The Theory Behind Stellar Mind AI's Application Tools for Space Biology

Quemada-Torres, Eulogio, Román-Sánchez, Alejandro, Montes-Pérez, Javier, Decena-Gímenez, Macorís, Dávila-Moreno, Candela

University of Málaga, Blvr. Louis Pasteur, 35, 29071 Málaga, Spain.

To cite this article: Quemada-Torres, E., Román-Sánchez, A., Montes-Pérez, J., Decena-Gímenez, M., Dávila-Moreno, C. 2025. The Theory Behind Stellar Mind AI's Application Tools for Space Biology. NASA Space Apps Challenge.

Abstract

We present a formal and fully theoretical specification of *Stellar Mind AI*, a space-biology knowledge engine featuring a graph-centric retrieval-augmented generation (GraphRAG) layer, an *Assay Finder* that maps natural-language intent to a normalized assay manifold, a decision-oriented *Gap Finder* based on interpretable coverage functionals, and a reserved *Mission Builder*. We define objects, invariants, and compact scoring principles grounded in spectral graph theory, submodular optimization, and exponential-family modeling. No implementation or API detail is discussed.

Keywords: Retrieval-Augmented Generation, Knowledge Graphs, Space Biology, Spectral Methods, Submodularity

Notation.. Let $G=(V,E)$ be a typed graph; L_H the normalized Laplacian of a subgraph H; $\lambda_2(L_H)$ its algebraic connectivity; π_q^{PPR} a seeded diffusion from query q; $\mathbb{1}\{\cdot\}$ the indicator. We write $\langle\cdot,\cdot\rangle$ for inner products, $\mathrm{KL}(\cdot\|\cdot)$ for Kullback–Leibler divergence.

1. Introduction

Space-biology knowledge is fundamentally relational. Transformer encoders model sequence-level semantics (Vaswani et al., 2017), while knowledge graphs canonize entities and typed relations (Hogan et al., 2021). Retrieval-augmented generation (RAG) composes both views (Manning et al., 2008; Gao and et al., 2023), and graph-based RAG (*GraphRAG*) elevates retrieval from documents to structured subgraphs, enabling global sensemaking and multi-hop attribution (Edge et al., 2024; Microsoft Research, 2024; Larson and Truitt, 2024; Peng and et al., 2024; Zhang and et al., 2025). We formalize a theoretical layer for four modules—GraphRAG, Assay Finder, Gap Finder, and a reserved Mission Builder—using spectral connectivity, random-walk diffusion, and submodular selection as first-class primitives (Chung, 1997; Andersen et al., 2006; Tong et al., 2006; Krause and Golovin, 2014).

2. GraphRAG: Subgraph Retrieval and Faithful Conditioning

Objects. From corpus $\mathscr{D}$ we induce a typed, heterogeneous graph $G=(V,E,\tau_V,\tau_E)$ with canonical labels and optional embeddings $\mathbf{z}_v,\mathbf{z}_e$ for textual evidence and mentions. Typed relations encode a constrained hypothesis class whose latent geometry can be regularized by knowledge–graph embedding priors (e.g., translational, complex, rotational constraints) to promote type-consistent neighborhoods and equivariant scoring (Bordes et al., 2013; Trouillon et al., 2016; Sun et al., 2019). GraphRAG assumes that global sensemaking emerges from *structure-aware retrieval* rather than bag-of-passages: queries target *subgraphs* that preserve multi-hop support and community-level semantics (Edge et al., 2024; Larson and Truitt, 2024; Microsoft Research, 2024).

Intent and constraints. A natural-language query q is mapped to an abstract intent $\mathscr{I}(q)=(\mathscr{C},\mathscr{M},\mathscr{F})$: (i) hard constraints $\mathscr{C}$ over entities/types; (ii) soft *motifs* $\mathscr{M}$ (relation templates, e.g., $\mathsf{perturbs}\circ\mathsf{expresses}$); (iii) semantic facets $\mathscr{F}$ (topic/domain vectors). These play the roles of feasibility, weak pattern matching, y *coverage* semántica respectivamente. Given $\mathscr{I}(q)$, retrieval selects a connected subgraph $H\subseteq G$ as the maximizer of the composite objective

$$
\begin{aligned}
H^\star \in \arg\max_{H\subseteq G} \lambda_1\,\Phi_{\mathsf{m}}(H;\mathscr{C},\mathscr{M},\mathscr{F}) + \lambda_2\,\Psi_{\mathsf{coh}}(H) \\
+\lambda_3\,\Xi_{\mathsf{cov}}(H) - \lambda_4\,R_{\mathsf{red}}(H), \qquad (1)
\end{aligned}
$$

subject to type-compatibility and connectivity. Each term captures a distinct theoretical desideratum:

- *Motif/constraint satisfaction* Φ_{m} scores typed pattern match with slack on facets, promoting subgraphs whose path-types agree with $\mathscr{M}$ and whose node/edge labels are feasible under $\mathscr{C}$.
- *Cohesion* $\Psi_{\mathsf{coh}}(H)=\lambda_2(L_H)$ uses the algebraic connectivity (second eigenvalue of the normalized Laplacian L_H). By Cheeger-type inequalities, large λ_2 implies low conductance and internally well-connected

subgraphs (spectral cuts avoid spurious bridges) (Chung, 1997; Shi and Malik, 2000; Ng et al., 2001).

