LINEAR LOBOTOMY FOR ALIEN GRAPH REPRESENTATIONS: PROBING, ABLATION, AND RECOVERY OF LINEAR FEATURES IN GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Which graph properties does a graph neural network (GNN) use, and are they the ones a domain expert would name? Instance-level explanations for GNNs are well developed, but tools that answer this model-level question are not. We call a graph-theoretic property that a model encodes and uses, although the literature on the task does not name it, an *alien graph representation* (AGR), and we make the notion testable by fixing an expert set S_2 per task before looking at any result. We then propose *linear lobotomy*, a pipeline that tests the Linear Representation Hypothesis (LRH) at three levels: linear probing decides whether a property is *encoded* (Weak), concept erasure with matched random controls decides whether it is *used* (Strong), and converged retraining of the classifier head decides whether it is *exclusively* encoded along the erased direction (Exclusive). We compare three erasure operators (LEACE, iterative nullspace projection and a rank-one ablation) and show that the choice of operator, not the property, drives the size of an amnesic drop unless the share of the representation it deletes is reported. Across five datasets, a synthetic motif benchmark, two molecular and social benchmarks and two clinical fMRI connectivity tasks, AGR candidates are the rule at the Weak level, few survive the Strong level, and exclusivity tracks the architecture rather than the property: on the synthetic benchmark a Graph Convolutional Network recovers fully from every erasure and a Graph Isomorphic Network never does. On MUTAG the model relies on branching and size counts rather than the ring counts a chemist would name; on the depression task six spectral, resistance and clique descriptors outside the clinical vocabulary are candidate used AGRs; on the social benchmark, the graph features align with the expert vocabulary.

1 INTRODUCTION

Which graph properties does a trained GNN use, and are they the ones a domain expert would name? As models surpass human performance in more domains (Kwa et al., 2025; Bubeck et al., 2023), human supervision loses leverage (Vaccaro et al., 2024), and the question of what a model has actually learned becomes a question about its internal representations, to be answered at several levels of analysis (Marr, 1984; He et al., 2024). For GNNs, which model structured data in neuroscience, chemistry and social systems, explanation methods are predominantly instance-based, returning salient subgraphs or node features for one prediction (Ying et al., 2019; Luo et al., 2020; Vu & Thai, 2020; Yuan et al., 2021; Schlichtkrull et al., 2021; Lucic et al., 2022; Yuan et al., 2023; Longa et al., 2025). Model-level explainers summarise a trained GNN with prototype graphs, logic rules or neuron-level concepts (Yuan et al., 2020; Saha et al., 2022; Azzolin et al., 2023; Xuanyuan et al., 2023). As Pelletreau-Duris et al. (2025) note, these methods scale poorly with graph size, are not task-agnostic, since each targets either node or graph classification, and explain the model as a whole rather than layer by layer. None of them reads the hidden representations in terms of graph properties, or uses the tools of mechanistic interpretability, probing and causal intervention, to test which of those properties the model actually uses.

Neural networks can rely on features that are predictive yet unintuitive to humans, as adversarial examples show (Ilyas et al., 2019). Such *alien representations* (Hubinger, 2019; Bereska & Gavves, 2024) are causally relevant to the prediction but absent from the vocabulary in which experts reason

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

about the task; they limit interpretability and are, at the same time, a possible source of discovery (Bereska & Gavves, 2024). Alien is thus relative to a domain, not to human understanding in general, and we use this to make the notion testable. Instead of probing the concepts a domain expert would propose, we probe the vocabulary of graph theory, a mathematical language defined for every graph and independent of any task, and test each property whether or not the literature on the task expects it. A property that the model uses but that this literature does not name is an *alien graph representation* (AGR): nameable in mathematics, unexpected in the domain. Detaching the probed vocabulary from expert priors is what allows the analysis to enrich the domain vocabulary rather than merely confirm it; directions that no mathematical vocabulary names remain outside its reach and call for unsupervised methods such as sparse dictionary learning (Cunningham et al., 2023). Following Pelletreau-Duris et al. (2025), we thus take graph-theoretic properties as the representational atoms of the analysis, and we adopt the Linear Representation Hypothesis (LRH), which posits that features are encoded as directions in activation space (Smolensky, 1986; Mikolov et al., 2013; Park et al., 2025; Nanda et al., 2023), on the evidence they report that diagnostic classifiers recover graph properties from GNN embeddings. The LRH has known limits where a model relies on nonlinear or multidimensional structure (Engels et al., 2025), so we test it rather than assume it.

The test has three levels, and we call the pipeline *linear lobotomy* (Figure 1). A linear probe (Belinkov, 2022) decides whether a property is *encoded*. Concept erasure (Ravfogel et al., 2020; Elazar et al., 2021; Belrose et al., 2023), scored against random ablations of the same size, decides whether the frozen classifier *uses* it (Geiger et al., 2021). Retraining the classifier head on the erased representation, to convergence and over seeds, decides whether the property was encoded *exclusively* along the erased direction. A property that passes a level while lying outside the expert set S_2 of its task is an alien graph representation (AGR) at that level. Our contributions are:

- **A testable definition of alien graph representations** (Section 2.4). An expert set S_2 is fixed from the literature for each task before any result is read, and for the clinical tasks validated by a neuroscientist (Appendix B), each property faces three nested tests with thresholds fixed in advance (Table 1), and an alien name that a linear model on S_2 reproduces is flagged as a proxy rather than counted. Every verdict is tabulated per property (Appendix C).
- **The linear lobotomy pipeline** (Section 2). It contrasts three erasure operators, a minimal and complete one (LEACE), a complete but destructive one (INLP) and a minimal but incomplete one (rank-one), and reports for INLP the share k/d of the representation it deletes, which decides whether its drops can be read at all (Table 9). Recovery is measured only after retraining has converged, since a short budget imitates exclusivity.
- **Evidence on five datasets from three domains** (Section 4). On the synthetic Grid-House benchmark, erasing the motif count is the intervention that costs the most in every model, and whether the model can do without it depends on the architecture (GCN recovers fully, GIN never). On MUTAG the model uses branching and size counts rather than ring counts; on IMDB-B nothing alien survives; on two fMRI tasks a minimal erasure finds nothing, and a subspace lesion exposes six candidate used AGRs on depression and none on autism.

2 LINEAR LOBOTOMY

2.1 THREE QUESTIONS, THREE TESTS

The LRH motivates three questions about a feature y and a frozen representation $\mathbf{H} \in \mathbb{R}^{n \times d}$ (here the pooled graph embedding). **(i)** Can y be linearly decoded from $\mathbf{H}$? We fit a linear probe $f(\mathbf{e}) = \mathbf{w}^\top \mathbf{e}$ and report its held-out coefficient of determination R^2 (Belinkov, 2022). **(ii)** Does the classifier rely on y? We erase the linear footprint of y and re-evaluate the frozen classifier, as in nullspace projection and amnesic probing (Ravfogel et al., 2020; Elazar et al., 2021). **(iii)** Can downstream layers compensate? We retrain the head on the erased representation. We call the three variants of the hypothesis *Weak*, *Strong* and *Exclusive*. The first two are well studied, through probing controls (Hewitt & Liang, 2019; Pimentel & Cotterell, 2021) and causal interventions (Ravichander et al., 2021; Geiger et al., 2021); the third has received little attention. Keeping the probe linear inherits the selectivity argument for linear probes, and the erasure step answers the standing objection that decodability does not entail use.

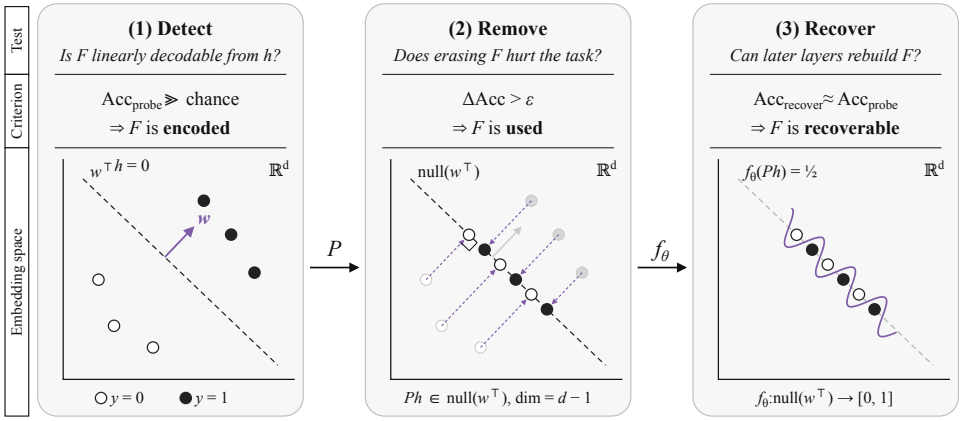


Figure 1: The linear lobotomy pipeline. A property dictionary S_1 is computed for every graph; a linear probe reads each property from the pooled embedding (Weak); an erasure operator removes its linear footprint and the frozen classifier is re-evaluated against random ablations of the same size (Strong); the head is retrained on the erased embedding until convergence (Exclusive). Properties outside the expert set S_2 that pass a level are alien graph representations at that level.

2.2 Concept erasure: a minimal edit and two alternatives

How the footprint is removed decides what a drop means, so we compare three operators. Our primary intervention is *least-squares concept erasure* (LEACE) (Belrose et al., 2023): a closed-form affine map that makes y linearly unpredictable while moving the representation as little as possible in the least-squares sense. For a scalar property it removes one direction in the whitened basis, so the edit is complete and minimal. The second operator is *iterative nullspace projection* (INLP): at iteration t a probe $\mathbf{w}^{(t)}$ is fitted on $\mathbf{H}^{(t)}$ and projected out, $\mathbf{H}^{(t+1)} = \mathbf{H}^{(t)}\big(\mathbf{I} - \mathbf{w}^{(t)}\mathbf{w}^{(t)\top}/\|\mathbf{w}^{(t)}\|^2\big)$, until the training R^2 falls below $\tau = 10^{-2}$ or plateaus. It is complete but not minimal: because it ignores the covariance between dimensions it deletes a k-dimensional subspace, and we report k/d, the share of the representation removed. Where k/d is small the intervention is a subspace-level lesion whose drops remain interpretable; where it approaches the rank of the representation, every property damages the classifier alike. In our runs k/d is 5–7% on the fMRI data, 40–62% on GridHouse and 84–98% on the two TU benchmarks (Table 9), so we read INLP results on the fMRI data and not elsewhere. The third operator is the naive *rank-one* ablation of the probe direction itself: minimal but not complete, since the property stays decodable from the residual (Table 8), so a drop shows that the classifier depends on that axis, not on the property.

For each operator the amnesic accuracy change is $\Delta = \mathrm{Acc}_{\mathrm{erased}} - \mathrm{Acc}_{\mathrm{baseline}}$ (negative when accuracy drops). Erasing k directions degrades a classifier whatever they are, so we compare Δ with a null distribution of M random k-dimensional ablations: for each trial we draw a Gaussian $\mathbf{R}^{(m)} \in \mathbb{R}^{d\times k}$, take its orthonormal factor $\mathbf{Q}^{(m)}$, apply $\mathbf{I} - \mathbf{Q}^{(m)}\mathbf{Q}^{(m)\top}$ to the test embeddings and score the frozen classifier. With μ_{rand} and σ_{rand} the mean and standard deviation of the random changes, $z = (\mu_{\mathrm{rand}} - \Delta)/\sigma_{\mathrm{rand}}$ measures how much more the targeted subspace carried than an equally sized random slice. Because the random ablations are scored by the same frozen classifier on the same embeddings, they also absorb generic side effects such as pushing embeddings off the training distribution: the comparison of interest is targeted versus random, not targeted versus zero. As a stricter robustness control we also draw random directions that remove the same embedding variance as the targeted erasure (z_{var}), and we report a one-sided McNemar test on the discordant test graphs with Holm correction over the properties tested.

