Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Graph Neural Network and Fingerprint Attributions of Antibacterial Activity: Atom-Level Agreement Varies Across Random Seeds Reference distributions, topology controls and multi-instance replication

Mohamed Saber
ITMO University
Omar Elazouni
ITMO University
Aya Darwish
ITMO University
Malikakhon Kurbonova
ITMO University
Mohamed Said Boulkrane
mboulkrane@itmo.ru

ITMO University

Research Article

Keywords: reproducibility, random seed, explainable artificial intelligence, molecular machine learning, attribution agreement, SHAP, graph neural networks, attention, antibacterial virtual screening, scaffold splitting, false discovery rate

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11390965/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Graph Neural Network and Fingerprint Attributions of Antibacterial Activity: Atom-Level Agreement Varies Across Random Seeds

Reference distributions, topology controls and multi-instance replication

Mohamed Saber Ahmad[1], Omar Elazouni[2], Aya Darwich [1], Malikakhon Kurbonova[1], Boulkrane Mohamed Said[1*]

[1] *Faculty of Biotechnologies, ITMO University, Lomonosova St. 9, St. Petersburg, 191002, Russia*
[2] *Faculty of Control Systems and Robotics, ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia*

[*]Correspondence author: mboulkrane@itmo.ru

Abstract

ML models with different inductive biases are routinely used to prioritise compounds for antibacterial screening, and attribution methods are used to explain them. Circular fingerprints consumed by tree ensembles and message-passing graph networks share neither featurisation nor optimisation, so agreement between their explanations would carry information that neither method alone provides. Whether two independently derived attribution schemes identify the same atoms of the same molecule, and whether such agreement survives retraining, has received little quantitative examination.

We curated a published *Staphylococcus aureus* screen (39,312 compounds, 512 actives) to 38,748 compounds by stereochemistry-aware deduplication and partitioned it by Bemis–Murcko scaffold into training (26,345; 123 actives), validation (4651; 31), and held-out (7752; 302; 3.90%) sets. Fingerprint Random Forest reached a held-out AUPRC of 0.1233 fitted on the training portion alone, and 0.1518 when refitted on training plus validation, with a 9.75-fold enrichment in its top 50; eight AttentiveFP networks differing only in random seed reached a median of 0.0834 (range 0.0587–0.0949). TreeSHAP attributions were verified for additivity (error 3.69×10^{-7}) and projected onto atoms through Morgan bit environments with a coverage normalisation that reduced their correlation with atom degree from +0.268 to +0.069; readout attention was obtained by a forward-pass replication verified against the model's own output (logit error 1.91×10^{-6}). Across 60 molecules the median per-molecule Spearman correlation between the two maps was +0.0890, exceeding a size-matched cross-molecule reference in 6 of 8 instances. The agreement is not topological: the attention–degree association varies in sign across instances (−0.2562 to +0.1497) while the |SHAP|–degree association is invariant at +0.0439. Across instances, the agreement spans −0.1067 to +0.1547, a range including both signs, and no prespecified antibiotic core survived Benjamini–Hochberg correction under either scheme.

Cross-model atom-level attribution agreement in this low-data regime is small, exceeds a demanding reference in most but not all training runs, and is not attributable to atom connectivity. Because its across-instance range includes both signs, a single training run can misstate both the magnitude and the direction of the agreement it reports. Attribution comparisons in comparable data regimes should be reported as distributions over independently trained instances rather than as point estimates.

Keywords: reproducibility; random seed; explainable artificial intelligence; molecular machine learning; attribution agreement; SHAP; graph neural networks; attention; antibacterial virtual screening; scaffold splitting; false discovery rate

Introduction

Antibacterial discovery is constrained by the low hit rate of physical screening. Models trained on primary screening data are now routinely used to reorder a library so that a fixed experimental budget is spent on a more promising fraction of it, and several prospective campaigns have reported experimentally confirmed antibacterials found this way [1–3]. The practical figure of merit is therefore early enrichment rather than global discrimination.

Models built on different molecular representations encode different inductive biases. Circular fingerprints enumerate local atom environments and are consumed by tree ensembles that partition a fixed,

sparse feature space [4–6]. Message-passing graph networks learn representations directly from the molecular graph and aggregate them with a learned readout [7–9]. Because the two families share neither featurization nor optimization, agreement between them is informative. Concordant behaviour is unlikely to arise from a shared representational artefact, and discordant behaviour marks where one family sees structure the other does not.

Explainability methods are widely applied to such models, most often SHAP for tree ensembles [10,11] and attention for graph networks [7]. Reviews nevertheless note that these quantities are model-specific and that their chemical validity is rarely tested [12,13], whether attention constitutes explanation is contested [14,15], and attribution methods differ substantially in fidelity even where ground truth exists [16]. Agreement between two independently derived attribution schemes offers a partial answer. If a fingerprint model and a graph model concentrate weight on the same atoms of the same molecule, the shared signal is at least not an artefact of either method alone. Three requirements make such a comparison interpretable, and they motivate the design used here.

The first is topology control. Both quantities are defined on molecular graphs and can track connectivity rather than chemistry. Circular environments are radial, so a high-degree atom accumulates attributional mass for purely structural reasons, and graph attention propagates over bonds and is similarly exposed.

The second is an appropriate reference. Comparing an observed correlation against zero presumes that unrelated attribution vectors are uncorrelated, which is false for smooth functions on molecules of similar size.

The third is replication. When a graph network is fitted to a few hundred actives and selected on a validation slice holding tens, the fitted parameters, and hence the attention maps, are not determined reproducibly. A single run then yields one draw from a distribution of unknown width.

We address all three together on a published antibacterial screen [1] and ask: (i) do a fingerprint model and a graph model rank the same compounds; (ii) within a shared molecule, do their attribution maps agree more than a size-matched cross-molecule reference allows; (iii) is any agreement separable from atom connectivity; (iv) does it reproduce across independently trained instances; and (v) does it extend to enrichment of recognised antibiotic cores under false-discovery control? We report the answers as measured, including where they are negative.

Contributions and Relation to Prior Work

Prior work on molecular attribution has advanced along three largely separate lines: benchmarking attribution methods against synthetic or curated ground truth [16,17]; assessing the internal consistency of a single attribution formalism across its own computational variants [18]; and qualitatively relating atom-level colourings to known pharmacophores [19]. What has not been established is whether two attribution schemes derived from model families that share no featurization, no architecture, and no training objective concentrate weight on the same atoms of the same molecule, and whether such a correspondence is a stable property of the modelling setup or an artefact of a single training run. This work addresses that question under an explicit reference distribution, an explicit topological control, and replication across independently trained model instances. Its five contributions are as follows.

1. Cross-family atom-level comparison on a common per-atom axis. Morgan bits index circular environments, not atoms, so fingerprint attributions and graph attention are not directly commensurable. We place both on a per-atom axis by distributing each bit's Shapley value across the atoms of the environments that set it and normalizing it by environment coverage (Equation (2)). The correction targets a bias intrinsic to the projection, and we measure its magnitude and its residual rather than assume them (Section 2.3).

2. A size-matched cross-molecule reference distribution. The reference pairs the attention map of one molecule with the fingerprint-attribution map of a different molecule of identical atom count (Equation (3)). Both sides therefore remain genuine attribution vectors of real molecules with identical dimension-

ality, and whatever generic structure the two schemes share is already contained in the reference. Exceeding it is evidence of molecule-specific correspondence rather than of shape (Section 2.4).

3. A sign-based connectivity control. In some instances, the two schemes' associations with atom degree point in opposite directions. A shared topological artefact cannot generate positive agreement between them in those instances, because connectivity would drive the two maps apart. This argument is stronger than a comparison of magnitudes, which would depend on where a threshold is placed (Section 2.5).

4. Attribution agreement reported as a distribution rather than a point estimate. Every graph-model-dependent quantity is computed over eight instances, differing only in initialization seed, and the resulting distributions, rather than any single run, are what the study reports. This is the principal contribution. In this data regime the across-instance range of the agreement includes both signs, and attention-based structural-core enrichment spans two to three orders of magnitude, so a single run can misstate both the magnitude and the direction of the quantity being claimed (Sections 2.6 and 2.7).

5. Substructure enrichment under a connected-subgraph reference with false-discovery control. Enrichment of prespecified antibiotic cores is tested against 2000 randomly grown *connected* subgraphs of equal atom count within the same molecule (Equation (4)). The connectivity requirement is what makes the test informative, because a reference of scattered atoms is exceeded trivially by any spatially smooth attribution. The outcome is negative and is reported as such (Section 2.7).

Table 1 places these design elements against the closest prior studies. The comparison is one of study design and not of performance, since the works differ in task, data, and endpoint, and their reported metrics are not commensurable. Read across the columns, the position of this study is specific. It is the cross-family case, evaluated where no attribution ground truth exists, and therefore the case in which the reference distribution and the replication protocol, rather than a ground-truth comparison, carry the inference.

Table 1. Design elements of the present study relative to the closest prior work.

