Additional file 1

Supplementary methods and additional analyses for: EpiPPI: Epigenetic Attention-Based Orientation of the Protein–Protein Interaction Network Across Stress Conditions

Swati Tiwari and Dolly Sharma

Department of Computer Science and Engineering, Shiv Nadar Institution of Eminence, Greater Noida, Uttar Pradesh, India.

Section and table numbering prefixed with S refers to this file; all unprefixed section, table, and equation numbering refers to the main manuscript.

S1 Scope

This file provides the parameter settings and procedural detail required to reproduce the EpiPPI workflow, together with additional analyses supporting the main manuscript. It covers network and path preprocessing, construction of the protein matrix K and the condition-specific query matrices Q, the supervision and optimization of the Attention-Based Guidance System (ABGS), the operators and constants of the epigenetic genetic algorithm (epiGA), the staged benchmark configuration, repeated-run orientation analysis, computational configuration, and the reproducibility checkpoints. Results already reported in the main manuscript are cross-referenced rather than repeated.

Protein identity is resolved by Entrez Gene identifier throughout. Two networks are used and are not combined: the main five-condition network derived from BioGRID release 5.0.253 (5,914 proteins; 146,426 interactions; 56,140 conflicted decision edges), and the benchmark network derived from BioGRID v2.0.51 (3,325 proteins; 11,290 interactions; 861 conflicted decision edges) used exclusively for the replication and ablation experiments of Section 3.1.

Table S1 Principal dimensions of the main and benchmark networks

Item	Main network	Benchmark network
BioGRID release	5.0.253	v2.0.51
Largest connected component	5,914 proteins	3,325 proteins
Undirected interactions	146,426	11,290
Conflicted decision edges	56,140	861
Fixed (non-conflicted) edges	90,286	10,429
Conditions analyzed	Five	Stage 3 uses five condition scores
GA runs	10 per condition	10 per stage and condition

S2 Workflow overview

The workflow separates condition-independent preprocessing from condition-specific optimization. The processed PPI network and the protein matrix K are computed once and shared across all five conditions; each condition contributes its own Q matrix and its own Probability Guidance Scores (Prgs).

1. Filter the BioGRID physical-interaction network and retain the largest connected component.
2. Harmonize expression and curated protein identifiers to Entrez Gene identifiers.
3. Construct the condition-independent K matrix and five condition-specific Q matrices.
4. Enumerate simple source–target paths of at most five edges.
5. Calculate directional path evidence, identify conflicted edges, and construct the chromosome.
6. Train one eight-head ABGS model jointly across the five conditions and derive protein- and edge-level Prgs.
7. Run epiGA ten times per condition, saving the chromosome, directed conflicted edges, and generation history.
8. Reconstruct the full directed network by combining optimized conflicted edges with fixed non-conflicted edges.

9. Calculate pathway, KEGG, edge-frequency, and Jaccard summaries from the reconstructed networks.

S3 Network, curated protein set, and path preprocessing

S3.1 Rationale for network filtering

Restriction to experimentally supported physical interactions keeps the orientation problem focused on observed physical contacts. Genetic interactions describe functional rather than direct physical relationships, predicted associations lack equivalent experimental support, and complex-level records do not necessarily identify a directly interacting pair. Restriction to the largest connected component ensures that all retained proteins occupy a single connected search space in which source–target paths can be evaluated.

The BioGRID v2.0.51 network is processed by the same procedure but kept entirely separate, so that differences in network size, edge content, and preprocessing are not mixed into the main five-condition analysis.

S3.2 Composition of the curated signaling set

The 38-protein gold-standard set was curated for this study from KEGG pathway annotations and supporting literature. Each protein is recorded with its Entrez Gene identifier, its signaling role, and a pathway category. Sources are upstream sensors or initiating kinases; targets are downstream effectors, transcription factors, or terminal kinases. The set covers the osmotic-stress (HOG) cascade, the mating and filamentous-growth MAPK cascades, the DNA-damage checkpoint, cAMP–PKA signaling, chromatin regulation, cell-cycle control, polarity, and secretory-pathway components, so that each of the five analyzed conditions is represented. Proteins are resolved by Entrez Gene identifier at every stage of the pipeline. The full composition is given in Table S2.