2.3 Recovery

To separate a displaced decision boundary from removed information we freeze the encoder and retrain the classifier head on the erased embeddings, either from a fresh initialisation or from the

Level	Passes when	Locked thresholds
Weak	encoded	held-out probe $R^2 \geq 0.10$
Strong	used	erasure converged (residual $R^2 \leq 0.05$); $\Delta < 0$; $z \geq z^* = \Phi^{-1}(1 - 0.05/m)$ against the dimension-matched control, m = properties tested ($z^* = 3.07$ for $m = 47$); z_{var} reported as robustness
Exclusive	not recoverable	Strong holds; every recovery run converged; no seed of any mode reaches $\rho = 100\%$; evaluated only for drops of at least three test graphs
S_2 proxy	—	a linear model on S_2 predicts the property with held-out $R^2 \geq 0.9$; excluded from the AGR count

Table 1: The protocol. A property outside S_2 is an AGR candidate at Weak and a used AGR at Strong; Exclusive characterises how it is encoded. All statistics are medians over five seeds.

original weights with the last n layers unfrozen, up to the full head. Recovery completeness

$$\rho = \frac{\text{Acc}_{\text{recovered}} - \text{Acc}_{\text{erased}}}{\text{Acc}_{\text{baseline}} - \text{Acc}_{\text{erased}}} \times 100\%$$

is high when the drop was a distribution shift and low when the information is gone. The measurement is only meaningful once retraining has converged: on Grid-House the same ablation gives $\rho = 91.5\%$ after 300 epochs, 97.7% after 1000 and 99.4% after 3000, so a fixed short budget reports exclusivity that longer training dissolves. Every recovery run therefore trains to a validation plateau (3000-epoch cap, patience 50 evaluations), is repeated over five seeds, and records a convergence flag; we report the best mode's median ρ with the seed range.

2.4 Operationalising alien graph representations

We distinguish S_1, the vocabulary of graph theory, probed without a priori knowledge of which properties matter (in practice the property dictionary of Appendix A); the expert set $S_2(\text{task}) \subset S_1$, the properties that the literature on the task leads us to expect (Appendix B, which derives the FC set from the connectome literature on autism and depression and reports its validation by a professor of neuroscience); and the latent space $\mathcal{H}$ of a GNN. Our research question is: *which properties in $S_1 \setminus S_2$ does a GNN encode and use, and how are they encoded, as measured by the three levels of the LRH?* Each property faces the three nested tests of Table 1, whose thresholds were fixed before the corrected runs were read. A property outside S_2 that passes Weak is an **AGR candidate**; one that also passes Strong is a **used AGR**, which is the key test; Exclusive is reported as a diagnostic of the LRH itself, not as a condition for being alien: full recovery shows that the label information is not confined to the erased linear subspace, which is a result about the hypothesis, not about the candidate. Properties in S_2 that pass the same tests are *expert-aligned*. Because the dictionaries are highly collinear (several properties are exact linear functions of the others, Table 20), an alien name is flagged as an S_2 *proxy* when a linear model on the S_2 properties predicts it with held-out $R^2 \geq 0.9$: its name is alien, its information is not, and it is never counted as a genuine AGR. Since S_1 contains only named properties, AGRs are a conservative, testable proxy for the broader notion of alien representations, which also covers directions that no vocabulary names.

3 Experimental setup

Datasets (Table 2). The five datasets span three levels of prior knowledge. Grid-House is a controlled proof of concept whose answer is known by construction; MUTAG and IMDB-B are standard graph-classification benchmarks (Morris et al., 2020) with an established expert vocabulary; and in functional connectivity the role of network topology in neurodevelopmental and psychiatric disorders is still poorly understood, because graph theory describes how brain regions specialise (segregation) better than how they communicate (integration) (Bullmore & Sporns, 2009; Farahani et al., 2019). There an alien representation could be a discovery rather than an artefact. *Grid-House* (Pelletreau-Duris et al., 2025; Agarwal et al., 2023) is a synthetic benchmark with a known solution. Each graph is a Barabási–Albert tree ($m = 1$, so the backbone has no cycle) to which

Dataset	Graphs (pos/neg)	Test	$\lvert S_1\rvert$	$\lvert S_2\rvert$ (+b.)	Model	Acc. (%)
Grid-House (synthetic motifs)	1987 (989/998)	398	10	3 (+1)	GCN, GIN, $\pm$wd	90.2–100
MUTAG (molecules)	188 (125/63)	38	16	2 (+5)	GIN_{L_2}	86.8 ± 5.4
IMDB-B (social)	1000 (500/500)	200	20	8 (+6)	GIN_{L_2}	70.3 ± 3.0
ASD (ABIDE, fMRI)	1099 (528/571)	55	47	12 (+9)	GCN	65.5
MDD (REST-meta-MDD, fMRI)	1604 (828/776)	81	47	12 (+9)	GCN	59.3

Table 2: Datasets, property dictionaries and models. $|S_1|$ counts the non-constant properties probed; $|S_2|$ the expert properties and, in brackets, borderline ones that are reported but never counted as alien (Appendix B). Accuracies are on the test split of the published checkpoints (Grid-House, fMRI) or mean $\pm$ s.d. over five seeds (TU).

motifs are attached by single bridge edges: a negative graph carries one house (one 4-cycle, one triangle) or one 3×3 grid (four 4-cycles), a positive graph one of each. Graphs have 20–29 nodes in both classes, constant node features, and isomorphic duplicates are removed. Every cycle lies inside a motif, so the 4-cycle count (1 or 4 versus 5), the cycle rank $|E| - |V| + 1$ (2 or 4 versus 6) and the summed square-clustering coefficient probed as `num_squares` (Lind et al., 2005) each determine the label exactly, whereas edge count and density do not (best single-threshold accuracy 65% and 68%). The task has several exact solutions; which one a model uses is an empirical question. *MUTAG* and *IMDB-B* are standard molecular and social benchmarks. *ASD* and *MDD* are resting-state fMRI cohorts from ABIDE (Di Martino et al., 2014) and REST-meta-MDD (Yan et al., 2019): 116 AAL regions, an undirected edge where the Pearson correlation between BOLD series exceeds 0.4 (520 ± 145 and 517 ± 118 edges), each node carrying its row of the full correlation matrix as features. Properties are computed on the binarised graph. Here the GNN is deployed as a scientific model in the sense of Cichy & Kaiser (2019): its representations are interrogated for structure in the data, and an erasure moves a property from *correlated* with the outcome to *used* by the mapping from structure to label.

Models and protocol. On Grid-House we train a **GCN** (Kipf & Welling, 2017) (four layers, 60 channels, max pooling, three-layer head) and a **GIN** (Xu et al., 2019) (three layers, 30 channels, mean pooling, two-layer head; GIN4 in the results), each without and with weight decay (10^{-5} for GCN_L2, 10^{-2} for GIN4_L2), Adam at 10^{-3}, on a fixed 80/20 split. GIN's injective sum aggregation makes it as discriminative as the 1-WL test, the upper bound for message passing; this bounds what it can represent, not what training finds, and excludes properties beyond 1-WL such as general substructure counts (Chen et al., 2020). On the fMRI data the GCN has five layers (128 channels, batch normalisation, max pooling, two-layer head), is trained for 450 epochs with weight decay 10^{-4}, and is evaluated on a single stratified 95/5 split, so these results carry no fold variability. On MUTAG and IMDB-B a GIN_{L_2} (three layers, 64 channels, mean pooling, two-layer head) is trained over five seeds on a stratified 80/20 split with validation-based selection. Properties are computed with `networkx` on the unweighted graph: 10 for Grid-House, 21 for the TU benchmarks and 50 for fMRI (47 non-constant; Appendix A). Probes are closed-form ridge regressions ($\lambda = 10^{-3}$) on the post-pooling embedding, fitted on the training graphs and scored on the test graphs. Random controls use 100 ablations per seed. Compute is listed in Appendix E.

4 Results

For each property we ask the three questions of Section 2.1: is it encoded (R^2), is it used (Δ, z) and is it recoverable (ρ). Per-property verdicts for every run are in Appendix C; Figure 2 plots encoding against use for all six models.

4.1 Grid-House: which cue, and can the model do without it

Encoded. All ten properties are linearly decodable, and the three cycle-structure measures lead in every variant: held-out $R^2 = 0.88$–0.93 for `cycle_rank`, 0.87–0.90 for `num_squares` and 0.87–0.91 for `num_4_cycles`, against 0.16–0.52 for size and distance descriptors, as the construction predicts.

Model	Readout	Acc.	num_squares				Properties passing		
		%	Δ pt	z_{dim}	z_{var}	ρ %	Strong	Excl.	AGR
GCN	Max, —	90.2	−34.9	2.43	0.28	101	0	0	0
GCN_L2	Max, wd 10^{-5}	96.7	−45.0	2.63	0.67	97	2	1	0
GIN4	Mean, —	100.0	−46.7	5.73	0.34	61	4	4	1
GIN4_L2	Mean, wd 10^{-2}	98.7	−50.8	19.00	0.55	63	10	10	5

Table 3: Grid-House under LEACE erasure (one direction, residual $R^2 \approx 0$), median over five seeds on the published checkpoints. Δ is the accuracy change after erasing num_squares; z_{dim} and z_{var} compare it with random ablations matched in dimension and in removed variance; ρ is the best retraining mode after convergence. The last three columns count, over the ten probed properties, those passing Strong, those partially exclusive, and the genuine AGRs (outside S_2 and not an S_2 proxy). Full tables: Appendix C, Tables 10–13.

Used. Erasing one direction with LEACE separates the properties (Table 3). num_squares costs 34.9 points on GCN and 50.8 on GIN4_L2, while the least-used property costs under two points on the same model. The ordering is the same in all four variants and the cycle-structure group leads, so the models use a cue aligned with cycle structure rather than with size. Two caveats keep this suggestive rather than conclusive: against the dimension-matched control the margins are clear for GIN ($z = 5.7$ and 19.0) but modest for GCN ($z = 2.4$–2.6), and against the variance-matched control no margin on any variant reaches $z = 1$. cycle_rank gives the largest drop almost everywhere, but a linear model on the motif counts predicts it perfectly ($R^2 = 1.00$), so it is an S_2 proxy, not a new cue. Beyond the motif counts, GIN4_L2 uses density, num_edges, avg_path_len and num_nodes ($z = 2.6$–13.6; the largest-component size is identical to num_nodes on these connected graphs) and GIN4 uses density; these are its genuine AGRs. The GCN variants use no alien property.

Recoverable. Retraining the head to a validation plateau over five seeds separates the architectures rather than the properties. The GCN recovers fully whatever is erased ($\rho = 97$–103%): its code is redundant, and at most one property (num_squares on GCN_L2, $\rho = 97\%$) sits marginally below full recovery. GIN does not recover ($\rho = 22$–63%, no seed reaching 100%), so ten of its properties are partially exclusive under weight decay and four without it. The contrast is not a budget artefact: every run converged, and on the same ablation a short budget would have reported $\rho = 91.5\%$ where convergence gives 99.4%.

What the architectures do differently. We offer an interpretation that fits these numbers but that we have not tested. The GCN's normalised aggregation with max readout appears to spread the class signal over several directions, so erasing any one is survivable; GIN's sum readout preserves neighbourhood cardinality (Xu et al., 2019) and behaves as if the cue were concentrated, so the same erasure costs more and cannot be retrained away. Weight decay sharpens whichever regime is present: it raises GCN accuracy by 6.5 points without creating a dependence, and on GIN multiplies the margin over the control from $z = 5.7$ to 19.0. Confirming the mechanism would need interventions on the aggregation itself.

4.2 Functional connectivity: a distributed code

The GCN reaches 65.5% on ASD (95% CI 52.3–76.6; majority class 52.7%) and 59.3% on MDD (48.4–69.3; majority 51.9%). Its node features are correlation rows, so this does not isolate topology, and we make no claim about what the edge weights would add.

Encoded. Across the 47 properties, held-out R^2 spans -0.32 to 0.78 (ASD) and -0.24 to 0.77 (MDD). Size and density proxies are the most decodable (num_edges, avg_degree, adjacency_energy, $R^2 = 0.73$–0.78); assortativity is not decodable on either task. Twenty of the 26 alien properties pass Weak on ASD and 24 on MDD: the representation contains a great deal of structure outside the clinical vocabulary.