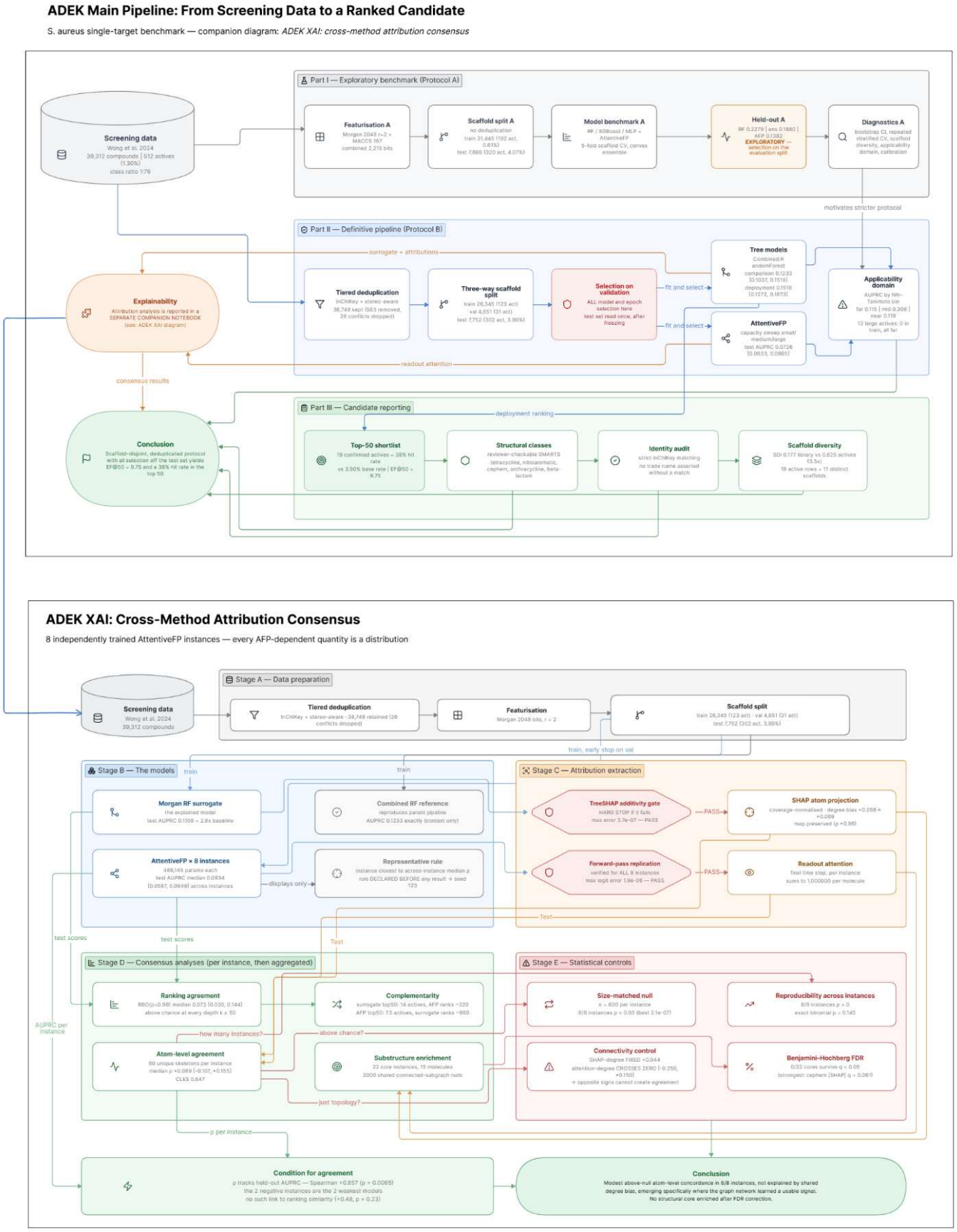
Study	Quantities compared	Level	Explicit reference	Topology control	Multi-run replication	Multiplicity control
Wong et al. 2024 [1]	GNN rationales	Substructure	n/a	n/a	n/a	n/a
Sanchez-Lengeling et al. 2020 [16]	Attribution methods vs. ground truth	Node / edge	Synthetic ground truth	n/a	Multiple trials	n/a
Rao et al. 2022 [17]	XAI methods vs. benchmarks and experts	Substructure	Curated ground truth	n/a	n/a	n/a
Jiménez-Luna et al. 2021 [19]	IG colourings vs. pharmacophores	Atom	n/a	n/a	n/a	n/a
Lamens & Bajorath 2025 [18]	Shapley variants against one another	Feature	n/a	n/a	n/a	n/a
Present work	Fingerprint SHAP vs. graph attention	Atom	Size matched cross-molecule	Sign-based, atom degree	8 instances	Benjamini–Hochberg

Materials and Methods

Software Architecture

The study is implemented as two independent notebooks that share a data definition but no runtime state: a main pipeline (upper panel of Fig. 1) and a self-contained explainability module (lower panel). The explainability module retrieves the data, rebuilds the deduplicated splits with the same code, fits the model it explains, and trains its own graph networks. It requires no cache, variable, or file produced by the main pipeline, so the attribution results can be reproduced without running the ranking pipeline.

Because both modules construct the partition independently, agreement between them is a check rather than an assumption. The Combined Random Forest reference in the explainability module reproduces the main pipeline's comparison-regime AUPRC to four decimal places, and both modules report identical curation counts (39,311 parsed to 38,748 retained; 563 removed across 510 merge groups; 26 label-conflict groups discarded).



tion on validation, held-out evaluation, and candidate reporting. (**Lower panel**) Self-contained explainability module: eight independently trained Attentive FP instances, two verification gates (red), consensus analyses, and statistical controls. Every quantity depending on the graph model is a distribution over the eight instances.

Table 2. Stages of the explainability module and where each is specified.

Stage	Function	Specified in
A. Data preparation	Tiered stereo-aware deduplication, Morgan featurization, scaffold-disjoint three-way split	Sections 4.2, 4.3
B. The models	Morgan RF surrogate, Combined RF reference, eight Attentive FP instances, representative-instance rule	Section 4.4
C. Attribution extraction	Two hard gates, each halting execution on failure, each feeding one attribution scheme	Section 4.5
D. Consensus analyses	Ranking agreement, complementarity, atom-level agreement, substructure enrichment	Sections 4.7, 4.8, 4.9
E. Statistical controls	Size-matched cross-molecule reference, replication across instances, connectivity control, Benjamini–Hochberg correction	Sections 4.8, 4.9, 4.12

Dataset and Curation

Primary screening data for *Staphylococcus aureus* were taken from the archived repository accompanying Wong et al. [1], comprising 39,312 compounds with binary growth-inhibition labels, of which 512 are active (1.30%, a class ratio near 1:75.8). Only this single-target label is used; the cytotoxicity data of the source study are not.

Structures were parsed with RDKit [30]. One SMILES string failed, leaving 39,311, and the largest fragment was retained from multi-fragment records. Deduplication applied a three-tier, stereochemistry-aware rule on InChIKeys:

1. Records sharing a full InChIKey were merged.
2. Records sharing a stereochemistry-free skeleton, in which at least one member had no defined stereocenters and the stereo-defined members resolved to a single InChIKey, were merged, treating the unannotated record as incompletely specified.
3. Records sharing a skeleton but carrying distinct, fully defined stereochemistry were retained (28,755 rows), because two-dimensional fingerprints cannot distinguish them and collapsing them would discard genuine chemical entities.

Groups with conflicting labels were discarded entirely rather than resolved by majority vote. The procedure removed 563 records across 510 merge groups and discarded 26 conflicting groups, leaving 38,748 compounds.

Representation and Scaffold-Disjoint Splitting

Morgan circular fingerprints at radius 2 folded to 2048 bits [4] and the 167-bit MACCS key set were computed; their concatenation (2215 bits) is the Combined representation, used for the deployed predictive model. The Morgan representation alone is used for the surrogate that undergoes attribution analysis, because Morgan bits carry an explicit mapping from bit to atom environment whereas MACCS keys do not.

Compounds were grouped by Bemis–Murcko scaffold [31] computed without stereochemistry, and whole scaffold groups were assigned atomically so that no scaffold appears in more than one split. Groups were accumulated into the held-out set in order of increasing size until 20% of compounds were allocated. A scaffold-disjoint validation slice of 15% of the remainder was then carved from the training portion by the same rule (Table 4).

The held-out set carries a higher positive rate (3.90%) than the training portion (0.47%) because actives in this library are more scaffold-diverse than inactive and therefore concentrate on the small scaffold groups routed to it. This is a deliberate extrapolation design. No AUPRC reported in this work is comparable across splits without its accompanying base rate; enrichment relative to the base rate of the set on which it was computed is the comparable quantity.

Two consequences are tracked explicitly. The training portion retains 123 actives, which is the binding constraint on the graph model. And the training portion contains no active compound of 70 or more heavy atoms, whereas the held-out set contains 13. This stratum was declared in the pipeline configuration before any result was inspected and is analyzed separately (Section 4.10).

Models and Multi-Instance Training

1. Random Forests: Random Forests [5] with 400 trees were fitted using scikit-learn [32]. The deployed model uses the Combined representation and is reported in two declared regimes: a comparison regime fitted on the training portion only, matching the data seen by the graph network, and a deployment regime refitted on training plus validation. The representation and learner were selected on the validation split against Morgan-only and gradient-boosted alternatives [6]. The surrogate uses Morgan features alone and is fitted on the training portion. Class weighting was deliberately omitted because it reduces AUPRC in this regime and invalidates the additivity guarantee of TreeSHAP.

2. Graph model: The graph model is Attentive FP [7] in PyTorch Geometric [33] on PyTorch [34], with hidden dimension 128, three message-passing layers, two readout time steps, and dropout 0.2, totaling 486,145 parameters. This capacity was selected on the validation split against alternatives of 24,993 and 95,041 parameters. Atom and bond features (41 and 7 dimensions) are derived from the data rather than fixed a priori.

Molecular graphs are built from RDKit molecule objects rather than from a canonical SMILES round trip, which guarantees that graph node i corresponds to RDKit atom i, and therefore that attention indices, atom-projected SHAP indices, and SMARTS match indices are directly comparable. This alignment is a precondition for every analysis in Sections 4.6 to 4.9.

Training used Adam at learning rate 3×10^{-4}, weight decay 10^{-5}, batch 128, cosine annealing over a 60-epoch budget, gradient-norm clipping at 5.0, and a positive-class-weighted binary cross-entropy loss, retaining the epoch of maximum validation average precision with early stopping after 15 epochs without improvement.

3. Multi-instance training: Eight instances were trained with identical data, architecture, optimizer, loss, epoch budget, and selection procedure, differing only in initialization seed (42, 1, 7, 13, 21, 99, 123, 2024). Every quantity depending on the graph model is computed per instance and reported as a distribution over the eight. Where a single instance is needed for per-molecule display, it is chosen by a rule declared before any result was inspected: the instance whose median atom-level agreement is closest to the across-instance median. With an even number of instances this rule is necessarily ambiguous, because the median lies midway between the two central values. Seeds 13 and 123 are equidistant from it at 0.0168, and the tie was resolved toward the lower value, selecting seed 123. The representative instance is used only for display, and no ensemble of the eight was formed or evaluated.

Attribution Extraction and Its Verification

Two verification gates precede any attribution being interpreted, and each halts execution on failure rather than emitting a warning. Both are drawn in red on Fig. 1.

1. TreeSHAP additivity gate: TreeSHAP [11] attributions for the positive class were computed for the Morgan surrogate. Because TreeSHAP is exact only when the attributions plus the base value reconstruct the model output, additivity was tested on 25 held-out molecules before any attribution was interpreted:

$$\max_{M} \left| \phi_0 + \sum_{b} \varphi_b(M) - \hat{p}(M) \right| = 3.69 \times 10^{-7} \qquad (1)$$

where ϕ_0 is the TreeSHAP base value, $\varphi_b(M)$ the Shapley value of bit b for molecule M, and $\hat{p}(M)$ the model output. Reporting the measured error rather than asserting that SHAP was used converts an assumption into a verified quantity. All attribution results describe this surrogate, whose held-out ranking correlates with the deployed Combined model at ρ = +0.820. They are not presented as explanations of the deployed model.

