Optimizing Protein Purification Strategies: Insights from LLM-Driven Mining of PDB Literature

J. Sivaraman

Department of Biological Sciences,

14 Science Drive 4,

National University of Singapore,

Singapore 117543

Abstract:

Successful structural studies, drug discovery, and biochemical assays depend on the efficient isolation and purification of target proteins. This process separates the protein of interest from other cellular components. While general approaches to protein expression and purification are similar, they vary in key aspects such as expression systems, buffer conditions, and fusion tags. As a result, the process is often time-consuming and involves trial and error. With over 81,000 Protein Data Bank (PDB)-related publications available as of October 2024, manually extracting relevant methods from the literature has become increasingly impractical. To address this challenge, we developed an automated tool powered by a large language model (LLM) to extract and classify key data from scientific articles without human input, saving time and effort. Our 2-step LLM pipeline with a 3-step prompting strategy significantly improved extraction accuracy. This work represents the largest database and literature analysis of its kind, offering numerous features. We validated the tool through case studies, including membrane protein purification and the evaluation of crosslinker and detergent preferences in cryo-EM sample preparation. Our findings provide valuable insights and a practical resource to guide protein expression and purification efforts, help reduce the trial-and-error process when working with novel or uncharacterized targets.

Reference

Chen Z, Sivaraman J. Using Large Language Model to Optimize Protein Purification: Insights from Protein Structure Literature Associated with Protein Data Bank. Adv Sci. 2025 Feb 20: e2413689.