The endpoint set used for path enumeration is larger than the supervision set and comprises 55 proteins: 31 sources, 28 targets, and four proteins occupying both roles. Of the 864 possible

non-self ordered source–target pairs, 648 contained at least one simple path of at most five edges and were retained for full enumeration. A pair without any path within the permitted depth cannot contribute to the path-based objective and is excluded at this screen.

The phase-1 screen counted paths for all 648 pairs. Full path sequences were subsequently written for 528 pairs, spanning 31 sources, 22 targets, four proteins in both roles, and 49 unique endpoints. Conflict detection, chromosome construction, and all fitness evaluation operate on this completed set.

Table S2 Composition of the 38 curated gold-standard signaling proteins

Sources		**Targets**	
Protein	Pathway	Protein	Pathway
SLN1	HOG	HOG1	HOG
SHO1	HOG	PBS2	HOG
MID2	HOG	MSN1	HOG, stress
MSB2	HOG	STE20	MAPK
STE50	MAPK	STE7	MAPK
RAS2	cAMP–PKA	FUS3	MAPK
GPR1	cAMP–PKA	STE5	MAPK
BCY1	cAMP–PKA	GPA1	MAPK
SIN3	chromatin	DIG1	MAPK
MEC1	DNA damage	DIG2	MAPK
TEL1	DNA damage	STE12	MAPK
RAD9	DNA damage	FUS1	MAPK
RAD24	DNA damage	FLO11	MAPK, filamentous growth
RGA1	CDC42 regulation	RAD53	DNA damage
RGA2	CDC42 regulation	DUN1	DNA damage
YCK1	stress	CDC42	polarity
YCK2	stress	SWI4	cell cycle
ACS2	chromatin acetylation	HUT1	ER–Golgi transport
RNH203	DNA repair		
ATG41	autophagy, stress		

Role denotes the assignment used for source–target path enumeration. Twenty-four of the 38 proteins additionally carry an explicit condition assignment, which supplies the anchor and condition-contrast loss terms; the remaining fourteen contribute through gs_flag, the triplet-ranking term and the edge-Prgs rule only. A machine-readable version of this table, including Entrez Gene identifiers, is released as `goldstandard_documented.csv` in the reproducibility repository.

S3.3 Path-enumeration procedure

For each retained source–target pair, paths are enumerated by stack-based iterative depth-first search. Each stack record holds the current node, the current path, and the set of visited nodes, so that every emitted path is simple; when the target is reached, the path is written and that branch is not extended further. Enumeration is executed with 20 worker processes and chunksize 1, and each pair is written to a separate file, `paths_<source>_<target>.csv`, with source, target, depth, and path_sequence columns. Existing pair files are skipped so that enumeration can be restarted without recomputation.

Intermediate proteins are retained explicitly in the stored sequence and depth is the number of edges. For example, a depth-two path is stored as `850325|851394|850330` and a depth-five path as `850325|851394|855750|854937|850647|850330`.

Simple paths are required so that a protein revisited within a route cannot contribute the same local interaction evidence more than once. Per-pair streaming is used because the enumerated path set is too large to hold as a single in-memory object.

S3.4 Hub filtering and path loading

Consistent with Section 2.3, a path is discarded when any intermediate node has degree greater than 300; source and target nodes are exempt from this test, because the filter is intended to control intermediate routing rather than remove the biological question represented by a curated endpoint pair. The threshold removes approximately the 50 most highly connected proteins of the largest connected component. Of the 11,304,109,036 paths enumerated across the 528 source–target pairs, 73,506,956 were retained, corresponding to 0.65% of the enumerated set. The value of 300 is a fixed empirical preprocessing choice and is not claimed to be an optimal biological cutoff.

The GA path loader processes path files in batches of 30 using six worker processes and CSV chunks of 100,000 rows. For each source–target file, $L_{\min}$ is estimated as the minimum depth observed in the first 100,000 rows. A retained path is represented only by its conflicted

chromosome edges and the directions those edges require; a path containing no conflicted decision edge cannot distinguish one chromosome from another and is excluded from fitness evaluation. The encoded path set is cached as `paths_cache_56k.pkl` and reused across all conditions and runs.

S3.5 Conflict-detection statistics

Applying the balance criterion of Equation (3) to the main network, 62,766 edges carried evidence in the canonical direction and 62,751 in the reverse direction, of which 61,773 carried evidence in both directions. Requiring balance ≥ 0.10 retained 56,140 decision edges. Their balance values ranged from 0.1000 to 0.5000, with mean 0.3257, first quartile 0.2362, median 0.3360, and third quartile 0.4217. The same procedure applied to the benchmark network produced 861 decision edges.