Used: a minimal edit finds nothing, a subspace lesion does. Removing one direction with LEACE costs at most 3.6 points on ASD and 1.2 on MDD, one or two test graphs, and no drop clears its control: no single property direction is load-bearing (Tables 16–17). Erasing the property's whole linear subspace is a different intervention, and on these datasets an admissible one: INLP needs seven to nine of the 128 dimensions ($k/d = 5$–7%, Table 9), so it removes a slice rather than the representation. Under that lesion the two tasks separate (Table 18). On MDD, six alien properties are decodable, clear the Bonferroni threshold against the dimension-matched control and are not S_2 proxies: `algebraic_connectivity` ($R^2 = 0.54$, $\Delta = -12.3$), `num_maximal_cliques` (0.52, -11.1), `cheeger_upper` (0.52, -9.9), `eff_resistance_max` (0.31, -9.9) and `pagerank_centralization` with its logarithm (0.23, -9.9), at $z = 3.3$–3.9. These are candidate used AGRs: spectral, resistance and clique descriptors that the clinical literature does not use. Those properties might be good candidates for further investigation and help neuroscientist understanding of the brain's functional connectivity. Ten further properties pass but are expert-aligned (`avg_path_len`, `avg_betweenness`, `small_world_index`), borderline (`diameter`, `radius`) or proxies of S_2 (`adjacency_energy`, `eff_resistance_mean`, `hkt_medium`, `orc_mean`). On ASD only `small_world_index` and `orc_min` clear the control, and neither is decodable on held-out graphs ($R^2 = -0.32$ and -0.13): INLP removed directions that fit the property on the training graphs only, so the drop is not evidence that the classifier uses it, and the first is in S_2 in any case. The ASD classifier reads its label off no property we can name.

Recoverable, and two cautions. Every MDD property that passes Strong recovers fully or better once the head is retrained ($\rho = 88$–156%): no fMRI property is exclusively encoded. Two cautions attach to the MDD result. Against the variance-matched control every margin falls below $z = 1$, and no property survives Holm correction over the 47 tested. With 81 test graphs one graph is 1.2 points and a 9.9-point drop is eight graphs; the LEACE effects sit at that resolution floor and every bootstrap interval for a drop contains zero. The fMRI results are therefore a cautionary case: a probe reports rich structure, a large ablation reports large effects, a minimal one reports almost nothing, and the sample is too small to separate the last two.

4.3 Molecular and social graphs

MUTAG is the clearest alien representation in this paper. The expert vocabulary for mutagenicity is ring structure, so S_2 holds the ring and circuit counts. The model instead leans on branching and size: erasing `num_claws` (3-stars) costs 21.1 points, `num_cliques` and `num_wedges` 28.9, and `diameter` 13.2, each clearing the dimension-matched control by more than 14 standard deviations, while the ring counts are decodable ($R^2 = 0.96$) but cost less (-10.5). Nine alien properties pass Strong and none is an S_2 proxy (Table 14). Their exclusivity is unresolved: retraining converged in only three or four of five seeds, and we report ρ only where it converged. On IMDB-B the picture is the opposite and reassuringly dull: the properties that matter are `density` and `avg_path_len` (-5.5 points, $z = 10.6$), both standard in social network analysis, and the three alien properties that pass Strong are S_2 proxies (Table 15). Where the expert vocabulary matches the task, the lobotomy finds no alien representation, which is the negative control the method needs.

4.4 Do alien graph representations survive the three tests?

Applying Table 1 to every run gives a graded answer (Table 7). *Candidates (Weak).* Alien candidates are everywhere: every property outside S_2 is decodable on Grid-House and the TU benchmarks, and 20 of 26 on ASD and 24 of 26 on MDD. Encoding structure that the expert vocabulary does not name is the norm. *Used AGRs (Strong).* Far fewer are used: nine genuine AGRs on MUTAG, five on GIN4_L2 and one on GIN4 for Grid-House, none for the GCN variants, none on IMDB-B, none on ASD, and six candidates on MDD under the subspace lesion. *Limits of the LRH (Exclusive).* Where retraining converged, exclusivity tracks the architecture: the GCN recovers from every erasure ($\rho = 97$–103%), whereas every Strong-level property of GIN4 and GIN4_L2 is partially exclusive ($\rho = 22$–63%). The linear direction is one route to the label in the GCN and the only linear route we could find in the GIN. A caution runs through all of this: against the variance-matched control no margin on any dataset reaches $z = 1.1$, so the used AGRs are supported by one control and not by the other; that control asks whether a drop is property-specific rather than larger than sampling noise, but it is sensitive to how the random directions are drawn.

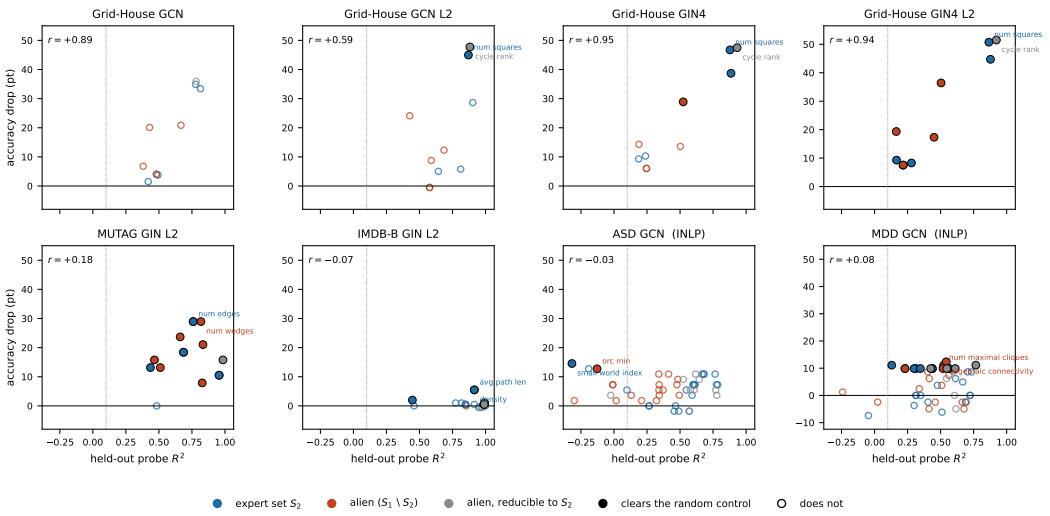


Figure 2: Encoding against use, one panel per trained model. Each point is a property: held-out probe R^2 (Weak) against the accuracy the classifier loses when it is erased (Strong), plotted upwards. Colour gives the expert set (in S_2, outside it, or an S_2 proxy); filled markers clear the dimension-matched control at the Bonferroni threshold. The dotted line is the decodability threshold and r the within-panel correlation. Grid-House and the TU benchmarks use LEACE; the fMRI panels use the subspace erasure, which there removes 5–7% of the representation. On Grid-House decodability predicts use ($r = +0.89$ to $+0.95$); elsewhere the relation is weak or absent ($r = +0.18$ on MUTAG, $|r| \leq 0.08$ on the social and clinical data), so a probe score says little about whether the model uses the property.

5 Discussion

Three regimes. The datasets expose three regimes of what a GNN encodes and uses. Where the expert vocabulary fits the task (IMDB-B), nothing alien survives. Where it does not (MUTAG), the model uses named properties a chemist would not list, and the lobotomy returns them as properties the model provably depends on, whether or not the literature expected them, which is what concept-based and neuro-symbolic methods need as input (Koh et al., 2020; Kim et al., 2018; Azzolin et al., 2023). Where the signal is weak and the sample small (fMRI), the answer depends on the operator: a minimal erasure finds nothing and a subspace lesion finds candidates. The MDD candidates, which point to global integration rather than local clustering, in line with reports of altered global topology in depression (Zhang et al., 2011), are hypotheses for network neuroscience to test on larger held-out cohorts, not findings. The property that matters is thus dataset-specific, while whether the model can do without it is architecture-specific.

What the pipeline adds to probing. Decodability and use are dissociable (Figure 2): outside the synthetic benchmark a probe score does not predict the cost of erasure, so probing alone would misreport what the model uses. A large amnesic drop is in turn necessary but not sufficient evidence of exclusive encoding: the same drop is fully recoverable in the GCN and not in the GIN, and a short retraining budget would have called both exclusive. An amnesic result is only as strong as the evidence that its erasure converged and that the retraining did, so the residual R^2, k/d and the convergence flag belong beside every drop.

Limitations. The used AGRs on MDD and the Grid-House GCN margins rest on the dimension-matched control and fail the variance-matched one; with 55 and 81 test graphs, no fMRI drop survives Holm correction. The fMRI classifier uses correlation rows as node features, so nothing here isolates topology. S_1 contains only named graph properties and S_2 omits some standard clinical measures (efficiency, rich-club, participation coefficients), so the expert side of the test is incomplete. The dictionaries are collinear, and the S_2-proxy flag, not the size of a drop, carries the attribution. The architectural mechanism we propose for the GCN/GIN contrast is untested.

6 Related work

Alien representations. Adversarial examples demonstrate that models can rely on features that are highly predictive yet inaccessible to human perception (Ilyas et al., 2019). Such *alien representations* (Bereska & Gavves, 2024; Hubinger, 2019) are internal variables causally relevant to a model's predictions but unaligned with human or domain-specific vocabularies. Recent literature suggests these representations may manifest as distributed geometric structures rather than discrete, nameable concepts (Olah et al., 2018; Elhage et al., 2022; Hubinger, 2019). We make a simpler claim: moving the probed vocabulary from the domain to a more abstract one, here graph theory, already suffices to find features that are nameable yet unaligned with the domain's intuitions. Beyond posing challenges for interpretability, alien representations may also support scientific discovery by capturing predictive structures outside predefined human abstractions (Bereska & Gavves, 2024; Cichy & Kaiser, 2019).

Graph representations and GNN explainability. GNNs are the dominant paradigm for learning on graph-structured data (Kipf & Welling, 2017; Joshi, 2025), with applications ranging from neuroscience and biology to recommendation systems and chip design (Mirhoseini et al., 2021; Dauparas et al., 2022). Existing GNN explainability methods are predominantly instance-based, focusing on salient nodes, edges, or subgraphs (Ying et al., 2019; Luo et al., 2020; Yuan et al., 2021; Schlichtkrull et al., 2021; Yuan et al., 2023; Longa et al., 2025). Model-level explainers summarise a trained GNN through generated prototype graphs or logic combinations of learned concepts (Yuan et al., 2020; Azzolin et al., 2023; Xuanyuan et al., 2023), but they do not test whether specific graph-theoretic properties are encoded in the representation and used by the classifier. More recent work studies the inductive biases of GNNs through graph-theoretic properties (Pelletreau-Duris et al., 2025), suggesting that latent graph representations increasingly encode task-relevant structural information across layers. Message-passing architectures are well understood computationally (Gilmer et al., 2017; Xu et al., 2019; Veličković et al., 2018); far less is known about how graph properties are geometrically organised within latent spaces, or which representations causally drive predictions (Geiger et al., 2021). Building on prior probing work for GNNs (Akhondzadeh et al., 2023; Pelletreau-Duris et al., 2025), we investigate whether latent directions correspond to graph-theoretic properties outside current expert vocabularies.

Probing and causal interventions. Probing methods assess whether specific properties are linearly decodable from neural representations (Belinkov, 2022). However, probing remains fundamentally correlational: a high probe score indicates that a model encodes a feature, but not that it is causally used for prediction. To address this limitation, intervention-based methods project representations onto the nullspace of learned probes to selectively remove encoded information (Ravfogel et al., 2020; Elazar et al., 2021). Building on these approaches, we study whether removing graph-theoretic directions affects downstream predictions and whether this removed information can subsequently be recovered through downstream layers.

7 Conclusion

Linear lobotomy asks of every named graph property whether a GNN encodes it, uses it and can do without it, and answers against an expert set fixed in advance. Alien candidates are the rule, used alien properties the exception, and exclusivity a property of the architecture rather than of the feature. The surviving properties are a deliverable in themselves: a vocabulary of graph-theoretic symbols that the model demonstrably uses, on which a concept bottleneck or a logic layer could be built and compared with the original model; the failure of the Exclusive test in the GCN suggests such a model would have to allow redundancy between concepts. More broadly, a GNN constrained to topology is a scientific model of the system it classifies, and the pipeline turns its representations into testable hypotheses about which structural features carry a macroscopic outcome.

