the EGFR boundary. Across all fourteen targets (development targets included), eight were positive: the five above plus BACE1, BTK, and EGFR (Table 1). EGFR crossed the threshold only in the combined family ($q_{\text{combined}} = 0.0498$) and not in the held-out family ($q_{\text{held-out}} = 0.0503$); this correction-family dependence is reported as a transparent boundary sensitivity, not as held-out positive evidence.

3.5 Skill is strongly target-dependent. Target S_{Δ} ranged from -0.041 to $+0.241$. Pair and scaffold counts showed no strong descriptive association with skill in this fourteen-target sample (S_{Δ} vs number of pairs $\rho = 0.26$; vs scaffolds $\rho = 0.24$), and the negative association with tie fraction ($\rho = -0.40$) is mechanistic, since ties carry no directional signal.

3.6 Ridge and random forest performed comparably. Across targets, the paired difference $S_{\Delta}(\text{ridge}) - S_{\Delta}(\text{RF})$ had median $+0.005$ (95% CI $[-0.032, +0.019]$; Wilcoxon $p = 0.76$; Fig. 4), consistent with much of the accessible stereochemical signal being additive in signed-fingerprint space; we do not claim formal statistical equivalence.

3.7 The signal-to-variability trend was suggestive but not confirmed. Across all fourteen targets, R_t correlated with skill ($\rho = 0.62$, $p = 0.018$ for S_{Δ} ; $\rho = 0.73$, $p = 0.003$ for Spearman). In the eleven held-out targets this weakened and did not reach significance (S_{Δ} : $\rho = 0.42$, $p = 0.20$; Spearman: $\rho = 0.53$, $p = 0.095$), and leave-one-out analysis showed the full-map correlation dropped to $+0.53$ when the highest-leverage development target (BACE1) was removed. We report the relationship as a descriptive, unconfirmed trend (Fig. 5).

3.8 A post-hoc counterexample. HDAC6 was FDR-positive despite a low median $|\Delta y|$ (0.30 ; $R_t = 0.67$): 43% of its pairs exceeded 0.45 log units and 28% exceeded 1.0 (upper quartile 1.18 , maximum 3.27). A median-only signal summary understates such a heavy upper tail, offering a post-hoc explanation for why the median-based R_t under-predicts skill for this target and a plausible reason the held-out association is weak.

3.9 The method does not manufacture predictability. Low-signal targets (e.g. JAK3) and the achiral control returned null or baseline skill, and four targets returned negative S_{Δ} ; in particular, KCNH2/hERG ($S_{\Delta} = -0.007$) is not a positive predictive target despite a low unadjusted p -value.

4. Discussion

Separating carry from use clarifies the interpretation. Because chirality is essentially fully carried at radius 3, the meaningful scientific quantity is use, and use is real but conditional: a subset of targets (five of eleven held-out; eight of fourteen overall) supports out-of-sample prediction of measured enantioselectivity under scaffold extrapolation, while others do not. The comparability of ridge and random forest suggests that much of the accessible component may be additive in signed-fingerprint space, which is encouraging for interpretability and simplicity; it does not establish formal statistical equivalence, nor should it be read as evidence that nonlinearity is never useful.

The most tempting narrative—that predictive use is governed by whether enantioselectivity exceeds measurement variability—was suggestive across the full map but did not confirm on held-out targets, and HDAC6 provides a concrete mechanism for why a median-based signal statistic is too coarse. This is the appropriately cautious conclusion: a hypothesis worth testing at larger scale with a tail-aware signal measure, not an established law. Future work should extend to more targets, adopt a tail-sensitive signal statistic, validate on an independent resource, and evaluate chirality-aware architectures [7, 8] on the same measured benchmark.