Every chromosome position therefore corresponds to an edge with retained path evidence in both directions; the decision set is ambiguous by construction. The 56,140 edges are sorted in descending order of balance score before encoding, so that the most ambiguous edges occupy the leading chromosome positions.

S3.6 Retention of low-confidence interactions

The composite edge confidence of Equation (5) is calculated from the evidence-code mapping of Table 2. Unlike the original PGMOGA preprocessing, which excluded interactions with $\mathrm{conf}(e) < 0.6$ before path enumeration, all 146,426 interactions of the largest connected component are retained, so that weakly supported but potentially condition-relevant interactions are not removed before path construction. Low-confidence interactions instead contribute proportionally smaller values to the path weight of Equation (2).

The confidence product and depth penalty of Equation (2) are used for preprocessing path weighting and conflict detection. In the epiGA fitness of Equation (13), the satisfied term is defined by mean edge Prgs and the unsatisfied term by mean edge Prgs together with the depth-decay factor; the confidence product is not carried into that evaluation.

S4 Condition-specific expression preprocessing

Table S3 Expression datasets and platforms used for the five conditions

Condition	GEO series	Platform	Representation
Heat Shock	GSE18	GPL54	log fold change relative to control
ER Stress	GSE18	GPL54	log fold change relative to control
Temp Steady State	GSE18	GPL54	log fold change relative to control
Osmotic Stress	GSE312455	GPL21656	log fold change relative to control
DNA Damage	GSE5301	GPL90	log fold change relative to control

Each dataset is converted to a condition-specific log-fold-change profile relative to its own unstressed or control state, and dataset-specific identifiers are harmonized to Entrez Gene identifiers before matching to the BioGRID network. Entrez harmonization provides a single identifier space shared by BioGRID proteins, expression records, GO-derived features, curated proteins, chromosome edges, and reconstructed networks.

Raw samples from the three platforms are not pooled into a single sample-level differential-expression model, and no cross-study batch correction is applied after the condition-specific log-fold-change values are calculated. Condition scale is handled within Q by the within-condition z-score, which standardizes each log-fold-change distribution so that conditions with different response ranges enter the shared ABGS model on a comparable relative scale.

Direct expression measurements were available for 431 of the 5,914 network proteins (7.3%), so that more than 92.7% of proteins rely on values obtained after preprocessing. Among the 38 curated proteins, 10 lacked direct measurements and were imputed by condition using the mean of available one-hop PPI neighbors, then available two-hop neighbors, and finally the condition mean.

S5 Additional detail on the feature matrices K and Q

S5.1 Protein matrix K

The 768-dimensional semantic block is generated from the concatenated GO Biological Process, Molecular Function, and Cellular Component text of each protein, including inherited parent terms. Each embedding is mean-centered and L_2-normalized across the protein set, so that cosine similarity between any two embeddings reduces to a dot product. The three topological features are transformed by $\log(1 + x)$ and min–max scaled to $[0, 1]$. The compartment block is a one-hot encoding over eight compartments derived from the most specific GO Cellular Component annotation; proteins without a Cellular Component annotation are assigned the zero vector. K is expression free and therefore identical across conditions, so that condition dependence enters only through Q.

S5.2 Query matrix Q

Table S4 Calculation of the seven Q features for each protein and condition

Index	Feature	Calculation
0	logFC	Condition-specific log fold change; 0.0 where unavailable
1	|logFC|	Absolute value of logFC
2	sign	Sign of logFC, giving -1, 0 or $+1$
3	z-score	(logFC − condition mean) / condition standard deviation, with the standard deviation floored at 10^{-6}
4	is_measured	$\tanh(n/10)$, where n is the number of PPI neighbors with available expression values after preprocessing; fixed at 0.90 for gold-standard proteins
5	comp_match	1 when the encoded compartment belongs to the condition-specific compartment set of Table S5; otherwise 0
6	gs_flag	1 for the 38 gold-standard Entrez identifiers; otherwise 0