AI USE STATEMENT

In this work we used generative AI tools for editing and polishing the text, and for assistance with writing and reviewing analysis and plotting scripts. We have not used generative AI tools for research ideation, for the design of the method or of the experiments, or for generating results, and the remaining required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work: every AI-assisted script was checked by an author against the stored run outputs, and every number in the paper is generated by a script from those outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The clinical experiments use two publicly released, de-identified resting-state fMRI consortia (ABIDE and REST-meta-MDD) under their data-use terms; no new human data were collected. The MDD candidates are reported as hypotheses about a classifier that is only marginally above chance and must not be read as diagnostic markers. We see no other ethical concern.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix E give the architectures, splits, optimisation and compute; Appendix A defines every probed property and asserts that the documented set equals the set the code computes; Appendix B fixes the expert sets; Appendix C tabulates every per-property statistic behind the claims, with the thresholds of Table 1. Code, the corrected pipeline and the run outputs are available in supplementary materials.

REFERENCES

Sophie Achard and Ed Bullmore. Efficiency and cost of economical brain functional networks. *PLoS computational biology*, 3(2):e17, 2007.

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, March 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-01974-x. URL https://www.nature.com/articles/s41597-023-01974-x. Number: 1 Publisher: Nature Publishing Group.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, and Aleksandar Bojchevski. Probing Graph Representations. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 11630–11649. PMLR, April 2023. URL https://proceedings.mlr.press/v206/akhondzadeh23a.html. ISSN: 2640-3498.

Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of GNNs via logic combination of learned concepts. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712, 2023. URL https://arxiv.org/abs/2303.12712.

Edward Bullmore and Olaf Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. doi: 10.1038/nrn2575.

Marco Catani, Derek K Jones, and Dominic H Ffytche. Perisylvian language networks of the human brain. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 57(1):8–16, 2005.

Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Radoslaw M. Cichy and Daniel Kaiser. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4):305–317, April 2019. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2019.01.009. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(19)30034-8.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/10.1126/science.add2187.

A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, Jun 2014. ISSN 1476-5578. doi: 10.1038/mp.2013.78. URL https://doi.org/10.1038/mp.2013.78.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, March 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00359. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00359/98091/Amnesic-Probing-Behavioral-Explanation-with.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs].

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Farzad V. Farahani, Waldemar Karwowski, and Nichole R. Lighthall. Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review. *Frontiers in Neuroscience*, 13:585, June 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00585. URL https://www.frontiersin.org/article/10.3389/fnins.2019.00585/full.

Dominic H Ffytche and Marco Catani. Beyond localization: from hodology to function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):767–779, 2005.

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

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Zhonghao He, Jascha Achterberg, Katie Collins, Kevin Nejad, Danyal Akarca, Yinzhu Yang, Wes Gurnee, Ilia Sucholutsky, Yuhan Tang, Rebeca Ianov, George Ogden, Chole Li, Kai Sandbrink, Stephen Casper, Anna Ivanova, and Grace W. Lindsay. Multilevel Interpretability Of Artificial Neural Networks: Leveraging Framework And Methods From Neuroscience, August 2024. URL http://arxiv.org/abs/2408.12664. arXiv:2408.12664 [cs].

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1275/.

Kirsten Hilger, Matthias Ekman, Christian J. Fiebach, and Ulrike Basten. Efficient hubs in the intelligent brain: Nodal efficiency of hub regions in the salience network is associated with general intelligence. *Intelligence*, 60:10–25, 2017. ISSN 1873-7935. doi: 10.1016/j.intell.2016.11.001. Place: Netherlands Publisher: Elsevier Science.

Evan Hubinger. Chris Olah's views on AGI safety. AI Alignment Forum, 2019. URL https://www.alignmentforum.org/posts/X2i9dQQK3gETCyqh2/chris-olah-s-views-on-agi-safety.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 125–136, 2019.

Chaitanya K. Joshi. Transformers are Graph Neural Networks, June 2025. URL http://arxiv.org/abs/2506.22084. arXiv:2506.22084 [cs].

Rex E Jung and Richard J Haier. The parieto-frontal integration theory (p-fit) of intelligence: converging neuroimaging evidence. *Behavioral and brain sciences*, 30(2):135–154, 2007.

Amirali Kazeminejad and Roberto C Sotero. Topological properties of resting-state fmri functional networks improve machine learning-based autism classification. *Frontiers in neuroscience*, 12: 1018, 2019.

Christopher L Keown, Michael C Datko, Colleen P Chen, Jose Omar Maximo, Afrooz Jahedi, and Ralph-Axel Müller. Network organization is globally atypical in autism: a graph theory study of intrinsic functional connectivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(1):66–75, 2017.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=CGNJL6CeV0.

Nicolas Langer, Andreas Pedroni, Lorena R. R. Gianotti, Jürgen Hänggi, Daria Knoch, and Lutz Jäncke. Functional brain network efficiency predicts intelligence. *Human Brain Mapping*, 33(6): 1393–1406, June 2012. ISSN 1097-0193. doi: 10.1002/hbm.21297.

Samuel JJ Leistedt, Nathalie Coumans, Martine Dumont, Jean-Pol Lanquart, Cornelis J Stam, and Paul Linkowski. Altered sleep brain functional connectivity in acutely depressed patients. *Human brain mapping*, 30(7):2207–2219, 2009.

Pedro G. Lind, Marta C. González, and Hans J. Herrmann. Cycles and clustering in bipartite networks. *Physical Review E*, 72(5):056127, 2005. doi: 10.1103/PhysRevE.72.056127.

Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Liò, Bruno Lepri, and Andrea Passerini. Explaining the explainers in graph neural networks: a comparative study. *ACM Computing Surveys*, 57(5):120:1–120:37, 2025. doi: 10.1145/3696444.

Anton Lord, Dorothea Horn, Michael Breakspear, and Martin Walter. Changes in community structure of resting state functional connectivity in unipolar depression. *no journal*, 2012.

Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v151/lucic22a.html.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 1984.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nova, Jiwoo Pak, Andy Tong, Kavya Srinivasa, William Hang, Emre Tuncer, Quoc V. Le, James Laudon, Richard Ho, Roger Carpenter, and Jeff Dean. A graph placement methodology for fast chip design. *Nature*, 594 (7862):207–212, June 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03544-w. URL https://www.nature.com/articles/s41586-021-03544-w.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020. URL https://arxiv.org/abs/2007.08663.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

Zachary P. Neal. How small is it? Comparing indices of small worldliness. *Network Science*, 5(1):30–44, March 2017. ISSN 2050-1242, 2050-1250. doi: 10.1017/nws.2017.5. URL https://www.cambridge.org/core/product/identifier/S2050124217000054/type/journal_article.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The Building Blocks of Interpretability. *Distill*, 3(3):e10, March 2018. ISSN 2476-0757. doi: 10.23915/distill.00010. URL https://distill.pub/2018/building-blocks.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=KXuYjuBzKo.

Tom Pelletreau-Duris, Ruud van Bakel, and Michael Cochez. Do Graph Neural Network States Contain Graph Properties? In *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, pp. 15–51. PMLR, October 2025. URL https://proceedings.mlr.press/v284/pelletreau-duris25a.html.

Tiago Pimentel and Ryan Cotterell. A Bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2869–2887. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.229/.

Zarina Rakhimberdina and Tsuyoshi Murata. Linear graph convolutional model for diagnosing brain disorders. In *Complex Networks and Their Applications VIII: Volume 2 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pp. 815–826. Springer, 2020.

Zarina Rakhimberdina, Xin Liu, and Tsuyoshi Murata. Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder. *Sensors*, 20(21):6001, 2020. URL https://sci-hub.ru/https://www.mdpi.com/1424-8220/20/21/6001. Publisher: MDPI.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Elizabeth Redcay, Joseph M Moran, Penelope L Mavros, Helen Tager-Flusberg, John DE Gabrieli, and Susan Whitfield-Gabrieli. Intrinsic functional network organization in high-functioning adolescents with autism spectrum disorder. *Frontiers in human neuroscience*, 7:573, 2013.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. doi: 10.1016/j.neuroimage.2009.10.003.

Jeffrey D Rudie, JA Brown, Devi Beck-Pancer, LM Hernandez, EL Dennis, PM Thompson, SY Bookheimer, and MJNC Dapretto. Altered functional and structural brain network organization in autism. *NeuroImage: clinical*, 2:79–94, 2013.

Sayan Saha, Monidipa Das, and Sanghamitra Bandyopadhyay. A Model-Centric Explainer for Graph Neural Network based Node Classification. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4434–4438, Atlanta GA USA, October 2022. ACM. ISBN 978-1-4503-9236-5. doi: 10.1145/3511808.3557535. URL https://dl.acm.org/doi/10.1145/3511808.3557535.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for NLP with differentiable edge masking. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In David E. Rumelhart, James L. McClelland, and PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1*, pp. 194–281. MIT Press, Cambridge, MA, 1986.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, December 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-02024-1. URL https://www.nature.com/articles/s41562-024-02024-1.

Martijn P. van den Heuvel and Hilleke E. Hulshoff Pol. Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, August 2010. ISSN 0924-977X. doi: 10.1016/j.euroneuro.2010.03.008. URL https://www.sciencedirect.com/science/article/pii/S0924977X10000684.

Martijn P. van den Heuvel, Cornelis J. Stam, René S. Kahn, and Hilleke E. Hulshoff Pol. Efficiency of functional brain networks and intellectual performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(23):7619–7624, June 2009. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.1443-09.2009.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Minh N. Vu and My T. Thai. PGM-Explainer: Probabilistic graphical model explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Han Xuanyuan, Pietro Barbiero, Dobrik Georgiev, Lucie Charlotte Magister, and Pietro Liò. Global concept-based interpretability for graph neural networks via neuron analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10675–10683, 2023. doi: 10.1609/aaai.v37i9.26267. Issue 9.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, Chang Cheng, Yu-Qi Cheng, Xi-Long Cui, Jia Duan, Yi-Ru Fang, Qi-Yong Gong, Wen-Bin Guo, Zheng-Hua Hou, Lan Hu, Li Kuang, Feng Li, Kai-Ming Li, Tao Li, Yan-Song Liu, Zhe-Ning Liu, Yi-Cheng Long, Qing-Hua Luo, Hua-Qing Meng, Dai-Hui Peng, Hai-Tang Qiu, Jiang Qiu, Yue-Di Shen, Yu-Shu Shi, Chuan-Yue Wang, Fei Wang, Kai Wang, Li Wang, Xiang Wang, Ying Wang, Xiao-Ping Wu, Xin-Ran Wu, Chun-Ming Xie, Guang-Rong Xie, Hai-Yan Xie, Peng Xie, Xiu-Feng Xu, Hong Yang, Jian Yang, Jia-Shu Yao, Shu-Qiao Yao, Ying-Ying Yin, Yong-Gui Yuan, Ai-Xia Zhang, Hong Zhang, Ke-Rang Zhang, Lei Zhang, Zhi-Jun Zhang, Ru-Bai Zhou, Yi-Ting Zhou, Jun-Juan Zhu, Chao-Jie Zou, Tian-Mei Si, Xi-Nian Zuo, Jing-Ping Zhao, and Yu-Feng Zang. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019. doi: 10.1073/pnas.1900390116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900390116.

Huzheng Yang, Xiaoxiao Li, Yifan Wu, Siyi Li, Su Lu, James S Duncan, James C Gee, and Shi Gu. Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pp. 799–807. Springer, 2019.

Ming Ye, Tianliang Yang, Peng Qing, Xu Lei, Jiang Qiu, and Guangyuan Liu. Changes of functional brain networks in major depressive disorder: a graph theoretical analysis of resting-state fmri. *PloS one*, 10(9):e0133775, 2015.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 9240–9251, 2019.

Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. XGNN: Towards Model-Level Explanations of Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 430–438, August 2020. doi: 10.1145/3394486.3403085. URL http://arxiv.org/abs/2006.02587. arXiv:2006.02587 [cs, stat].

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12241–12252. PMLR, 2021. URL https://proceedings.mlr.press/v139/yuan21c.html.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5782–5799, 2023. doi: 10.1109/TPAMI.2022.3204236.

Junran Zhang, Jinhui Wang, Qizhu Wu, Weihong Kuang, Xiaoqi Huang, Yong He, and Qiyong Gong. Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder. *Biological psychiatry*, 70(4):334–342, 2011.

Kaizhong Zheng, Shujian Yu, Baojuan Li, Robert Jenssen, and Badong Chen. BrainIB: Interpretable Brain Network-based Psychiatric Diagnosis with Graph Information Bottleneck, May 2023. URL http://arxiv.org/abs/2205.03612. arXiv:2205.03612 [cs, eess].