5. Limitations

Only fourteen targets met the frozen data-quality bar, so the held-out family (eleven) is small and the signal-to-variability test is underpowered. **The eligibility criteria enrich for data-rich, well-studied targets, so the qualifying set is not representative of all proteins;** results should not be extrapolated beyond comparable targets. **Assay-level Δy medians aggregate across distinct assays for a given pair, which may blur genuine assay-to-assay differences.** The variability reference derives from provenance-distinct, not provably independent, repeats and is used only secondarily. The analysis is restricted to IC50, confidence-9, single-protein, binding data—clean but narrow—and **was not validated against an independent external bioactivity database.** No experimental validation was performed; predictions concern *measured* potency differences, not causal efficacy. EGFR's positivity is correction-family dependent. Full null arrays were not stored (only summaries), though they are exactly regenerable from the recorded seeds.

6. Conclusions

Chirality-aware radius-3 Morgan fingerprints carry stereochemistry almost completely, so the scientifically meaningful question is use rather than representational capacity. Under a prospectively frozen, scaffold-extrapolation protocol, that use is real but conditional: five of eleven held-out targets—eight of fourteen across the complete map—supported significant out-of-sample prediction of measured enantiomer potency differences, ridge and random-forest models performed comparably (indicating that much of the accessible signal is additive), and achiral controls and several targets returned null or negative skill. The elevated low-radius collision rate was a neighborhood-radius effect, not a hashing artifact. The signal-to-variability relationship was suggestive across the full map but did not confirm on the held-out targets; it remains a tail-aware hypothesis for larger, externally validated testing rather than an established law.

Abbreviations

BH
Benjamini–Hochberg
CI
confidence interval
ECFP
extended-connectivity fingerprint
FDR
false discovery rate
IC50
half-maximal inhibitory concentration
InChI
IUPAC International Chemical Identifier
MAE
mean absolute error
QSAR
quantitative structure–activity relationship
RF

random forest

R_t

signal-to-variability index ($\text{median}|\Delta y| / 0.45$)

S_Δ

normalized zero- Δ skill score.

Declarations

Ethics approval and consent to participate

Not applicable. This is a computational study using publicly available bioactivity data (ChEMBL 37); it involves no human participants, human data, or animals.

Consent for publication

Not applicable.

Availability of data and materials

All analysis code, frozen result artifacts, target metadata, and checksum files are openly available in the project repository <https://github.com/gnarlylace51842/eudysmic-ratio-map>, with a persistent archive deposited on Zenodo (v1.0.1: <https://doi.org/10.5281/zenodo.21478868>). The repository release corresponding to this manuscript is tag **v1.0.1**. Every shipped file's SHA-256 is listed in HASH_MANIFEST.txt; the prospectively specified, cryptographically frozen protocol and artifact hashes are in PROTOCOL_FROZEN.json and FROZEN_ARTIFACTS.json; immutable target identities are in target_metadata_14.csv/.json. Two reproduction pathways are documented in REPRODUCTION.md: a **fast path** that regenerates every figure, table, and reported number from the shipped processed data (cohort_pairs.csv and cohort_results.json) in minutes on a standard workstation, and a **full path** that re-derives the cohort from the external ChEMBL 37 database. Empirical null distributions are exactly regenerable from the recorded random seeds (orientation 7, nulls 202) and cohort_pairs.csv. The analysis code is released under the MIT License; processed data derive from ChEMBL 37 (EMBL-EBI; dataset doi:10.6019/CHEMBL.database.37) and are redistributed under CC BY-SA 3.0. The complete ChEMBL 37 SQLite database (~28 GB) is not redistributed but is publicly available from EMBL-EBI.

Competing interests

The author declares that they have no competing interests.

Funding

No funding was received for this work. The research was conducted independently by the author; the listed affiliation is the author's school, which did not fund, supervise, or institutionally endorse the study.

Authors' contributions

D.A. is the sole author: designed the study, developed the frozen protocol and analysis code, performed all analyses, and wrote the manuscript. The author read and approved the final manuscript.

Acknowledgements

The author thanks the EMBL-EBI ChEMBL team for maintaining the ChEMBL database and the RDKit open-source community.

Use of large language models

During this work the author used Anthropic's Claude (accessed through the Claude Code command-line interface; Claude Opus 4.x-class models) as an AI assistant. Claude assisted with software development and debugging of the analysis pipeline; methodological refinement and audit design (including null-model calibration, the development/held-out separation, and stress-testing of the signal-to-variability analysis); execution of the analyses and reproducibility checks; verification and formatting of references; correction of target-identity metadata; structuring and drafting of the manuscript and supplement; document conversion; and assembly of the reproducibility package. The author defined the research question and framing, made and approved all scientific decisions and claims, and takes full responsibility for the content and integrity of the manuscript. The AI tool is not an author and bears no responsibility for the work.