Table S5 Condition-to-compartment mapping used by the comp_match feature

Condition	Compartments treated as matched
Heat Shock	cytoplasm; nucleus; endoplasmic reticulum
ER Stress	endoplasmic reticulum; Golgi apparatus; plasma membrane
Temp Steady State	cytoplasm; nucleus
Osmotic Stress	plasma membrane; cytoplasm; vacuole
DNA Damage	nucleus

S6 ABGS supervision and training

S6.1 Construction of the expression labels

For each condition, the 90th and 10th percentiles of the absolute expression response are calculated over the processed expression table. Proteins at or above the 90th percentile receive label 1 and proteins at or below the 10th percentile receive label 0; intermediate proteins do not contribute to the binary cross-entropy term. The loss uses a positive-class weight of $\max(N_{\text{negative}}/N_{\text{positive}}, 1)$.

The extreme deciles supply high- and low-response supervision without forcing a binary label onto the middle 80% of proteins, whose response is less clearly separated. Absolute response is used so that strong up- and down-regulation both indicate a condition-responsive protein, and class weighting prevents the smaller labeled class from contributing less solely because of its size.

S6.2 Definition of the five loss components

The weighted loss of Equation (10) has the following components.

- *Binary cross-entropy*: class-balanced cross-entropy on the top- and bottom-decile expression labels, averaged across available conditions.
- *Triplet-ranking loss*: $\max(0,\ \overline{\text{Prgs}}_{\text{silent}} - \overline{\text{Prgs}}_{\text{active GS}} + 0.15)$, averaged across eligible conditions.
- *Condition-contrast loss*: for each mapped gold-standard protein, penalizes a non-related

condition when its Prgs is not at least 0.15 below the protein's primary-condition Prgs.

- *Anchor loss*: mean squared error between Prgs and 0.87 for the gold-standard protein–condition assignments defined in `GS_CONDITION_MAP`, of which 24 of the 38 curated proteins receive an explicit condition assignment.
- L_2 *regularization*: sum of squared model parameters, weighted by 10^{-6}.

S6.3 Optimizer settings

Table S6 ABGS training configuration

Item	Value
Optimizer	Adam
Initial learning rate	3×10^{-4}
Schedule	CosineAnnealingLR over at most 2,500 epochs
Final scheduled learning rate	1×10^{-5}
Gradient clipping	global norm 1.0
Early-stopping patience	200 epochs
Minimum improvement	1×10^{-5} in total training loss
Dropout	0.1
Learnable temperature	initialized at 1.2
Trainable parameters	402,970
Model selection	parameter state at stopping

All loss components are calculated from the same jointly fitted model, and early stopping monitors the total training loss. No separate validation split is used, and the trained weights are applied to all five conditions.

S6.4 Protein-to-edge Prgs mapping

The endpoint-weighting rule summarized in Section 2.7 is implemented as follows. A missing endpoint Prgs defaults to 0.5, and the edge score is calculated by one of three rules according to the gold-standard status of the endpoints:

- Both endpoints gold standard: $\max(s_{\mathrm{dir}}, 0.5) \times [\mathrm{Prgs}(u) + \mathrm{Prgs}(v)]/2$, where the direction score s_{dir} is twice the absolute deviation of the forward evidence ratio from 0.5.
- One endpoint gold standard: $0.7\,\mathrm{Prgs}(\text{GS endpoint}) + 0.3\,\mathrm{Prgs}(\text{other endpoint})$.

- Neither endpoint gold standard: $\min[\mathrm{Prgs}(u), \mathrm{Prgs}(v)]$.

The tiered rule gives greater influence to a curated endpoint when exactly one endpoint is curated, requires both endpoints to be supported when neither is curated, and combines endpoint relevance with directional path evidence when both are curated. The resulting edge score is guidance for the search and is not an experimentally measured edge direction.

S7 epiGA parameter settings

Table S7 Constants used by the main-network epiGA

Parameter	Value	Role
Population size N_{pop}	50	chromosomes per generation
Generations G	200	evolutionary iterations
Runs	10	independent runs per condition
Mutation rate p_{mut}	0.01	per unlocked bit
Elite size E	5	highest-fitness chromosomes retained
Selection pool N_{sel}	20	uniform parent selection from the top 20
Lock threshold τ_{lock}	0.70	edge Prgs above which a bit is locked
Initialization threshold τ_{bias}	0.15	minimum endpoint Prgs difference for directed initialization
Reprogramming patience	20	consecutive non-improving generations
Global seed	42	Python, NumPy, and PyTorch initialized before the run loop

Locked positions retain their initial values at population creation; once a global best is found, its locked positions are copied into the bookmark and restored in subsequently generated children. Because the chromosome encoding is unchanged, locking alters the accessibility of a position to the evolutionary operators rather than removing an edge from the representation. Random generators are initialized once with seed 42 before the run loop, so that successive runs of a condition consume a continuing random stream.