- *Coverage* $\Xi_{\mathsf{cov}}(H) = \sum_{v \in V(H)} \pi_q^{\mathrm{PPR}}(v)$ aggregates a seeded diffusion (personalized PageRank / random-walk-with-restart) from query-induced seeds, thus favoring regions with high stationary mass near relevant anchors (Andersen et al., 2006; Tong et al., 2006).

- *Redundancy penalty* R_{red} discourages near-duplicates and improves diversity, instantiated either as a facility-location repulsion (submodular) or as a DPP energy over features/paths with kernel K on $V(H)$ (Kulesza and Taskar, 2012; Krause and Golovin, 2014).

Why spectral + diffusion?. Spectral cohesion (λ_2) controls the isoperimetric profile of H; diffusion captures *locality* relative to q. Together they select subgraphs that are (i) *internally tight* (few boundary edges), (ii) *externally well-anchored* near query seeds, and (iii) *semantically on-pattern* via Φ_{m}. This mirrors local partitioning results where PPR "finds" low-conductance sets around a seed (Andersen et al., 2006), while RWR provides relevance scores to rank candidate neighborhoods (Tong et al., 2006).

Algorithmic relaxations and guarantees. The exact problem (1) couples connectivity, pattern constraints y diversidad, lo que es NP-hard. Dos relajaciones cubren casos prácticos:

(R1) Relajación espectral. Optimizar Rayleigh quotients $\mathbf{u}^\top L_H \mathbf{u}$ con restricciones suaves de pertenencia produce vectores propios que, al umbralizar, dan cortes de baja conductancia con garantías clásicas (Chung, 1997; Ng et al., 2001; Shi and Malik, 2000). Se puede incorporar Φ_{m} ponderando L (p.ej., aristas compatibles) o prefiltrando el soporte.

(R2) Relajación submodular. Si $\Phi_{\mathsf{m}} + \Xi_{\mathsf{cov}}$ es (aprox.) monótona submodular y R_{red} es un matroide (o se usa un surrogate submodular de DPP), el greedy bajo presupuesto/knapsack obtiene factor $(1 - 1/e)$ (Krause and Golovin, 2014). La conectividad se repara a posteriori vía augmentación de Steiner mínima sobre el corte seleccionado, o se impone *a priori* con restricciones de árbol en expansión.

Faithful conditioning and invariants.. Let $\mathscr{P}(H)$ denote typed *minimal support paths.* Conditioning is constrained by three invariants (compact form to fit a column):

$$
\begin{aligned}
&\text{(I) } \forall c \in \mathscr{C}_{\mathrm{at}}\ \exists p \in \mathscr{P}(H):\ p \models c && \text{(provenance closure)}\\
&\text{(II) } \mathscr{N} \circ \mathscr{N} = \mathscr{N} \ \text{ on } V(H), E(H) && \text{(idempotent normalization)}\\
&\text{(III) } \max_{|\mathrm{ctx}| \le B} \{\mathrm{DPP(ctx)} + \mathrm{Cov(ctx)}\} && \text{(diversity packing)}
\end{aligned}
$$

(I) forces every atomic claim to be backed by a typed path; (II) avoids drifting entity canonicalization across decoding steps; (III) formalizes context selection as a small-B packing that balances diversity (repulsion) and coverage, consistent with DPP and submodular objectives used in diverse summarization (Kulesza and Taskar, 2012; Krause and Golovin, 2014; Yu and et al., 2024). These constraints align with attribution/evaluation practices in RAG and with the "local-to-global" design of GraphRAG (Edge et al., 2024; Larson and Truitt, 2024).

Communities and local-to-global sensemaking.. Global queries (e.g., "main themes") are ill-posed for paragraph retrieval but natural for graph *communities.* Let $\{C_k\}$ be clusters (spectral or modularity-based) (Newman, 2006; Blondel et al., 2008; Ng et al., 2001). GraphRAG precomputes summaries $s_k = \mathscr{S}(C_k)$ and, at query time, weights them by $\lambda_2(L_{C_k})$ (internal coherence) and $\pi_q^{\mathrm{PPR}}(C_k)$ (query affinity), then composes partial answers $\to$ final synthesis (Edge et al., 2024). This *two-level* pipeline explains empirically observed gains on "global" QFS-style questions while preserving attribution to H's paths (Edge et al., 2024; Microsoft Research, 2024).

Relation to standard RAG.. Classical RAG scores independent chunks; GraphRAG scores *structured* neighborhoods. Replacing TF–IDF/dot-product retrieval by (1) yields (i) fewer boundary hallucinations (vía λ_2 y caminos tipados), (ii) mejor cobertura multihop (difusión sembrada), y (iii) *diversidad controlada* del contexto (DPP/submodular), lo cual facilita la evaluación de fe y cobertura (Gao and et al., 2023; Yu and et al., 2024).

