Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Master Thesis

Ligand Transplantation and Optimization in Protein Models

Submitted by

Marius Rüve
Student number 7657327

in the study program

Master of Science Bioinformatics
at the Center for Bioinformatics Hamburg
and the Department of Informatics
of the MIN Faculty

September 9, 2025

1. Examiner: Prof. Dr. Matthias Rarey
2. Examiner: Prof. Dr. Andrew Torda

Abstract

Todo: Fill a 200–250 word abstract summarizing problem, method (FoldFusion), key results (quantitative), limitations, and contributions.

Todo: Ensure advisor names/titles and matriculation number are correct and match university records.

Todo: Decide on final chapter structure: currently Introduction, Methods, Results, Conclusion. Remove references to non-existent chapters (Background, Discussion) or add those chapters.

Todo: Move final figures into thesis/figures as PDF/SVG (preferred) or 300 DPI PNG; reference them in Methods/Results.

Todo: Add an Appendix section with exact reproducibility details: environment (Python version, key packages), commands to run pipeline, git commit hash, dataset versions, and config.toml snapshot.

Todo: Run full LaTeX compile and resolve all warnings: no undefined refs/citations, consistent figure/table numbering, proper hyphenation.

Test text

Contents

1 Chapter 1. Introduction

Proteins are the primary effectors of cellular function, and understanding their three-dimensional (3D) structures is essential for mechanistic insight, target identification, and rational design in biotechnology and drug discovery. For decades, experimental structure determination by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) has provided atomic-level information, but at significant cost and with notable coverage gaps. Recent breakthroughs in machine learning, most prominently *AlphaFold* and *RoseTTAFold*, have transformed the landscape by enabling accurate in silico structure prediction directly from amino-acid sequences, dramatically increasing the number of available models across many proteomes [1, 2, 3]. These advances have catalyzed a shift from "structure scarcity" to a new phase in which the central challenges concern structural context, functional annotation, and downstream usability of predicted models.

In this thesis, a distinction is made between *experimental structural models* (determined by X-ray, NMR, or cryo-EM and deposited in the PDB [4]) and *predicted models* (e.g., AlphaFold, RoseTTAFold). Experimental structures frequently capture proteins in a *holo* state, i.e., with bound small molecules such as metal ions, cofactors, or ligands; by contrast, predicted models are provided as *apo* proteins without non-polymer entities by design. The term *ligands* is used to refer broadly to small molecules in binding sites, including organic cofactors (e.g., heme, FAD, NAD(P)H) and mono- or polyatomic ions (e.g., Zn^{2+}, Mg^{2+}). Because many proteins require such moieties to stabilize their fold or to realize catalysis, the absence of these entities in predicted structures can obscure function, binding-site chemistry, and downstream computability. Moreover, predicted models include per-residue confidence scores (e.g., pLDDT in AlphaFold) that correlate with

local reliability and often flag flexible or poorly constrained regions, reinforcing the need to reason about context before analysis [1, 4, 5].

1.1. Motivation

Predicted single-chain protein models typically represent the polypeptide backbone and side chains under canonical residue chemistry. However, many proteins only adopt their native fold or functional state in the presence of small molecules: metal ions that stabilize architecture, organic cofactors that mediate catalysis, and physiologically relevant ligands such as ATP, heme, or NAD(P)H. In predicted models, these moieties are absent by design, which complicates functional interpretation, binding-site analysis, and computational follow-ups such as molecular docking or molecular dynamics. Moreover, the conformational state of a predicted model is often not annotated, and flexible regions may be modelled with lower confidence, further widening the gap between prediction and experimental use in structure-guided workflows.

As a concrete example, protein kinases and many ATPases require Mg^{2+}-ATP in the active site to position catalytic residues and neutralize charge; in an apo predicted model, the nucleotide pocket may appear collapsed or incorrectly polarized, hampering binding-site recognition and leading to unrealistic docking poses. Similarly, heme enzymes without the porphyrin cofactor or zinc-finger domains without Zn^{2+} can appear destabilized or ambiguous with respect to their functional geometry. Restoring these entities from homologous experimental structures often reveals the correct local architecture and plausible interaction networks, enabling more faithful functional hypotheses and more realistic computational experiments [5].

1.2. Problem Statement

A central question is how to systematically enrich predicted protein structures with plausible, biochemically sensible small molecules and ions to make them more useful for functional reasoning and downstream computation, without requiring de novo quantum mechanics or extensive experimental input for each target. The central idea explored in this thesis is to exploit the wealth of biophysical knowledge already contained in experimentally determined structures deposited in the Protein Data Bank (PDB) [4] by transferring (“transplanting”) ligands and cofactors from suitable homologues into corresponding predicted models.