References

1. Rogers D, Hahn M, Extended-Connectivity Fingerprints (2010) J Chem Inf Model 50(5):742–754. 10.1021/ci100050t
2. Schneider N, Lewis RA, Fechner N, Ertl P (2018) Chiral Cliffs: Investigating the Influence of Chirality on Binding Affinity. ChemMedChem. ;13(13):1315–1324. 10.1002/cmdc.201700798
3. Comajuncosa-Creus A, Lenes A, Sánchez-Palomino M, Dalton D, Aloy P (2024) Stereochemically-aware bioactivity descriptors for uncharacterized chemical compounds. J Cheminform 16:70. 10.1186/s13321-024-00867-4
4. Golbraikh A, Muratov E, Fourches D, Tropsha A (2014) Data Set Modelability by QSAR. J Chem Inf Model 54(1):1–4. 10.1021/ci400572x
5. Luque Ruiz I, Gómez-Nieto MÁ (2018) Regression Modelability Index: A New Index for Prediction of the Modelability of Data Sets in the Development of QSAR Regression Models. J Chem Inf Model 58(10):2069–2084. 10.1021/acs.jcim.8b00313
6. Tynes M, Gao W, Burrill DJ, Batista ER, Perez D, Yang P, Lubbers N (2021) Pairwise Difference Regression: A Machine Learning Meta-algorithm for Improved Prediction and Uncertainty Quantification in Chemical Search. J Chem Inf Model 61(8):3846–3857. 10.1021/acs.jcim.1c00670
7. Adams K, Pattanaik L, Coley CW Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations. In: International Conference on Learning Representations (ICLR); 2022 (poster;

OpenReview id hm2tNDdgaFK). Peer-reviewed conference paper (no publisher DOI); preprint 10.48550/arXiv.2110.04383

8. Gaiński P, Koziarski M, Tabor J, Śmieja M (2023) ChiENN: Embracing Molecular Chirality with Graph Neural Networks. In: Machine Learning and Knowledge Discovery in Databases: Research Track (ECML PKDD 2023). Lect Notes Comput Sci 14171:36–52. 10.1007/978-3-031-43418-1_3
9. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magariños MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR (2024) The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52(D1):D1180–D1192. 10.1093/nar/gkad1004(Describes ChEMBL release 33.)
10. ChEMBL Team EMBL-EBI (May 2026) ChEMBL Database, Release 37. 10.6019/CHEMBL.database.37. (Dataset used in this study.)
11. Bemis GW, Murcko MA (1996) The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem 39(15):2887–2893. 10.1021/jm9602928
12. Landrum G et al RDKit: Open-source cheminformatics software, version 2026.03.3. Zenodo. 10.5281/zenodo.20446949 (version 2026.03.3; concept DOI 10.5281/zenodo.591637)
13. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC International Chemical Identifier. J Cheminform 7:23. 10.1186/s13321-015-0068-4
14. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc B 57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x

Figures

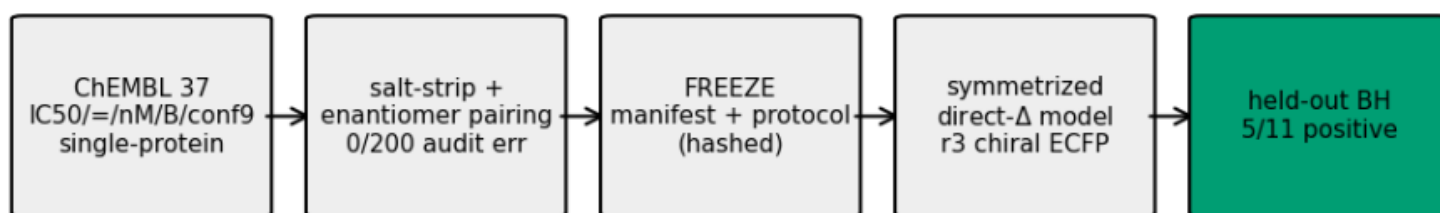


Figure 1

Study workflow. Pipeline from ChEMBL 37 extraction and same-assay enantiomer pairing through frozen modeling, with explicit separation of the three development targets from the eleven held-out confirmatory targets.