A LOCAL AND GLOBAL GRAPH PROPERTIES

Property	Sets	Definition	Computation
Size and density			
avg_degree	FC TU	Mean number of neighbours per node	$2m/n$
degeneracy	FC	Largest k for which a k-core exists, i.e. how deeply nested the graph is	$\max_v \mathrm{core}(v)$
max_degree	TU	Largest number of neighbours of any node	$\max_v d_v$
Clustering and mixing			
avg_clustering	FC TU	Mean local clustering coefficient: how often a node's neighbours are themselves joined	average_clustering
Subgraph counts			
num_claws	TU	Number of 3-stars	$\sum_v \binom{d_v}{3}$
num_squares	GH	Summed square-clustering coefficient. Despite the name this is *not* a 4-cycle count: it is a normalised per-node measure of square closure	$\frac{1}{4}\sum_v c_4(v)$ (Lind et al., 2005)
num_tailed_triangles	TU	Triangles with a pendant edge	$\sum_v t_v(d_v-2)$
num_tri_shared_edge	TU	Pairs of triangles sharing an edge	$\sum_{uv\in E}\binom{\lvert N(u)\cap N(v)\rvert}{2}$
num_triangles	GH FC TU	Number of triangles	$\sum_v t_v/3$
num_wedges	TU	Number of paths of length two (2-stars)	$\sum_v \binom{d_v}{2}$
Resistance and diffusion			
eff_resistance_max	FC	Largest effective resistance over node pairs	from L^+ on the largest component
eff_resistance_mean	FC	Mean effective resistance over node pairs, i.e. how many alternative routes connect them	from L^+ on the largest component
Curvature			
frc_mean	FC	Mean augmented Forman–Ricci edge curvature	$4-d_u-d_v+3t_{uv}$
frc_min	FC	Minimum Forman–Ricci curvature	$4-d_u-d_v+3t_{uv}$
frc_std	FC	Spread of Forman–Ricci curvature over edges	$4-d_u-d_v+3t_{uv}$
orc_mean	FC	Mean Ollivier–Ricci edge curvature (lazy walk, $\alpha=0.5$): whether neighbourhoods overlap	W_1 by linear programme
orc_min	FC	Minimum Ollivier–Ricci curvature, the most bottleneck-like edge	W_1 by linear programme
orc_std	FC	Spread of Ollivier–Ricci curvature over edges	W_1 by linear programme
Centralisation			
avg_betweenness	FC	Mean normalised betweenness: how often nodes lie on shortest paths	betweenness_centrality
avg_betweenness_log	FC	Signed $\log(1+x)$ of the mean betweenness	$\mathrm{sgn}(x)\log(1+\lvert x\rvert)$
betweenness_centralization	FC	Concentration of betweenness on the most central node	$\sum_v(b_{\max}-b_v)/\binom{n-1}{2}$
betweenness_centralization_log	FC	Signed $\log(1+x)$ of the betweenness centralisation	$\mathrm{sgn}(x)\log(1+\lvert x\rvert)$
pagerank_centralization	FC	Concentration of PageRank on the most central node ($\alpha=0.85$)	$\sum_v(p_{\max}-p_v)/(n-1)$
pagerank_centralization_log	FC	Signed $\log(1+x)$ of the PageRank centralisation	$\mathrm{sgn}(x)\log(1+\lvert x\rvert)$

Table 4: Local properties: node- or edge-level quantities aggregated to a single value per graph. Each entry gives the definition, the computation used and, in small capitals, the datasets that probe it (*Sets*): GH Grid-House (10 properties), FC the two functional-connectivity panels (50, of which 47 are non-constant), TU MUTAG and IMDB-B (21). All properties are computed with networkx on the unweighted graph, and each is a single graph-level scalar. n, m and c denote the number of nodes, edges and connected components, d_v the degree of v, t_v the triangles through v, A the adjacency matrix and L the Laplacian.

We are using the Small-World Index, $SWI = \left(\frac{L-L_l}{L_r-L_l}\right) \times \left(\frac{C-C_r}{C_l-C_r}\right)$ in our experiment because it provides a more balanced and robust measure of small-world properties. Unlike the Small-World Quotient: $Q = \frac{C/C_r}{L/L_r}$, which can be sensitive to network size and degree, SWI normalizes both the clustering coefficient and average path length with respect to both random and lattice reference graphs. This dual normalization approach ensures that SWI is less prone to false positives or negatives, making it a more reliable metric for our analysis (Neal, 2017).

Property	Sets	Definition	Computation
Size and density			
density	GH FC TU	Fraction of node pairs that are joined	$2m/(n(n-1))$
largest_component	TU	Number of nodes in the largest connected component (TU name)	largest connected component
largest_component_size	GH FC	Number of nodes in the largest connected component	largest connected component
num_edges	GH FC TU	Number of edges	m
num_nodes	GH FC TU	Number of nodes. On FC, ROIs left isolated by thresholding are dropped	n
Distance			
avg_path_len	GH FC TU	Mean shortest-path length, on the largest component	mean over pairs, largest comp.
diameter	FC TU	Longest shortest path, on the largest component	diameter
girth	TU	Length of the shortest cycle	shortest cycle in cycle_basis
radius	FC	Smallest node eccentricity, on the largest component	radius
Clustering and mixing			
assortativity	FC TU	Degree correlation across edges: do high-degree nodes attach to each other	degree correlation over edges
hom_density_K3	FC	Homomorphism density of a triangle	$6T/n^3$
transitivity	FC	Global clustering: triangles over connected triples	$3T/\#$triples
Subgraph counts			
circuit_rank	TU	Circuit rank (TU name for the cycle rank)	$m-n+c$
clique_number	FC	Size of the largest clique, capped at 5 in the implementation, hence constant on FC	enumerate_all_cliques, $k \le 5$
cycle_rank	GH	Cycle rank (circuit rank): independent cycles of the graph	$m-n+c$
num_3_cliques	FC	Number of 3-cliques (identical to the triangle count)	enumerate_all_cliques
num_4_cliques	FC	Number of 4-cliques	enumerate_all_cliques
num_4_cycles	GH FC TU	Exact number of simple 4-cycles	$\frac{1}{8}\sum_{i\neq j}(A^2)_{ij}\big((A^2)_{ij}-1\big)$
num_5_cliques	FC	Number of 5-cliques	enumerate_all_cliques
num_cliques	GH TU	Number of maximal cliques	find_cliques
num_cycles	TU	Number of independent cycles (equals the cycle rank on these graphs)	len(cycle_basis)
num_maximal_cliques	FC	Number of maximal cliques (FC name)	find_cliques
Mesoscale			
modularity	FC	Newman modularity of a greedy community partition: how separable the graph is into communities	greedy CNM, unweighted Q
Spectral			
adjacency_energy	FC	Adjacency energy	$\sum_i \lvert\lambda_i(A)\rvert$
algebraic_connectivity	FC TU	Second-smallest Laplacian eigenvalue: how hard the graph is to disconnect (zero when disconnected)	$\lambda_2(L)$
cheeger_lower	FC	Lower Cheeger bound on the conductance of the best cut	$\lambda_2(L_{\mathrm{norm}})/2$
cheeger_upper	FC	Upper Cheeger bound on the conductance of the best cut	$\sqrt{2\lambda_2(L_{\mathrm{norm}})}$
laplacian_energy	FC	Laplacian energy	$\sum_i \lvert\mu_i - 2m/n\rvert$
spectral_gap_adj	FC	Gap between the two largest adjacency eigenvalues	$\lambda_1(A)-\lambda_2(A)$
spectral_radius	FC	Largest adjacency eigenvalue, a measure of overall connection strength	$\lambda_{\max}(A)$
Resistance and diffusion			
commute_time_mean	FC	Mean commute time of a random walk	$2m\,\bar{R}$
hkt_long	FC	Heat-kernel trace at $t=10$: global diffusion	$\sum_i e^{-t\mu_i}$
hkt_medium	FC	Heat-kernel trace at $t=1$	$\sum_i e^{-t\mu_i}$
hkt_short	FC	Heat-kernel trace at $t=0.1$: local diffusion	$\sum_i e^{-t\mu_i}$
Topology			
betti_0	FC	Number of connected components of the clique complex (undefined without gudhi; constant here)	clique complex, dim ≤ 2
betti_1	FC	Number of independent loops of the clique complex (constant here)	clique complex, dim ≤ 2
Small-world			
small_world_index	FC	Small-world index against a random and a lattice reference (Neal, 2017)	$\frac{L-L_\ell}{L_r-L_\ell}\cdot\frac{C-C_r}{C_\ell-C_r}$
small_world_sigma	FC	Small-world coefficient σ against one random reference	$(C/C_r)/(L/L_r)$, $G(n,m)$, seed 42

Table 5: Global properties: quantities defined for the graph as a whole. Columns, dataset codes and notation as in Table 4.

B EXPERT SETS

Task	S_2 (expert-aligned)	Borderline (reported, never counted as AGR)
Grid-House	num_squares, num_triangles, num_4_cycles (the motif counts that define the task)	num_cliques
ASD, MDD	num_edges, density, avg_degree, avg_path_len, avg_clustering, transitivity, modularity, small_world_sigma, small_world_index, assortativity, avg_betweenness (and its logarithm)	diameter, radius, num_triangles, num_3_cliques, hom_density_K3, degeneracy, betweenness_centralization (and its logarithm), largest_component_size
MUTAG	circuit_rank, num_cycles (ring and circuit counts: the chemist's vocabulary for mutagenicity)	num_edges, num_nodes, avg_path_len, girth, max_degree
IMDB-B	density, num_edges, num_nodes, avg_degree, max_degree, avg_clustering, num_triangles, num_cliques (the descriptors of social network analysis)	avg_path_len, assortativity, diameter, largest_component, num_claws, num_wedges

Table 6: Expert sets S_2, fixed from the literature on each task before probing: Grid-House from its generator (Pelletreau-Duris et al., 2025); MUTAG from the structure–activity study that introduced it (**?**); IMDB-B (**?**) from the standard descriptors of social network analysis (**?**); FC from clinical network neuroscience (Bullmore & Sporns, 2009; Rubinov & Sporns, 2010; Rudie et al., 2013; Zhang et al., 2011). Every other computed property is in $S_1 \setminus S_2$. Global and local efficiency, rich-club coefficients and participation coefficients are standard expert measures that our panel does not compute, so the expert side of the FC test is incomplete.

How the FC expert set was built. The FC expert set follows the graph-theoretic literature on resting-state connectomes (Bullmore & Sporns, 2009; Rubinov & Sporns, 2010; Farahani et al., 2019). Studies of individual differences in general intelligence already relate functional networks to small-world organisation, characteristic path length and the centrality of hub regions (van den Heuvel et al., 2009; van den Heuvel & Hulshoff Pol, 2010; Langer et al., 2012; Hilger et al., 2017; Jung & Haier, 2007), and disrupted connections between specialised regions are linked to atypical integration across distributed networks (Catani et al., 2005; Ffytche & Catani, 2005). GNNs classify ASD, MDD and other disorders from such networks with promising accuracy (Rakhimberdina et al., 2020; Rakhimberdina & Murata, 2020; Yang et al., 2019; Zheng et al., 2023), but which network properties they rely on remains open (Zheng et al., 2023).

ASD. Graph-theoretic resting-state studies of autism (Redcay et al., 2013; Rudie et al., 2013; Di Martino et al., 2014; Keown et al., 2017; Kazeminejad & Sotero, 2019) report local over-connectivity, in particular within sensory and association cortices, together with reduced long-range connectivity. In network terms, modularity, the clustering coefficient and local efficiency are reduced while global efficiency is increased, with shorter average path lengths (Farahani et al., 2019), and betweenness centrality and local connections increase in the prefrontal areas of adolescents with ASD (Redcay et al., 2013). An expert would therefore expect an ASD classifier to rely on betweenness centrality, clustering, characteristic path length and small-worldness, which balances local clustering against global integration.

MDD. Studies of depression report a loss of small-world organisation (Ye et al., 2015; Achard & Bullmore, 2007) and a reorganisation of community structure (Zhang et al., 2011; Leistedt et al., 2009; Lord et al., 2012): global and local clustering increase, and modularity increases, with fewer inter-modular and more intra-modular edges, which may reflect reduced communication between the default mode and cognitive control networks (Ye et al., 2015). An expert would therefore expect an MDD classifier to rely on clustering, modularity and small-world measures rather than on counts of small motifs such as triangles or squares.

