Towards a public repository of designed proteins

Karolina Lopes Barbosa[1], Benedita Pereira[2], Manuel N. Melo[2], Diana Lousa[2], and Miguel Rocha[1]

[1] Centro de Engenharia Biológica, Escola de Engenharia da Universidade do Minho, Braga, Portugal
[2] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal

Abstract. Protein Design is a field of research that aims to create synthetic proteins with new functionalities that can be applied in fields such as the development of new drugs and vaccines, and a range of other biotechnological applications. However, this type of data remains decentralized and spread across repositories, publications, and other databases, making it challenging to efficiently integrate and extract meaningful insight. This project focuses on developing a public-access database of computationally designed proteins, following FAIR guidelines [9], and focusing on interoperability. The first stage of this project is to establish a conceptual model through MySQL Workbench to refine the database structure and ensure logical organization. The second stage relies on running test queries to verify data retrieval, integrity constraints, and confirm the foreign key dependencies, which should help optimize the schema, document the entity relationships. This should help optimize the schema, clarify entity relationships and data flow, and address issues like data redundancy, missing links, and inefficiencies. The third stage is to set up the database environment, develop scripts to insert, retrieve, and manage data, and implement data ingestion mechanisms to handle CSV, JSON, or direct database input. Finally, we will optimize query performance and implement error-handling mechanisms for data processing.

1 Introduction

All natural proteins are the product of evolution. However, because of the limitations of evolutionary processes, new molecular activities only slowly emerge through natural evolution, and many desirable activities may not have been found. The design of proteins makes the expansion of the natural molecular mechanism possible by creating synthetic proteins with new functionalities that can be applied in fields such as the development of new drugs and vaccines [5], and a range of other biotechnological applications.

Many techniques are used to construct these proteins, and for this purpose, parameters such as activity, binding affinity, and stability of proteins are measured [10]. This process generates vast amounts of computational and experimental data.

Furthermore, with recent scientific advances in protein structure prediction, driven by artificial intelligence (AI) tools such as AlphaFold[6], which was recognized with the Chemistry Nobel Prize, shows that the Protein Design field has been significantly developing. This progress suggests a growing tendency for even more data to be generated. However, this data remains decentralized and spread across repositories, publications, and other databases, making it challenging to efficiently integrate and extract meaningful insight. It is noticed that the scientific community needs a standardized and centralized repository in line with the FAIR guidelines [9] to facilitate access to these data for analysis and comparison. A centralized database would not only provide a deeper understanding of the field but also accelerate the development of new algorithms and the use of machine learning to enhance computational protein design and generate more refined, efficient design models.

2 Objectives

Acknowledging the necessity of having a centralized repository for computationally designed proteins, this project aims to develop a public-access database of computational protein designs, following FAIR guidelines [9], and focusing on interoperability. The initial stage aims to identify state-of-the-art design efforts and public repositories that have been established for designed proteins so far, to define needs and standards for this type of data. Later, the project aims to perform the implementation of the repository with a focus on data ingestion.

3 State of the Art

Open-access databases such as The Protein Data Bank [2] and Uniprot [1] exist because the scientific community recognized that having a centralized protein database would be extremely valuable for organizing and structuring all the information generated by analyzing these biomolecules [8]. Although these databases are valuable resources for conventional proteins, their scope is limited when it comes to designed proteins. UniProt, for instance, cannot be used for synthetic proteins, and while the PDB stores experimental data, it is focused solely on structure determination and does not capture additional information.

There are databases related to designed proteins. Listing some examples, the Protein Design Archive [4] has been recently published, and serves as a repository to store and share experimentally determined designed protein structures. Its focus is on accessibility and reliability of protein design, however, it lacks data on experimental results other than structural data. ModelArchive [7] is a database focused only on archiving computational structure models. Protabank [8], a protein engineering database, has features that allow the display of a wide range of properties, including but not limited to activity, binding, stability, folding, and solubility. Additionally, it stores data in a relational database, integrating information from computational approaches such as docking, molecular modeling, statistical models, as well as experimental techniques like yeast and

phage display, circular dichroism, among others. Since the purpose of this tool aligns with our project, analyzing its database schema represented in figure 1 will provide valuable insight into how designed protein data can be structured and integrated.

