

# MouseInteract: 小鼠蛋白质相互作用 网络的功能模块发现与可视化分析

## 一、全局网络拓扑：度分布的幂律特征分析

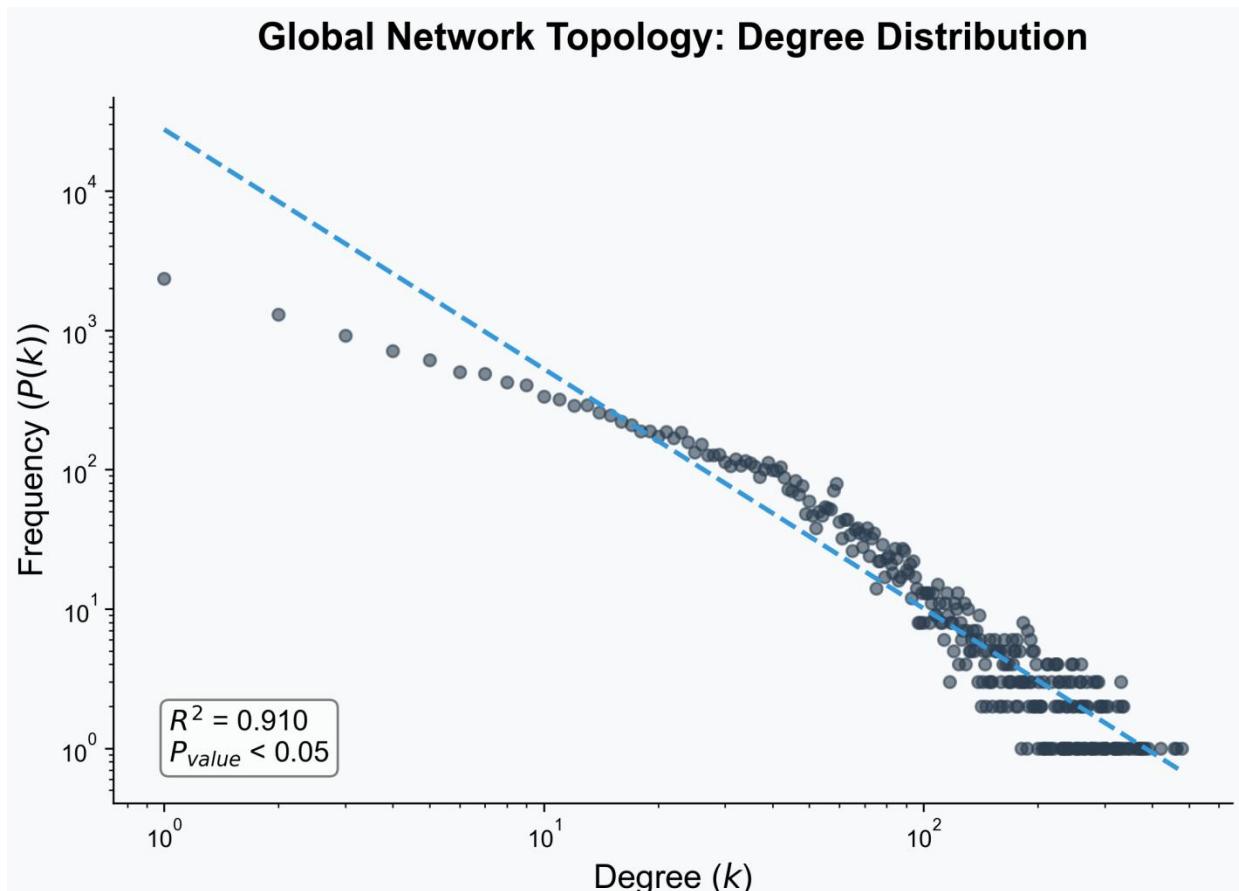


图 1 全局网络拓扑：度分布

### 1.1 研究背景与分析目的

在系统生物学中，蛋白质相互作用网络（PPI）并非随机生成的图结构，而是一个具有高度组织性的复杂系统。分析全局网络的拓扑结构是理解生物系统鲁棒性（Robustness）和脆弱性（Vulnerability）的第一步。图 1 展示的是基于 STRING 数据库（*Mus musculus*）构建的全基因组 PPI 网络的度分布（Degree Distribution）。

度 (Degree,  $k$ ) 定义为一个蛋白质节点与其他蛋白质直接连接的数量。度分布  $P(k)$  则反映了网络中拥有特定连接数的节点所占的比例。我们的核心目标是验证小鼠 PPI 网络是否符合“无尺度网络 (Scale-free Network)”的特征，即是否存在少数承载绝大部分连接的“枢纽蛋白 (Hub Proteins) ”。

## 1.2 实验方法与技术路径

根据配套脚本 图 1.py 的逻辑，该分析经历了以下严谨的步骤：

1. **数据高压过滤:** 从原始的 STRING 链路数据中筛选出 `combined_score > 700` 的记录。在 STRING 体系中，700 分代表“高置信度 (High Confidence) ”，这有效剔除了通过低质量文本挖掘产生的噪音，确保后续分析基于具有强生物学证据的相互作用。
2. **网络构建:** 利用 NetworkX 库将数万个蛋白质及其相互作用转化为数学图模型  $G$ 。
3. **对数转换与统计:** 由于生物网络中节点的度数差异极大（从 1 到数百），直接绘图会导致数据堆叠在原点。因此，我们将度  $k$  和频率  $P(k)$  进行  $\log_{10}$  转换，将幂律关系  $P(k) \sim k^{-\gamma}$  转化为线性关系  $\log P(k) \sim -\gamma \log k$ 。
4. **回归拟合:** 采用 `scipy.stats.linregress` 进行最小二乘法拟合，计算判定系数  $R^2$  和显著性  $P$  值。

## 1.3 图像细节的微观解剖

- **坐标系与刻度:** 横轴为  $\text{Degree } (k)$ ，纵轴为  $\text{Frequency } (P(k))$ ，均采用 1 到  $10^2$  以上的对数刻度。这种坐标设计能够清晰展示横跨多个数量级的数据分布。
- **散点分布 (深色半透明圆点) :**

**左侧高频区:** 在  $k < 10$  的区域，圆点密集且频率极高。这表明网络中绝大多数蛋白质（数以千计）只与极少数邻居发生作用，它们执行的是相对孤立或特异性的生物功能。

**右侧长尾区:** 随着  $k$  值增加到  $10^2$  (即连接数超过 100)，点数骤减但依然存在。

这些位于长尾末端的节点即为“枢纽蛋白”。

- **蓝色虚线（拟合趋势）**：这条斜向下的直线是对观测数据的数学抽象。其倾斜程度（斜率）即为幂律指数 $\gamma$ 。
- **统计信息框**：左下角标注 $R^2 = 0.910$ ，这是一个极高的拟合优度，证明了模型的解释力； $P_{value} < 0.05$  则说明这种分布模式在统计学上是极其显著的，而非随机巧合。

## 1.4 核心发现与生物学意义

1. **无尺度特征的确认：**  $R^2 = 0.910$  强有力地证明了小鼠 PPI 网络是一个典型的无尺度网络。这意味着生命系统在进化过程中为了保证信息传递的高效性，选择了这种高度集中的拓扑模式。
2. **系统的鲁棒性：** 由于“枢纽蛋白”只占极少数，随机的基因突变或蛋白质损伤更有可能发生在那些低度数的节点上。网络对这种随机“故障”具有极强的容错能力，不会导致整个细胞系统的瘫痪。
3. **脆弱的“阿喀琉斯之踵”：** 反之，如果受到攻击（如病毒入侵或关键致癌突变）的对象是那些位于长尾右侧的枢纽节点（如 TP53、AKT1 等），整个蛋白质交互网络可能会迅速崩解。
4. **演化动力学：** 这种幂律分布通常源于“优先连接（Preferential Attachment）”机制，即在进化中，新产生的蛋白质更倾向于与那些已经具有很多连接的成熟蛋白质建立联系。