From the literature to S_2. The FC expert set of Table 6 collects these measures (clustering and transitivity, characteristic path length, modularity, the two small-world indices and mean betweenness) together with degree, density and assortativity, the basic descriptors of the connectomics toolbox (Bullmore & Sporns, 2009; Rubinov & Sporns, 2010); one set serves both tasks. Properties closely tied to these measures but not the ones the literature reports are borderline and never counted as alien: diameter and radius as relatives of path length, triangle counts and triangle homomorphism density as the raw material of clustering, betweenness centralisation as the whole-graph counterpart of mean betweenness, degeneracy, and the size of the largest component. Global and local efficiency, which this literature also uses, are not computed by our panel. The FC expert set was reviewed and validated by a university professor of neuroscience, whose name is withheld for anonymous review.

The other expert sets. The Grid-House set needs no external validation: its generator defines the task, so the motif counts that determine the label are known by construction. The MUTAG and IMDB-B sets are aligned with the literature on those benchmarks, ring and circuit counts for mutagenicity, after the structure–activity analysis of the original MUTAG study, in which fused aromatic rings are a predictor of mutagenic potency (**?**), and for IMDB-B (**?**) the size, density, degree and clustering descriptors of social network analysis (**?**), but were not reviewed by a domain expert. Asking a chemist to review the MUTAG set and a social scientist to review the IMDB-B set, as we did for FC, is left for future work.

C AGR VERDICT PER PROPERTY

Table 7 counts the alien properties passing each LRH level in every run, and Table 8 compares the three erasure operators on the same runs. The per-run tables that follow give every property: its held-out R^2, the accuracy change after erasure, the margin over the dimension-matched control, the recovery ρ and the resulting verdict. All values are medians over five seeds; Grid-House and the two FC datasets use the published checkpoints, MUTAG and IMDB-B models are retrained with validation-based selection.

Dataset	Model	$\lvert S_1 \setminus S_2 \rvert$	Weak	Strong	Excl.	AGR	Genuine AGRs
ASD	GCN	26	20	0	0	0	—
MDD	GCN	26	24	0	0	0	—
Grid-House	GCN	6	6	0	0	0	—
Grid-House	GCN_L2	6	6	1	0	0	—
Grid-House	GIN4	6	6	2	2	1	`density`
Grid-House	GIN4_L2	6	6	6	6	5	`density`, `num_edges`, `avg_path_len`, `num_nodes`
IMDB-B	GIN_L2	6	6	3	0	0	—
MUTAG	GIN_L2	9	9	9	2	9	`diameter`, `num_claws`, `avg_degree`, `num_wedges`, `density`

Table 7: Alien graph representations per run, median over five seeds. $S_1 \setminus S_2$ counts the properties outside the expert set; Weak, Strong and Excl. count those passing each level; AGR counts the Strong-level ones that are not linearly reducible to S_2 (held-out $R^2 < 0.9$ from the expert properties). Exclusivity is reported only for runs whose retraining converged.

Dataset	Operator	Directions removed	Residual R^2	Mean Δ (pt)
Grid-House	LEACE	1.0	−0.00	−20.4
	INLP	22.2	0.00	−46.2
	rank-one	1.0	0.53	−17.0
ASD	LEACE	1.0	−0.01	−0.3
	INLP	8.7	−0.01	−5.9
	rank-one	1.0	0.34	−3.2
MDD	LEACE	1.0	−0.02	0.8
	INLP	7.4	−0.02	−4.1
	rank-one	1.0	0.38	1.3
MUTAG	LEACE	1.0	−0.02	−17.3
	INLP	61.8	−0.01	−30.9
	rank-one	1.0	0.73	−0.0
IMDB-B	LEACE	1.0	−0.00	−1.1
	INLP	47.5	−0.00	−13.8
	rank-one	1.0	0.81	0.0

Table 8: The three erasure operators on the same runs and the same classifiers, averaged over properties and seeds. LEACE removes one direction and leaves the property undecodable; INLP reaches the same residual R^2 only by deleting a large part of the representation, after which every property damages the classifier alike; the rank-one ablation removes one direction but leaves the property decodable. The choice of operator, not the property, drives the size of the drop.

THE SUBSPACE LESION ON THE FMRI DATA

Tables 16 and 17 give the LEACE verdicts, under which no fMRI property passes Strong. Table 18 gives the same panel under INLP, which on these datasets deletes 5–7% of the representation (Table 9) and is the reading used for the MDD candidates; each seed draws 100 dimension-matched and 100 variance-matched random ablations on the published split.

Dataset	Model	d	k (median)	k range	k/d	Reading
ASD	GCN	128	9	5–11	7%	subspace-level lesion
MDD	GCN	128	7	2–10	5%	subspace-level lesion
Grid-House	GIN4_L2	30	12	11–13	40%	large slice
Grid-House	GIN4	30	12	12–13	40%	large slice
Grid-House	GCN_L2	60	30	23–30	50%	near-total erasure
Grid-House	GCN	60	37	29–38	62%	near-total erasure
IMDB-B	GIN_L2	64	54	25–62	84%	near-total erasure
MUTAG	GIN_L2	64	63	49–64	98%	near-total erasure

Table 9: How much of the representation iterative nullspace projection deletes to make a property undecodable, median over properties and seeds. d is the embedding width and k the number of directions removed. The ratio decides whether an INLP drop can be read as evidence about the property: on the functional-connectivity data INLP removes a twentieth of the representation and its drops remain property-specific, whereas on Grid-House and the two TU benchmarks it removes between two fifths and almost all of it, after which every property damages the classifier alike. LEACE removes one direction everywhere ($k/d \leq 3\%$) and reaches the same residual R^2. We therefore report INLP results for the FC datasets only.

Property	S_2	R^2	Δ (pt)	z_{dim}	ρ %	Weak	Strong	Verdict
cycle_rank	no	0.78	−35.9	2.52	103	yes	no	AGR-Weak, S2 proxy
num_squares	yes	0.78	−34.9	2.43	101	yes	no	expert-Weak
num_4_cycles	yes	0.81	−33.4	2.29	101	yes	no	expert-Weak
density	no	0.67	−20.9	1.15	104	yes	no	AGR-Weak
avg_path_len	no	0.43	−20.1	1.09	98	yes	no	AGR-Weak
num_edges	no	0.38	−6.8	−0.12	96	yes	no	AGR-Weak
num_nodes	no	0.48	−4.0	−0.40	56	yes	no	AGR-Weak
largest_component_size	no	0.48	−4.0	−0.40	56	yes	no	AGR-Weak
num_cliques	borderline	0.49	−3.8	−0.43	100	yes	no	borderline-Weak
num_triangles	yes	0.42	−1.5	−0.67	117	yes	no	expert-Weak

Table 10: AGR verdict, gridhouse GCN. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.58$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	z_{dim}	ρ %	Weak	Strong	Verdict
cycle_rank	no	0.88	−47.7	2.83	99	yes	yes	AGR-Strong, S2 proxy
num_squares	yes	0.87	−45.0	2.63	97	yes	yes	expert-Exclusive
num_4_cycles	yes	0.90	−28.6	1.45	86	yes	no	expert-Weak
avg_path_len	no	0.43	−24.1	1.13	96	yes	no	AGR-Weak
density	no	0.69	−12.3	0.28	98	yes	no	AGR-Weak
num_edges	no	0.59	−8.8	0.02	40	yes	no	AGR-Weak
num_triangles	yes	0.81	−5.8	−0.20	100	yes	no	expert-Weak
num_cliques	borderline	0.64	−5.0	−0.25	75	yes	no	borderline-Weak
num_nodes	no	0.58	0.5	−0.66	–	yes	no	AGR-Weak
largest_component_size	no	0.58	0.5	−0.66	–	yes	no	AGR-Weak

Table 11: AGR verdict, gridhouse GCN_L2. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.58$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	$z_{\dim}$	ρ %	Weak	Strong	Verdict
cycle_rank	no	0.93	−47.5	5.83	63	yes	yes	AGR-Exclusive, S2 proxy
num_squares	yes	0.88	−46.7	5.73	61	yes	yes	expert-Exclusive
num_4_cycles	yes	0.89	−38.7	4.66	53	yes	yes	expert-Exclusive
density	no	0.52	−28.9	3.37	52	yes	yes	AGR-Exclusive
num_edges	no	0.19	−14.3	1.43	30	yes	no	AGR-Weak
avg_path_len	no	0.50	−13.6	1.33	57	yes	no	AGR-Weak
num_triangles	yes	0.24	−10.3	0.90	0	yes	no	expert-Weak
num_cliques	borderline	0.19	−9.3	0.77	11	yes	no	borderline-Weak
num_nodes	no	0.25	−6.0	0.34	0	yes	no	AGR-Weak
largest_component_size	no	0.25	−6.0	0.34	0	yes	no	AGR-Weak

Table 12: AGR verdict, gridhouse GIN4. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.58$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	$z_{\dim}$	ρ %	Weak	Strong	Verdict
cycle_rank	no	0.92	−51.5	19.28	36	yes	yes	AGR-Exclusive, S2 proxy
num_squares	yes	0.87	−50.8	19.00	63	yes	yes	expert-Exclusive
num_4_cycles	yes	0.88	−44.7	16.71	42	yes	yes	expert-Exclusive
density	no	0.50	−36.4	13.57	26	yes	yes	AGR-Exclusive
num_edges	no	0.16	−19.3	7.11	22	yes	yes	AGR-Exclusive
avg_path_len	no	0.45	−17.3	6.34	35	yes	yes	AGR-Exclusive
num_cliques	borderline	0.17	−9.3	3.30	30	yes	yes	borderline-Exclusive
num_triangles	yes	0.28	−8.3	2.92	15	yes	yes	expert-Exclusive
num_nodes	no	0.22	−7.5	2.63	17	yes	yes	AGR-Exclusive
largest_component_size	no	0.22	−7.5	2.63	17	yes	yes	AGR-Exclusive

Table 13: AGR verdict, gridhouse GIN4_L2. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.58$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	$z_{\dim}$	ρ %	Weak	Strong	Verdict
diameter	no	0.51	−13.2	17.10	35	yes	yes	AGR-Strong
num_claws	no	0.83	−21.1	16.94	75	yes	yes	AGR-Strong
avg_degree	no	0.98	−15.8	16.94	50	yes	yes	AGR-Exclusive(partial)
num_wedges	no	0.82	−28.9	14.14	75	yes	yes	AGR-Strong
num_edges	borderline	0.76	−28.9	14.14	62	yes	yes	borderline-Strong
density	no	0.66	−23.7	14.14	59	yes	yes	AGR-Exclusive(partial)
num_cliques	no	0.76	−28.9	14.14	62	yes	yes	AGR-Strong
avg_path_len	borderline	0.44	−13.2	12.88	49	yes	yes	borderline-Strong
algebraic_connectivity	no	0.47	−15.8	12.88	3	yes	yes	AGR-Strong
assortativity	no	0.83	−7.9	12.88	33	yes	yes	AGR-Strong
circuit_rank	yes	0.96	−10.5	11.34	38	yes	yes	expert-Strong
num_nodes	borderline	0.69	−18.4	11.34	60	yes	yes	borderline-Strong
largest_component	no	0.69	−18.4	11.34	60	yes	yes	AGR-Strong
num_cycles	yes	0.96	−10.5	11.34	38	yes	yes	expert-Strong
girth	borderline	0.48	−0.0	0.31	75	yes	n/a	borderline-Weak
max_degree	borderline	–	−0.0	0.25	–	no	no	not encoded