2. Forward-pass replication gate: The PyTorch Geometric implementation of Attentive FP returns only the prediction and exposes no accessor for the readout attention coefficients, so the forward pass was replicated to capture them. Because a replicated forward pass is a claim about equivalence, the replication returns the logit computed along its own path and this was asserted against the model's native output for every instance. The maximum discrepancy across all eight instances and 25 molecules was 1.91×10^{-6} against a pre-set tolerance of 10^{-4}.

Two further properties were confirmed: the coefficient vector has exactly one entry per atom, and the coefficients sum to unity within each molecule. That constraint fixes the terminology used throughout. These are SoftMax aggregation weights combining atom embeddings into molecular embedding, referred to as readout attention and never as attributions. Coefficients from the final readout time step were used.

Because both gates are assertions rather than logged checks, a future library release that alters the readout will cause the analysis to fail loudly rather than silently produce plausible coefficients from a different computation.

Atom-Level Projection of Fingerprint Attributions

Morgan bits index circular environments, not atoms, so the two attribution schemes are not directly commensurable. The Shapley value of each active bit was distributed evenly across the atoms of every environment that set it, and each atom's accumulated value was then divided by the number of environments covering that atom [35]:

$$\tilde{\varphi}(a) = \frac{1}{c(a)} \sum_{b \in \mathcal{B}(M)} \sum_{\substack{e \in \mathcal{E}(b) \\ a \in \mathcal{A}(e)}} \frac{\varphi_b}{|\mathcal{A}(e)|}, c(a) = |\{e : a \in \mathcal{A}(e)\}| \quad (2)$$

where $\mathcal{B}(M)$ is the set of bits active for molecule M, $\mathcal{E}(b)$ the environments setting bit b, $\mathcal{A}(e)$ the atoms of environment e, φ_b the Shapley value of bit b, and $c(a)$ the number of environments covering atom a.

This correction targets a dependence intrinsic to the projection. An atom of high degree lies within more circular environments and would otherwise accumulate more attributional mass for connectivity reasons alone. The magnitude of that dependence, and the effect of the correction, were measured over 150 held-out molecules rather than assumed. The measured values are reported in Section 2.3 and are what allow the connectivity control of Section 4.8 to be evaluated against a known rather than an assumed baseline.

Both signed and unsigned projections were computed. The unsigned projection is used for comparison with attention, which is non-negative by construction, so the analyses concern which atoms carry attributional weight and not the direction of their effect. This quantity is a projection of fingerprint-level attributions onto atoms and is referred to as such. It is not a native atom-level explanation.

Ranking Agreement and Complementarity

Before comparing atoms, it is necessary to establish whether the two models agree on molecules. A global rank correlation answers the wrong question, being dominated by the large majority of compounds that both models correctly place at the bottom of the ranking. Two top-weight measures are used instead, both computed separately for each instance so that reported figures carry their across-instance variation.

1. Rank-biased overlap: RBO [21] with persistence parameter $p = 0.98$, which concentrates weight in approximately the top 50 positions, was computed between the surrogate ranking and each instance's Attentive FP ranking over the full held-out set ($n = 7752$). Spearman correlations over all compounds, and over actives only, are reported alongside.

2. Overlap by depth: For each depth k, the observed size of the intersection of the two top-k sets was compared against the expectation under independent uniform selection, k^2/n, and the Jaccard index of the two sets was recorded. This baseline assumes the two rankings are independent; they correlate weakly (Spearman +0.060 to +0.169), so it is an approximation, and the depth-50 result is well above the margin that approximation introduces. At depths where the underlying intersection is zero or one compound the ratio does not support a stable estimate. Those depths are reported to have their full across-instance range and are not interpreted.

3. Complementarity cross-reference: For each instance, the actives appearing in the surrogate's top 50 were located in the Attentive FP ranking and vice versa. The median rank assigned by the other model, and the number falling inside the other model's top 500, are reported. This analysis establishes the premise for the atom-level comparison. Because the two models prioritize different compounds, their behavior within a shared compound is a separate question.

Atom-Level Agreement, Reference Distribution, and Connectivity Control

1.Comparison: For each analyzed molecule the coverage-normalized $|\tilde{\varphi}|$ map and the readout attention map were compared by Spearman rank correlation across atoms, which is invariant to monotone scaling, and were min–max scaled for the top-quintile Jaccard step. Agreement is also reported as the Jaccard index of the two top-quintile atom sets, whose chance expectation is 0.1111 for two independently selected sets each comprising 20% of a molecule's atoms.

2. Molecule pool: The fixed pool comprises 60 molecules drawn from the head of the surrogate ranking and deduplicated by stereochemistry-free InChIKey skeleton, so stereoisomers of one compound cannot contribute as independent observations. The pool is defined from the surrogate ranking alone and is therefore identical across the eight instances, which isolates variability in the attention maps themselves. An adaptive pool interleaving the two rankings was evaluated in parallel; its composition varies with the instance, overlapping the fixed pool by 47–57%.

3. Size-matched cross-molecule reference: The reference pairs the attention map of one molecule with the $|\tilde{\varphi}|$ map of a different molecule having the same atom count:

$$\mathcal{N}^{(s)} = \left\{ \begin{matrix} \mathrm{Spearman}\left(\alpha^{(s)}(M_i),\ \left|\tilde{\varphi}(M_j)\right|\right) : \\ i \neq j,\ n_{\mathrm{atoms}}(M_i) = n_{\mathrm{atoms}}(M_j) \end{matrix} \right\} \qquad (3)$$

where $\alpha^{(s)}$ denotes the readout attention of instance s. Atom-count matching makes both sides genuine attribution vectors of real molecules with identical dimensionality, so whatever generic structure the two schemes share is already contained in the reference. Fifteen atom-count groups with at least two members were available, covering 42 of the 60 molecules and admitting 102 distinct ordered pairs. Observed and reference distributions were compared by a one-sided Mann–Whitney U test [36], reported together with the common-language effect size. Because the ordered pairs are the independent units of this comparison, the test is evaluated on the distinct pairs rather than on repeated draws from them.

4. Connectivity control: For every molecule the Spearman correlation of each attribution map with atom degree was also computed. The control is evaluated as a relative comparison operating through sign

rather than magnitude. Where the two maps degree associations point in opposite directions, a shared topological artefact cannot produce positive agreement between them, since connectivity would drive them apart. Because the $|\tilde{\varphi}|$ side does not depend on the graph model, its degree correlation is invariant across instances by construction, and this invariance is used as a check that the two sides of the analysis were correctly isolated. The number of instances satisfying the sign condition is reported with the result, and the control is treated as uninformative rather than as evidence in the instances where it does not apply.

Connected-Subgraph Enrichment and Multiplicity Control

Structural cores were located with explicit, reviewer-checkable SMARTS patterns for β-lactam, cephem, penam, carbapenem, quinolone, sulfonamide, nitroaromatic, and anthraquinone motifs, plus a macrolactone detector keyed on a ring-oxygen/carbonyl motif within rings of at least twelve atoms. For each located core c the enrichment ratio is

$$E(c) = \frac{\overline{\psi}_{a \in c}}{\overline{\psi}_{a \notin c}} \qquad (4)$$

where ψ is the attribution map under test and the bars denote means over the atoms inside and outside the core within the same molecule. A ratio of 1.0 denotes indifference, and a value below 1.0 indicates that the core received less weight than the remainder of the molecule.

The reference is 2000 randomly grown connected subgraphs of the same atom count within the same molecule. The connectivity requirement is what makes the test informative. Any spatially smooth attribution would exceed a reference of randomly scattered atoms merely because neighboring atoms carry similar values, whereas requiring connected fragments asks specifically whether the identified core is favored over any comparable connected fragment of the same molecule. Because the reference depends only on the molecule and the core size and not on the attribution under test, it was generated once per (molecule, size) pair and reused for $|\tilde{\varphi}|$ and for all eight instances, so every comparison is paired against an identical reference. Empirical one-sided p-values have a resolution floor of $1/(2000 + 1) = 5.00 \times 10^{-4}$.

Twenty-two core instances, arising from 15 molecules, were tested per attribution scheme under the Benjamini–Hochberg procedure [37] at a false-discovery rate of 0.05. For $|\tilde{\varphi}|$, which does not depend on the graph model, the procedure was applied once to the instance-invariant p-values. For attention, which varies across instances, the procedure was applied separately within each of the eight instances and the number of instances in which a core reached $q < 0.05$ is reported. Combining the eight p-values by their median would not yield a valid p-value, since the median of eight uniform variates is not itself uniform. The per-instance criterion also matches the question the study asks of the attention maps, which is whether a finding reproduces. The core definitions overlap by construction, since a cephem contains a β-lactam ring, so the tests are positively dependent rather than independent. Benjamini–Hochberg remains valid under positive dependence, but the count of significant tests is not a count of independent findings.

Applicability-Domain Stratification

Each held-out compound was characterized by its maximum Tanimoto similarity to any training compound, computed on Morgan fingerprints at radius 2 folded to 2048 bits, and the held-out set was stratified into three bins (< 0.40, 0.40–0.60, ≥ 0.60). Because the bins carry different positive rates, bin-wise AUPRC values are not directly comparable across bins. Enrichment relative to each bin's own base rate is the comparable quantity, and both are reported together with the number of actives per bin.

The 13 held-out actives of 70 or more heavy atoms were analyzed as the pre-declared stratum of Section 4.3. Their nearest-neighbor similarity distribution, their rank under each model, and their recovery within the top 100 and top 500 are reported.

This analysis is retrospective stratification of a labelled held-out set. No out-of-domain detection rule, abstention threshold, or conformal predictor is applied at prediction time, and no such capability is claimed.

Candidate Reporting

The top 50 compounds of the deployment-regime ranking were annotated by the SMARTS structural-class assignment of Section 4.9 and by their nearest-neighbor Tanimoto similarity. Because these are held-out test compounds with known labels rather than prospective leads, they are reported as confirmed actives recovered within the top 50 and not as candidates.

Two redundancy corrections are applied to any hit-rate figure. Ranked positions occupied by stereoisomers of a single compound, and by one compound appearing under more than one record, are identified. The number of distinct Bemis–Murcko scaffolds among the recovered actives is therefore reported alongside the number of active rows.