Design checklist (theoretical).. *Typed feasibility* (respecto a $\mathscr{C}$); *Motif slack* calibrado en Φ_{m}; *Cohesion floor* $\lambda_2(L_H) \ge \varepsilon$; *Coverage budget* via PPR mass $\sum_{v \in V(H)} \pi_q^{\mathrm{PPR}}(v) \ge \eta$; *Diversity budget* $|\mathrm{ctx}| \le B$ con repulsión DPP o facility-location. Este "perfil" abstrae implementaciones y deja claros los grados de libertad del modelo (Andersen et al., 2006; Tong et al., 2006; Kulesza and Taskar, 2012; Krause and Golovin, 2014).

3. Assay Finder: Exponential-Family Projection on a Normalized Manifold

Manifold and normalization. Let the assay manifold be

$$\mathscr{X} = \mathscr{O} \times \mathscr{T} \times \mathscr{C} \times \mathscr{A},$$

where $\mathscr{O}$ (organism), $\mathscr{T}$ (parent tissue), $\mathscr{C}$ (condition; coarse Spaceflight vs. Ground/Analog), and $\mathscr{A}$ (assay type) are finite controlled sets. A normalization operator $\mathscr{N}$ acts on raw metadata strings and induces an equivalence relation $\sim_{\mathscr{N}}$ whose classes are the canonical tokens. We require

$$\mathscr{N} \circ \mathscr{N} = \mathscr{N} \quad \text{and} \quad x \sim_{\mathscr{N}} x' \Rightarrow \text{ same axes in } \mathscr{X},$$

i.e., idempotence and axis-consistent canonicalization. In practice, $\mathscr{N}$ is grounded by anatomy/biomedical ontologies on $\mathscr{T}$ and $\mathscr{O}$ (e.g., parent-tissue lifts and taxon canonicalization) (Mungall and et al., 2012; Smith and et al., 2007). Let $\Omega \subseteq \mathscr{X}$ be the observed (nonempty) subset.

Semantic encoding and exponential family. A request q is mapped to a natural parameter $\theta = \theta(q) \in \mathbb{R}^d$ and to sufficient statistics $\mathbf{f}: \mathscr{X} \to \mathbb{R}^d$ (token indicators and semantic features). We define the exponential family

$$\pi_\theta(x) = \frac{1}{Z(\theta)} \exp\big(\langle \theta, \mathbf{f}(x) \rangle\big), \qquad Z(\theta) = \sum_{x \in \mathscr{X}} e^{\langle \theta, \mathbf{f}(x) \rangle},$$

with standard properties of minimal, regular exponential families on finite domains (Brown, 1986; Wainwright and Jordan, 2008). A *concept set* is the superlevel region

$$\mathscr{A}_q(\tau) = \{x \in \mathscr{X} : \pi_\theta(x) \ge \tau\}, \quad \tau \in (0,1).$$

We assume a feasibility polytope $\mathscr{M} \subset \Delta(\mathscr{X})$ enforcing ontology/legal constraints (e.g., type compatibility). A raw intent distribution μ (e.g., a soft prior over $\mathscr{X}$) is projected as an I-projection

$$\hat{\pi} = \arg\min_{\pi \in \mathscr{E} \cap \mathscr{M}} \mathrm{KL}(\mu \| \pi), \quad \mathscr{E} = \{\pi_\theta : \theta \in \mathbb{R}^d\},$$

which is equivalent to maximum-entropy under moment constraints (Jaynes, 1957; Csiszár, 1975; Amari and Nagaoka, 2000).

Principles. (P1) $\mathscr{N}$ idempotent; (P2) *support discipline:* if selection is restricted to the observed mask (i.e., we enforce $\mathrm{supp}(\pi) \subseteq \Omega$ via $\mathscr{M}$), then the projection introduces no mass outside Ω; otherwise, forward KL does not penalize assigning mass where μ has zeros, so additional support may appear (Csiszár, 1975). (P3) *auditability*: membership $x \in \mathscr{A}_q(\tau)$ is justified by typed paths in G (cf. Sec. 2).

Existence/uniqueness and moment matching.. Since $\mathscr{E}$ is a regular minimal exponential family on finite $\mathscr{X}$, the objective is strictly convex in π and the feasible set $\mathscr{E} \cap \mathscr{M}$ is convex; thus $\hat{\pi}$ exists and is unique whenever μ admits a feasible moment vector $\mathbb{E}_\mu[\mathbf{f}]$ in the relative interior of the convex moment set induced by $\mathscr{M}$ (Brown, 1986; Wainwright and Jordan, 2008). Moreover, $\hat{\theta}$ satisfies the moment conditions

$$\mathbb{E}_{\hat{\pi}}[\mathbf{f}] = \mathbb{E}_\mu[\mathbf{f}] \quad (\text{maximum-entropy}/I\text{-projection}).$$

Grouping, deduplication, comparability. We formalize three operations, expressed purely over $(\mathscr{X}, \Omega)$:

(G1) Grouping by technology. Let $g : \mathscr{X} \to \mathscr{A}$ be the projection $g(o,t,c,a) = a$. For any distribution ρ on $\mathscr{X}$ (e.g., $\rho = \hat{\pi}$), define group-mass

$$\Gamma(a) = \sum_{x \in \mathscr{X} : g(x) = a} \rho(x), \quad a \in \mathscr{A}.$$

(G2) Deduplication. Let $\equiv$ be an equivalence relation on Ω capturing duplicates at the assay granularity (e.g., same canonical assay-name within dataset). The canonical representative map $D : \Omega \to \Omega/\equiv$ is idempotent and order-preserving for any scoring that is constant on classes.

