ChatGPT

人源-鼠源抗原结构对数据集构建计划

本计划旨在系统性地构建**人源抗体-人源抗原**复合结构与对应**鼠源抗原**结构/模型的配对数据集。整个流程分阶段进行，每阶段明确关键步骤、使用的数据源或工具，并分析潜在难点与应对策略。

1. 阶段一：收集人源抗体-人源抗原复合结构数据（SAbDab）

关键步骤：
1. **筛选SAbDab数据库：** 使用SAbDab（结构抗体数据库）筛选出抗体重链、轻链均为人（Homo sapiens），且抗原来源也是人类蛋白的抗体-抗原复合物结构 [1]。
2. **获取结构列表：** 将符合条件的条目（PDB ID）收集成初始列表，包括每个复合物中抗原链的标识符。确保抗原类型为蛋白质而非肽段、小分子等。
3. **提取元数据：** 对每个入选PDB条目，记录实验分辨率等质量信息，以及抗原链的序列和物种注释（可从SAbDab直接获取或通过PDB提供的注释）。必要时使用RCSB/PDBe API获取每条链的物种和序列信息 [2]。

潜在难点及对策：
- **难点：物种注释准确性。** 某些PDB/SAbDab条目的物种标注可能存在错误或缺失（例如7WRL的抗体物种曾标注错误 [3]）。
对策： 交叉核对PDB原始记录中的链分类和Uniprot映射，确保抗原链确为人源。如果发现物种标注冲突，可手动检验序列同源性或参考文献修正。
- **难点：数据提取批量化。** SAbDab网页界面筛选手动导出效率低。
对策： 利用SAbDab提供的批量下载或编写脚本调用其API/数据库导出功能（如“Get all structures”结果再过滤）。也可采用SAbDab衍生资源（如SAAINT-DB）的汇总文件来获取候选列表 [4] [1]。
- **难点：抗原链识别。** 部分PDB包含多个非抗体链，需辨识哪条是抗原。
对策： SAbDab已对抗原链有标记，可直接使用其注释。若不可靠，则根据链序列类型（免疫球蛋白折叠的是抗体链，其余为抗原）或链与抗体的接触情况来判定。

主要资源/工具：
- **SAbDab数据库：** 提供规范化的抗体结构及复合物注释 [1]。
- **PDB/RCSB API：** 获取PDB条目链的序列、物种等详情（例如 `rcsb.org/fasta/entry/<PDBID>` 获取序列）。
- **相关平台：** Thera-SAbDab（治疗抗体数据库）可作为补充，用于定位以人源蛋白为靶标的抗体结构条目。

2. 阶段二：提取人源抗原的抗体结合表位

关键步骤：
1. **结构解析确定表位：** 针对每个复合物结构，确定抗原上与抗体直接相互作用的残基集合，即抗体表位区域。通常定义为抗原中所有**与抗体原子距离在4–5Å以内**的残基。
2. **自动识别接口残基：** 编写脚本遍历抗体重链、轻链与抗原链之间的原子距离，标记满足距离阈值的抗原残基（包括可能的分段不连续表位）。
3. **获取表位序列片段：** 根据识别的表位残基序号，从抗原序列中提取相应片段或记录残基列表。若表位由多个分散片段组成，则分别记录各片段序列。

潜在难点及对策：
- **难点：识别精度与一致性。** 不同复合物可能需统一判定标准，如采用原子距离阈值可能影响表位边界判定。
对策： 采用文献通用标准（如4Å接触距离）保证一致 [4]。可考虑结合SAbDab自带的表位注释（如有）验证结果一致性，或使用SAAINT-parser等工具验证界面残基识别的准确性 [4]。

- **难点：表位片段的处理。** 若表位残基不连续，如何表示和后续比对。
对策： 保留残基索引列表和所属片段序列，后续在比对同源时分别处理。对于结构比对，可基于空间邻近将表位视为一个整体区域。
- **难点：自动化批处理。** 手工逐个通过分子可视化软件（如PyMOL）确认表位费时。
对策： 借助BioPython的PDB模块或MDAnalysis库脚本化处理，实现批量分析接口残基。经过小规模手动验证后，大规模应用于全部条目。