Table 14: AGR verdict, mutag GIN_L2. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.73$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	$z_{\dim}$	ρ %	Weak	Strong	Verdict
avg_path_len	borderline	0.92	−5.5	10.62	95	yes	yes	borderline-Exclusive
density	yes	0.92	−5.5	10.62	95	yes	yes	expert-Exclusive
assortativity	borderline	0.45	−2.0	5.50	183	yes	yes	borderline-Strong
num_4_cycles	no	0.99	−1.0	2.85	250	yes	yes	AGR-Strong, S2 proxy
num_tri_shared_edge	no	0.99	−1.0	2.85	225	yes	yes	AGR-Strong, S2 proxy
num_tailed_triangles	no	0.99	−0.5	2.85	200	yes	yes	AGR-Strong, S2 proxy
num_triangles	yes	0.99	−0.5	2.85	220	yes	yes	expert-Strong
num_cliques	yes	0.77	−1.0	2.37	–	yes	no	expert-Weak
diameter	borderline	0.82	−1.0	1.41	133	yes	no	borderline-Weak
largest_component	borderline	0.85	−0.5	1.15	50	yes	no	borderline-Weak
max_degree	yes	0.85	−0.5	1.15	50	yes	no	expert-Weak
num_nodes	yes	0.85	−0.5	1.15	50	yes	no	expert-Weak
num_claws	borderline	0.92	−0.5	1.01	280	yes	no	borderline-Weak
algebraic_connectivity	no	0.85	−0.0	0.46	177	yes	no	AGR-Weak
avg_degree	yes	1.00	−0.0	−0.14	140	yes	no	expert-Weak
avg_clustering	yes	0.46	−0.0	−0.18	–	yes	no	expert-Weak
num_wedges	borderline	0.97	0.5	−1.06	140	yes	no	borderline-Weak
num_edges	yes	0.95	0.5	−1.69	–	yes	no	expert-Weak
num_cycles	no	0.96	0.5	−1.69	–	yes	no	AGR-Weak, S2 proxy
circuit_rank	no	0.96	0.5	−1.69	–	yes	no	AGR-Weak, S2 proxy

Table 15: AGR verdict, imdb-b GIN_L2. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 2.81$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	$z_{\dim}$	ρ %	Weak	Strong	Verdict
transitivity	yes	0.46	−3.6	1.57	–	yes	no	expert-Weak
assortativity	yes	−0.19	−3.6	1.57	–	no	no	not encoded
small_world_index	yes	−0.32	−1.8	0.70	–	no	no	not encoded
num_4_cycles	no	0.48	−1.8	0.70	–	yes	no	AGR-Weak
laplacian_energy	no	0.52	−1.8	0.70	–	yes	no	AGR-Weak, S2 proxy
orc_mean	no	0.57	−1.8	0.70	–	yes	no	AGR-Weak, S2 proxy
spectral_radius	no	0.62	−1.8	0.70	–	yes	no	AGR-Weak, S2 proxy
degeneracy	borderline	0.55	−1.8	0.70	–	yes	no	borderline-Weak
num_5_cliques	no	−0.30	−1.8	0.70	–	no	no	not encoded
num_4_cliques	no	0.41	−1.8	0.70	–	yes	no	AGR-Weak
num_3_cliques	borderline	0.68	−1.8	0.70	–	yes	no	borderline-Weak
hom_density_K3	borderline	0.67	−1.8	0.70	–	yes	no	borderline-Weak
num_triangles	borderline	0.68	−1.8	0.70	–	yes	no	borderline-Weak
avg_clustering	yes	0.49	−1.8	0.70	–	yes	no	expert-Weak
frc_std	no	0.65	−1.8	0.70	–	yes	no	AGR-Weak, S2 proxy
small_world_sigma	yes	0.61	−1.8	0.70	–	yes	no	expert-Weak
eff_resistance_mean	no	0.58	−0.0	−0.20	–	yes	no	AGR-Weak, S2 proxy
frc_min	no	0.48	−0.0	−0.20	–	yes	no	AGR-Weak
avg_betweenness	yes	0.60	−0.0	−0.20	–	yes	no	expert-Weak
frc_mean	no	0.34	−0.0	−0.20	–	yes	no	AGR-Weak
betweenness_centralization	borderline	0.27	−0.0	−0.20	–	yes	no	borderline-Weak
avg_betweenness_log	yes	0.61	−0.0	−0.20	–	yes	no	expert-Weak
orc_std	no	0.34	−0.0	−0.20	–	yes	no	AGR-Weak
orc_min	no	−0.13	−0.0	−0.20	–	no	n/a	not encoded
betweenness_centralization_log	borderline	0.27	−0.0	−0.20	–	yes	no	borderline-Weak
hkt_medium	no	0.61	−0.0	−0.20	–	yes	no	AGR-Weak, S2 proxy
hkt_short	no	0.77	−0.0	−0.20	–	yes	no	AGR-Weak, S2 proxy
adjacency_energy	no	0.78	−0.0	−0.20	–	yes	no	AGR-Weak, S2 proxy
num_maximal_cliques	no	0.49	−0.0	−0.20	–	yes	no	AGR-Weak
avg_path_len	yes	0.60	−0.0	−0.20	–	yes	no	expert-Weak
density	yes	0.77	−0.0	−0.20	–	yes	no	expert-Weak
avg_degree	yes	0.78	−0.0	−0.20	–	yes	no	expert-Weak
num_edges	yes	0.78	−0.0	−0.20	–	yes	no	expert-Weak
spectral_gap_adj	no	0.02	−0.0	−0.20	–	no	no	not encoded
modularity	yes	0.57	−0.0	−0.20	–	yes	no	expert-Weak
radius	borderline	0.59	−0.0	−0.20	–	yes	no	borderline-Weak
pagerank_centralization_log	no	−0.01	1.8	−1.09	–	no	no	not encoded
largest_component_size	borderline	0.10	1.8	−1.09	–	no	no	not encoded
pagerank_centralization	no	−0.01	1.8	−1.09	–	no	no	not encoded
eff_resistance_max	no	0.34	1.8	−1.09	–	yes	n/a	AGR-Weak
diameter	borderline	0.47	1.8	−1.09	–	yes	no	borderline-Weak
commute_time_mean	no	0.37	1.8	−1.09	–	yes	no	AGR-Weak
algebraic_connectivity	no	0.32	1.8	−1.09	–	yes	no	AGR-Weak
hkt_long	no	0.13	1.8	−1.09	–	yes	no	AGR-Weak
cheeger_lower	no	0.34	1.8	−1.09	–	yes	no	AGR-Weak
cheeger_upper	no	0.21	1.8	−1.09	–	yes	no	AGR-Weak
num_nodes	no	−0.02	1.8	−1.09	–	no	n/a	not encoded

Table 16: AGR verdict, fc-asd GCN. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 3.07$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

Property	S_2	R^2	Δ (pt)	z_{dim}	ρ %	Weak	Strong	Verdict
adjacency_energy	no	0.77	−1.2	1.30	–	yes	no	AGR-Weak, S2 proxy
density	yes	0.67	−1.2	1.30	–	yes	no	expert-Weak
avg_degree	yes	0.70	−1.2	1.30	–	yes	no	expert-Weak
largest_component_size	borderline	0.61	−1.2	1.30	–	yes	n/a	borderline-Weak
orc_std	no	0.46	−1.2	1.30	–	yes	no	AGR-Weak
avg_clustering	yes	0.40	−1.2	1.30	–	yes	no	expert-Weak
commute_time_mean	no	0.34	−1.2	1.30	–	yes	no	AGR-Weak
num_edges	yes	0.73	−1.2	1.30	–	yes	no	expert-Weak
num_maximal_cliques	no	0.52	−1.2	1.30	–	yes	no	AGR-Weak
num_3_cliques	borderline	0.72	−0.0	0.63	–	yes	no	borderline-Weak
frc_min	no	0.51	−0.0	0.63	–	yes	no	AGR-Weak
hkt_medium	no	0.61	−0.0	0.63	–	yes	no	AGR-Weak, S2 proxy
laplacian_energy	no	0.66	−0.0	0.63	–	yes	no	AGR-Weak
algebraic_connectivity	no	0.54	−0.0	0.63	–	yes	no	AGR-Weak
modularity	yes	0.48	−0.0	0.63	–	yes	no	expert-Weak
num_nodes	no	0.54	−0.0	0.63	–	yes	no	AGR-Weak, S2 proxy
radius	borderline	0.35	−0.0	0.63	–	yes	no	borderline-Weak
hom_density_K3	borderline	0.69	−0.0	0.63	–	yes	no	borderline-Weak
diameter	borderline	0.30	−0.0	0.63	–	yes	no	borderline-Weak
num_triangles	borderline	0.72	−0.0	0.63	–	yes	no	borderline-Weak
eff_resistance_max	no	0.31	1.2	−0.02	–	yes	no	AGR-Weak
avg_betweenness	yes	0.44	1.2	−0.02	–	yes	no	expert-Weak
orc_min	no	−0.24	1.2	−0.02	–	no	no	not encoded
betweenness_centralization	borderline	0.31	1.2	−0.02	–	yes	no	borderline-Weak
hkt_long	no	0.41	1.2	−0.02	–	yes	no	AGR-Weak
pagerank_centralization	no	0.23	1.2	−0.02	–	yes	no	AGR-Weak
hkt_short	no	0.59	1.2	−0.02	–	yes	no	AGR-Weak, S2 proxy
avg_betweenness_log	yes	0.44	1.2	−0.02	–	yes	no	expert-Weak
eff_resistance_mean	no	0.55	1.2	−0.02	–	yes	no	AGR-Weak, S2 proxy
frc_std	no	0.62	1.2	−0.02	–	yes	no	AGR-Weak, S2 proxy
cheeger_upper	no	0.52	1.2	−0.02	–	yes	no	AGR-Weak
cheeger_lower	no	0.56	1.2	−0.02	–	yes	no	AGR-Weak
betweenness_centralization_log	borderline	0.31	1.2	−0.02	–	yes	no	borderline-Weak
pagerank_centralization_log	no	0.23	1.2	−0.02	–	yes	no	AGR-Weak
small_world_sigma	yes	0.30	1.2	−0.02	–	yes	no	expert-Weak
spectral_radius	no	0.61	1.2	−0.02	–	yes	no	AGR-Weak, S2 proxy
degeneracy	borderline	0.51	1.2	−0.02	–	yes	no	borderline-Weak
avg_path_len	yes	0.57	1.2	−0.02	–	yes	no	expert-Weak
num_4_cliques	no	0.68	2.5	−0.62	–	yes	no	AGR-Weak
frc_mean	no	0.40	2.5	−0.62	–	yes	no	AGR-Weak
num_5_cliques	no	0.41	2.5	−0.62	–	yes	no	AGR-Weak
transitivity	yes	0.35	2.5	−0.62	–	yes	no	expert-Weak
num_4_cycles	no	0.69	2.5	−0.62	–	yes	no	AGR-Weak
orc_mean	no	0.42	3.7	−1.22	–	yes	no	AGR-Weak, S2 proxy
assortativity	yes	−0.05	3.7	−1.22	–	no	no	not encoded
spectral_gap_adj	no	0.02	3.7	−1.22	–	no	no	not encoded
small_world_index	yes	0.13	3.7	−1.22	–	yes	no	expert-Weak

Table 17: AGR verdict, fc-mdd GCN. Median over five seeds; LEACE erasure (one direction, residual $R^2 \approx 0$). Weak: $R^2 \geq 0.1$. Strong: the drop clears the dimension-matched control at $z \geq 3.07$ (Bonferroni, $\alpha = 0.05$). ρ is the best retraining mode, median over seeds, measured after convergence; a property is partially exclusive only when no seed recovers fully.