(G3) Comparability flags. Define binary functionals on $x = \langle o,t,c,a \rangle$:

$$\chi_{\mathrm{Sf}}(x) = \mathbb{1}\{c = \mathrm{Sf}\}, \quad \chi_{\mathrm{Gnd}}(x) = \mathbb{1}\{c = \mathrm{Gnd}\},$$

and a "both" flag at the (o,t,a) level:

$$\chi_{\mathrm{Both}}(o,t,a) = \mathbb{1}\Big\{\exists c \neq c' : \langle o,t,c,a \rangle, \langle o,t,c',a \rangle \in \Omega\Big\}.$$

These flags are measurable in Ω and invariant under $\mathscr{N}$.

Prioritization on the manifold. Let $\kappa : \mathscr{X} \to \mathbb{N}$ count instances in Ω (cf. Gap Finder). Define compact, interpretable signals on x:

$$\eta_1(x) = \chi_{\mathrm{Sf}}(x), \quad \eta_2(x) = \chi_{\mathrm{Gnd}}(x), \quad \eta_3(x) = \chi_{\mathrm{Both}}(o,t,a),$$

$$\eta_4(x) = \min\Big(1, \frac{\kappa(\langle o,t,\mathrm{Gnd},a \rangle)}{3}\Big)$$

$$\eta_5(x) = \min\Big(1, \frac{\kappa(\langle o,t,\mathrm{Sf},a \rangle)}{3}\Big)$$

and a neighborhood density on the product space,

$$\eta_6(x) = \frac{1}{Z_\rho} \sum_{y \in \mathscr{N}_\rho(x)} \kappa(y),$$

with $\mathscr{N}_\rho$ a product-metric neighborhood and Z_ρ a normalizer. For $\alpha \in \mathbb{R}^6_{\geq 0}$, the *assay-score*

$$R(x) = \langle \alpha, \eta(x) \rangle \quad \text{and} \quad R^*(a) = \sum_{x : g(x) = a} \hat{\pi}(x) R(x)$$

provide (i) *cell-level* prioritization within $\mathscr{X}$ and (ii) *technology-level* prioritization via R^*. By construction, R is monotone in each η_j and stable under $\mathscr{N}$.

Stability and calibration.. (S1) Normalization stability. If $x \sim_{\mathscr{N}} x'$, then $\mathbf{f}(x) = \mathbf{f}(x')$ and $\eta_j(x) = \eta_j(x')$, hence $\pi_\theta(x) = \pi_\theta(x')$ and $R(x) = R(x')$.

(S2) Threshold calibration. The level-sets $\mathscr{A}_q(\tau)$ are nested: $\tau_1 \leq \tau_2 \Rightarrow \mathscr{A}_q(\tau_2) \subseteq \mathscr{A}_q(\tau_1)$. For any group a, the group-mass $\Gamma(a)$ is a right-continuous, piecewise-constant, nonincreasing function of τ on finite $\mathscr{X}$ (i.e., changes only at finitely many breakpoints).

(S3) Feasibility preservation. If μ respects $\mathscr{M}$ and $\mathscr{N}$ (e.g., masking to Ω and canonical tokens), then $\hat{\pi}$ respects both by construction (projection onto $\mathscr{E} \cap \mathscr{M}$); in particular, no mass is assigned to infeasible axes or non-canonical tokens.

Explainability and audit. Let $\eta(x)$ be the feature vector and $R(x) = \sum_j \alpha_j \eta_j(x)$. The *principal reason* for a selection at x is $\arg\max_j \alpha_j \eta_j(x)$. Because η_j are counts/flags over (o,t,c,a) or local neighborhoods, explanations reduce to reporting sufficient statistics and typed support paths in G validating $x \in \mathscr{A}_q(\tau)$ (cf. Sec. 2). This yields compact, auditable rationales with no implementation dependence.

Optional diversity at presentation-time.. If one wishes to select a small panel $\mathscr{S} \subset \mathscr{A}_q(\tau)$ of size K for display, define

$$F(A) = \sum_{x \in A} R(x) + \gamma \mathrm{Div}(A), \quad |A| \leq K,$$

where Div is a monotone submodular diversity (e.g., facility-location on a similarity kernel over $\mathscr{X}$). Then the greedy algorithm achieves a $(1 - 1/e)$ approximation to $\max_{|A| \leq K} F(A)$ (Nemhauser et al., 1978; Lin and Bilmes, 2011), and DPP-based Div induces repulsion among near-duplicates (Kulesza and Taskar, 2012).