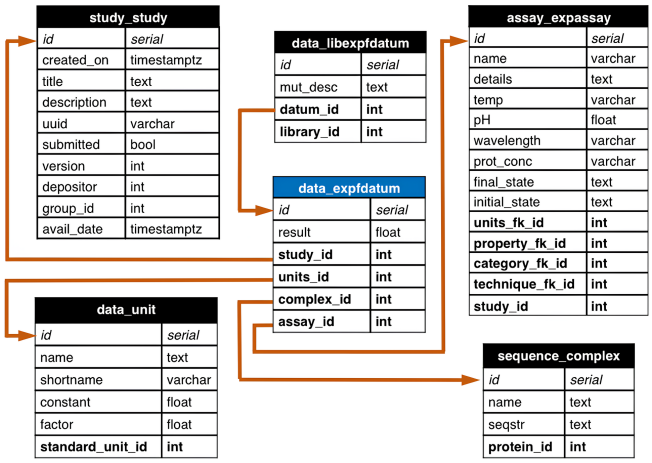


Fig. 1. ProtaBank's database schema [8]

The interpretation of the schema represented in Figure 1 indicates that the table "data_expfdatum" stores numerical assay results and has foreign keys linking to the tables "study_study", "data_unit", "assay_expassay", and "sequence_complex". The table "study_study" holds information about the submitted study and does not have any foreign keys. The table "data_unit" specifies the measurement units and references another table via a foreign key, which stores what is supposed to be a foreign key to a table "standard_unit_id" not displayed in the schema. The table "assay_expassay" describes the experimental assay used and includes foreign keys such as "units_fk_id", "property_fk_id", "category_fk_id", "technique_fk_id", and "study_id". The table "sequence_complex" represents the protein mutant sequence and seems to have "protein_id" as a foreign key. For data belonging to a mutant library, a foreign key links it to the library table (data_libexpfdatum).

Additionally, the relationships between entities and some associated tables could not be identified due to the limitations of the presented schema.

While Protabank provides a structured approach to storing designed protein data, it was observed to have unclear commitments to data availability, and since the database is owned by a private company and some functionalities are exclusively available to sponsors, its openness is questionable. In addition, this tool has not been used since 2022, and it seems that the data inputted in there was mostly added by the maintainers themselves. Recognizing that a program

like this already exists, building upon its existing framework will be valuable in constructing this database, while also identifying areas for improvement to enhance its scope.

3.1 Methodology

Research and Database schema

The initial phase of this project focuses on a literature review regarding Protein Design and its relevance in modern science. The goal is to consolidate knowledge on the types of computational and experimental data that are generated in Protein Design, as well as the approaches used. This step is crucial for identifying the key data that should be stored in the database.

To determine the necessary data that should be integrated into the database, the plan is to build an initial schema designed to guide its construction, identifying the relevant information, defining how the data should be structured, and establishing relationships between different entities. Building on this foundation, the project will develop a conceptual model through MySQL Workbench to refine the database structure and ensure logical organization.

The database schema represented in Figure 2 is designed to organize the data according to the table arrangement shown below in the same figure. Gathering information to represent "design_id" 1 [3] and "design_id" 2 [10], the idea is to exemplify how data entries are presented, to illustrate the format and organization of information within the database. The data entries exemplified provide a more visual understanding of how the schema is structured.

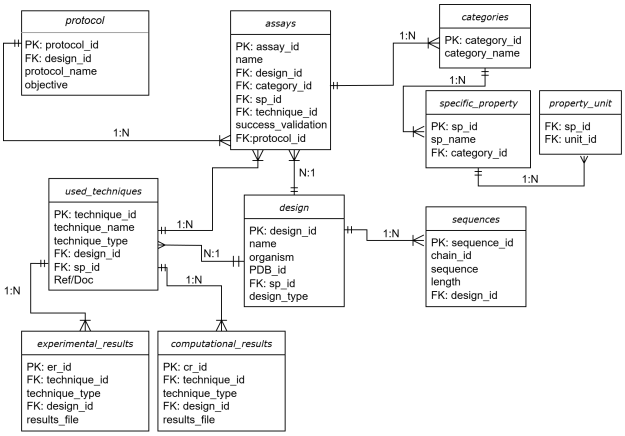
The "design" table serves as the central entity, storing essential information about the designed protein, including its name, expression system, classification, PDB ID, associated publication, and reference link. Its primary key (design_id) acts as a foreign key across multiple tables, enabling queries that retrieve details linked to each design.