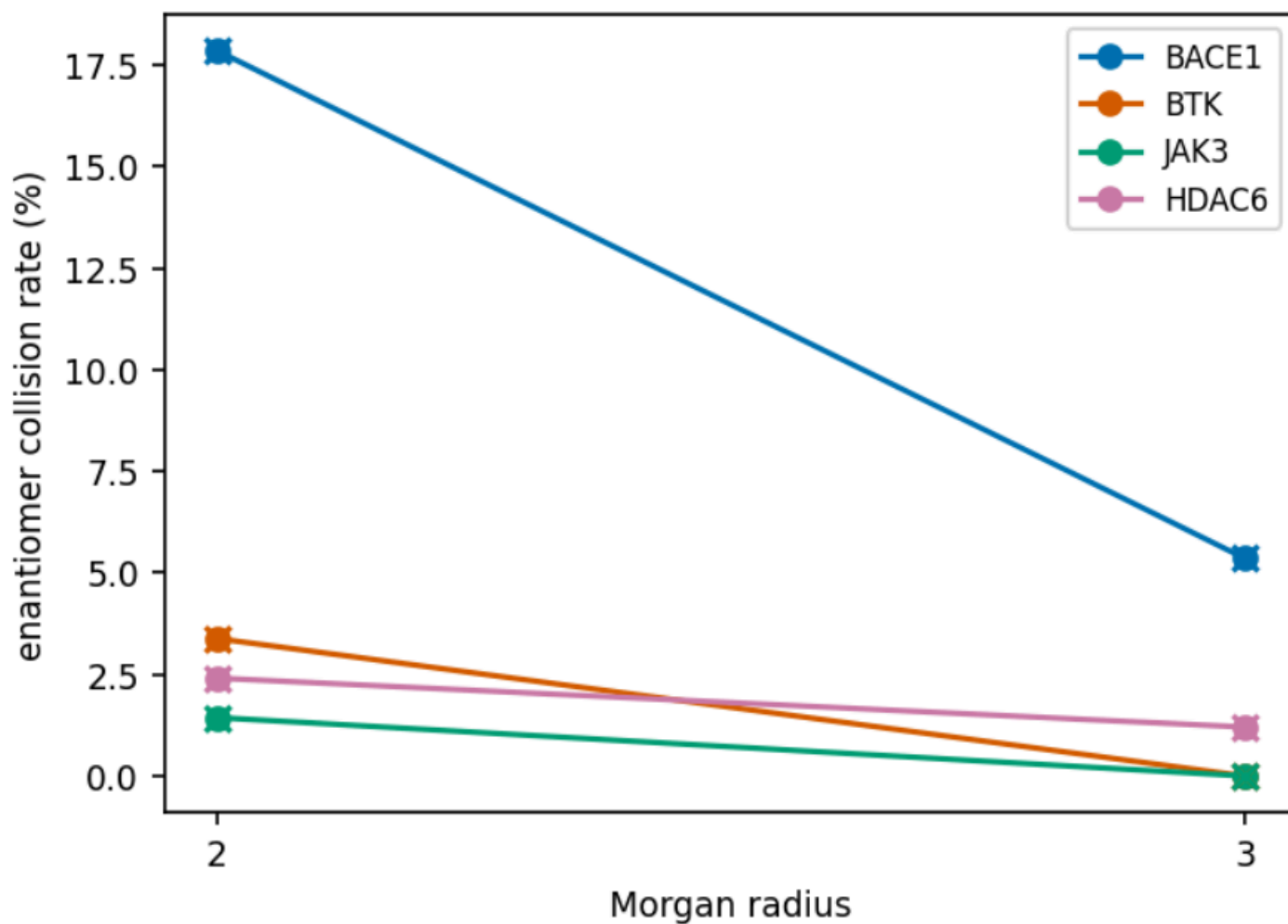


Figure 2

Fingerprint collisions are a radius effect, not a hashing artifact. Enantiomer collision rate versus Morgan radius and fingerprint configuration; binary, count, and unhashed identifiers coincide across 1,024–8,192 bits, and the rate falls sharply from radius 2 to radius 3.