4. Gap Finder: Coverage Functionals and Interpretable Ranking

Axes, measure, and coverage. Let $\mathscr{X} = \mathscr{O} \times \mathscr{T} \times \mathscr{C} \times \mathscr{A}$ be the product space (organism, parent tissue, condition $\{\mathrm{Sf}, \mathrm{Gnd}\}$, assay). Normalization $\mathscr{N}$ is idempotent and axis-consistent (cf. Sec. 3). Let $\kappa : \mathscr{X} \to \mathbb{N}$ be a multiplicity functional (count of distinct observed instances; $\kappa \equiv 0$ off Ω). Fix $m \in \mathbb{N}$; define the covered set and the *gap set*

$$\mathscr{COV} = \{x : \kappa(x) \geq m\}, \qquad \mathscr{G} = \mathscr{X} \setminus \mathscr{COV}.$$

We endow $\mathscr{X}$ with a product metric $d(x,y) = \sum_j w_j d_j(x_j, y_j)$ (discrete on categorical axes), and with the counting measure $\mu_\kappa(A) = \sum_{x \in A} \kappa(x)$.

Neighborhoods and feasibility. For $\rho > 0$ let $\mathscr{N}_\rho(x) = \{y \in \mathscr{X} : d(x,y) \leq \rho\}$ and define a normalized density

$$\nu_\rho(x) = \frac{1}{Z_\rho} \sum_{y \in \mathscr{N}_\rho(x)} \kappa(y), \qquad Z_\rho = \sum_{y \in \mathscr{X}} \kappa(y).$$

Assay feasibility is encoded by $\varphi : \mathscr{A} \to [0,1]$; set $\vartheta(x) = \varphi(a)$ for $x = \langle o,t,c,a \rangle$.

Signals (compact Greek block). For $x = \langle o,t,c,a \rangle \in \mathscr{G}$ define

$$\begin{aligned}
\gamma(x) &= \min\Big(1, \frac{\kappa(\langle o,t,\mathrm{Gnd},a \rangle)}{3}\Big) \mathbb{1}\{c = \mathrm{Sf}\}, \\
\mu(x) &= \mathbb{1}\{\exists a' \neq a : \kappa(\langle o,t,c,a' \rangle) \geq 1\}, \\
\phi(x) &= \mathbb{1}\{\exists c' \neq c : \kappa(\langle o,t,c',a \rangle) \geq 1\}, \\
\sigma(x) &= \mathbb{1}\{\exists o' \neq o : \kappa(\langle o',t,c,a \rangle) \geq 1\}, \\
\nu_\rho(x) &= \frac{1}{Z_\rho} \sum_{y \in \mathscr{N}_\rho(x)} \kappa(y), \qquad \vartheta(x) = \varphi(a).
\end{aligned}$$

(Abbreviations: Sf = Spaceflight; Gnd = Ground/Analog.)

Scoring functional and properties. Let $\alpha \in \mathbb{R}^6_{\geq 0}$ and $\beta \geq 0$. Define the compact score

$$\begin{aligned}
S(x) = {} & \alpha_1 \gamma(x) + \alpha_2 \mu(x) + \alpha_3 \phi(x) \\
& + \alpha_4 \sigma(x) + \alpha_5 \nu_\rho(x) + \alpha_6 \vartheta(x) - \beta \,\mathrm{Red}(x).
\end{aligned}$$

Here $\mathrm{Red}(x)$ is a redundancy proxy (e.g., kernel similarity to previously selected gaps). Then:

- *Monotonicity in signals.* S is nondecreasing in each positive signal (by linear construction).
- *Normalization stability.* If $x \sim_{\mathscr{N}} x'$, then $\gamma, \mu, \phi, \sigma, \nu_\rho, \vartheta$ coincide, hence $S(x) = S(x')$.
- *Bounded sensitivity (local).* For x,y with $d(x,y) \leq \rho$, the variation of ν_ρ is bounded by $\frac{\max_z \kappa(z)}{Z_\rho}$ times the discrepancy in neighborhood overlap; on finite domains this induces local stability of S.

Diversity-aware selection (budgeted). Given $K \in \mathbb{N}$, select a panel $A \subseteq \mathscr{G}$, $|A| \leq K$, by

$$F(A) = \sum_{x \in A} S(x) + \gamma \mathrm{Div}(A),$$

where Div promotes dispersion:

$$\begin{aligned}
\text{Facility-location: } \mathrm{Div}(A) &= \sum_{y \in \mathscr{G}} \max_{x \in A} k(x,y), \\
\text{DPP (log-MAP): } \mathrm{Div}(A) &= \log\det(K_A),
\end{aligned}$$

with a PSD kernel $K = (k(x,y))$ over $\mathscr{G}$. If Div is monotone submodular, the cardinality-constrained greedy achieves $(1 - 1/e)$ (Nemhauser et al., 1978; Krause and Golovin, 2014). For DPPs, greedy-MAP selections encourage explicit repulsion and diversity (Kulesza and Taskar, 2012).

Connectivity repair (optional).. If connectivity is desired in projections of A (e.g., by tissue), apply a minimal post-augmentation via a Steiner-style repair on the co-occurrence graph over $\mathscr{X}$; this keeps the score close to $F(A)$ in practice, and the repair can leverage approximation algorithms for Steiner trees (Vazirani, 2003; Robins and Zelikovsky, 2000).