## 二、关键节点（Hubs）识别

### 2.1 研究目的

本部分旨在通过拓扑分析识别小鼠蛋白质相互作用网络（PPI network）中的起关键连接或转运作用的蛋白质（Hubs）。通过计算中心性指标，量化各个蛋白质在网络中的重要程度，希望能够回答以下几个问题：

- (1) 在给定的置信度网络中，哪些蛋白质拥有最多的直接相互作用，从而可能处于生

生物学通路的核心位置？（2）是否存在某些蛋白质虽然连接数不是最多，但却处于不同功能模块之间的关键节点上，起着控制信息流动的关键作用？（3）度中心性与介数中心性最高的蛋白质列表是否存在重叠？（例如，TP53 是否既是节点又是桥接点？）

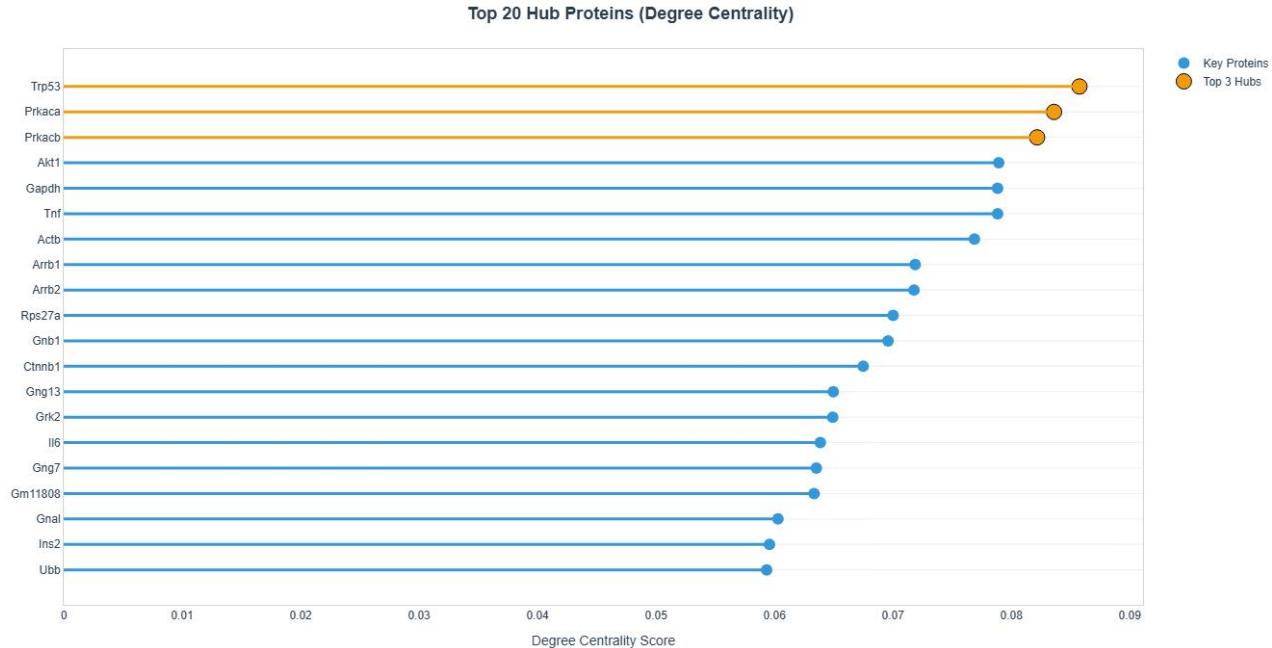


图 2 蛋白质关键节点棒棒糖图

## 2.2 数据来源与文件说明

数据来源为 STRING database v12.0，物种：小鼠 Mus musculus (10090)。下载入口为 STRING 官方下载页面（用户提供链接）。

本方向主要使用以下文件：

(1) 10090.protein.links.v12.0.txt.gz: 蛋白质网络数据。

作用：这是构建网络拓扑结构的核心文件。分析代码将读取该文件中的 protein1 (节点 A)、protein2 (节点 B) 以及 combined\_score (综合评分)，以此绘图 (Graph)，并基于此计算节点的度中心性和介数中心性，从而量化蛋白质在网络中的重要性。

(2) 10090.protein.info.v12.0.txt.gz: STRING 蛋白列表。

作用：数据包括蛋白 ID、名称和描述。通过代码直接计算出的 Top 20 关键节点通常对应 STRING 内部 ID (如 10090.ENSMUSP00000000001)，需要通过此文件映射为常用的缩写 (如 Tp53, Akt1)，以便在棒棒糖图 (Lollipop Chart) 的 Y 轴标签中直观展示。

## 2.3 方法与流程

本部分遵循“数据 → 处理 → 算法 → 可视化”的流程构建图 2。

(1) 构图与计算：基于预清洗的数据（“combined\_score ≥ 400”），使用 “NetworkX” 构建无向图。计算所有节点的度中心性 (Degree Centrality)，该指标反映了蛋白质在网络中直接互作伙伴的数量，是衡量“枢纽”地位的核心指标。(2) 排序与截取：将所有节点按度中心性从高到低排序，提取前 20 个节点 (Top 20) 作为关键候选蛋白进行展示。

可视化编码，严格按照以下规范绘制的棒棒糖图：(1) 颜色编码：对前 3 的核心蛋白进行高亮显示。Top 3 核心枢纽：使用高亮色（“#F39C12 / Orange”），节点尺寸放大 (Size 16)，以强调网络中最重要的三个蛋白。Top 4-20：使用主色调（“#3498DB / Scientific Blue”），节点尺寸适中 (Size 12)。(2) 坐标轴优化：X 轴表示度中心性分值，强制设定线性刻度（“dtick=0.02”）并去除网格线（“showgrid=False”），保持背景洁净。(3) 交互提示：显示具体排名和中心性数值。

图表读图说明：(1) Y 轴：关键蛋白质名称（已映射为 Gene Name）。(2) 棒长/位置：向右延伸越长 (X 轴数值越大)，代表该蛋白在网络中的连接度越高，生物学功能可能越核心。(3) 颜色差异：橙色突出显示了网络中前三名核心蛋白。

## 2.4 主要结论

### 2.4.1 棒棒图解析

在 Top 20 关键节点中，Trp53 与其后的 Prkaca 和 Prkacb 形成了明显的第一梯队，尤其是 Trp53，其度中心性 (Degree Centrality) 接近 0.09，呈现出断崖式的领先优势。这表明在当前构建的小鼠蛋白质相互作用网络中，信息流极其依赖这少数几个核心节点，网络的鲁棒性可能高度依赖于这些节点的完整性。如果移除 Trp53，整个网络的连通性可能会遭受最严重的破坏。

其次，观察 Top 20 列表的详细构成，可以发现网络骨架主要由信号转导激酶主导。除了前三名中的 Prkaca/b，列表中还包含了 Akt1 (调控细胞生存与代谢的关键激酶)、Grk2 以及多个 G 蛋白亚基 (Gnb1, Gng13, Gng7)。这种分布模式强烈暗示了该 PPI 网络的核心骨架是围绕着细胞内的主要信号级联反应构建的。与此同时，列表混入了 Gapdh (代谢