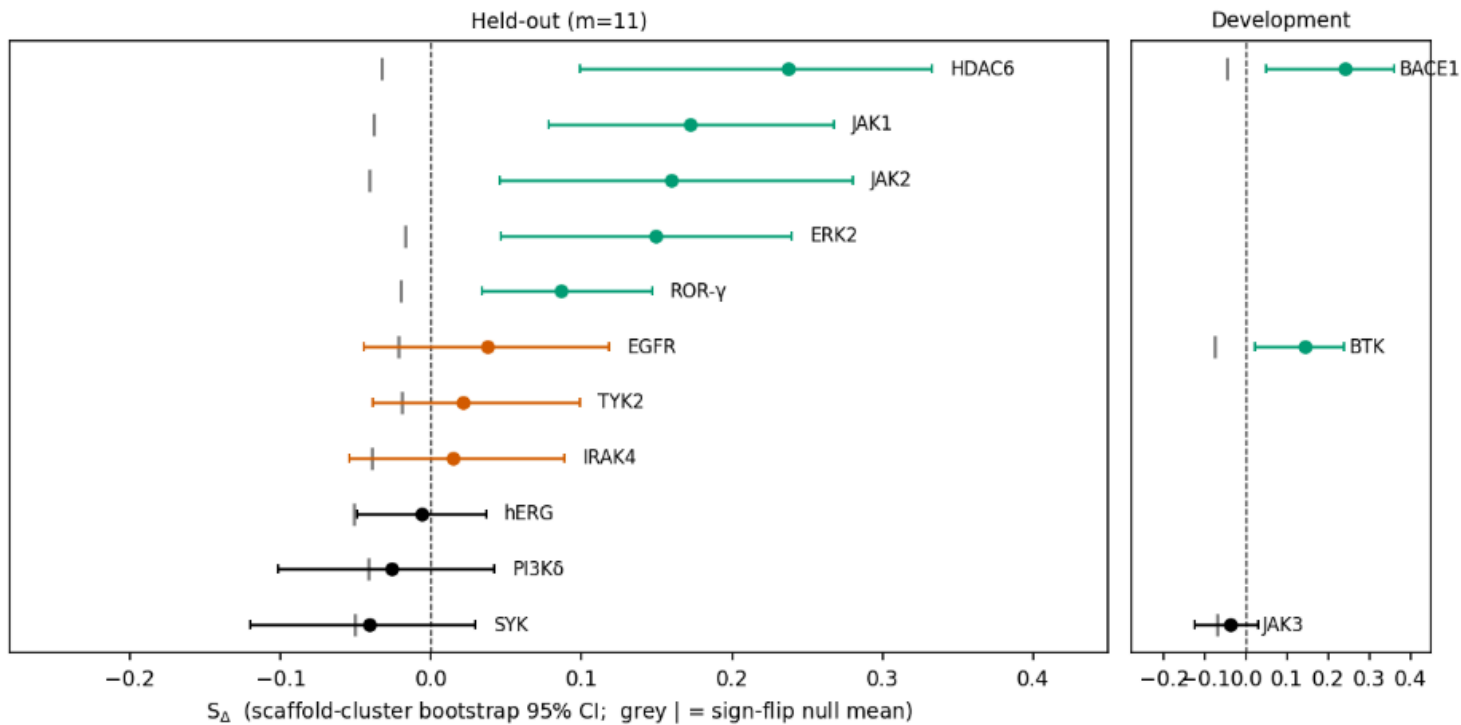


Figure 3

Held-out target forest plot. Per-target S_{Δ} under scaffold extrapolation with scaffold-cluster-bootstrap 95% confidence intervals and sign-flip null markers; the three development targets are shown in a separate panel.