	ASD							MDD						
Property	R^2	k	Δ (gr.)	z_{dim}	z_{var}	p_{H}	$\rho_{\max}$	R^2	k	Δ (gr.)	z_{dim}	z_{var}	p_{H}	$\rho_{\max}$
num_nodes	-0.02	10	-3.6 (-2)	-0.0	-1.0	1.00	n/a	0.54	9	-6.2 (-5)	2.3	-1.7	1.00	60
num_edges	0.78	10	-7.3 (-4)	1.0	0.2	1.00	75	0.73	8	-8.6 (-7)	2.9	0.4	1.00	129
density	0.77	8	-10.9 (-6)	2.3	1.3	1.00	67	0.67	8	-4.9 (-4)	1.8	0.5	1.00	175
avg_degree	0.78	10	-7.3 (-4)	1.0	0.2	1.00	50	0.70	8	-8.6 (-7)	2.9	0.5	1.00	129
largest_component_size	0.10	10	-5.5 (-3)	0.5	-0.7	1.00	33	0.61	9	-6.2 (-5)	2.3	-1.7	1.00	60
avg_path_len	0.60	11	-7.3 (-4)	1.0	-0.1	1.00	125	0.57	9	-9.9 (-8)	**3.3**	0.1	1.00	113
diameter	0.47	10	-0.0 (+0)	-1.0	-1.5	1.00	n/a	0.30	9	-9.9 (-8)	**3.3**	0.0	1.00	113
radius	0.59	10	-3.6 (-2)	-0.0	-0.8	1.00	n/a	0.35	9	-9.9 (-8)	**3.3**	0.1	1.00	113
avg_clustering	0.49	7	+1.8 (+1)	-1.4	-1.3	1.00	n/a	0.40	5	+2.5 (+2)	-0.2	-0.5	1.00	n/a
transitivity	0.46	8	+1.8 (+1)	-1.5	-1.6	1.00	n/a	0.35	7	-0.0 (+0)	0.4	-0.3	1.00	n/a
assortativity	-0.19	7	-12.7 (-7)	2.9	1.2	1.00	71	-0.05	4	+7.4 (+6)	-1.9	-1.8	1.00	n/a
num_triangles	0.68	7	-10.9 (-6)	2.3	1.7	1.00	83	0.72	7	-0.0 (+0)	0.6	0.1	1.00	n/a
num_4_cycles	0.48	7	-9.1 (-5)	1.8	1.4	1.00	80	0.69	7	+2.5 (+2)	-0.2	-0.5	1.00	n/a
hom_density_K3	0.67	7	-10.9 (-6)	2.3	1.6	1.00	83	0.69	7	+2.5 (+2)	-0.2	-0.5	1.00	n/a
num_maximal_cliques	0.49	10	-3.6 (-2)	-0.0	-0.8	1.00	n/a	0.52	9	-11.1 (-9)	**3.6**	0.5	1.00	89
num_3_cliques	0.68	7	-10.9 (-6)	2.3	1.7	1.00	83	0.72	7	-0.0 (+0)	0.6	0.1	1.00	n/a
num_4_cliques	0.41	7	-10.9 (-6)	2.3	2.0	1.00	83	0.68	7	+4.9 (+4)	-0.9	-1.0	1.00	n/a
num_5_cliques	-0.30	7	-1.8 (-1)	-0.3	-0.5	1.00	n/a	0.41	7	+4.9 (+4)	-0.9	-1.1	1.00	n/a
modularity	0.57	11	+1.8 (+1)	-1.5	-1.8	1.00	n/a	0.48	9	-3.7 (-3)	1.6	-1.4	1.00	133
degeneracy	0.55	7	-5.5 (-3)	0.8	0.7	1.00	67	0.51	5	+6.2 (+5)	-1.3	-1.4	1.00	n/a
spectral_radius	0.62	7	-10.9 (-6)	2.3	1.7	1.00	83	0.61	6	-0.0 (+0)	0.6	0.2	1.00	n/a
spectral_gap_adj	0.02	6	-1.8 (-1)	-0.3	-0.2	1.00	n/a	0.02	5	+2.5 (+2)	-0.2	-0.5	1.00	n/a
algebraic_connectivity	0.32	10	-1.8 (-1)	-0.5	-1.2	1.00	n/a	0.54	8	-12.3 (-10)	**3.9**	1.0	1.00	110
adjacency_energy	0.78	11	-3.6 (-2)	-0.0	-0.9	1.00	n/a	0.77	9	-11.1 (-9)	**3.7**	0.6	1.00	122
laplacian_energy	0.52	7	-9.1 (-5)	1.8	1.4	1.00	60	0.66	6	+2.5 (+2)	-0.1	-0.5	1.00	n/a
cheeger_lower	0.34	10	-3.6 (-2)	-0.0	-1.1	1.00	n/a	0.56	7	-7.4 (-6)	2.8	-0.9	1.00	150
cheeger_upper	0.21	9	-1.8 (-1)	-0.4	-1.4	1.00	n/a	0.52	7	-9.9 (-8)	**3.6**	0.1	1.00	125
eff_resistance_mean	0.58	10	-5.5 (-3)	0.5	-0.4	1.00	100	0.55	10	-9.9 (-8)	**3.3**	0.1	1.00	113
eff_resistance_max	0.34	10	-7.3 (-4)	1.0	-0.1	1.00	75	0.31	8	-9.9 (-8)	**3.3**	0.2	1.00	125
commute_time_mean	0.37	10	-5.5 (-3)	0.5	-0.7	1.00	167	0.34	7	-2.5 (-2)	1.3	-2.8	1.00	n/a
hkt_short	0.77	10	-5.5 (-3)	0.5	-0.1	1.00	33	0.59	9	-8.6 (-7)	2.9	0.3	1.00	157
hkt_medium	0.61	10	-5.5 (-3)	0.5	-0.4	1.00	133	0.61	10	-9.9 (-8)	**3.3**	0.1	1.00	113
hkt_long	0.13	10	-3.6 (-2)	-0.0	-0.9	1.00	n/a	0.41	8	-6.2 (-5)	2.2	-1.2	1.00	160
orc_mean	0.57	10	+1.8 (+1)	-1.5	-1.5	1.00	n/a	0.42	8	-9.9 (-8)	**3.3**	0.6	1.00	125
orc_min	-0.13	5	-12.7 (-7)	3.3	1.3	1.00	86	-0.24	3	-1.2 (-1)	1.2	0.5	1.00	n/a
orc_std	0.34	9	-5.5 (-3)	0.5	-0.3	1.00	67	0.46	7	+2.5 (+2)	-0.2	-2.3	1.00	n/a
frc_mean	0.34	9	-10.9 (-6)	2.0	0.5	1.00	133	0.40	8	-8.6 (-7)	2.9	-0.0	1.00	129
frc_min	0.48	8	-7.3 (-4)	1.2	0.4	1.00	50	0.51	8	-3.7 (-3)	1.5	-0.1	1.00	200
frc_std	0.65	7	-9.1 (-5)	1.8	1.4	1.00	80	0.62	5	+4.9 (+4)	-0.9	-1.1	1.00	n/a
avg_betweenness	0.60	10	-7.3 (-4)	1.0	-0.2	1.00	125	0.44	9	-9.9 (-8)	**3.3**	0.1	1.00	138
betweenness_centralization	0.27	10	-0.0 (+0)	-1.0	-1.7	1.00	n/a	0.31	7	-0.0 (+0)	0.5	-3.9	1.00	n/a
pagerank_centralization	-0.01	8	-7.3 (-4)	1.2	-0.0	1.00	75	0.23	8	-9.9 (-8)	**3.3**	0.2	1.00	88
avg_betweenness_log	0.61	10	-7.3 (-4)	1.0	-0.2	1.00	125	0.44	9	-9.9 (-8)	**3.4**	0.1	1.00	138
betweenness_centralization_log	0.27	10	-0.0 (+0)	-1.0	-1.7	1.00	n/a	0.31	7	-0.0 (+0)	0.5	-3.9	1.00	n/a
pagerank_centralization_log	-0.01	8	-7.3 (-4)	1.2	-0.0	1.00	75	0.23	7	-9.9 (-8)	**3.6**	0.2	1.00	88
small_world_sigma	0.61	7	-7.3 (-4)	1.3	0.6	1.00	75	0.30	5	+3.7 (+3)	-0.6	-0.8	1.00	n/a
small_world_index	-0.32	6	-14.5 (-8)	3.4	2.5	0.91	100	0.13	6	-11.1 (-9)	**4.5**	0.8	1.00	156

Table 18: Functional connectivity under the subspace erasure (INLP), the reading behind Section 4.2. GCN, published checkpoints and published 95/5 split; every entry is the median over five seeds of the erasure and its controls. R^2: held-out probe R^2 before erasure. k: directions removed. Δ: change in test accuracy in points (test graphs in brackets; negative means accuracy lost). z_{dim} and z_{var}: margin over 100 random ablations matched in dimension and in removed variance; bold marks $z_{\mathrm{dim}} \geq z^* = 3.07$ with $R^2 \geq 0.1$. p_{Holm}: one-sided exact McNemar test, Holm-corrected over the 47 properties. $\rho_{\max}$: best of the retraining modes, median over five seeds, reported only for drops of at least three graphs (†: some run did not converge). Test sets: ASD 55 graphs (baseline 65.5%), MDD 81 graphs (59.3%).

D GRID-HOUSE: DATASET AND PROPERTIES

Graphs have $n \in [20, 29]$ nodes; a single-motif graph has $m - n \in \{1, 3\}$ and a two-motif graph $m - n = 5$. `num_squares` is $\frac{1}{4}\sum_v c_4(v)$, the summed square-clustering coefficient (Lind et al., 2005), not a 4-cycle count; we therefore also probe the exact count `num_4_cycles` and the cycle rank. Table 19 gives the correlations between all ten properties and with the label: the three cycle-structure measures are the only ones that determine the class, and the cycle rank is a linear function of the motif counts, which is why the verdict counts it as a restatement of the expert vocabulary rather than an alien representation.

	nodes	edges	dens.	path	cliq.	tri.	sq.	LCC	c.rank	C4	label
num_nodes	1.00										
num_edges	0.86	1.00									
density	−0.90	−0.58	1.00								
avg_path_len	0.25	0.49	−0.02	1.00							
num_cliques	0.86	0.97	−0.60	0.46	1.00						
num_triangles	−0.07	0.03	0.13	0.08	−0.23	1.00					
num_squares	−0.00	0.49	0.39	0.52	0.45	0.10	1.00				
largest_comp.	1.00	0.86	−0.90	0.25	0.86	−0.07	−0.00	1.00			
cycle_rank	−0.01	0.49	0.41	0.54	0.43	0.17	0.97	−0.01	1.00		
num_4_cycles	0.00	0.49	0.38	0.53	0.50	−0.09	0.96	0.00	0.97	1.00	
label	−0.04	0.42	0.40	0.48	0.26	0.58	0.85	−0.04	0.90	0.76	1.00

Table 19: Pearson correlations between the Grid-House properties and with the class label, over all 1987 graphs (isomorphic duplicates removed). `num_squares` is the summed square-clustering coefficient, `c.rank` the cycle rank $|E| - |V| + 1$ and C4 the true number of 4-cycles; the last two were added to the dictionary for the corrected runs reported here. `num_nodes` and the largest component coincide because every graph is connected.

Property	VIF	Reading
Grid-House (10 properties)		
`cycle_rank`	∞	perfectly redundant
`largest_component_size`	∞	perfectly redundant
`num_4_cycles`	∞	perfectly redundant
`num_cliques`	∞	perfectly redundant
`num_edges`	∞	perfectly redundant
`num_nodes`	∞	perfectly redundant
`num_triangles`	∞	perfectly redundant
`density`	47.6	severe collinearity
`num_squares`	22.7	severe collinearity
`avg_path_len`	1.6	acceptable
Functional connectivity, ASD (47 properties)		
perfectly redundant	9	of 47 properties
VIF > 10	30	
MUTAG (16 non-constant properties)		
perfectly redundant	9	of 16 properties
VIF > 10	5	
IMDB-B (20 non-constant properties)		
perfectly redundant	8	of 20 properties
VIF > 10	10	

Table 20: Variance inflation factors of the property dictionaries. A property with an infinite factor is an exact linear function of the others, so no probe can separate it from them, and a drop after erasing it cannot be attributed to that property alone. This is why the verdict reports an S_2-proxy flag rather than treating every alien name as a distinct concept, and why the matched random control, rather than the size of the drop, carries the evidence.

E COMPUTE AND RECOVERY MODES

Experiments were run on an HPC cluster via SLURM on a single NVIDIA A100 GPU (80 GB), 8 CPU cores and 32 GB of RAM per job. Training each Grid-House and fMRI model took several hours of wall-clock time; all results on these datasets use the resulting published checkpoints. The analysis itself (probing, the three erasure operators, both random controls and the five-seed recovery, for every property) takes under 30 minutes per configuration and seed, and the MUTAG and IMDB-B jobs, which also train their model, take under 10 minutes. Recovery modes: for the Grid-House GCN (three-layer head) the last layer, the last two, or all three are retrained from the original weights, or the whole head from a fresh initialisation; for the Grid-House GIN and the fMRI and TU models (two-layer heads), the last layer or both, from the original or from fresh weights. Every mode trains to a validation plateau (3000-epoch cap, patience 50 evaluations) over five seeds, and ρ is the best mode's median.