酶/内参)、Actb(细胞骨架)以及Tnf/Il6(炎症因子)，说明该网络并未局限于单一功能，而是广泛覆盖了细胞代谢、结构支撑以及免疫应答等基础生命活动。

#### 2.4.2 生物学结论推断

细胞核心生命活动的网络映射Top 3 Hubs的生物学身份与它们在网络拓扑中的地位高度吻合，负责整合各类压力信号并决定细胞是修复DNA还是走向凋亡，这种功能在拓扑结构上就表现为与大量上下游蛋白的密集连接。同样，Prkaca/Prkacb作为cAMP信号通路的核心执行者，需要对代谢、基因表达和细胞骨架进行多重调控。

Ubb(泛素B)出现在榜单中，且排名靠前，这揭示了蛋白质降解系统在网络中的基础性作用。泛素化是一种普遍存在的蛋白质修饰方式，几乎细胞内所有的蛋白质在生命周期的某个阶段都可能被泛素标记并降解。因此，泛素蛋白在物理上会与成千上万种底物发生短暂或稳定的接触。在PPI网络中，这使得Ubb天然成为一个结构性的Hub。它的高连接度反映了“蛋白质周转”是维持该网络稳态的必需机制，而非特定的信号传导事件。

### 2.5 后续验证与优化

图中出现了异常节点Gm11808，后续必须执行严格的验证步骤。首先在NCBI Gene或UniProt数据库中查询该ID的详细注释，确认其是否为假基因。如果确认是假基因或注释质量低下的预测蛋白，应当在数据预处理阶段将其进行剔除，然后重新运行代码。

目前的分析完全基于度中心性(Degree)，然而，网络中还存在一类桥接蛋白，后续需要补充计算介数中心性(Betweenness Centrality)。这类蛋白可能自身的直接连接数并不多，但它们占据了连接不同功能模块(如“免疫模块”与“代谢模块”)的必经之路。识别出这些高介数节点，对于寻找能够阻断信号跨通路传播的关键药物靶点往往比单纯寻找节点更具价值。

为了验证Top 20结果的稳健性，后续进行阈值敏感性测试，建议将阈值设置为400、700和900，重新绘制棒棒糖图。如果Trp53和Prkaca在高阈值下依然稳居前三，说明它们的核心地位是由高质量的实验证据支撑的；如果某些节点在高阈值下消失，则说明其连接可靠性较低。

### 三、社区检测与聚类（Louvain/Leiden）

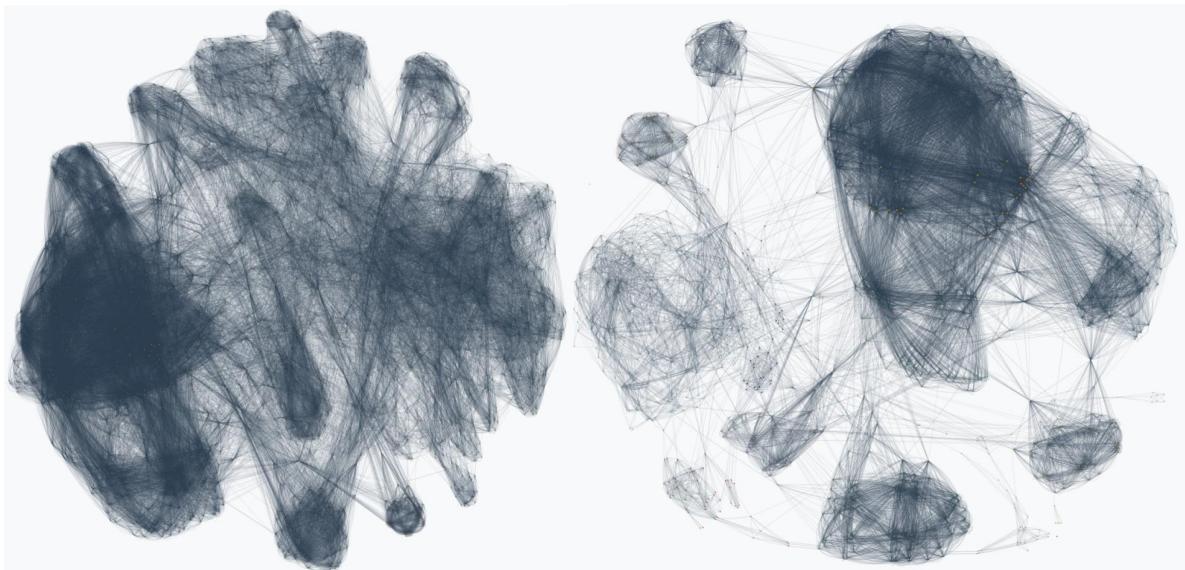


图 3 基于 STRING v12.0 小鼠 (10090) 蛋白互作网络的社区检测可视化 (combined\_score  $\geq 700$ )。节点颜色表示 Louvain 识别的社区模块，橙金色节点为连接度排名前 20 的关键枢纽蛋白 (Top hubs)；该图用于揭示网络的模块化结构及潜在跨模块连接特征。

#### 3.1 研究目的

本部分基于 STRING 数据库构建小鼠 (*Mus musculus*, 10090) 蛋白质相互作用网络 (PPI network)，通过社区检测 (Community Detection) 识别网络中的潜在功能模块/蛋白复合体结构，并评估网络是否呈现明显的模块化特征。具体希望回答以下问题：

- 1) 在较高置信度互作阈值下，网络能否分解为若干内部连接紧密的模块（社区）？
- 2) 网络中是否存在连接度显著更高的关键枢纽蛋白 (hubs)？
- 3) 不同模块之间是否通过少量桥接节点/桥接边相互连接，从而形成跨模块耦合结构？

#### 3.2 数据来源与文件说明

数据来源为 STRING database v12.0，物种：小鼠 *Mus musculus* (10090)。下载入口为 STRING 官方下载页面 (用户提供链接)。本方向使用/涉及文件如下 (并说明其作用): 10090.protein.links.v12.0.txt.gz：PPI 边列表，包含 protein1、protein2 与综合评分 combined\_score。本图网络结构 (节点、边、权重) 主要由该文件构建。

10090.protein.info.v12.0.txt.gz：节点注释信息文件，提供 STRING 蛋白 ID 对应的 preferred\_name（常用符号/简称）及描述信息，用于可视化时 tooltip（悬停提示）显示蛋白信息。10090.protein\_aliases.v12.0.txt.gz（可选）：名称/ID 映射文件，用于将 STRING 内部 ID 映射到其他数据库标识（如 UniProt/Ensembl/gene name），便于后续结果解释与跨库对齐。10090.protein.links.detailed.v12.0.txt.gz（可选）：分证据通道得分文件，可用于后续“证据来源分析”（方向四）或解释边的来源可靠性，本图（方向三）不强制使用。

### 3.3 方法与流程（可复现）

本部分遵循“数据 → 处理 → 算法 → 可视化”的流程构建 Figure 3。边筛选与加权构图：1) 从 10090.protein.links.v12.0.txt.gz 读取互作边(protein1, protein2, combined\_score)。2) 采用阈值过滤：仅保留 combined\_score  $\geq 700$  的边（本图标题为 cutoff=700；阈值可在 400/700/900 等间调整以做稳健性分析）。3) 构建 无向加权图：节点：蛋白（STRING protein id），边：互作关系，权重：weight = combined\_score / 1000（标准化到 0–1，供加权社区检测使用）

可视化规模控制：避免网络过大导致卡顿，由于全网络可能非常大，为保证交互可视化可运行，通常采取：先取最大连通子图（Largest Connected Component, LCC）；若仍过大，则取度数最高的前 N 个节点构成诱导子图（例如  $N \approx 1500\text{--}3000$ ）用于绘图。该步骤主要服务于交互性能与可读性，但会影响模块细节与 hub 排名，应在局限性中说明。

社区检测：对构建后的加权网络使用 Louvain 社区检测算法进行聚类，输出每个节点的社区标签（community id）。Louvain 的目标是最大化网络模块度（modularity），使得同一社区内部连接更密集、社区之间连接更稀疏，从而识别潜在功能模块。