The "sequences" table, identified by sequence_id, captures multiple sequences generated for a given design, including attributes such as chain ID and sequence length, while maintaining a foreign key reference to design_id.

The "assays" table, structured around assay_id, records detailed assay information, such as assay category, specific measured properties, technique used, and assay success. It also includes foreign keys linking it to the "protocol", "design", "used_techniques", and "categories" tables, facilitating structured queries on assay-related data.

The "categories" table (category_id) classifies techniques based on their purpose, such as assessing activity, stability/folding, or other functional aspects. The "specific_properties" table further refines these classifications by detailing the specific properties within each category. For instance, under stability/folding, recorded properties include melting temperature, equilibrium constant, and rate of folding. The "property_unit" table acts as an intermediary between "specific_properties" and "units", ensuring flexibility by allowing multiple unit options for each specific property.

The "used_techniques" table, with technique_id as its primary key, tracks computational and experimental methods applied in a design. For example, design_id "2", corresponding to the wFAP1.1 structure, which used the computational techniques AlphaFold and molecular docking. This table also serves as a foreign key for both the "experimental_results" and "computational_results" tables, which store the result files of the implemented techniques.



design

design_id	name	expression	classification	pdb_id	Publication	Ref link
1	Protein Gß1	Escherichia coli BL21(DE3)	membrane protein	P06654	Choi EJ, Mayo SL. Generation and analysis of proline mutants in protein G. Protein Eng Des Sel. 2006	https://pubmed.ncbi.nlm.nih.gov/1
2	wFAP1.1 structure	Escherichia coli	de novo protein	8W6F	Zhu, J., Liang, M., Sun, K. et al. De novo design of transmembrane fluorescence-activating proteins	https://doi.org/10.1038/s41586-025-08598-8

assays

assay_id	name	design_id	category_id	sp_id	technique_id	success_val	protocol_id
1	Tm	1	1 (Stability/ Folding)	1 (Melting Temperature	1 (CD)	true	1
2	RIF docking	2	3 (Binding)	4 (Docking Score)	3 (Molecular Docking)	true	2

used_techniques

technique_i	technique_name	technique_type	design_id
1	Circular Dichroism (CD)	Experimental	1
2	AlphaFold2	Computational	2
3	Molecular Docking	Computational	2

sequences

sequence id	chain_id	sequence	length	design_id
1	A,B	MDEERLKEILI	170	2

specific_property

sp_id	category_id	sp_name
1	1	Melting Temperature
2	1	Equilibrium Constant (K)
3	1	Rate of Folding (kF)
4	3	Docking Score

category

category_id	category_name
1	Stability /Folding
2	Activity
3	Binding
4	Expression
5	Solubility

Fig. 2. Schema that aims to structure how data produced by protein design is stored in the database.

Schema Validation and Database Deployment

This stage focuses on reviewing the schema and structure to confirm table relationships, followed by creating and populating the database with sample data. A concrete aspect on which ProtaBank can be supplemented is by adding data on the origin of protein designs, whether they were created de novo through computational methods or developed through incremental improvements of existing designs. Additionally, tracking the lineage of these designs, including the original design they are based on, where applicable, would provide valuable context.

The next step relies on running test queries to verify data retrieval, integrity constraints, and confirm the foreign key dependencies. This should help optimize the schema, clarify entity relationships and data flow, and address issues like data redundancy, missing links, and inefficiencies. After that, the aim is to set up the database environment, develop scripts to insert, retrieve, and manage data, and implement data ingestion mechanisms to handle CSV, JSON, or direct database input. Then, finally, we will optimize query performance and implement error-handling mechanisms for data processing.