主要资源/工具：
- **PDB结构分析工具：** BioPython，PyMOL API，或者OPIG提供的结构分析脚本。
- **数据库参考：** SAbDab/AACDB等数据库可能提供抗原接触残基列表供参考核对。
- **免疫学数据库：** IEDB（Immune Epitope Database，免疫表位数据库）可用于核对已知线性表位是否包含在所识别区域内，以确保合理性。

3. 阶段三：寻找对应的鼠源抗原序列同源物

关键步骤：
1. **确定人源抗原序列：** 获取阶段二中每个抗原链的氨基酸序列（如果PDB中的抗原为片段，可尝试映射回全长Uniprot序列）。记录表位残基在该序列中的位置范围。
2. **同源序列搜索：** 使用上述人源抗原序列，在**鼠属（Mus musculus）**中搜索同源蛋白序列。常用方法包括：
- 在Uniprot中查找人源蛋白的鼠源直系同源蛋白（若已知基因名称，可直接检索对应小鼠蛋白条目）。
- 利用BLAST或MMseqs2将人源抗原序列比对鼠源蛋白数据库（限定物种taxonomy:10090），寻找最高同一性匹配序列。
- 如果抗原序列是人特有但含有保守域，亦可仅比对关键表位片段以寻找鼠源蛋白中存在的保守区域。
3. **选定候选鼠源抗原：** 基于比对结果，确定最可能的鼠源对应蛋白序列。优先选择全长同源蛋白；若无全长同源，则考虑是否存在包含该表位区域的鼠源蛋白片段。记录鼠源蛋白的UniProt ID或基因名。

潜在难点及对策：
- **难点：一对多或同源性不足。** 人源抗原可能对应鼠源多个同源序列（如基因家族成员），或序列同一性偏低导致无法明确唯一对应。
对策： 优先利用已知直系同源（ortholog）关系数据库（如NCBI HomoloGene、Ensembl compara）确定**一一对应**的鼠蛋白。如果BLAST出现多个近似匹配，则考虑物种保守性最高者，或借助基因功能注释选择最合理的候选。必要时参考文献确认目标蛋白在小鼠中的对应物。
- **难点：抗原片段映射。** 若PDB中抗原仅为蛋白片段（如单个结构域），找到的鼠源序列是全长蛋白，需要确认表位区域在鼠序列中的对应位置。
对策： 对齐人源与鼠源序列，定位人表位残基对应的鼠源序列位置，确保这些位置没有缺失或过度变异。对于关键氨基酸如结合热点，应检查它们在鼠序列中是否保留（同源或保守替代）。
- **难点：自动化程度。** 需批量处理多个序列的跨物种比对。
对策： 使用脚本调用BLAST+本地库或通过Biopython的接口批量查询。也可利用UniProt的API，通过人蛋白ID直接查询对应的小鼠UniProt条目（UniProt提供物种同源代表性映射）。

主要资源/工具：
- **序列数据库：** UniProt（包括物种同源注释）、NCBI蛋白库（FASTA）、Ensembl基因组数据库。
- **比对工具：** NCBI BLAST/PSI-BLAST（支持限定物种）或 EMBL-EBI的 MUSCLE/Clustal Omega 服务，用于找出人-鼠序列最佳匹配。
- **IEDB免疫表位数据库：** 可查询目标抗原的已知表位是否在小鼠中也有免疫反应记录，侧面验证所选同源蛋白确有类似表位功能。