可视化编码与交互（统一 UI/UX）：本图采用 Pyvis 输出交互式网络 HTML，并按照统一规范进行视觉编码：节点颜色 = 社区（community）（统一调色板，如 Tableau10/Viridis）。关键枢纽蛋白（Top hubs）：按节点度数排序，Top 20 用 Orange/Gold (#F39C12) 强调。tooltip（悬停提示）：至少包含 Protein Symbol、描述/功能说明、Degree（连接度）等信息。支持滚轮缩放、拖拽平移，背景与字体按团队统一规范设置。

### 3.4 图 3 的读图说明

节点(点)表示蛋白质;边(线)表示 STRING 支持的互作关系(通过 combined\_score 表示置信度);同色区域/团簇表示 Louvain 识别的社区模块:模块内部边更密,代表互作更紧密的潜在功能单元;橙金色节点(Top 20 hubs)表示连接度最高的关键枢纽蛋白:位于模块内部中心:可能是模块核心枢纽;位于模块连接处:可能承担跨模块桥接作用。

### 3.5 结果与含义

在 combined\_score  $\geq 700$  的阈值下,网络整体仍呈现明显的模块化趋势:图中可见多个相对聚集的“团块/区域”,提示互作并非随机均匀分布,而是由若干内部连接更紧密的结构单元构成。与此同时,阈值 700 相对宽松,使得网络边更为密集、跨模块连接更多,模块边界在视觉上更容易“粘连”,呈现一个大型连通结构中包含多个亚模块的形态。这说明:在保证较高置信度的同时,网络仍保留较多互作信息,适合用于观察整体模块结构与候选关键节点分布。

### 3.6 可视化发现与讨论

发现 1: 网络呈现明显模块化,但模块之间仍存在较多连接。含义/推测: cutoff=700 下保留了较多中高置信边,模块之间可能存在共享蛋白或过程耦合,也可能引入部分间接关联导致模块边界变“粘”。下一步验证:提高阈值(如 900)并对比模块数量、模块度与社区稳定性;对每个社区做 GO/KEGG 富集以确认功能一致性。

发现 2: Top 20 hubs 分散分布于多个区域,而非集中于单一模块。含义/推测:网络可能呈“多核心结构”,不同模块各有关键枢纽蛋白,提示后续解释应以“模块内核心”而非“全局单点控制”为主。下一步验证:按社区分别统计模块内 Top hubs,形成“模块核心蛋白清单”,结合富集结果与文献进一步解释。

发现 3: 存在少数潜在跨模块连接的“通道/桥接结构”。含义/推测:桥接节点/桥接边可能承担不同功能模块的连接与信息传递,具有结构重要性;也可能是阈值较低导致的跨模块边增加。下一步验证:计算介数中心性(betweenness)识别桥接节点,并在更高阈值或更严格证据(如实验/数据库证据)条件下检验桥接是否仍存在。

### 3.7 局限性与注意点

- 1) 阈值敏感性：社区划分与网络形态对 `combined_score` 阈值敏感；700 更完整但更密集，900 更稀疏但模块边界更清晰，应进行多阈值稳健性对比。
- 2) 可视化子网偏差：若为了性能对节点数做了截断（如仅取 LCC 或 Top N 节点），模块数量与 `hub` 排名会受影响；需要在报告中明确“图为可视化子网结果”。
- 3) 算法随机性/参数影响：Louvain/Leiden 可能受随机种子与权重设置影响，建议固定 `random_state`，并用 Leiden 或多次运行评估社区稳定性。

## 四、相互作用证据分布

### 4.1 研究目的

STRING 的蛋白互作(PPI)并非只来自“实验验证”，而是综合了多类证据通道(`evidence channels`)，例如实验、数据库注释、共表达、文本挖掘等。方向四的目标是：

- 1) 统计在 `combined_score`  $\geq 700$  的高置信互作网络中，不同证据通道对“边”的贡献比例（整体层面）；
- 2) 针对网络中的关键枢纽蛋白（Top hubs），刻画其互作关系主要由哪些证据通道支撑（节点层面）；
- 3) 为后续解释与稳健性检验提供依据：例如结果是否过度依赖文本挖掘，或是否具有较强实验支撑。

### 4.2 数据来源与文件说明

数据来源：STRING database v12.0，物种：*Mus musculus*（小鼠，10090）。

本方向主要使用以下文件（并说明用途）：`10090.protein.links.detailed.v12.0.txt.gz`（核心文件）含每条互作边的各证据通道子得分（例如 `experimental`、`database`、`textmining`、`coexpression`、`neighborhood`、`cooccurrence`、`fusion` 等）以及 `combined_score`。用于统计证据通道覆盖率与关键蛋白证据画像。`10090.protein.links.v12.0.txt.gz`（辅助）含边列表与 `combined_score`，用于快速构图、计算度数（`degree`）并筛选关键蛋白（Top hubs）。`10090.protein.info.v12.0.txt.gz`（辅助）用于将 STRING 蛋白 ID 映射为常用名称（`preferred_name`），便于图例与报告表达。`10090.protein_aliases.v12.0.txt.gz`（可选）用于

跨库映射（UniProt/Ensembl/Gene symbol），若后续要做功能解释或富集分析建议加入。

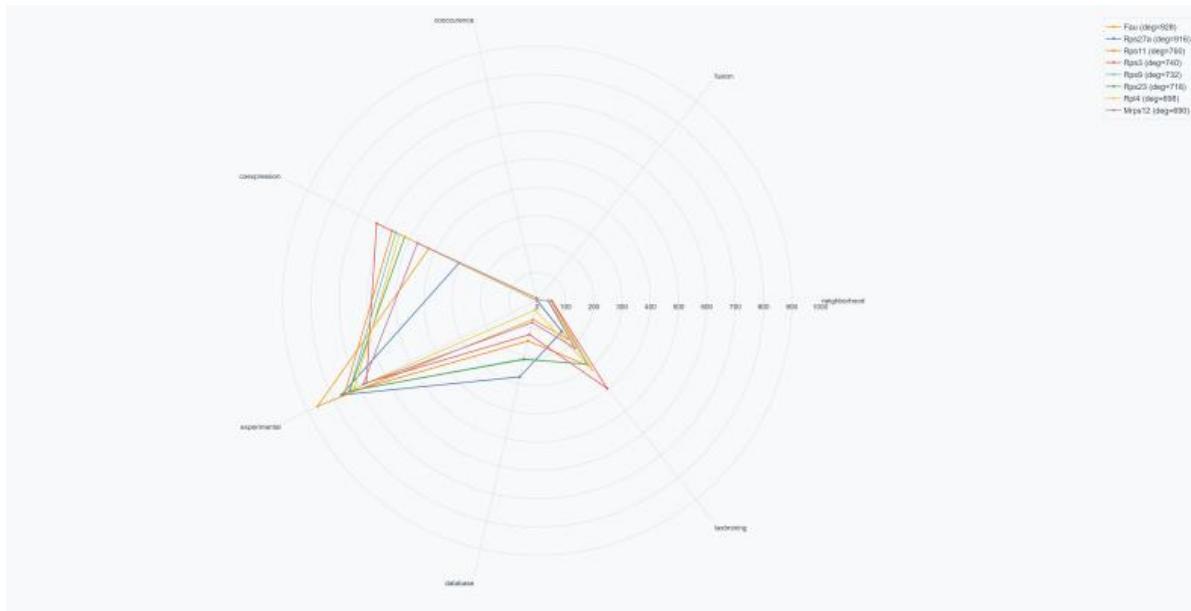


图 4A. STRING v12.0 小鼠（10090）PPI 网络在  $\text{combined\_score} \geq 700$  条件下的证据通道覆盖率统计。柱状图展示各证据通道中“得分非零”的边所占比例（Non-zero edge ratio %），用于刻画网络整体由哪些证据来源支撑。