The neighborhood block used by the recombination operator preserves a graph-local group of incident edges, which is more consistent with PPI structure than splitting the chromosome at arbitrary bit positions, since adjacent chromosome entries do not necessarily represent adjacent proteins. Mutation is restricted to unlocked positions so that it supplies exploration without overriding the lock mask.

S8 Orientation of edges outside the chromosome

Only the 56,140 conflicted edges are optimized. The remaining 90,286 interactions are oriented by three deterministic rules: an edge with balance below 0.10 follows its majority weighted direction; an edge observed in only one path direction retains that direction; and an edge absent from all enumerated paths follows canonical Entrez-identifier ordering.

This separation confines evolutionary computation to ambiguous edges, since majority-supported and single-direction edges already carry a path-evidence orientation and need not consume chromosome variables. Canonical Entrez ordering supplies a reproducible storage direction only where the enumerated path set provides no directional evidence, and is not a biologically inferred direction.

S9 Configuration of the staged benchmark

All three benchmark stages reported in Section 3.1 use the BioGRID v2.0.51 network, the same 861-bit chromosome, ten runs per stage, population size 100, 50 generations, and mutation rate 0.001. The stages differ in path weighting and in the source of biological guidance, as summarized in Table S8.

Table S8 Objective and guidance definitions for the three benchmark stages

Stage	Path weight	Selection fitness	Reported metric	Biological guidance
1: PGMOGA replication	original cached path weight	satisfied-path count − weighted unsatisfied paths	sum of weights of satisfied paths	PGMOGA operators
2: Epigenetic	cached weight$^{1/\text{depth}}$	same formula using normalized weights	sum of normalized weights of satisfied paths	random 20% locking with the four epigenetic operators
3: ABGS multi-head	as Stage 2	as Stage 2	as Stage 2	ABGS-guided locking

Stages 2 and 3 therefore share the same normalized path-weighted objective and are directly comparable, isolating the contribution of ABGS-guided locking as the 10.8% improvement

reported in Section 3.1. Stage 1 retains the original PGMOGA path weighting, so the 55.8% and 72.6% values are cumulative stage-wise differences that include the change in path weighting. Benchmark fitness does not use Prgs inside the objective; ABGS enters the benchmark only through locking guidance, whereas the Prgs-weighted objective of Equation (13) applies to the main condition-specific pipeline. The Stage 3 value summarizes the five condition means, and the reported dispersion is the sample standard deviation across those means.

S10 Additional analyses

S10.1 Lock coverage and repeated-run orientation agreement

The lock mask is determined by the condition-specific edge Prgs and is identical across the ten runs of a condition, covering between 8.6% and 10.6% of the 56,140 chromosome positions. To characterize run-to-run behavior beyond the objective values reported in Table 5, all 50 saved directed-edge files were compared over the same 56,140 decision edges.

Table S9 Lock coverage and within-condition orientation agreement across ten runs

Condition	Locked edges	Pairwise agreement (mean ± SD)	Unlocked agreement	Unanimous	At least 9/10
Heat Shock	5,843 (10.4%)	50.2200% ± 0.2082%	50.2536%	0.3295%	2.4760%
ER Stress	4,849 (8.6%)	50.2072% ± 0.1979%	50.2368%	0.3153%	2.4653%
Temp Steady State	5,961 (10.6%)	50.2743% ± 0.2106%	50.3046%	0.3242%	2.5864%
Osmotic Stress	5,315 (9.5%)	50.2163% ± 0.2161%	50.2367%	0.2993%	2.4118%
DNA Damage	4,976 (8.9%)	50.3129% ± 0.2318%	50.3402%	0.3545%	2.6042%

SD standard deviation. Percentages in the second column are relative to the 56,140 conflicted decision edges.

Across the 225 within-condition run pairs, mean directional agreement was 50.2462% (SD 0.2152%); across 1,000 between-condition run pairs it was 50.1827% (SD 0.2080%), a difference of 0.063 percentage points. Because these comparisons reuse individual runs, they are descriptive rather than independent observations. Condition-specific 5/5 tie rates ranged from 23.76% to 24.22%, and among non-tied consensus edges, opposite-condition consensus ranged from 48.36% to 49.24%.