Explainability as sensitivity decomposition. Let $s(x) = (\gamma, \mu, \phi, \sigma, \nu_\rho, \vartheta)$ and $S(x) = \langle \alpha, s(x) \rangle - \beta \,\mathrm{Red}(x)$. *Principal reason*: $\arg\max_j \alpha_j s_j(x)$. *Detail*: report $s(x)$ and, when applicable, underlying statistics (e.g., counts in $\mathscr{N}_\rho(x)$). Auditability is completed with typed paths in G that justify the signals (cf. Sec. 2).

Calibration and scope.. Let $\mathscr{X}_{\mathrm{sc}} \subseteq \mathscr{X}$ be the *scope* induced by filters (or the observed support). Then: (i) κ and ν_ρ are computed over $\mathscr{X}_{\mathrm{sc}}$; (ii) S is invariant to extensions outside $\mathscr{X}_{\mathrm{sc}}$ that do not modify κ on $\mathscr{N}_\rho(x)$; (iii) increasing m shrinks $\mathscr{COV}$ and expands $\mathscr{G}$, with S stable when $\gamma, \mu, \phi, \sigma$ remain unchanged.

Relation to Assay Finder.. Assay Finder defines $\hat{\pi}$ over $\mathscr{X}$ (Sec. 3). An optional compatible prior rescales S as $\tilde{S}(x) = \hat{\pi}(x) S(x)$, focusing the panel on conceptually relevant cells without altering monotonicity or normalization stability.

5. Mission Builder: Multiobjective Planning with Evidence-Constrained Synthesis

Objects and evidence. Let $\mathscr{O}$ be a finite set of mission objectives, $\mathscr{P}$ a finite set of phases (e.g., PDR→CDR→Ops), and $\mathscr{R}$ a resource index (mass, fuel, crew, power, ...). A *mission architecture* is a tuple

$$\mathbf{m} = (\Theta, \Pi, \mathscr{T}, \rho, \xi),$$

where $\Theta \subseteq \mathscr{O}$ (selected objectives), Π is a phasewise policy, $\mathscr{T}$ is a precedence-respecting timeline, $\rho \in \mathbb{R}^{|\mathscr{R}|}_{\geq 0}$ is the aggregated resource vector, and $\xi \in [0,1]$ is an aggregated risk level. GraphRAG (Sec. 2) contributes *typed evidence* $\mathsf{E} = \{(c_i, w_i)\}_{i=1}^M$ with weights $w_i \in [0,1]$ that support Θ and key decisions in Π via typed shortest paths in G.

Timeline and precedence. Let $(\mathscr{P}, \preceq)$ be a precedence DAG. A *schedule* is a map $t : \mathscr{P} \to \mathbb{R}_{\geq 0}$ such that

$$p \preceq p' \Rightarrow t(p) + d(p) \leq t(p'),$$

where $d(p)$ is the canonical duration of p. Define aggregate slack $\Delta = \sum_{p \in \mathscr{P}} \max\{0, t_{\max}(p) - t(p) - d(p)\}$ (Pinedo, 2016).

Resources and feasibility. Let $r_p \in \mathbb{R}^{|\mathscr{R}|}_{\geq 0}$ be phase consumption and $r = \sum_p r_p$ the total; let $\bar{r} \in \mathbb{R}^{|\mathscr{R}|}_{\geq 0}$ be limits.

$$\text{Feasibility:} \qquad r \leq \bar{r} \quad \text{(componentwise)}.$$

For each objective $\theta \in \Theta$, specify minimal requirements $\underline{r}(\theta)$ and a set of enabling phases $\mathscr{P}(\theta)$ (as in RCPSP-style feasibility) (Brucker et al., 1999; Herroelen and Leus, 2005).

Risk model (coherent). Let Ω be a finite scenario set with probability $\mathbb{P}$. Each $p \in \mathscr{P}$ induces a nonnegative loss $L_p(\omega)$;

total $L(\omega) = \sum_p L_p(\omega)$. Define risk by a coherent risk measure, e.g., CVaR_α:

$$\xi = \mathrm{CVaR}_\alpha\big(L\big) = \min_{\tau\in\mathbb{R}} \tau + \frac{1}{1-\alpha}\,\mathbb{E}\big[(L-\tau)_+\big],$$

with coherence per Artzner et al. (1999) and convex representation per Rockafellar and Uryasev (2000, 2002). For stochastic planning background, see Shapiro et al. (2009).

Utility and evidence constraints. Let $u : \mathscr{O} \to \mathbb{R}_{\geq 0}$ denote objective utility and $g : \mathscr{P} \to \mathbb{R}_{\geq 0}$ denote gains for critical phases. The structural utility is

$$U(\Theta,\Pi) = \sum_{\theta\in\Theta} u(\theta) \;+\; \sum_{p\in\mathscr{P}} g(p)\mathbb{1}\{\Pi \text{ executes } p\}.$$

Evidence E imposes *consistency*: there exists a set $\mathscr{Q} \subseteq \mathscr{P}(G)$ of typed paths in G such that

$$\forall\theta\in\Theta\;\exists q\in\mathscr{Q}:\; q\models\theta, \qquad \sum_{(c_i,w_i)\in\mathsf{E}:\,c_i\in q} w_i \;\geq\; \tau_{\mathrm{ev}},$$

and analogously for decisions in Π (justification of instruments/resources). This enforces architecture-level provenance-closure (cf. Sec. 2).