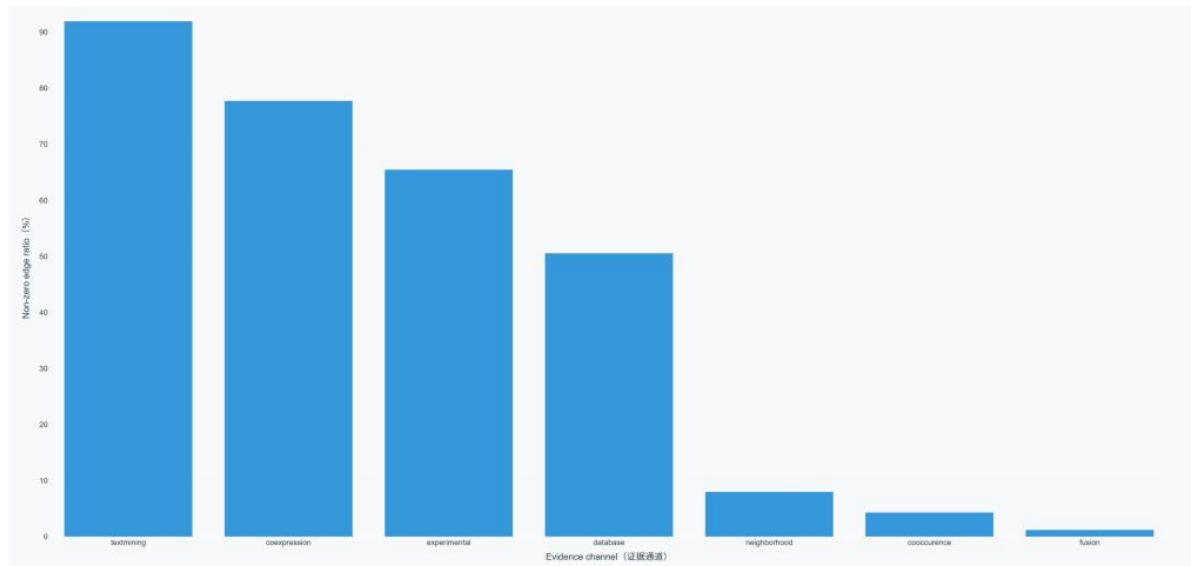


图 4B. cutoff=700 条件下关键枢纽蛋白（Top hubs）的证据画像雷达图。每条折线表示一个关键蛋白在不同证据通道上的支撑强度（0–1000 量级的汇总得分，如平均通道得分），用于比较不同 hub 的互作关系主要由哪些证据类型驱动。

## 4.3 方法与流程

本方向采用“数据 → 过滤 → 统计 → 可视化”的流程。

网络与边筛选：1) 读取 `links.detailed` (或 `links + detailed` 联合) 得到互作边及其 `combined_score`；  
2) 设置阈值：仅保留 `combined_score ≥ 700` 的边（与你图中 `cutoff=700` 一致）；3) 构建无向网络，  
用于后续的“度数 (degree)”统计与关键蛋白筛选。

图 4A 用于证据通道覆盖率（整体层面）：对所有保留边，逐个证据通道统计其“非零得分边比例”  
(覆盖率)：对于某通道（如 `experimental`），统计满足 `experimental > 0` 的边数 / 总边数 × 100%，  
得到每个通道的覆盖率柱状图 (y 轴为 Non-zero edge ratio %)。直观含义：该通道在高置信网络里“参  
与支持”了多少比例的互作边。

图 4B 用于关键蛋白证据画像（节点层面）：1) 按 `cutoff=700` 的网络计算每个节点的度数 (degree)，  
选取 Top hubs (图中展示了 `Fau`、`Rps27a`、`Rps11`、`Rps3`、`Rps9`、`Rps23`、`Rpl4`、`Mrps12` 等，并在图例  
标注 `degree`)。2) 对每个关键蛋白，提取与其相连的所有边，在每个证据通道上对得分进行汇总，形  
成“证据画像”。本项目的雷达图数值范围为 0–1000，常见做法是对该蛋白相邻边的通道得分取平均值  
(或加权平均) 来表示“该蛋白互作在该通道上的总体支撑强度”。3) 使用雷达图展示不同关键蛋白在  
各证据通道上的得分轮廓，用于比较它们“靠什么证据被连接起来”。

## 4.4 图形结果解读与含义

图 4A (Evidence Channel Coverage, `cutoff=700`) 从柱状图可以直接读出：在 `cutoff=700` 的高置信  
边集中，不同证据通道的“覆盖率”存在明显差异。整体趋势为：`textmining` 覆盖率最高 (约 90%+)：  
说明大量高置信互作在 STRING 中可从文献/文本挖掘通道获得非零支持。`coexpression` 覆盖率较高  
(约 70–80%)：共表达证据对网络贡献显著，提示许多互作对在转录水平具有协同变化信号。

`experimental` 覆盖率中等偏高 (约 60–70%)：表明相当一部分边具有实验支持，但并非所有边都  
由实验直接驱动。`database` 覆盖率约一半 (约 50%)：提示有不少互作受到数据库整理/通路注释的支  
撑。`neighborhood` / `cooccurrence` / `fusion` 覆盖率很低 (个位数甚至更低)：这些属于“基因组上下文”类  
证据，在哺乳动物中通常较弱/较少见 (更常用于细菌/古菌的操纵子、基因邻近等情景)，因此在小鼠  
网络中占比偏低是合理现象。

该网络（cutoff=700）整体上并不是“纯实验网络”，而是一个多证据整合网络，且在本阈值下文本挖掘与共表达证据非常突出。因此后续做生物学解释时需要警惕：一部分连接可能反映“文献共现/研究热点”带来的偏倚；但同时也存在较明显的实验与数据库支撑，说明网络并非完全由弱证据驱动。

图 4B（Key Proteins Evidence Profile, cutoff=700）雷达图展示了多个关键枢纽蛋白在不同证据通道上的“证据轮廓”。从图形形态可以概括为：1) 关键蛋白的轮廓整体相似，但存在通道强弱差异多数关键蛋白在 experimental、coexpression、textmining 三个通道上形成较明显的“外扩”，而在 neighborhood / fusion / cooccurrence 上接近零，和 Figure 4A 的整体覆盖率结论一致。2) “实验/共表达很强”的 hub 更像网络的稳健核心连接，若某蛋白在 experimental 和 coexpression 上得分较高，意味着其互作关系更可能同时具备实验支撑与表达协同信号。对这类 hub，后续可优先作为“稳健关键节点”进入功能模块解释或富集分析。3) “文本挖掘更突出”的 hub 需要注意文献偏倚部分蛋白在 textmining 维度上相对更突出时，可能提示其连接更易受到研究热度、文献共现的影响。对此类蛋白，后续可做“仅实验/仅数据库子网”复核其 hub 地位是否稳定。注意：仅凭雷达图不应直接断言某蛋白“具体功能是什么”。若要进行功能解释，需要结合后续 GO/KEGG 富集、通路数据库或文献证据做验证。

## 4.5 可视化发现与讨论

发现 1：textmining 覆盖率最高，说明高置信边中大量含文献/文本证据。含义/推测：网络结构可能部分反映研究热点与文献偏倚；某些连接“被报道得多”会提高其 textmining 支撑。下一步验证：构建 experimental-only 子网（仅保留 experimental>0 的边）或 database-only 子网，对比 hub 排名与社区结构是否稳定。