4 Workplan

- Conduct a literature review on protein design and protein design data repositories.
- Analyze existing data models and repository structures in the field.
- Identify the key data that should be stored in the database.
- Design database schema
- Review the schema structure and confirm table relationships.
- Create and insert sample data representing real-world protein design data.
- Run test queries to verify data retrieval, integrity constraints, and foreign key dependencies.
- Identify and resolve issues such as redundancy, missing relationships, or inefficiencies.
- Optimize the schema based on findings and document entity relationships and data flow.
- Set up the database environment (local or server-based).
- Develop scripts for inserting, retrieving, and managing data.
- Fill the database with data obtained by us in the scope of protein design campaigns, as well as with literature data, to test its capabilities

References

1. Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E.H., Bye-A-Jee, H., Carpentier, D., Denny, P., Fan, J., Garmiri, P., da Costa Gonzales, L.J., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Luo, J., Lussi, Y., Marin, J.S.M., Raposo, P., Rice, D.L., Santos, R., Speretta, E., Stephenson, J.,

Totoo, P., Tyagi, N., Urakova, N., Vasudev, P., Warner, K., Wijerathne, S., Yu, C.W.H., Zaru, R., Bridge, A.J., Aimo, L., Argoud-Puy, G., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Neto, T.M.B., Blatter, M.C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., de Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Mercier, P.L., Lieberherr, D., Masson, P., Morgat, A., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Sundaram, S., Sveshnikova, A., Wu, C.H., Arighi, C.N., Chen, C., Chen, Y., Huang, H., Laiho, K., Lehvaslaiho, M., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Y., Zhang, J.: Uniprot: the universal protein knowledgebase in 2025. Nucleic Acids Research **53**, D609–D617 (1 2025). https://doi.org/10.1093/nar/gkae1010

2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank (1 2000). https://doi.org/10.1093/nar/28.1.235
3. Choi, E.J., Mayo, S.L.: Generation and analysis of proline mutants in protein g. Protein Engineering, Design and Selection **19**, 285–289 (6 2006). https://doi.org/10.1093/protein/gzl007
4. Chronowska, M., Stam, M.J., Woolfson, D.N., Costanzo, L.F.D., Wood, C.W.: The protein design archive (pda): insights from 40 years of protein design. Nature Biotechnology (3 2025). https://doi.org/10.1038/s41587-025-02607-x
5. Goverde, C.A., Pacesa, M., Goldbach, N., Dornfeld, L.J., Balbi, P.E., Georgeon, S., Rosset, S., Kapoor, S., Choudhury, J., Dauparas, J., Schellhaas, C., Kozlov, S., Baker, D., Ovchinnikov, S., Vecchio, A.J., Correia, B.E.: Computational design of soluble and functional membrane protein analogues. Nature **631**, 449–458 (7 2024). https://doi.org/10.1038/s41586-024-07601-y
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with alphafold. Nature **596**, 583–589 (8 2021). https://doi.org/10.1038/s41586-021-03819-2
7. Tauriello, G., Waterhouse, A.M., Haas, J., Behringer, D., Bienert, S., Garello, T., Schwede, T.: Modelarchive: A deposition database for computational macromolecular structural models. Journal of Molecular Biology (2025). https://doi.org/10.1016/j.jmb.2025.168996
8. Wang, C.Y., Chang, P.M., Ary, M.L., Allen, B.D., Chica, R.A., Mayo, S.L., Olafson, B.D.: Protabank: A repository for protein design and engineering data. Protein Science **27**, 1113–1124 (6 2018). https://doi.org/10.1002/pro.3406
9. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J.V.D., Mulligen, E.V., Velterop,

J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: Comment: The fair guiding principles for scientific data management and stewardship. Scientific Data **3** (3 2016). https://doi.org/10.1038/sdata.2016.18

10. Zhu, J., Liang, M., Sun, K., Wei, Y., Guo, R., Zhang, L., Shi, J., Ma, D., Hu, Q., Huang, G., Lu, P.: De novo design of transmembrane fluorescence-activating proteins. Nature (2025). https://doi.org/10.1038/s41586-025-08598-8