An identity audit was performed by strict InChIKey matching against a reference set. This procedure establishes only whether a record was returned. An unresolved query cannot be distinguished from a request failure, and no compound is described as chemically novel anywhere in this work.

Scaffold diversity was quantified for the library and for the actives as the ratio of unique Bemis–Murcko scaffolds to compounds carrying a scaffold, on the parsed set before deduplication. Acyclic molecules yield an empty scaffold and are excluded from both counts, which is why the denominators fall below the corresponding set sizes.

Statistics and Software

Area under the precision–recall curve is the primary metric under this imbalance [20], with AUROC reported alongside. Enrichment factor at depth 50 is the hit rate in the top 50 ranked held-out compounds divided by the held-out base rate [23]. Intervals are stratified bootstrap intervals over 2000 resamples, stratified by class label, under the single global seed., and the held-out set was evaluated once per final model.

Analyses used Python 3 with NumPy 2.1.3, pandas 2.2.3, scikit-learn 1.6.1, SciPy 1.16.3, stats models 0.14.6, RDKit 2026.03.5 [30], shap 0.52.0, PyTorch 2.11.0+cu128, and PyTorch Geometric 2.8.0. post1, on a single CUDA device. A single seed governs all pseudo-random procedures other than the eight declared initialization seeds.

Where a count of positive instances is summarized, the exact one-sided binomial probability is reported alongside the across-instance median and the common-language effect size, so that magnitude and count are read together.

A note on the multi-instance design

The eight instances differ in initialization seed alone. Data, architecture, optimizer, loss, epoch budget, and selection procedure are held fixed by construction. Any variation observed across the eight is therefore attributable to the initialization seed, and statements of the form “the quantity varies across random seeds” are licensed by the experimental design rather than inferred post hoc. One-sided tests are used throughout because each hypothesis specifies a direction in advance: agreement above the reference, enrichment inside the core, and a positive count of instances.

Table 3. Software stack.

Layer	Components
Cheminformatics	RDKit 2026.03.5 for parsing, fragment selection, InChIKey generation, Morgan and MACCS fingerprints, Bemis–Murcko scaffolds, SMARTS matching, and similarity maps
Fingerprint models	scikit-learn 1.6.1 (Random Forest); XGBoost, Combined and Morgan-only variants, compared on validation
Graph model	PyTorch 2.11.0+cu128 with PyTorch Geometric 2.8.0. post1; Attentive FP, 486,145 parameters, eight instances
Attribution	shap 0.52.0 (TreeSHAP) with an enforced additivity gate; custom replicated forward pass for readout attention

Layer	Components
Statistics	SciPy 1.16.3 (Spearman, Mann–Whitney, exact binomial); statsmodels 0.14.6 (Benjamini–Hochberg)
Numeric and output	NumPy 2.1.3, pandas 2.2.3, Matplotlib; per-instance CSV exports and a machine-readable JSON summary

Notation

Symbol	Meaning
M	Molecule
a	Atom
b	Active Morgan bit
e	Circular environment (center atom, radius)
$\mathcal{B}(M)$	Set of bits active for molecule M
$\mathcal{E}(b)$	Environments setting bit b
$\mathcal{A}(e)$	Atoms of environment e
φ_b	Shapley value of bit b
ϕ_0	TreeSHAP base value
$\hat{p}(M)$	Model output for molecule M
$\tilde{\varphi}(a)$	Coverage-normalized atom-level projection
$c(a)$	Number of environments covering atom a
$\alpha^{(s)}$	Readout attention of instance s
$\mathcal{N}^{(s)}$	Size-matched cross-molecule reference distribution for instance s
ψ	Generic attribution map under test
$E(c)$	Within-molecule enrichment ratio of core c
S	Number of independently trained graph-model instances ($S = 8$)

Results and Discussion

Results

Curation, Partition, and Held-Out Performance

Curation reduced 39,311 parsed structures to 38,748 compounds carrying 456 actives (Table 4). Deduplication removed actives disproportionately: 56 of 512, or 10.9% of all actives, against 1.43% of the library. An active was therefore about 7.6 times more likely to be removed than an average row. Known drugs recur in screening libraries as stereoisomers and duplicate entries far more often than inactive compounds do.

On the held-out set the Combined Random Forest reached an AUPRC [20] of 0.1233 (95% CI 0.1037–0.1519) in the comparison regime and 0.1518 (0.1272–0.1873) in the deployment regime, against a base rate of 0.0390 (Table 3, Fig. 2). The Morgan-only surrogate reached 0.1108 with an AUROC of 0.7252, its ranking correlating with the Combined model at $\rho = +0.820$.

The eight AttentiveFP instances reached a median held-out AUPRC of 0.0834 (range 0.0587–0.0949) and a median AUROC of 0.6259 (0.5949–0.6715). Held-out performance therefore spans a 1.6-fold range across models differing only in initialization. The regimes are not interchangeable. The like-for-like comparison against the graph model is the comparison-regime value of 0.1233, and the deployment value is quoted only where the additional training data are declared. At the median the graph model attains 2.1 times the held-out base rate, so it retains a weak but non-zero signal rather than none.

The ratio of best validation AUPRC to held-out AUPRC has a median of 3.14 and ranges from 2.22 to 3.89. It exceeds 2 in every instance, so the generalization gap is systematic in direction rather than selection noise confined to one run, although its size varies by a factor of 1.75. Early stopping occurred between epochs 27 and 60, so the training budget was not limiting.

Table 4. Dataset composition and scaffold-disjoint partition.

Stage / split	Compounds	Actives	Rate
Raw dataset [1]	39,312	512	1.30%
Parsed by RDKit	39,311	512	1.30%
Retained after deduplication	38,748	456	1.18%
Removed across 510 merge groups	563	n/a	n/a
Label-conflict groups discarded	26 groups	n/a	n/a
Training	26,345	123	0.47%
Validation	4651	31	0.67%
Held out	7752	302	3.90%
Large actives, training	n/a	0	n/a
Large actives, held out	n/a	13	n/a

Large actives are compounds of ≥ 70 heavy atoms.

Table 5. Held-out predictive performance.

Model	AUPRC	95% CI	AUROC
Combined RF, comparison regime	0.1233	0.1037–0.1519	n/a
Combined RF, deployment regime	0.1518	0.1272–0.1873	n/a

Model	AUPRC	95% CI	AUROC
Morgan RF surrogate	0.1108	n/a	0.7252
AttentiveFP, median of 8 instances	0.0834	0.0587–0.0949 [1]	0.6259

[1] *Range across instances, not a confidence interval; the two quantities answer different questions. The comparison regime is fitted on the training portion alone, matching the data seen by the graph model; the deployment regime adds the validation split. Held-out base rate 0.0390.*

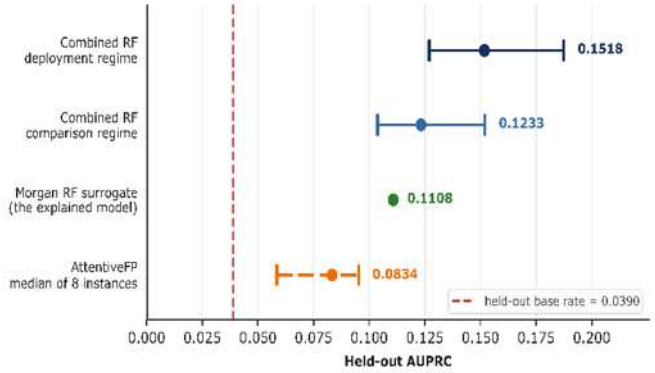


Fig. 2 Held-out AUPRC by model. Solid bars are 95% stratified bootstrap intervals. The dashed bar for AttentiveFP is the range across eight instances and the point their median. The two quantities answer different questions and are drawn differently for that reason.

Ranking Agreement and Model Complementarity

Rank-biased overlap [21] between the surrogate and each AttentiveFP instance has a median of 0.0731 across instances (range 0.0350–0.1443), indicating limited global agreement between the two molecular rankings. Spearman correlation over all compounds ranges from +0.060 to +0.169 across instances. Over actives only it ranges from −0.036 to +0.077 and includes zero in every instance, so the two models do not agree on the relative ordering of the compounds that matter.

Overlap by depth nevertheless exceeds chance in every instance from depth 50 onwards and decays monotonically. The median ratio observed to chance overlap is 17.1 at depths 50, 9.7 at 100, 3.2 at 250, 2.1 at 500, and 1.3 at 2000 (Fig. 6). At depths of 25 or fewer the ratio spans 0 to 77.5 across instances because the underlying overlap is zero or one compound, and those depths do not support a stable estimate. Absolute set overlap remains small over the interpretable range, with a median Jaccard index of 0.054–0.072 for depths 50–500.

The complementarity cross-reference makes the pattern concrete. Of the 14 actives (median across instances) in the surrogate's top 50, AttentiveFP places 8.5 within its own top 500 at a median rank of 320. Of the 7.5 actives in AttentiveFP's top 50, the surrogate places 2.5 within its top 500 at a median rank of 960. The two models nominate largely distinct candidate sets while agreeing well above chance on the region of chemical space worth examining. That is the premise for what follows. Because they prioritize different compounds, their behaviour within a shared compound is a separate question.

The Projection and Its Topological Dependence

Over 150 held-out molecules the number of Morgan environments covering an atom correlates with atom degree at a median Spearman ρ of +0.857, confirming that the projection carries a substantial topological dependence (Table 4). Before correction the projected |SHAP| magnitude correlates with atom degree at +0.268. After coverage normalization, it drops to +0.069, while the corrected and uncorrected maps remain highly concordant (+0.959). The correction therefore removes approximately three quarters of the degree dependence while leaving the attribution pattern substantially intact. Quantifying the residual is what allows the control of Section 2.5 to be evaluated against a known rather than an assumed baseline.

Table 6. Topological dependence of the atom-level projection and the effect of coverage normalization.

Quantity	Spearman ρ	Reading
Coverage counts vs. atom degree	+0.857	Magnitude of the underlying dependence
Raw $\|\varphi\|$ vs. atom degree	+0.268	Before correction
Coverage-normalised $\|\tilde{\varphi}\|$ vs. atom degree	+0.069	After correction
Raw vs. normalised map	+0.959	Attribution pattern preserved

Medians over n = 150 held-out molecules.