Multiobjective synthesis (scalarized). With weights $\lambda = (\lambda_U, \lambda_\Delta, \lambda_\rho, \lambda_\xi)$, define

$$J(\mathbf{m}) = \lambda_U\, U(\Theta,\Pi) + \lambda_\Delta\,\Delta - \lambda_\rho\,\|r\|_1 - \lambda_\xi\,\xi.$$

Mission Builder solves

$$\max_{\mathbf{m}}\; J(\mathbf{m}) \quad \text{s.t.} \quad \begin{cases} \text{precedences and durations,} \\ r\leq\bar{r},\;\; \theta\in\Theta \Rightarrow \mathscr{P}(\theta)\subseteq\Pi, \\ \text{consistency with } \mathsf{E},\;\; \xi\leq\bar{\xi}. \end{cases}$$

Pareto fronts can be traced by varying λ (weighted-sum scalarization) or via ε-constraint formulations (Miettinen, 1999; Haimes et al., 1971). (Weighted sums recover supported efficient points under standard convexity assumptions.)

Context packing for rationale (compact). Let $\mathrm{ctx} \subseteq \mathsf{E}$ be a *budgeted* evidence portfolio to accompany the architecture. Select ctx by

$$\max_{|\mathrm{ctx}|\leq B}\; \mathrm{DPP}(\mathrm{ctx}) + \mathrm{Cov}(\mathrm{ctx};\Theta,\Pi),$$

where DPP induces repulsion and Cov is a coverage functional over objectives/phases (Kulesza and Taskar, 2012; Nemhauser et al., 1978; Krause and Golovin, 2014). This yields compact, nonredundant rationales within budget B.

Phase policies (Markovian abstraction). Let $\mathscr{S}$ be a high-level state space with $s_{t+1} = F(s_t, a_t, \omega_t)$ and $a_t \in \mathscr{A}_t$ (phase decisions). A policy $\Pi = (\pi_t)$ is Markovian if $a_t = \pi_t(s_t)$. For finite horizons, a stochastic formulation is

$$\max_{\Pi,\,t(\cdot)}\; \mathbb{E}\left[\sum_t r_t(s_t,a_t) - \lambda_\xi L(\omega)\right]$$
$$\text{s.t.}\quad \text{precedences}, \quad r\leq\bar{r}, \quad \text{consistency with } \mathsf{E}.$$

in the spirit of constrained MDPs and risk-aware planning (Puterman, 1994; Shapiro et al., 2009). The realized policy is "compiled" to $\mathbf{m}$ by extracting achieved Θ, $\mathscr{T}$, and ρ.

Feasibility cuts and repair. If a candidate violates evidence or resources, add cuts of the form

$$\sum_{p\in\mathscr{P}(\theta)} z_p \;\geq\; 1 \quad (\theta\in\Theta), \qquad \sum_p a_{pr} z_p \;\leq\; \bar{r}_r \;(r\in\mathscr{R}),$$

with $z_p \in \{0,1\}$ execution indicators (valid inequalities/Benders-style logic cuts) (Geoffrion, 1972). Precedence repairs can be posed as minimal augmentations on the temporal DAG (using Steiner-style heuristics); this is a pragmatic post-step leveraging approximation algorithms, rather than introducing new guarantees (Vazirani, 2003; Robins and Zelikovsky, 2000).

Compact invariants (column-friendly).

$$\begin{aligned} &\text{(I)}\ \forall\theta\in\Theta\;\exists q\in\mathscr{Q}:\; q\models\theta && \text{(provenance)} \\ &\text{(II)}\ r\leq\bar{r},\;\; \xi\leq\bar{\xi},\;\; t \text{ respects } (\mathscr{P},\preceq) && \text{(feasibility)} \\ &\text{(III)}\ \max_{\mathrm{ctx}}\{\mathrm{DPP}+\mathrm{Cov}\} \text{ s.t. } |\mathrm{ctx}|\leq B && \text{(rationale diversity)} \end{aligned}$$

(II) separates resources, risk, and time; (III) aligns explanation with diversity/coverage principles.

Manager-oriented views (scores). Define a *mission score*

$$\mathscr{S}(\mathbf{m}) = \beta_1\,U + \beta_2\,\Delta - \beta_3\,\|r\|_1 - \beta_4\,\xi,$$

and an optional *gap-closure gain* $\sum_{x\in\mathscr{G}} \tilde{S}(x)\mathbb{1}\{\text{closed by } \mathbf{m}\}$ (cf. Sec. 4). Both are traceable: $\mathscr{S}$ decomposes by objectives/phases, and each term is justified by ctx and typed paths in G.

References

Amari, S.-i., Nagaoka, H., 2000. Methods of Information Geometry. AMS & Oxford University Press.

Andersen, R., Chung, F., Lang, K., 2006. Local graph partitioning using pagerank vectors. In: FOCS.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9 (3), 203–228. DOI: 10.1111/1467-9965.00068

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. In: NeurIPS.