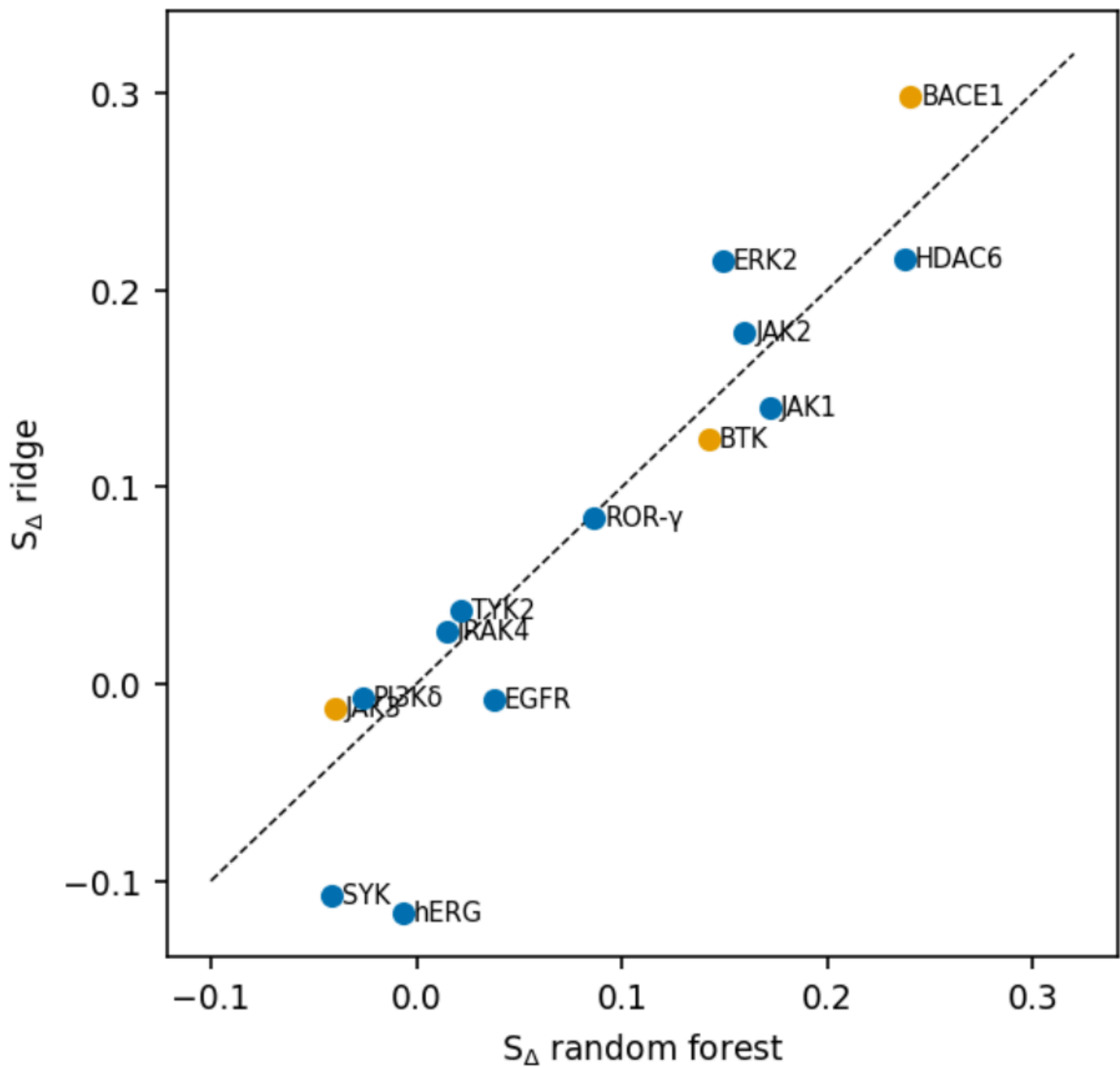


Figure 4

Ridge versus random forest. Paired per-target comparison of S_{Δ} for ridge regression and the random forest, showing comparable observed skill.