This analysis provides the quantitative basis for the statement in Section 3.3 that the 93.4% of directed edges differing among the five selected networks may reflect both condition effects and run-to-run stochasticity. Objective values are stable across runs while individual edge orientations are not, which is expected in a 56,140-bit search space containing many near-equivalent high-scoring chromosomes and in an objective that includes no edge-consensus term. Edge-level conclusions are therefore expressed as recovery frequencies across runs, as done for MEC1→RAD53 and RAD53→DUN1 in Table 5.

S10.2 Scope of the pathway and KEGG validation

The six pathway edges of Table 8 and the eight KEGG edges of Table 9 are defined after network reconstruction and are not used inside the objective function, so they are not directly optimized by the genetic algorithm. Each of the eight KEGG edges involves at least one curated gold-standard protein, five involve two, and eight of the ten unique proteins belong to the 38-protein supervision set. GO annotations also enter the model through the semantic block of K.

These checks therefore assess biological consistency within the curated knowledge used by the study rather than generalization to an independent, held-out direction set. Within that scope, the highest concordance under DNA Damage and the condition-dependent recovery of RAD53→DUN1 indicate that the reconstructed orientations track the expected condition-specific signaling logic.

S11 Computational configuration

The enumerated path directory occupied 620.78 GB and the encoded path cache `paths_cache_56k.pkl` occupied 3.972 GB.

The reported experiments use fixed values for the conflict balance threshold (0.10), the intermediate-node degree cutoff (300), population size (50), generations (200), mutation rate (0.01), lock threshold (0.70), initialization threshold (0.15), elite size (5), and reprogramming

Table S10 Execution configuration by pipeline stage

Stage	Configuration
Path enumeration	Intel Core i9 13th generation, 24 cores / 32 logical processors, 63.7 GB RAM; 20 worker processes; chunksize 1; output streamed to external storage
GA path loading	six processes; 30 files per batch; 100,000-row CSV chunks; persistent `paths_cache_56k.pkl`
ABGS training	PyTorch; CUDA when available, otherwise CPU
Main epiGA	PyTorch; CUDA when available, otherwise CPU; path tensors moved to the selected device
Benchmark stages	CUDA device

patience (20). No sensitivity sweep over these values was performed.

S12 Limitations

The following considerations complement the limitations discussed in Section 4 and define the scope within which the reported results should be interpreted.

- ABGS is trained without a held-out validation split, and early stopping monitors total training loss. Generalization of the guidance scores to proteins outside the training distribution was not separately quantified.
- The 38-protein curated set contributes to Q through gs_flag, to ABGS through the ranking, contrast, and anchor terms, and to the pathway and KEGG checks. GO annotations also enter K through the semantic block. Concordance values are therefore internal biological-consistency measures rather than independent external validation, and no GO-masked ablation was performed.
- Direct expression measurements are available for 431 of 5,914 proteins (7.3%); the condition-specific signal for the remainder depends on neighborhood imputation and on shared protein features.
- Expression data derive from three microarray platforms. Condition-specific log-fold-change values are standardized within condition by z-score, without cross-study batch correction.

- Path enumeration is limited to simple paths of at most five edges, and paths through intermediate nodes of degree greater than 300 are excluded. All conclusions apply to the retained path set.
- The balance threshold (0.10), degree cutoff (300), lock threshold (0.70), and initialization threshold (0.15) are fixed empirical choices; no sensitivity analysis over these values was performed.
- Edges absent from all enumerated paths are oriented by canonical Entrez ordering. This convention is deterministic and reproducible but carries no biological information, and its downstream effect was not separately tested.
- Repeated runs converge to comparable objective values while differing at the level of individual edge orientations, and the objective contains no edge-stability or consensus term. Edge-level conclusions are therefore reported as recovery frequencies across runs rather than as single-run orientations.
- The BioGRID physical-interaction network does not represent every phosphorylation or kinase–substrate relation, which limits the set of curated directions that can be tested. No independent directional resource, such as SIGNOR, YEASTRACT, Reactome, or a phosphoproteomic dataset, was evaluated.
- The 55.8% and 72.6% benchmark values are cumulative differences that include the change in path weighting; the directly comparable measure of ABGS guidance is the 10.8% Stage 2 to Stage 3 improvement. None of these values represents an increase in the raw number of satisfied paths.
- Complete measured runtime, peak memory, and scaling curves were not recorded for all stages.
- All conclusions apply to the *Saccharomyces cerevisiae* BioGRID interactome and to the five analyzed stress conditions.