Brown, L. D., 1986. Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. IMS.

Brucker, P., Drexl, A., Möhring, R., Neumann, K., Pesch, E., 1999. Resource-constrained project scheduling: Notation, classification, models, and methods. European Journal of Operational Research 112 (1), 3–41. DOI: 10.1016/S0377-2217(98)00204-5

Chung, F. R. K., 1997. Spectral Graph Theory. American Mathematical Society.

Csiszár, I., 1975. I-divergence geometry of probability distributions and minimization problems. The Annals of Probability 3 (1), 146–158. DOI: 10.1214/aop/1176996454

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J., 2024. From local to global: A graph rag approach to query-focused summarization. arXiv:2404.16130.

Gao, Y., et al., 2023. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

Geoffrion, A. M., 1972. Generalized benders decomposition. Journal of Optimization Theory and Applications 10 (4), 237–260. DOI: 10.1007/BF00934810

Haimes, Y. Y., Lasdon, L. S., Wismer, D. A., 1971. On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE Transactions on Systems, Man, and Cybernetics SMC-1 (3), 296–297.
DOI: 10.1109/TSMC.1971.4308298

Herroelen, W., Leus, R., 2005. Project scheduling under uncertainty: Survey and research potentials. European Journal of Operational Research 165 (2), 289–306.
DOI: 10.1016/j.ejor.2004.04.002

Hogan, A., Blomqvist, E., Cochez, M., et al., 2021. Knowledge graphs. ACM Computing Surveys 54 (4), 1–37.
DOI: 10.1145/3447772

Jaynes, E. T., 1957. Information theory and statistical mechanics. Physical Review 106 (4), 620–630.
DOI: 10.1103/PhysRev.106.620

Krause, A., Golovin, D., 2014. Submodular function maximization. In: Tractability: Practical Approaches to Hard Problems. Cambridge University Press.

Kulesza, A., Taskar, B., 2012. Determinantal Point Processes for Machine Learning. Now Publishers.

Larson, J., Truitt, S., February 2024. Graphrag: Unlocking llm discovery on narrative private data.

Lin, H., Bilmes, J., 2011. Class-subset selection via submodular augmentation. In: Proc. NIPS.

Manning, C. D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

Microsoft Research, 2024. Project graphrag. Project page.

Miettinen, K., 1999. Nonlinear Multiobjective Optimization. Springer.
DOI: 10.1007/978-1-4615-5563-6

Mungall, C. J., et al., 2012. Uberon: an integrative multi-species anatomy ontology. Genome Biology 13 (1), R5.

Nemhauser, G. L., Wolsey, L. A., Fisher, M. L., 1978. An analysis of approximations for maximizing submodular set functions. Mathematical Programming 14, 265–294.
DOI: 10.1007/BF01588971

Newman, M. E. J., 2006. Modularity and community structure in networks. PNAS 103 (23), 8577–8582.

Ng, A. Y., Jordan, M. I., Weiss, Y., 2001. On spectral clustering: Analysis and an algorithm. In: NeurIPS.

Peng, B., et al., 2024. Graph retrieval-augmented generation: A survey. arXiv:2408.08921.

Pinedo, M. L., 2016. Scheduling: Theory, Algorithms, and Systems, 5th Edition. Springer.
DOI: 10.1007/978-3-319-26580-3

Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley.

Robins, G., Zelikovsky, A., 2000. Improved steiner tree approximation in graphs. SIAM Journal on Discrete Mathematics 19 (1), 122–134.
DOI: 10.1137/S0895480100375002

Rockafellar, R. T., Uryasev, S., 2000. Optimization of conditional value-at-risk. Journal of Risk 2 (3), 21–41.

Rockafellar, R. T., Uryasev, S., 2002. Conditional value-at-risk for general loss distributions. Journal of Banking & Finance 26 (7), 1443–1471.
DOI: 10.1016/S0378-4266(02)00271-6

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2009. Lectures on Stochastic Programming: Modeling and Theory. SIAM.
DOI: 10.1137/1.9780898718751

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. TPAMI 22 (8), 888–905.

Smith, B., et al., 2007. The obo foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25 (11), 1251–1255.

Sun, Z., Deng, Z.-H., Nie, J.-Y., Tang, J., 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In: ICLR.

Tong, H., Faloutsos, C., Pan, J.-Y., 2006. Fast random walk with restart and its applications. In: ICDM. pp. 613–622.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G., 2016. Complex embeddings for simple link prediction. In: ICML.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS).

Vazirani, V. V., 2003. Approximation Algorithms. Springer.
DOI: 10.1007/978-3-662-04565-7

Wainwright, M. J., Jordan, M. I., 2008. Graphical models, exponential families, and variational inference. In: Foundations and Trends in Machine Learning. Vol. 1. pp. 1–305.
DOI: 10.1561/2200000001

Yu, H., et al., 2024. Evaluation of retrieval-augmented generation: A survey. arXiv:2405.07437.

Zhang, Q., et al., 2025. A survey of graph retrieval-augmented generation for llms. OpenReview.