4. 阶段四：获取鼠源抗原的结构或模型

关键步骤：
1. **查找实验结构：** 检索PDB中是否已有**鼠源抗原**的实验解析结构（晶体或冷冻电镜等）。检索方式包括：按UniProt ID或序列搜索PDB，限定物种为Mus musculus。如果找到：
- 若存在完整蛋白结构，记录PDB ID及链。
- 若仅有局部结构（例如某结构域）且覆盖抗体表位区域，则也可接受。记录该结构片段的信息。
2. **获取预测模型：** 如果无实验结构，可从公开模型数据库获取鼠源抗原的模型：
- **AlphaFold DB：** 该数据库提供了包括小鼠在内的多种生物全蛋白组的高精度结构预测 [5] 。通过鼠源蛋白的UniProt ID，在AlphaFold数据库中下载其预测结构模型（通常是全长）。
- **SWISS-MODEL Repository：** 作为补充来源，获取基于同源模板的模型。如果AlphaFold模型存在低可信度区域，可比对SWISS-MODEL的同一蛋白模型以供参考 [6] 。
3. **模型筛选与处理：** 如果同时获得多个模型/结构：
- 优先选择覆盖抗体表位区域且质量高的结构。优先级顺序：高分辨率实验结构 > AlphaFold高置信度模型 > 同源建模模型。
- 若模型为全长而抗原在PDB中是片段，可截取模型中对应片段用于后续分析，以减少无关区域干扰。
- 记录所采用的鼠源抗原结构的来源（PDB ID或模型数据库ID）和覆盖范围。

潜在难点及对策：
- **难点：实验结构缺失。** 许多小鼠蛋白未有解析结构。
对策： 充分利用AlphaFold DB，因其已涵盖>200万蛋白结构预测，提供了小鼠全蛋白组的结构数据 [5] 。AlphaFold模型精度通常较高，对于有明确同源的人蛋白，其关键区域预测可靠性也较好 [7] [8] 。
- **难点：模型可信度评估。** AlphaFold模型在无模板区域可能可信度低（pLDDT分值低），需要确认抗体表位区域的预测可靠度。
对策： 查看AlphaFold模型的pLDDT/color图谱，确认表位对应片段是否高分（例如>70）。如果关键表位区域预测不确定且存在其他数据源（如部分实验结构），可考虑只采用实验结构片段或高置信度区域。SWISS-MODEL仓库提供了同源建模质量评分（QMEAN等）可辅助判断 [6] 。
- **难点：结构序列对齐。** 获取的鼠源结构（实验或预测）有时与人源抗原序列长度不同（插入缺失）。
对策： 基于前一阶段的序列比对结果，将鼠源结构的序列与人源抗原序列对齐，确保结构坐标与人源表位残基位置对应。必要时对鼠源结构进行简单比对校正（如使用PyMOL或TM-align对人/鼠抗原结构进行初步叠合，以验证对应关系）。

主要资源/工具：
- **PDB数据库：** RCSB/PDBe检索用于寻找小鼠蛋白结构（可利用序列相似性搜索或按UniProt AC查询已有结构条目）。
- **AlphaFold Protein Structure DB：** 获取小鼠蛋白的AlphaFold预测结构 [5] （例如通过其UniProt Accession）。
- **SWISS-MODEL Repository：** 获取小鼠蛋白的同源模型及质量评注 [6] 。
- **结构处理工具：** PyMOL、UCSF Chimera等用于截取结构片段、结构比对；或BioPython结构模块用于自动化处理模型和对齐。

5. 阶段五：验证人源-鼠源抗原表位同源性

关键步骤：
1. **序列同源验证：** 将人源抗原的表位序列与鼠源抗原对应序列片段进行比对，计算表位区域的序列同一性/相似性。确认在该区域至少存在显著同源（关键残基保留）。可设定阈值，例如表位区序列身份率≥X%或保守替换率≥Y%。
2. **结构比较验证：** 如果两者结构（实验或模型）均可用，对比人源抗原表位结构与鼠源抗原结构在对应区域的三维构象：
- 叠合两种抗原的表位区域主链原子，计算RMSD以量化结构差异。