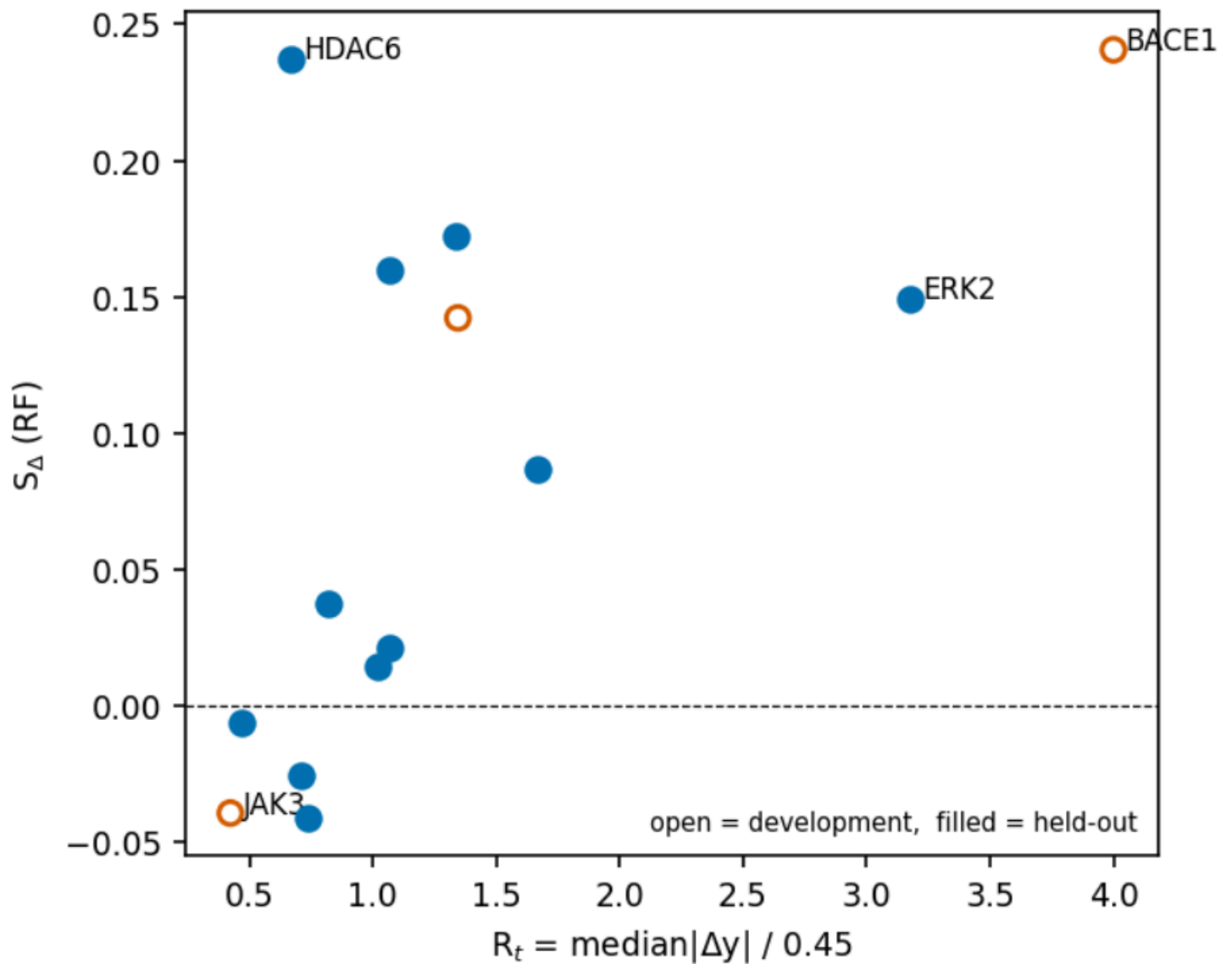


Figure 5

Signal-to-variability versus skill. Scatter of the per-target index R_t against S_Δ , distinguishing development and held-out targets, with HDAC6 annotated as the heavy-upper-tail counterexample. Null distributions and extended robustness analyses are in the Supplement.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ASHRAFSUPPLEMENT.docx](#)