The residual degree dependence of the projection (+0.069) was measured on 150 held-out molecules, whereas the agreement analysis uses the 60-molecule pool; the two populations differ, and the residual is therefore an estimate for held-out molecules generally rather than for the pool specifically.

Atom-Level Agreement Against a Size-Matched Reference

Across the 60-molecule fixed pool the median per-molecule Spearman correlation between readout attention and coverage-normalised |SHAP| is +0.0890, with a range of −0.1067 to +0.1547 and a standard deviation of 0.0899 across the eight instances (Table 7, Fig. 3). The correlation is positive in 6 of 8 instances, and in the same 6 the observed distribution exceeds that instance's own size-matched reference at $p < 0.05$. The eight per-instance tests are reported descriptively and are not corrected across instances. The median common-language effect size is 0.647: a randomly chosen molecule shows higher attention–|SHAP| agreement than a randomly chosen cross-molecule pair about two thirds of the time. Because each molecule's maps are reused across reference pairs, the 102 pairs are not independent and the per-instance p-values are correspondingly anti-conservative. With 60 observed values against that reference, a Mann-Whitney test also detects displacements well below the size at which a correlation of this magnitude would carry chemical meaning. The effect sizes rather than the p-values are therefore the quantities to read.

Two statistics answer different questions and are reported together. The per-instance Mann–Whitney tests ask whether the 60 correlations within an instance are displaced above that instance's own reference, and answer affirmatively in six. The exact one-sided binomial probability of six or more positives from eight is 0.145, which does not reach conventional significance. That is expected for a statistic discarding both magnitude and the per-instance evidence. Neither statistic alone establishes universality across instances, and they are reported jointly for that reason.

In the two instances where the observed median is not above the reference it lies below it (seed 42: −0.1067 against a reference median of −0.0566; seed 99: −0.0220 against +0.0587). In those instances, matching attention and |SHAP| maps within the same molecule performs worse than pairing them across molecules, which is a stronger negative than the count of six alone conveys.

The median Jaccard index of top-quintile atoms is 0.1250 against a chance expectation of 0.1111, a ratio of 1.12, with three instances below chance. This measure is at chance and is not treated as evidence. It nevertheless carries the most direct chemical reading, since it asks whether the two schemes select the same important atoms. An alternative account consistent with all reported values is therefore that the two maps share a weak monotone alignment across the bulk of low-attribution atoms while disagreeing at the

top. This study does not exclude that account. The rank correlation over the whole attribution map remains the appropriate primary measure, because thresholding at 20% discards the information carried by the remaining atoms.

The adaptive pool gives a median of +0.1236 over a wider range (−0.1787 to +0.6032), consistent with the fixed-pool result while carrying the extra variance of changing pool membership.

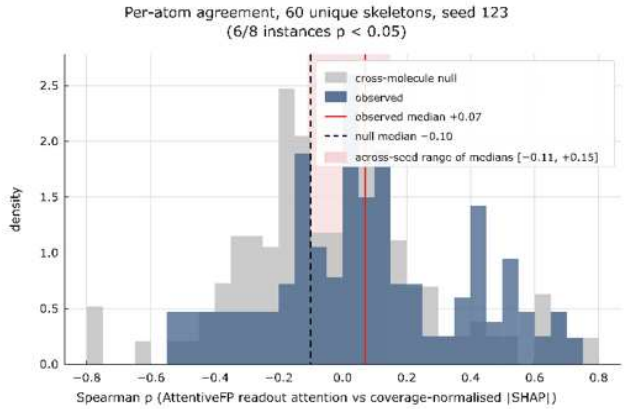


Fig. 3 Per-molecule agreement against a size-matched cross-molecule reference, representative instance (seed 123), n = 60 skeleton-deduplicated molecules. Grey: reference, constructed from 102 distinct ordered molecule pairs matched on atom count. Blue: observed. Solid line: observed median; dashed line: reference median; shaded band: range of medians across all eight instances. Observed and reference distributions overlap substantially. The result is a displacement in central tendency across a population of molecules rather than a separation of two groups, and the band includes negative values.

The Agreement Is Not Topological

The correlation of readout attention with atom degree varies in sign across the eight instances, from −0.2562 to +0.1497, while the correlation of coverage-normalised |SHAP| with atom degree is invariant at +0.0439 in every instance (Table 7, Fig. 4). The invariance of the second quantity is expected by construction and is a check that the two sides of the analysis were correctly isolated.

The control operates through signs rather than magnitude. It applies to the three instances in which the two-degree associations point in opposite directions. These are seeds 1, 13, and 123, at −0.0230, −0.0635, and −0.2562 against a fixed +0.0439. In those three, a shared topological artefact cannot produce positive agreement between the two maps, because connectivity would drive them apart, and all three nonetheless agree positively (+0.1327, +0.1058, and +0.0722). In the remaining five instances the attention–degree association is positive or near zero and the control is uninformative rather than violated. The argument is stronger than a comparison of magnitudes, which would depend on where a threshold is placed, and where it applies it establishes that the concordance is not attributable to the residual projection bias of Section 2.3.

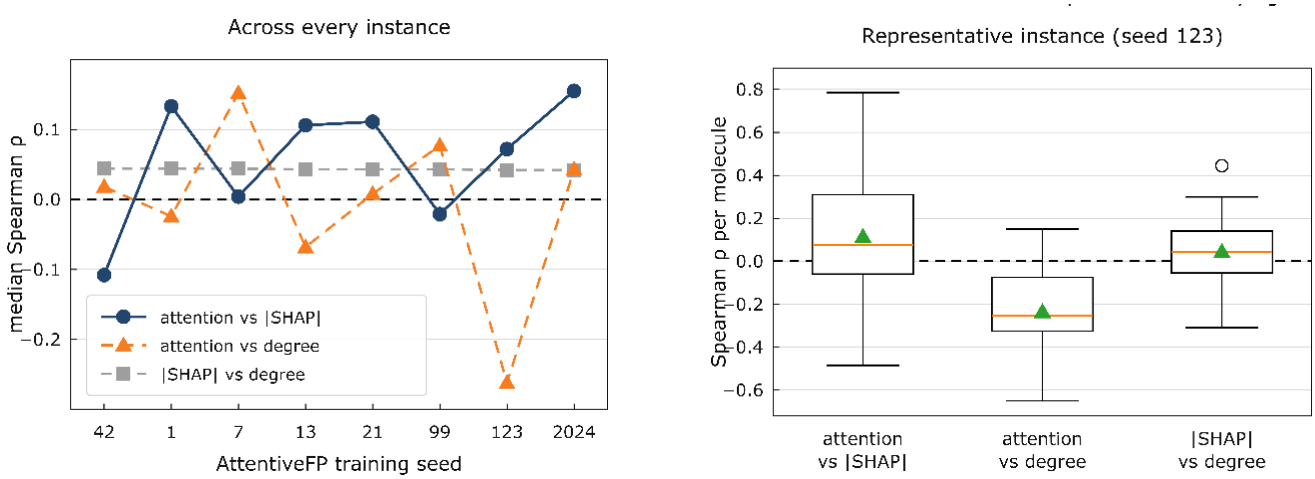


Fig. 4 Connectivity control. (Left) Per-molecule Spearman correlations for the representative instance: attention versus |SHAP|, attention versus atom degree, and |SHAP| versus atom degree (n = 60; boxes span the interquartile range, orange line the median, green triangle the mean). (Right) The same three medians for each of the eight instances. The attention–degree trace crosses zero while the |SHAP|–degree trace is flat by construction, the latter a built-in check that the two sides were isolated. The first box spans negative values, so the concordance is a population-level tendency.

Reproducibility Across Random Seeds

Across models differing only in initialization seed, held-out performance spans a 1.6-fold range (0.0587–0.0949) and atom-level agreement spans −0.1067 to +0.1547, a range that includes both signs. Attention-based structural-core enrichment varies more widely still, by two to three orders of magnitude for every core (Section 2.7). A single training run yields one draw from these distributions, and its position within them is not recoverable from the run itself.

Across the eight instances, atom-level agreement correlates with held-out AUPRC at Spearman +0.857 (one-sided $p = 0.0065$, n = 8; Fig. 5. The two instances with negative agreement, seeds 42 and 99, are the two weakest models (0.0587 and 0.0693), while the two strongest (0.0948 and 0.0949) show the highest agreement (+0.1327 and +0.1547). Agreement with the surrogate's ranking shows no comparable relationship (+0.476, $p = 0.233$).

This relationship rests on eight points, was examined after the per-instance results were available, and is one of two correlations tested. It is reported as an exploratory observation requiring confirmation on a larger replicate set, not as an established dependency, and it is not part of the study's primary evidence.

Table 7. Atom-level attribution agreement, per instance.

Seed	AUPRC	Median ρ	Jaccard	Molecules with ρ > 0	Reference median	*p*	CLES	ρ(attn, deg)
42	0.0587	−0.1067	0.1250	37%	−0.0566	0.931	0.442	+0.0188
1	0.0948	+0.1327	0.1250	75%	−0.0463	1.11×10^{-6}	0.685	−0.0230
7	0.0864	+0.0048	0.0769	50%	−0.0390	0.0288	0.574	+0.1497
13	0.0818	+0.1058	0.1111	65%	+0.0240	7.72×10^{-4}	0.624	−0.0635
21	0.0851	+0.1110	0.1548	68%	−0.0581	2.09×10^{-7}	0.698	+0.0076

Seed	AUPRC	Median ρ	Jaccard	Molecules with ρ > 0	Reference median	*p*	CLES	ρ(attn, deg)
99	0.0693	−0.0220	0.0955	48%	+0.0587	0.991	0.408	+0.0835
123	0.0727	+0.0722	0.1429	67%	−0.0981	7.33×10^{-6}	0.670	−0.2562
2024	0.0949	+0.1547	0.1339	77%	+0.0088	2.48×10^{-7}	0.697	+0.0514
Median	**0.0834**	**+0.0890**	**0.1250**	n/a	**−0.0427**	n/a	**0.647**	**+0.0132**

ρ (|SHAP|, degree) = +0.0439 in every instance, invariant by construction. Jaccard chance expectation is 0.1111. p is the one-sided Mann–Whitney value against that instance's own reference, evaluated on the 102 distinct ordered pairs. Seed 123 is the representative instance (Section 4.4).