发现 2：experimental 与 coexpression 覆盖率也较高，说明网络仍具备较强的客观支撑。含义/推测：在 cutoff=700 下，网络并非纯“弱证据”拼接，仍具有较多可重复的实验/表达层面信号。下一步验证：对 Top hubs 的相邻边统计“实验通道得分分布”，并检查在更高阈值（如 900）下结论是否一致（稳健性分析）。

发现 3：基因组上下文类通道（neighborhood/fusion/cooccurrence）占比极低。含义/推测：这类证据对哺乳动物 PPI 支撑有限，因此后续解释不应把这些通道作为主要论据来源。下一步验证：若研究对象转为细菌/真菌网络，这三类通道可能显著上升；可在跨物种对比时作为“数据特征差异点”讨论。

## 4.6 局限性与注意点

1) STRING 证据通道并非独立：同一边可能同时被多通道支持，覆盖率不能简单相加解释为“总可靠性”。2) textmining 易受文献偏倚影响：覆盖率高不等价于“全都经过实验确认”，需要用 experimental/database 子网复核结论。3) 阈值敏感性：cutoff=700 相对保留更多边，可能提升跨模块连接与弱证据边比例；建议与 cutoff=900 做对照，报告稳定性。

## 五、功能富集分析：核糖体相关功能的景观图

### 5.1 研究背景与分析目的

在理解了全局拓扑后，我们需要对网络中特定功能模块的生物学意义进行“语义化”解释。图 5 展示了针对一组特定基因（以“核糖体 ribosomal”为关键词筛选）进行的 GO 生物过程（GO Biological Process）富集分析结果。这旨在揭示这组蛋白在小鼠体内参与了哪些核心生命活动。

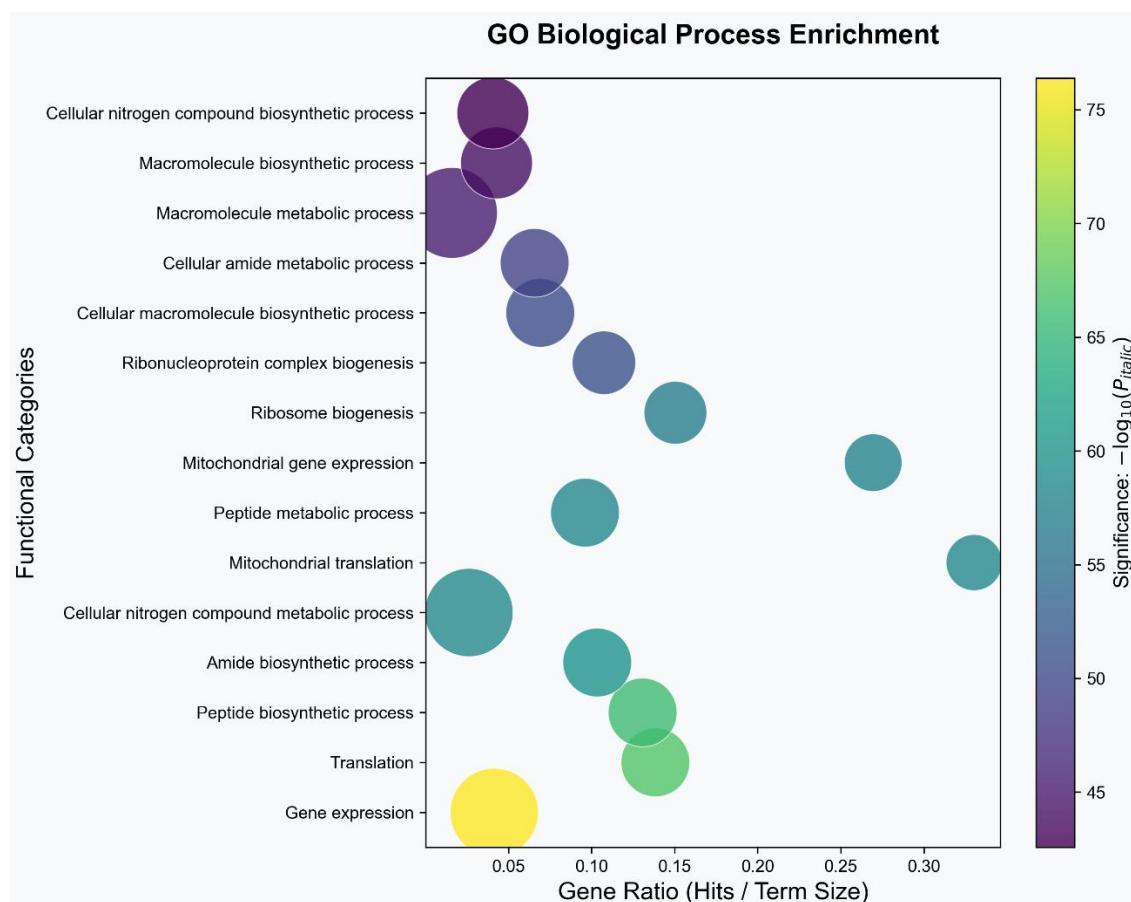


图 5 GO 生物过程富集分析

## 5.2 实验方法与技术路径

根据 图 5.py 的源码，该分析流程如下：

1. 精准靶向采样：从 protein.info 文件中利用正则表达式检索 annotation 列中包含 "ribosomal" 的蛋白，并提取前 100 个作为测试集（Input Gene List）。
2. 超几何背景检验：使用 scipy.stats.hypergeom 算法。简单来说，如果背景中有 20000 个基因，其中 200 个属于“翻译”功能，而我们的测试集中有 50 个基因属于“翻译”，那么这种“巧合”发生的概率极低。
3. 多指标映射可视化：

Gene Ratio 计算： $\frac{\text{测试集中属于该 Term 的基因数}}{\text{该 Term 在全基因组中的总基因数}}$ 。

显著性计算：取 P 值的负对数  $-\log_{10}(P)$ 。

## 5.3 图像细节的微观解剖

- 纵轴（Functional Categories）：列出了排名前 15 的显著富集条目。条目排序通常兼顾了 P 值和生物学层级。
- 横轴（Gene Ratio）：表示该通路被“命中”的深度。
- 气泡色彩（Viridis 色带）：

亮黄色气泡：位于图表底部（如 Gene expression），代表  $-\log_{10}(P)$  值极高（接近 75），即显著性最强。这不仅是因为命中数量多，更是因为其在统计学上具有压倒性的必然性。

紫色/深蓝色气泡：位于图表顶部，代表显著性相对较低，但仍处于统计阈值内。

- 气泡大小（Count）：气泡的直径与命中基因的绝对数量成正比。例如，Gene expression 和 Cellular nitrogen compound metabolic process 的气泡巨大，说明这组蛋白中绝大多数都参与了这些宏观过程。

## 5.4 核心发现与生物学发现

- 翻译逻辑的层级展开:** 最底层的 Gene expression 具有最高的显著性和最大的规模。随后是更具体的 Translation 和 Ribosome biogenesis。这证明了我们筛选的“ribosomal”蛋白不仅是核糖体的结构组分，还深度参与了核糖体的组装加工过程。
- 线粒体功能的意外富集:** 注意到 Mitochondrial translation 和 Mitochondrial gene expression 条目的出现。这说明在小鼠蛋白集中，胞质核糖体蛋白与线粒体核糖体蛋白之间存在功能上的强耦合，或者我们的筛选捕获了这两类关键蛋白。
- 生物合成的高效率:** Peptide biosynthetic process 的 Gene Ratio 较高（约 0.10-0.15），意味着这组蛋白在多肽合成这一具体环节中占据了核心比例，体现了核糖体作为“蛋白质工厂”的专一性。
- 代谢与合成的平衡:** Cellular nitrogen compound biosynthetic process 的高显著性暗示了核糖体蛋白在含氮物质代谢中的基石作用，不仅是消耗氨基酸，更是维持细胞氮平衡的关键环节。