- 观察关键侧链取向是否类似（尤其是直接参与抗体结合的残基）。
3. **同源性判定：** 根据以上分析，判定该**人-鼠抗原对**是否满足“表位区域序列或结构同源性”的要求：
- 若满足，同源性高且结构相似，则纳入数据集。
- 若序列差异过大且结构比对RMSD显著（表明表位结构不保守 [9] ），则剔除该对，避免不可靠的对应关系。
4. **记录验证结果：** 对每对保留的数据，记录人源和鼠源抗原表位序列的比对结果（例如identity%、保守位点标识）、结构RMSD值等，作为数据集注释，方便日后按同源性阈值筛选或分析。

潜在难点及对策：
- **难点：如何量化“同源性”。** 不同抗原对的表位长度和保守程度不同，统一阈值可能不适用于所有情况。
对策： 综合考虑**序列**和**结构**两方面：例如主要以序列同一性为准，结构比对作为辅证。如果序列相似度略低但结构RMSD很小，也可认为满足同源要求（说明尽管序列变异，结构仍模仿了相同表位构造 [10] [9] ）。最终阈值可根据已知范例调整，使筛选结果符合直觉预期。
- **难点：结构比对的可行性。** 部分鼠源抗原仅有模型，无对应表位复合物结构，直接比较其表位构象可能有不确定性。
对策： 允许使用AlphaFold模型进行对比，但重点放在主链骨架形态。利用已知同源蛋白结构的经验——同源序列通常形成相似的局部折叠 [9] ——只要序列同源且模型未表现出异常折叠，即可认为结构上相容。对于关键位点，如AlphaFold模型显示表位某Loop无定义，则在记录中注明不确定性或考虑剔除。
- **难点：自动评估。** 序列和结构比较需要自动化以处理大量数据。
对策： 使用程序执行序列比对（如Biopython的pairwise2或ClustalOmega命令行）并计算同源性指标；使用现有工具计算结构RMSD（如TM-align、BioPython Superimposer）。将结果与预设标准比较，实现批量过滤。人工抽查边缘案例确保流程可靠。

主要资源/工具：
- **序列比对工具：** Clustal Omega、MAFFT 或 Biopython模块用于表位序列对比。
- **结构比对工具：** TM-align、Dali服务器，或PyMOL的 `align` 命令等，用于计算表位区域的结构RMSD。
- **免疫学数据库：** IEDB提供跨物种的表位实验数据，可用于核对筛选结果（例如若某人源表位在小鼠中有抗体响应记录，则支持我们的同源表位选择）。
- **相关研究参考：** 跨物种表位模拟分析工具（如Epitopedia [10] ）强调了只有序列同源时跨物种表位的三维结构才可能高度相似 [9] ；这验证了我们以同源序列为基础筛选结构对的合理性。

6. 阶段六：数据集整合与整理发布

关键步骤：
1. **数据汇总：** 将通过验证的所有“人源抗原 vs. 鼠源抗原”配对记录汇总成数据集。每条记录包括：人源抗体-抗原复合物的标识（PDB ID及链）、抗原表位序列/位点；对应的鼠源抗原结构标识（PDB ID及链或模型ID）、对应序列片段位点；以及序列同源百分比、（若有）结构RMSD等注释信息。
2. **格式化存储：** 选择适当的格式保存数据集，例如表格（CSV/Excel）或JSON。表格列包括上述各项，使后续检索过滤方便。也可构建一个关系数据库（如SQL）或knowledge base存储，便于复杂查询（比如按抗原名称聚合）。
3. **质量控制：** 最终浏览数据集，检查是否有重复条目或明显异常（例如同一人源抗原对应多个鼠源条目、或反之——如有则判断是否合并或都保留）。对每个配对的来源进行引用注释（PDB doi、AlphaFold模型版本等)。
4. **发布与共享：** 准备数据集的文档，说明构建方法和字段含义。可将数据提交至公共平台（如Antibody Society的数据库、Github仓库或资料库）供他人使用，并注明**不得使用用户端AI建模再造结构**的约束（即本数据集中鼠源结构均来自已有公开资源）。如果适用，可在Antibody Society或相关会议分享数据集，以供抗体工程领域使用。