S13 Reproducibility checkpoints

Table S11 Implementation files and their roles

File	Role
`01_biobert_embed.py`	generates the 768-dimensional GO-text embeddings for the semantic block of K
`02_k_matrix.py`	assembles the $5{,}914 \times 779$ protein matrix K
`03_q_matrix.py`	constructs the five $5{,}914 \times 7$ condition-specific Q matrices
`04_confidence.py`	aggregates BioGRID evidence codes into per-edge confidence
`05_dfs.py`	enumerates simple paths of at most five edges with restart-safe per-pair output
`06_conflict.py`	calculates directional path evidence and constructs the conflicted-edge chromosome
`09_abgs_v4_final.py`	defines the eight-head ABGS model, the five-component loss, and the tiered edge Prgs
`11_epiga_all_conditions.py`	runs the 56,140-bit epiGA and saves per-run histories and edge directions
`12_reconstruction_all_conditions.py`	reconstructs the five condition-specific directed networks
`13_pathways.py`	calculates pathway and KEGG directional concordance
`pgmoga.py`	Stage 1 PGMOGA replication
`epigenetic_only.py`	Stage 2 normalized-weight epigenetic benchmark
`multihead_all_condition.py`	Stage 3 ABGS-guided benchmark across five conditions

Table S12 Persisted data checkpoints

Checkpoint	Content
`K_matrix_779_biobert.csv`	condition-independent protein matrix K
`combined_expression_logfc_imputed.csv`	processed expression consumed by the Q builder
`Q_matrix_v2.pkl` / `Q_matrix_v2.csv`	five condition-specific seven-feature matrices
`conflicted_edges.pkl`	56,140 conflicted edges with forward and reverse path evidence
`prgs_all_conditions_v4.pkl`	protein-level Prgs for all five conditions
`prgs_edges_all_conditions_v4.pkl`	edge-level Prgs for all five conditions
`abgs_v4_weights.pt`	trained ABGS weights
`paths_cache_56k.pkl`	retained paths encoded over the conflicted decision edges
`epiga_result_<condition>_56k_v4_run<r>.pkl`	chromosome, parameters, validation output, and history for one run
`directed_edges_<condition>_56k_v4_run<r>.csv`	directed conflicted edges for one run
`epiga_history_<condition>_56k_v4_run<r>.csv`	generation-level history for one run
`directed_full_<condition>.csv`	reconstructed directed network for one condition
`kegg_validation.csv`	per-condition KEGG directional concordance
`Goldstandard.csv`	38-protein Entrez-mapped curated source and target set
`phase1_summary_k5.csv`	phase-1 path-positive source–target screen

The code, configuration files, software environment, and reproduction workflow are available in the EpiPPI reproducibility repository at `https://github.com/st872-hue/EpiPPI-reproducibility`. The processed datasets, saved epiGA run outputs, reconstructed condition-specific directed networks, and validation results are available through Zenodo at `https://doi.org/10.5281/zenodo.21407573`.

S14 Future validation

The analyses in this file identify three directions for further validation. First, consensus-based reconstruction across larger replicate sets, together with explicit edge-stability regularization or constraints, would allow a high-confidence subnetwork to be defined by repeated-run recovery rather than by a single selected run. Second, GO-masked, expression-free, and parameter-sensitivity analyses would separate the contributions of semantic knowledge, expression guidance, and the fixed preprocessing thresholds. Third, independent directional resources or condition-specific experimental data would be required to test generalization beyond the curated knowledge used in the present validation.

Abbreviations

ABGS	Attention-Based Guidance System
epiGA	Epigenetic genetic algorithm
ER	Endoplasmic reticulum
GA	Genetic algorithm
GEO	Gene Expression Omnibus
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
logFC	Log fold change
MAPK	Mitogen-activated protein kinase
PGMOGA	Pseudo-guided multi-objective genetic algorithm
PPI	Protein–protein interaction
Prgs	Probability Guidance Score

SD	Standard deviation