## 六、阈值敏感性分析

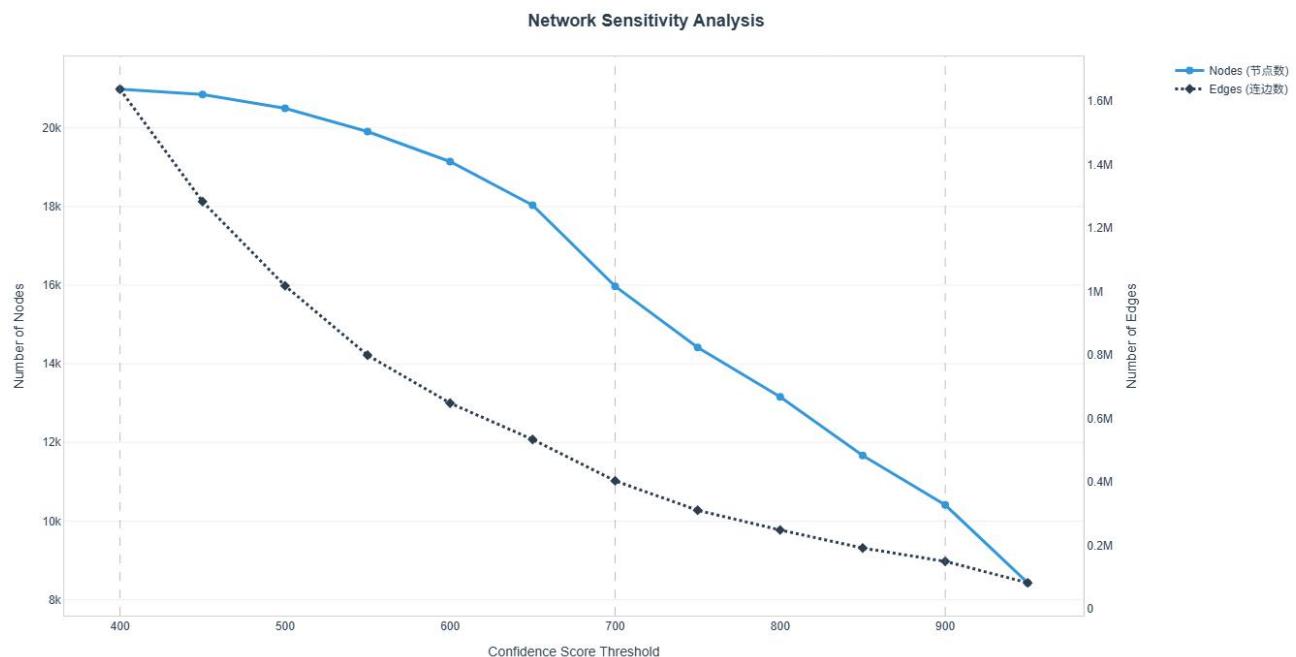


图 6 GO 阈值敏感性分析图

## 6.1 研究目的

本部分旨在探究不同置信度阈值对 PPI 网络拓扑规模与结构完整性的影响。由于 STRING 数据库包含从“低置信度”到“极高置信度”的多种相互作用，选择合适的阈值通常需要在“覆盖率（网络规模）”与“可靠性（假阳性率）”之间进行权衡。具体希望回答以下问题：（1）网络衰减趋势：阈值选取 400、700、900，网络中的节点数（Nodes）和边数（Edges）呈现怎样的衰减曲线？（2）核心骨架的稳定性：在高阈值 (>700 或 >900) 过滤下，保留下来的网络骨架是否仍具有足够的规模进行后续的模块化分析？（3）最佳阈值决策：通过双轴折线图观察，寻找一个既能保留足够多生物学信息，又能剔除大部分噪声边的“平衡点”阈值。

## 6.2 数据来源与文件说明

数据来源为 STRING database v12.0，物种：小鼠 *Mus musculus* (10090)。下载入口为 STRING 官方下载页面（用户提供链接）。

本方向主要使用文件如下：

(1) 10090.protein.links.v12.0.txt.gz

描述：蛋白质网络数据（全网络，蛋白质之间的评分链接）。

作用：本分析的输入数据源。代码将循环读取此文件，设定不同的置信度阈值（例如从 400 到 900，步长 100），统计在每个阈值过滤下剩余的“节点数量”和“连边数量”。这些统计数据直接用于绘制双轴折线图，以观察网络规模随置信度提高而产生的衰减趋势。

(2) 10090.protein.info.v12.0.txt.gz

描述：STRING 蛋白列表，包括其显示名称和描述。

作用：辅助验证。在统计节点数量时，可作为全量蛋白的参考集合，确保统计的节点 ID 均在标准注释范围内（虽然本方向主要依赖 links 文件进行计数，但在严谨的数据清洗流程中，此文件可用于核对节点合法性）

## 6.3 方法与流程

本部分遵循“数据 → 处理 → 算法 → 可视化”的流程构建图 6。多阈值迭代统计：

(1) 动态过滤：设定置信度阈值区间为[400, 950]，步长为 50。(2) 循环计算：在每个阈值切片下（如“ $score \geq 450$ ”，“ $\geq 500$ ”...），分别统计当前网络的拓扑规模；(3) 节点数：网络中存在的唯一蛋白质数量。(4) 连边数：当前置信度要求的互作关系数量。

双轴可视化设计为了在同一张图中展示量级差异巨大的两个变量，采用双 Y 轴折线图：(1) 左 Y 轴：映射节点数。使用实线+圆形标记（“lines+markers”），颜色为主色调（Light Blue）。(2) 右 Y 轴（次轴）：映射连边数。使用虚线+菱形标记（“dash=dot”），颜色为深色调（Dark Blue），并在 Layout 中明确开启“secondary\_y=True”。

## 6.4 主要结论

1) 双轴折线图解析：图表中最显著的特征是两条曲线的下降速率截然不同。代表连边数（Edges）的黑色虚线呈现出急剧的指数级衰减，从阈值 400 处的约 160 万条边，迅速下降到 threshold 700 处的约 40 万条，跌幅超过 75%。相比之下，代表节点数（Nodes）的蓝色实线下降得更为平缓，在 threshold 700 时依然保留了约 1.6 万个节点（初始值的 ~75%）。这种现象表明，网络中存在大量的低置信度边，去除这些边并不会导致大量蛋白质从网络中孤立脱落。

观察 X 轴上的 700 刻度，这是一个关键的转折区间。在此之前（400-600），连边数量大幅冗余，虽然覆盖面广但包含大量噪声。在 700 后，节点数量开始加速下降（曲线斜率变大），意味着继续提高阈值将开始误判真实的蛋白质节点，导致网络覆盖度受损。因此，Score = 700 在统计学上构成了一个理想的平衡点：既剔除了绝大部分（约 75%）可能为假阳性的连边，又保留了绝大部分（约 75%）的蛋白质节点。

当阈值达到极高的 900 时，网络依然保留了约 1 万个节点，但连边数仅剩约 20 万。此时的网络极其稀疏，剩余的连接通常是经过多重实验验证（Experimental）或强数据库匹配（Database）的，这部分网络极其可靠，但其连通性可能较差，网络可能已经由一个巨大的连通图破碎成许多孤立的小团簇。

2) 网络特性与策略推断: 数据的衰减模式间接验证了生物网络的鲁棒性 (Robustness)。如果网络是随机连接的, 随机移除边会导致节点迅速脱落。然而, 在移除 75% 的连边后, 网络依然维持了 75% 的节点, 这符合无标度网络 (Scale-free Network) 的特征: 网络由少量拥有海量连接的 Hub 支撑, 大部分非 Hub 节点仅有少量连接。低置信度的边往往连接的是外围节点, 去除它们不会破坏网络的核心骨架。