潜在难点及对策：
- **难点：数据量和平衡。** 数据集规模取决于有多少人源抗原有对应鼠源同源物。可能出现某些抗原（如人特有蛋白）无对应，小鼠无条目，从而数据量减少。
对策： 提前统计筛选命中率，必要时放宽某些标准以涵盖更多数据（例如略降低序列同一性阈值，但前提是结

构预测可靠）。同时关注数据集中不同抗原类别的分布，确保不被少数大型抗原垄断。
- **难点：元数据丰富性。** 下游用户可能需要更多信息（如抗体类型、抗原功能、抗体是否交叉反应等）。
对策： 在数据集中增加有益字段：例如抗体是否来源于Therapeutic（可从Thera-SAbDab查得）、抗原基因名、人鼠蛋白的Uniprot ID、抗体是否已知跨种反应。利用Antibody Society等提供的疗法抗体数据来注释哪些抗体靶点有已知鼠源替代抗体。
- **难点：发布平台选择。** 如果数据集较大或需定期更新，选择托管平台和版本控制是挑战。
对策： 可将数据提交到**免疫学数据库**（如IEDB附属资源）或预印发表。也可借助Springer Protocols等文献平台发表方法论文，或在GitHub维护数据和脚本。Springer的相关实验方案库中关于表位映射的数据构建方法 [11] [12] 可以作为记录引用，使本计划透明可重复。

主要资源/工具：
- **数据发布平台：** Antibody Society数据库（用于分享抗体相关数据的社区平台）、GitHub/GitLab（公开数据集和代码）、乾符科学数据库等。
- **文档撰写资源：** Springer Protocols等文献资源，可用于撰写发布数据集的详细方法说明，确保他人能够依据本计划复现或更新数据集。
- **后续支持：** 持续关注SAbDab更新和AlphaFold DB更新，以定期刷新数据集内容；关注IEDB等数据库新增的跨物种表位数据，验证和丰富本数据集的生物学意义。

通过上述分阶段计划，我们将在不借助用户自建AI模型的前提下，利用公开数据库和工具，系统化地构建**人源-鼠源抗原结构对**数据集。该数据集将为后续分析（如抗体跨物种识别研究、表位保守性评估等）提供宝贵的基础资源 [10] [9]。数据构建流程中的每一步都考虑了潜在技术难点并提出了解决策略，确保计划具有可执行性和可靠性。最终，我们将产出一个结构良好、信息丰富的数据集，并通过恰当的平台共享，为生物医药研究提供支持。 [5] [1]

---

[1] [2] [4] SAAINT-DB: a comprehensive structural antibody database for antibody modeling and design | Acta Pharmacologica Sinica
https://www.nature.com/articles/s41401-025-01608-5?error=cookies_not_supported&code=6b1e8316-74b9-422a-9a0b-f38e5afcb165

[3] AACDB: Antigen-Antibody Complex Database - eLife
https://elifesciences.org/reviewed-preprints/104934/reviews

[5] [7] [8] AlphaFold Protein Structure Database
https://alphafold.ebi.ac.uk/

[6] SWISS-MODEL Repository
https://swissmodel.expasy.org/repository

[9] [10] Epitopedia: identifying molecular mimicry between pathogens and known immune epitopes - ScienceDirect
https://www.sciencedirect.com/science/article/pii/S2667119023000034

[11] [12] Structure-Based Epitope Profiling with the Structural Profiling of Antibodies to Cluster by Epitope 2 (SPACE2) Algorithm | Springer Nature Experiments
https://experiments.springernature.com/articles/10.1007/978-1-0716-4591-8_16?error=cookies_not_supported&code=368ddd38-18e2-4b31-af9e-89e59ebd9248