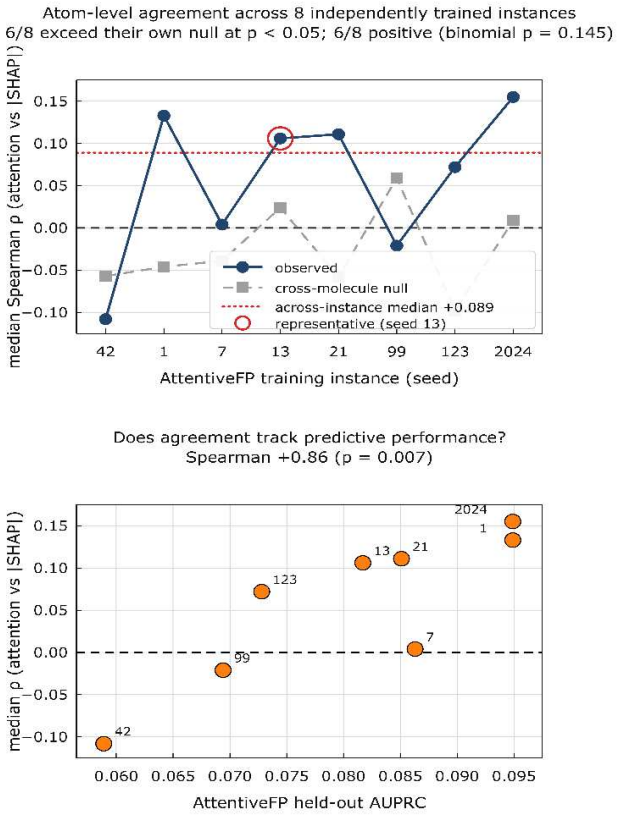


Fig. 5 Reproducibility across independently trained instances. (**Upper**) Median per-molecule agreement per instance (solid) against the median of that instance's own reference (dashed), with the across-instance median dotted. Six of eight lie above zero and above their own reference at $p < 0.05$. (**Lower**) Median agreement against held-out AUPRC, annotated by seed (Spearman +0.857, $p = 0.0065$, n = 8). The circled point marks seed 13, which ties with seed 123 under the median-proximity rule of Section 4.4. Seed 123 is the representative instance used for display elsewhere.

No Structural Core Survives False-Discovery Control

Twenty-two core instances were located across 15 of the 60 pool molecules: 8 β-lactam, 7 cephem, 6 nitroaromatic, and 1 anthraquinone (Table 6). Under the |SHAP| map, median enrichment ratios are 2.79, 4.20, 2.39, and 0.90, respectively, and 7 of 22 instances reach $p < 0.05$ before correction. Under readout attention the ratios are 1.11, 0.97, 1.85, and 1.77, and no instance reaches $p < 0.05$ before correction. Attention-based enrichment does not reproduce. The across-instance range spans two to three orders of magnitude for every core, from 0.008 to 2.151 for β-lactam and from 0.141 to 8.116 for nitroaromatic. Ratios below unity mean the core received less attentional weight than a comparable connected fragment of the same molecule. No core is significant in all eight instances, and the single-instance values a one-

run analysis would report are arbitrary draws from these ranges. The |SHAP| side is invariant and directionally consistent, with the three multi-member cores enriched 2.4- to 4.2-fold, but the individual tests are weak.

After Benjamini–Hochberg correction, no core reaches significance under either scheme. Under $|\tilde{\varphi}|$ the adjusted values range from 0.061 to 0.593 across the four core types. Under attention the procedure was applied within each instance separately, and no core reached $q < 0.05$ in any of the eight. The lowest adjusted value, 0.061 for the cephem cores under $|\tilde{\varphi}|$, falls short of the threshold in a family of 22 underpowered tests and carries no evidential weight beyond that. We report this negative outcome directly. The test set is small, comprising 22 instances from 15 molecules across four overlapping core definitions, so the analysis has limited power, and the absence of controlled enrichment is not evidence that these cores are unimportant to the models. Nor does it overturn Section 2.4. The two analyses ask different questions: whether the schemes agree with each other across all atoms, and whether either concentrates on a prespecified substructure. Cross-method concordance does not require that the agreed-upon atoms coincide with classical pharmacophoric cores.

A single molecule illustrates both halves of that statement. The compound with the highest agreement in the representative instance's pool ($\rho = +0.78$; Fig. 7) is a nitrofuran-hydrazone bearing a morpholine-substituted acetamide. Both maps concentrate on the nitro-substituted furan ring and the hydrazone linker while de-emphasising the distal morpholine, a convergence produced by methods sharing no featurization and no training objective. Yet nitroaromatic cores did not survive false-discovery correction, and at roughly ten times that instance's median agreement the molecule is an illustration rather than a typical case. Agreement between the two schemes and concentration on a recognised pharmacophore are different claims, and only the first is supported here.

Applicability Domain and Chemotype-Selective Enrichment

Held-out compounds were stratified by their maximum Tanimoto similarity to any training compound (median 0.354 across the held-out set). Read by raw AUPRC the stratification is non-monotone, because the bins carry different positive rates. Read by enrichment relative to each bin's own base rate it is monotone, rising from 3.34-fold in the far bin to 13.40-fold in the near bin (Table 8). Prevalence correction does not merely rescale the comparison here; it changes the ordering of the bins. Both quantities are therefore reported together with the number of actives per bin.

Table 8. Connected-subgraph enrichment of structural cores with false-discovery correction.

Core	n	Attention: median ratio	Attention: across-instance range	Attention: instances with q < 0.05	|SHAP|: ratio	|SHAP|: range	|SHAP|: *q*
β-Lactam	8	1.111	0.008–2.151	0 of 8	2.788	Invariant	0.227
Cephem	7	0.967	0.010–2.283	0 of 8	4.203	Invariant	0.061
Nitroaromatic	6	1.855	0.141–8.116	0 of 8	2.393	Invariant	0.241
Anthraquinone	1	1.768	0.338–5.104	0 of 8	0.904	Invariant	0.593

Ratios are within-molecules mean attribution inside the core over mean attribution outside it, against 2000 connected subgraphs of equal atom count in the same molecule (Equation (4)). Attention values are across-instance medians; the |SHAP| side does not depend on the graph model and is therefore invariant across instances. Uncorrected one-sided p-values for |SHAP| are 0.145, 0.015, 0.183, and 0.566 in row order; no attention test reaches $p < 0.05$ uncorrected. The 22 tests arise from 15 molecules and nested core definitions and are positively dependent. The anthraquinone row rests on a single molecule.

The AttentiveFP instances remain below 0.09 in every bin and gain nothing from proximity to the training chemistry. The distance-dependence of fingerprint-model reliability is a long-established property of

quantitative structure–activity models [22]. It is reported here to characterize the scope of the present result, not as a finding.

Table 9. Held-out performance stratified by nearest-neighbour Tanimoto similarity to the training set.

Bin	Compounds	Actives	Base rate	RF AUPRC	RF EF@50	AttentiveFP AUPRC [1]
Far (< 0.40)	4678	224	4.79%	0.1145	3.34	0.0821
Mid (0.40–0.60)	1638	63	3.85%	0.3078	11.44	0.0781
Near (≥ 0.60)	1436	15	1.04%	0.1187	13.40	0.0419

[1] *Representative instance. The near bin contains 15 actives, so its estimates rest on very few positives and should be read as indicative.*
AUPRC is not comparable across bins because the base rates differ; the enrichment column is the comparable quantity.

The 13 held-out actives of ≥ 70 heavy atoms lie at nearest-neighbour similarities of 0.130–0.340 (median 0.207), entirely within the far bin, against a training portion containing no active of comparable size. Neither model recovers any of them within the top 100. Within the top 500 Random Forest recovers 7 (best rank 182, median 449) and AttentiveFP none (best rank 4561, median 7663 of 7752). The graph model therefore places this class of true actives near the bottom of its own ranking. The failure is located in distance from the training chemistry rather than in molecular size as such.

Enrichment is chemotype-selective in a way the aggregate figure conceals (Table 8). None of the 25 macrolides, 18 quinolones, 17 penams, 7 sulfonamides, 6 ansamycins, 5 glycopeptides, or 4 lincosamides among the held-out actives appears anywhere in the top 50, and only one of the 23 compounds carrying a cephem core does. Conversely, 4 of the 6 nitroaromatics are present. The composition of the top 50 is 38 compounds carrying no recognised core, 5 tetracyclines, 4 nitroaromatics, 1 anthracycline, 1 β-lactam, and 1 cephem. The classes that are missed are concentrated among the larger compounds with no near neighbour in the training set, which links this observation to the coverage gap above. That link is an explanation consistent with the measured quantities rather than a tested hypothesis.

Table 10. Recovery of recognised structural classes among held-out actives.

Class	Among 302 held-out actives	Present in top 50
Not assigned to a listed class	173	38
Macrolide	25	0
Cephem	23	1
Quinolone	18	0
Penam	17	0
Sulfonamide	7	0
Nitroaromatic	6	4
Ansamycin	6	0
Glycopeptide	5	0
Lincosamide	4	0

Classes are assigned by the explicit SMARTS definitions of Section 4.9 and are neither exhaustive nor mutually exclusive; the ten most frequent assignments are listed. The second column counts compounds carrying the core anywhere in the top 50 of the deployment ranking, whether or not that compound is active.

Within the top 50 of the deployment ranking, 19 compounds are confirmed active, a hit rate of 38.0% (bootstrap interval 22.0–50.0%) and an enrichment factor [23] of 9.75 at depth 50 against the 3.90% base rate. In the comparison regime the corresponding figures are 12 actives and an enrichment factor of 6.16 (3.59–9.75). The 19 active rows correspond to 11 distinct Bemis–Murcko scaffolds. Ranks 26 to 31 are six stereoisomers of a single compound sharing an identical score of 0.3090 and ranks 49 and 50 are one compound appearing under two records at an identical score of 0.192917. The first confirmed active appears at rank 11. The recovered actives lie at nearest-neighbour similarities of 0.433–0.730, all within the interpolation region.

An identity audit by strict InChIKey matching returned no exact match to the reference set for any of the 50 compounds. This establishes only that no record was returned. An unresolved query cannot be distinguished from a request failure, and no compound is described as chemically novel anywhere in this work.

Scaffold diversity, computed on the parsed set before deduplication, is 0.6245 for actives (316 unique scaffolds among 506 compounds carrying one) against 0.1767 for the library (6837 among 38,702), a 3.5-fold difference. Acyclic molecules yield an empty scaffold and are excluded from both counts. This asymmetry is what routes actives preferentially into the small scaffold groups allocated to the held-out set.