从 400 到 700 的区间内, 连边数减少了 120 万条, 而节点仅减少了 5000 个。这意味着这 120 万条边大多是“冗余连接”或“噪声”。在生物学上, STRING 的低分值 (<400) 通常来源于文本挖掘 (Textmining) 或跨物种同源预测 (Co-expression/Homology), 这些证据往往不如实验证据可靠。该图证明了在进行具体功能模块分析之前, 进行严格的阈值过滤 (如设定 cutoff=700) 是极其必要且安全的, 它能显著提升后续分析的信噪比。

3) 后续验证与优化建议: 目前的图表只统计了节点总数, 没有展示网络的“破碎程度”。后续计算每个阈值下的最大连通子图占比。观察在 700 或 900 时, LCC 是否依然包含 80% 以上的节点。如果阈值到了 900, LCC 突然跌到 20%, 说明网络已经崩解成碎片, 无法进行全局拓扑分析。这能帮助你确定“不可逾越”的最高阈值上限。

分别提取阈值 400、700、900 下的 Top 10 Hub 列表如果 Trp53 和 Prkaca 在三个列表中始终霸榜, 说明它们是真正的核心节点; 如果某些 Hub 在 900 时消失了, 说明它们的地位主要靠低置信度的边堆砌而来, 其重要性可能被高估。

对 400 网络和 700 网络的聚类结果分别进行 KEGG 富集分析, 700 网络富集出的通路 P 值应该更显著, 且通路描述更具体。如果 700 网络丢失了某些关键通路 (如“Wnt 信号通路”完全消失), 则说明 700 的阈值可能过高, 切断了该通路的特征连边。

## 七、亚网络交互: 核糖体蛋白复合体的精细结构

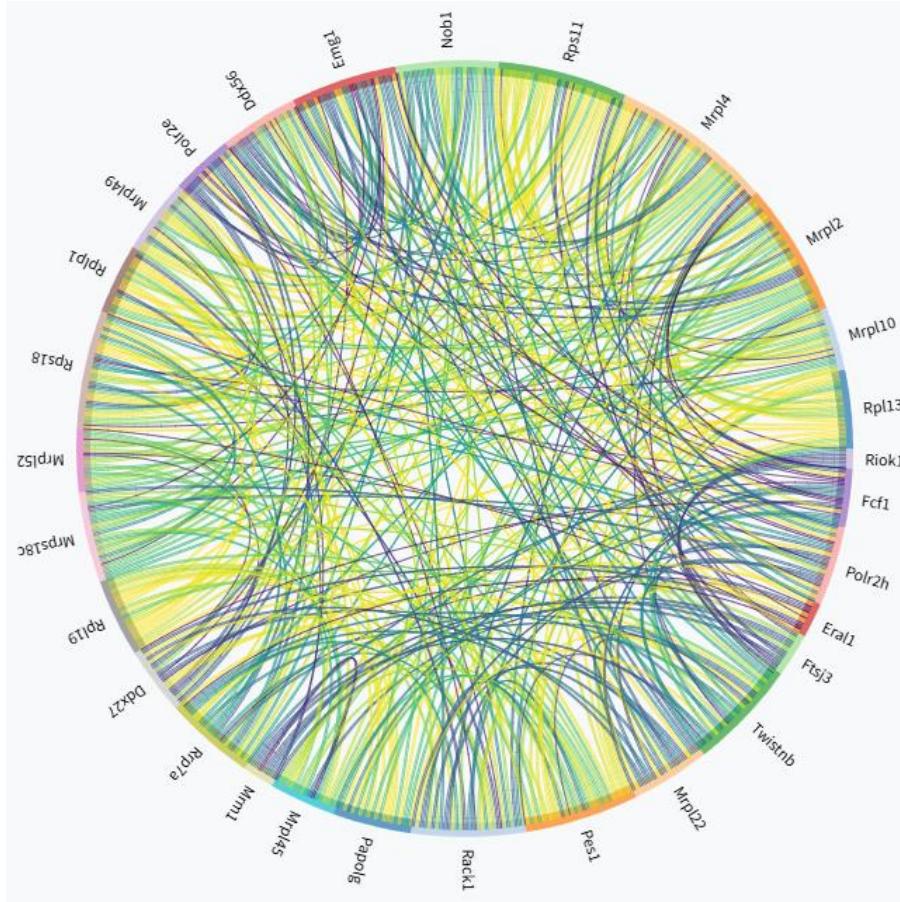


图 7 相互作用亚网络：核糖体蛋白

## 7.1 研究背景与分析目的

宏观统计（图 1）和功能分类（图 5）完成后，最后一步是观察这些蛋白是如何物理接触并形成复合体的。图 7 采用交互式弦图（Chord Diagram）展示了前 30 个核糖体相关蛋白之间的相互作用强度，为理解蛋白质机器的物理组装提供了直观证据。

## 7.2 实验方法与技术路径

根据 图 7.py 的逻辑，该图的生成非常复杂：

1. **诱导子图提取：**从全网中切取出仅包含这 30 个蛋白的“小世界”。
2. **权重编码：**将 STRING 的 combined\_score 映射到弦的颜色和透明度上。分值越高，弦的颜色越趋向于亮绿色/黄色（Viridis 映射），代表交互证据越可靠。
3. **交互式渲染：**基于 Holoviews 和 Bokeh 引擎，采用圆周布局（Circular Layout），将蛋白节点均匀分布在圆环上，内部用复杂的弧线（弦）连接。

### 3.3 图像细节的微观解剖

1. **节点（外环刻度）：**可以看到 Rps11（核糖体小亚基蛋白 11）、Mrpl4（线粒体核糖体大亚基蛋白 4）、Rack1（受体活化 C 激酶 1）等。
2. **弦的密集度：**圆心区域被密集的弧线完全覆盖。这种“高密度连接”在 PPI 亚网络中非常罕见，通常仅出现在核糖体、蛋白酶体或 RNA 聚合酶等高度稳定的蛋白质复合体中。
3. **颜色梯度的含义：**金黄色/亮绿色弦：例如 Rps11 周边的连线。这表示这些蛋白之间有极强的物理交互证据（如晶体结构验证、共免疫沉淀实验）。深紫色/蓝色细线：代表这些蛋白虽然属于同一功能簇，但其直接接触的证据稍弱，可能是通过中间蛋白间接作用。
4. **节点标签分布：**节点按名称首字母或连接度在圆周排列。例如，Mrpl 系列（线粒体核糖体）和 Rps/Rpl 系列（胞质核糖体）交织在一起。

### 3.4 核心发现与生物学发现

1. **核糖体复合体的“全连接”特性：**该图展现了一个极其紧密的模块化结构。这在生物学上对应了核糖体大小亚基组装成完整功能单位时的物理状态。
2. **关键节点的识别：****Rack1 的核心地位：**在图中 Rack1 表现出非常密集的向心连线。在实际生物学中，Rack1 被认为是核糖体上的一个支架蛋白，负责招募多种信号分子（如 PKC），将细胞信号转导与蛋白质翻译直接挂钩。**Nob1 与 Emg1：**这两个节点的连线较亮，反映了它们在核糖体 RNA 加工和成熟过程中的协同作用。
3. **跨部门协作（线粒体与胞质）：**图中 Mrpl 与 Rps 蛋白之间存在的连线是一个值得关注的发现。这暗示了在小鼠细胞中，可能存在某种协调机制（如共表达或共同进化压力）使得线粒体翻译机器与胞质翻译机器在功能网络中紧密关联。
4. **交互的置信度分层：**通过弦的颜色层次，我们可以清楚地分辨出哪些是“核心结构骨架”（高分亮弦），哪些是“外周辅助因子”（低分暗弦），这为后续的蛋白敲除实验提供了优先级参考。