Ranking Agreement in Depth

Ranking Agreement in Depth

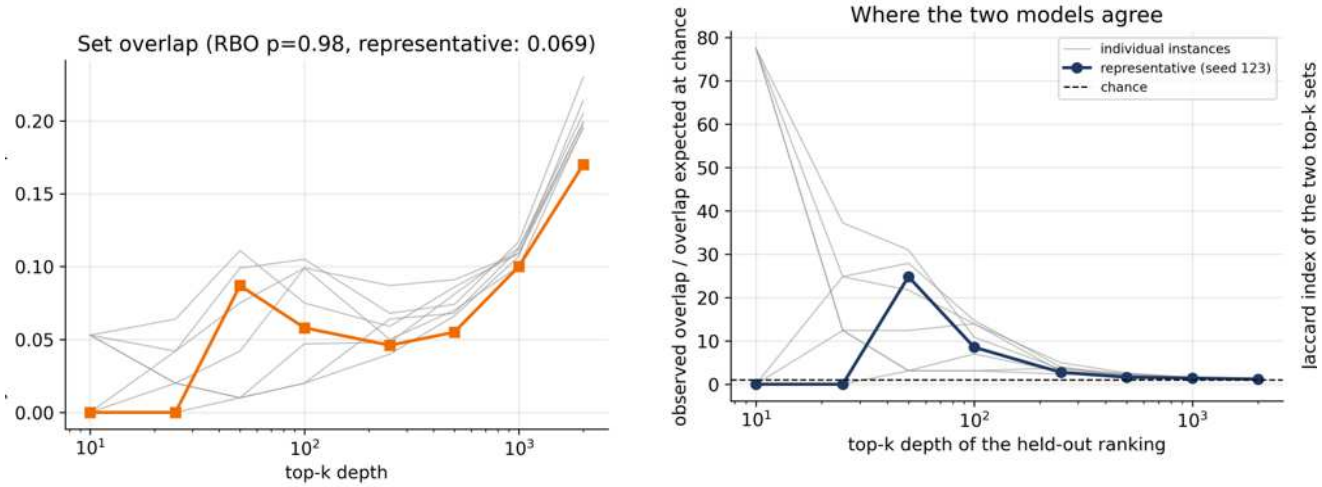


Fig. 6 Ranking agreement between the Morgan surrogate and AttentiveFP as a function of screening depth. (Upper) Observed top-k overlap divided by the chance expectation k^2/n (n = 7752) on a logarithmic depth axis. Grey traces are the eight instances, the overlaid trace is the representative instance (seed 123), and the dashed line mark's chance. (Lower) Jaccard index of the two top-k sets. The median across instances is 17.1× chance at depth 50, 9.7× at 100, 3.2× at 250, 2.1× at 500, and 1.3× at 2000. Overlap is above chance at every depth from 50 onwards in every instance and decays monotonically as the signal is diluted by the inactive majority. The grey envelope is the substance of the panel: at depths of 25 or fewer the ratio spans 0 to 77.5 across instances because the underlying overlap is zero or one compound, whereas from depth 250 it narrows to about a factor of two.

Global agreement between the two rankings is limited (rank-biased overlap median 0.0731; absolute Jaccard 0.054–0.072 over depths 50–500), so the two models do not prioritize the same compounds even though they highlight overlapping regions within a shared compound. That tension is what makes the atom-level analysis of Section 2.4 non-trivial.

Representative Molecular Attribution Maps

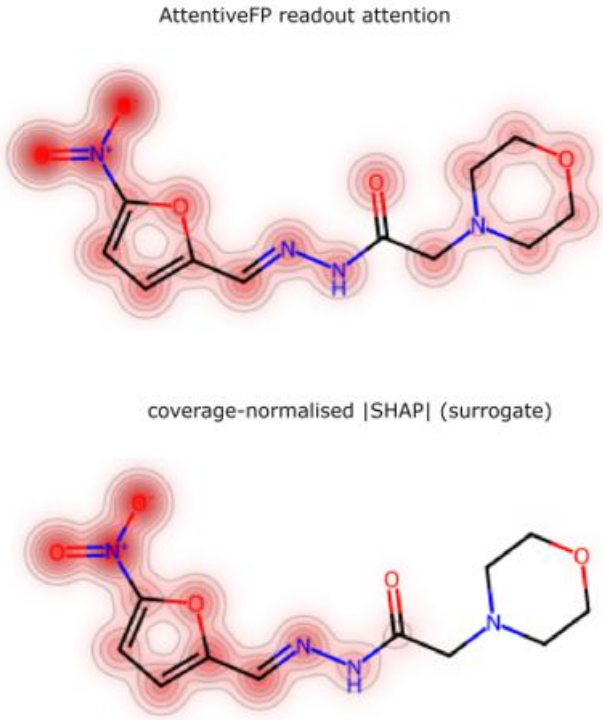


Fig. 7 Representative molecular attribution maps. Test row 38, representative instance (seed 123), confirmed active, ρ = +0.78. (**Upper**) AttentiveFP readout attention. (**Lower**) Coverage-normalised atom-level |SHAP| from the Morgan surrogate. The compound is a nitrofuran-hydrazone bearing a morpholine-substituted acetamide. Both maps concentrate on the nitro-substituted furan ring and the hydrazone linker and de-emphasize the distal morpholine.

Three limits apply to this figure. The panels use independent colour scales and are comparable by spatial pattern only, never by colour magnitude. Colour encodes attribution value, not lipophilicity, charge, or any physicochemical property. And this is the strongest agreement in the pool of a single instance, against a pool median of +0.0722 for that instance, roughly ten times the median, so it is an illustration and not a typical case. Although the nitrofuran moiety is a recognised antibacterial pharmacophore, nitroaromatic cores did not survive false-discovery correction (Table 7, $q = 0.414$ for attention), and one molecule does not establish what 22 formal tests failed to establish.

Discussion

Within a molecule, readout attention and the coverage-normalized projection of fingerprint SHAP values agree at an across-instance median Spearman correlation of +0.0890, exceeding a size-matched cross-molecule reference at $p < 0.05$ in 6 of 8 independently trained instances. The effect is modest and should be described as such. Its credibility rests less on the magnitude than on the design of the reference, which pairs real attribution vectors of real molecules matched on atom count and therefore already contains the generic structure the two schemes share.

The connectivity control adds a distinct piece of evidence. In the three instances where the two schemes' degree associations point in opposite directions, no shared topological artefact can generate positive agreement between them, since connectivity would drive them apart, and all three agree positively. That argument is stronger than a comparison of magnitudes, and it justifies quantifying the projection bias, because a measured residual lets the control be evaluated against a known baseline. It is also narrower than a global claim. In the other five instances the control is uninformative.

The multi-instance protocol changed what could be concluded. Held-out performance spans a 1.6-fold range across instances differing only in initialization, atom-level agreement spans −0.1067 to +0.1547, a range including both signs, and attention-based core enrichment spans two to three orders of magnitude. A single training run yields a draw whose position in that distribution is not recoverable from the run itself. Explainability analyses of low-data molecular models that report a single run may therefore misstate not only the strength but the direction of attribution agreement. Seed dependence of model explanations is not confined to this setting: it has been reported for attention, gradient-based and surrogate explanations in text classification [24], for graph explainers whose feature importances vary across architectures, hyperparameters and seed values [25], and for attribution methods in electrocardiogram classifiers, where self-consistency across random initializations was only moderate [26]. In molecular machine learning, distinct feature-importance distributions have likewise been obtained for the same compounds under different Shapley variants and kernel functions [18,27]. The present result extends that literature to cross-family atom-level agreement in a low-data molecular setting, where the range across instances includes both signs, and it is why every graph-model-dependent quantity here is reported as a distribution. Whether that variability is structured rather than arbitrary is an open question this study raises but cannot settle. Concordance was higher in the instances where the graph network attained higher held-out performance (Spearman +0.857, $p = 0.0065$), with no comparable relationship to ranking similarity (+0.476, $p = 0.233$), which would locate the effect in the internal representation rather than in output agreement. With eight instances, an observation examined after the fact and one of two correlations tested, this is a hypothesis for confirmation on a larger replicate set, not a result.

The substructure analysis returned to a null result under false-discovery control, for two distinct reasons. On the attention side, enrichment does not reproduce, so no stable estimate exists to be tested. On the |SHAP| side, the estimates are invariant and directionally consistent, with β-lactam, cephem, and nitroaromatic cores enriched 2.4- to 4.2-fold relative to comparable connected fragments, but the individual tests are underpowered, the strongest reaching $q = 0.061$. Cephem under |SHAP| is the natural candidate for a targeted, preregistered test on an independent set. These results are not evidence that classical cores are unimportant to the models. They indicate that this sample cannot establish enrichment at conventional error control.

The aggregate screening figures conceal a systematic pattern that matters operationally. An enrichment factor of 9.75 at depth 50 describes a ranking that recovers no macrolide, quinolone, penam, sulfonamide, or glycopeptide among the top 50, while over-representing tetracyclines and nitroaromatics, and the recovered actives lie entirely within the interpolation region of the training chemistry. Reported alone, the aggregate number would misdescribe what the model can and cannot find.

This study sits at a specific point in the literature. The source study used ensembles of graph networks with substructure rationales on the same library [1]. The present work makes no prospective claim and examines the internal consistency of attribution across model families instead, so the retrospective 9.75-fold enrichment is a triage result rather than a discovery claim [2,3]. Graph attribution benchmarks evaluate methods against synthetic or curated ground truth and report that current methods perform poorly on realistic tasks [16,17]. A closely related study trained AttentiveFP on a skin-sensitization endpoint and scored atom-level attributions from seven methods against SMARTS-encoded reactive centers under a label-permutation control, concluding that most apparent attribution–mechanism correlation reflects structural priors of the attribution method rather than learned chemistry [28]. That work and the present one reach compatible conclusions from opposite directions: [28] measures attributions against a known mechanistic ground truth, whereas here no ground truth exists and the reference distributions and controls carry the inference. Comparisons among variants of a single attribution formalism have found little

agreement between alternative explanations of molecular activity models and recommend that feature-importance analyses include a consistency assessment [18]. The present work extends that question across model families and to the atom level. On whether attention is explanation [14,15], the results support the cautious position: the coefficients are constrained to sum to unity per molecule, are demonstrably unstable across training runs in this regime, and are treated throughout as aggregation weights. We did not identify prior work combining an enforced additivity check, a verified forward-pass replication, a size-matched cross-molecule reference, a sign-based connectivity control, and multi-instance replication under false-discovery control in one molecular workflow. This is the outcome of a targeted search, not an exhaustive priority claim.

For molecular discovery the practical reading is narrow. On this dataset and protocol, the fingerprint model provides usable early enrichment while the graph model retains only a weak signal, and where the two agree at the atom level the agreement is modest, above a demanding reference, and not explained by topology. None of these licenses a mechanistic interpretation. An attribution states where a model places weight, not how a molecule acts.

Limitations

The dataset is a single retrospective library against one organism, and no prospective validation was performed, so all conclusions concern model behaviour rather than confirmed activity, a retrospective split of any construction provides only a bounded estimate of prospective performance [29]. The training portion retains only 123 actives after scaffold-disjoint partitioning. This is the binding constraint on the graph model and the proximate cause of the across-instance variability reported throughout. The split routes the smallest scaffold groups into the held-out set, raising its positive rate to 3.90% against 0.47% in training. That is a deliberate extrapolation design, but one that makes no AUPRC comparable across splits without its base rate [24]. Capacity selection rests on a validation slice of 31 actives, which the validation-to-held-out ratio (median 3.14, above 2 in all eight instances) shows supports epoch selection but not an unbiased performance estimate.

The explained model is a Morgan-only surrogate rather than the deployed Combined model (rankings correlate at $\rho = +0.820$), and its atom-level values are a projection of bit-level attributions carrying a residual topological dependence of +0.069. The projection is unsigned, so the analyses address which atoms carry weight and not the direction of their contribution. Readout attention is a SoftMax weight summing to unity within a molecule and cannot express that two atoms both matters greatly.

Eight instances bound every across-instance statement, including the eight-point correlation between agreement and performance. The reference distributions are finite: 102 distinct ordered pairs for the cross-molecule comparison and 2000 connected subgraphs per enrichment test, the latter imposing a resolution floor of 5.00×10^{-4} on the empirical *p*-values. The 60 analyzed molecules come from the head of the surrogate ranking rather than the held-out set at large, so the agreement estimate describes compounds the surrogate ranks highly and its generalization to held-out molecules in general is untested. SMARTS core definitions are explicit but not exhaustive: 45 of the 60 pool molecules carry no recognised core and are untested by that analysis. The applicability-domain analysis is a retrospective stratification of a labelled set and implements no detection rule at prediction time.

Conclusions

On a deduplicated, scaffold-disjoint benchmark of 38,748 antibacterial screening compounds, a fingerprint Random Forest delivered usable early enrichment (comparison-regime AUPRC 0.1233; deployment-regime AUPRC 0.1518 with an enrichment factor of 9.75 at depth 50 against a 3.90% base rate), while an AttentiveFP graph network retained only a weak signal at approximately 2.1 times the base rate (median held-out AUPRC 0.0834 across eight instances). The aggregate enrichment figure is chemotype-selective. No macrolide, quinolone, penam, sulfonamide, or glycopeptide among the held-out actives appears in the top 50, and the recovered actives lie entirely within the interpolation region of the training chemistry.

Readout attention and a coverage-normalised projection of fingerprint SHAP values agree within molecules at an across-instance median Spearman correlation of +0.0890, exceeding a size-matched cross-molecule reference in 6 of 8 independently trained instances. The agreement is modest and is not attributable to shared topological bias. It does not establish that the agreed-upon atoms are mechanistically relevant. No prespecified antibiotic core reached false-discovery-controlled significance under either scheme, and top-quintile atom overlap is at chance.

The agreement is not reproducible from a single training run. Across eight instances differing only in initialization, its range includes both signs and attention-based structural-core enrichment spans two to three orders of magnitude, so a single run can misstate both the magnitude and the direction of the quantity being reported.

The transferable contribution is a set of controls applicable to any cross-model attribution comparison: an enforced additivity check before attributions are interpreted, a verified replication of the forward pass before attention is extracted, a size-matched cross-molecule reference, a sign-based connectivity control, and replication across independently trained instances. Confirmation on a larger replicate set, extension to further attribution schemes, model families, and targets, and prospective testing of ranked candidates are the next steps.

Supplementary Materials

The following are available: per-instance CSV exports of all agreement, enrichment, and ranking statistics; a machine-readable JSON summary recording configuration, verification outcomes, and software versions; and the two Jupyter notebooks described in Section Software Architecture.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The primary screening data are those published by Wong et al. [1] and are available in the archived repository accompanying that study. Two Jupyter notebooks reproduce every result reported here: a main pipeline covering curation, splitting, model development, held-out evaluation, and candidate ranking; and a self-contained explainability notebook covering attribution extraction and validation, cross-model agreement, reference distributions, controls, and multiple-testing correction. The second requires no artefact from the first, and the two independently reproduce identical curation counts and a common reference AUPRC of 0.1233. All figures and tables are emitted by these notebooks with machine-readable summaries recording configuration, per-instance results, verification outcomes, and software versions.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Russian Science Foundation [grant no. 326152].

Authors' contributions

MS and OE. Resources, conceived the study, methodology, software, designed and implemented the computational pipeline, performed the analysis, prepared the figures and tables, and wrote the manuscript. AD and MK. Resources, wrote the main manuscript text, reviewed and edited the manuscript. BMS. Conceptualisation of the study, validation, formal analysis, resources, review and editing, visualization, supervision, project administration, funding acquisition.

All authors wrote and approved the final manuscript.

Acknowledgements

The authors thank ITMO University for computational resources.

References

1. Wong, F.; Zheng, E.J.; Valeri, J.A.; Donghia, N.M.; Anahtar, M.N.; Omori, S.; Li, A.; Cubillos-Ruiz, A.; Krishnan, A.; Jin, W.; et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* 2024, *626*, 177–185. https://doi.org/10.1038/s41586-023-06887-8
2. Stokes, J.M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N.M.; MacNair, C.R.; French, S.; Carfrae, L.A.; Bloom-Ackermann, Z.; et al. A deep learning approach to antibiotic discovery. *Cell* 2020, *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021
3. Liu, G.; Catacutan, D.B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J.C.; Chiappino-Pepe, A.; Syed, S.A.; Fragis, M.; Rachwalski, K.; et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat. Chem. Biol.* 2023, *19*, 1342–1350. https://doi.org/10.1038/s41589-023-01349-8
4. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 2010, *50*, 742–754. https://doi.org/10.1021/ci100050t
5. Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32. https://doi.org/10.1023/A:1010933404324
6. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785
7. Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* 2020, *63*, 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959
8. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; Gomes, J.; Geniesse, C.; Pappu, A.S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* 2018, *9*, 513–530. https://doi.org/10.1039/C7SC02664A
9. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 2019, *59*, 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237
10. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
11. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2020, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9
12. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* 2020, *2*, 573–584. https://doi.org/10.1038/s42256-020-00236-4
13. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 2019, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
14. Jain, S.; Wallace, B.C. Attention is not explanation. In *Proceedings of NAACL-HLT*, Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556. https://doi.org/10.18653/v1/N19-1357
15. Wiegreffe, S.; Pinter, Y. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, Hong Kong, China, 3–7 November 2019; pp. 11–20. https://doi.org/10.18653/v1/D19-1002
16. Sanchez-Lengeling, B.; Wei, J.; Lee, B.; Reif, E.; Wang, P.; Qian, W.; McCloskey, K.; Colwell, L.; Wiltschko, A. Evaluating attribution for graph neural networks. In *Advances in Neural Information Processing Systems 33*, Virtual, 6–12 December 2020; pp. 5898–5910.
17. Rao, J.; Zheng, S.; Lu, Y.; Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns* 2022, *3*, 100628. https://doi.org/10.1016/j.patter.2022.100628

18. Lamens, A.; Bajorath, J. Comparing explanations of molecular machine learning models generated with different methods for the calculation of Shapley values. Mol. Inform. 2025, 44, e202500067. https://doi.org/10.1002/minf.202500067
19. Jiménez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *J. Chem. Inf. Model.* 2021, *61*, 1083–1094. https://doi.org/10.1021/acs.jcim.0c01344
20. Saito, T.; Rehmsmeier, M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 2015, *10*, e0118432. https://doi.org/10.1371/journal.pone.0118432
21. Webber, W.; Moffat, A.; Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 2010, *28*, 20. https://doi.org/10.1145/1852102.1852106
22. Sheridan, R.P.; Feuston, B.P.; Maiorov, V.N.; Kearsley, S.K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1912–1928. https://doi.org/10.1021/ci049782w
23. Truchon, J.-F.; Bayly, C.I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. J. Chem. Inf. Model. 2007, 47, 488–508. https://doi.org/10.1021/ci600426e.
24. Madhyastha, P.; Jain, R. On model stability as a function of random seed. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 929–939.
25. Cvetković, D.; Mitrovic, M. Assessing the explainability of graph neural networks in random graphs classification task. J. Complex Netw. 2026, 14, cnag006. https://doi.org/10.1093/comnet/cnag006
26. Arends, B.; van Amsterdam, W.V.; van der Harst, P.; van Smeden, M.; van Es, R.; van de Leur, R.R. Signal or noise? Evaluating commonly used attribution methods for explaining deep neural networks in electrocardiogram classification. Eur. Heart J. Digit. Health 2026, 7, ztag038. https://doi.org/10.1093/ehjdh/ztag038
27. Roth, J.P.; Bajorath, J. Machine learning models with distinct Shapley value explanations decouple feature attribution and interpretation for chemical compound predictions. Cell Rep. Phys. Sci. 2024, 5, 102110. https://doi.org/10.1016/j.xcrp.2024.102110
28. Jeon, B. Systematic validation of graph neural network explanations against adverse outcome pathway reactive centers for skin sensitization. J. Cheminform. 2026, 18, 4. https://doi.org/10.1186/s13321-026-01249-8
29. Sheridan, R.P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J. Chem. Inf. Model. 2013, 53, 783–790. https://doi.org/10.1021/ci400084k
30. Landrum, G. RDKit: Open-Source Cheminformatics. Available online: https://www.rdkit.org (accessed on [date]).
31. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 1996, 39, 2887–2893. https://doi.org/10.1021/jm9602928
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 2011, *12*, 2825–2830.
33. Fey, M.; Lenssen, J.E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, LA, USA, 6 May 2019; arXiv:1903.02428.
34. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
35. Riniker, S.; Landrum, G.A. Similarity maps A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminform.* 2013, *5*, 43. https://doi.org/10.1186/1758-2946-5-43

36. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 1947, *18*, 50–60. https://doi.org/10.1214/aoms/1177730491
37. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 1995, 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.