1.3. Prior Work and Opportunity

A closely related line of work demonstrates that large-scale ligand transplantation from experimentally solved homologous structures into AlphaFold models is both feasible and useful. For example, the *AlphaFill* resource applies sequence and structure similarity to transplant common ligands, cofactors, and metal ions from curated experimental models, validating quality with metrics such as local RMSD and a transplant-clash score. At scale, AlphaFill reported over twelve million transplants across nearly one million AlphaFold models, exposing binding sites, restoring essential cofactors (e.g., heme, Zn^{2+}, Mg^{2+}), and enabling hypothesis generation about function [5]. While this demonstrates the promise of homology-driven enrichment, there remains a practical need for open, lightweight, and extensible pipelines that researchers can adapt to bespoke datasets, integrate into modern data/ML workflows, and evaluate end-to-end on specific biological questions.

This opportunity is addressed with a binding-site-guided, modular pipeline that differs from prior resources in several ways. First, candidate donor structures are prioritized via binding-site detection to focus alignment on functionally relevant pockets (using *DoGSite3* [6]). Second, site-centric structural alignment is performed to position donor sites relative to the predicted target (via *SIENA* [7]). Third, transplanted ligands are optionally refined with a simple pose optimization/scoring step (*JAMDAScorer* [8]) to reduce clashes and improve geometry. Finally, each placement is annotated with quality indicators, *Local RMSD* in the ligand environment and a *Transplant Clash Score (TCS)*, to support downstream triage and decision-making. These choices emphasize transparency, configurability, and ease of integration with scripting- and batch-oriented research workflows.

1.4. Thesis Aim

This thesis presents a proof-of-concept pipeline that automates ligand and cofactor transplantation from the PDB into AlphaFold structures to produce *context-enriched* models suitable for exploratory analysis and computational follow-up.[1] At a high level, the pipeline (i) identifies homologous donor structures with relevant non-polymer entities, (ii) performs global and local structural alignments to position candidate ligands, (iii) applies simple but effective filters to prioritize biochemically sensible placements, and (iv) emits enriched models and metadata for quality control and downstream use.

[1] A public code repository accompanies this work; see Chapter 2 for implementation details and reproducibility considerations.

Throughout, *Local RMSD* (a measure of structural agreement in the protein environment around the ligand) and a *Transplant Clash Score* (a measure of steric overlap between ligand and protein) are reported as lightweight indicators of placement quality; full definitions appear in Chapter 2. The pipeline is evaluated on a representative set of UniProt targets assembled for this study, and outcomes are summarized in Chapter 3.

1.5. Contributions

The main contributions of this thesis are:

1. **A modular, open pipeline for ligand transplantation.** The pipeline implements an end-to-end workflow that ingests predicted structures, retrieves homologous experimental entries, and transplants non-polymer entities (ligands, cofactors, metal ions) with provenance tracking and reproducible configuration.
2. **Quality indicators and metadata for downstream trust.** The pipeline annotates each transplant with alignment measures and simple clash checks, enabling users to stratify placements by confidence and decide when refinement is warranted.
3. **Empirical evaluation on representative targets.** The evaluation illustrates use cases where enrichment adds value for functional interpretation (e.g., revealing cofactor requirements or likely substrate preferences) and for computation (e.g., seeding docking with realistic binding-site chemistry).
4. **Engineering for reproducibility and integration.** The codebase is designed for batch execution, scripted analysis, and integration with common structural bioinformatics tools, easing adoption in research settings.

1.6. Research Questions

The following questions guide this thesis:

- **RQ1:** To what extent can homolog-based ligand transplantation reliably restore biochemically plausible small-molecule context in predicted structures?
- **RQ2:** Which alignment and filtering criteria most influence transplant quality, and how should they be configured in practice?

- **RQ3:** How does structural enrichment affect downstream tasks such as docking preparation, site annotation, or hypothesis generation about protein function?

1.7. Scope and Limitations

The pipeline focuses on non-polymer ligands and metal ions commonly represented in the PDB. It does not perform full flexible-receptor docking or quantum refinement, and it does not attempt to model post-translational modifications or glycans. As with any homology-based approach, transplant reliability depends on the availability and quality of structurally similar donors, as well as on the conformational compatibility between donor and acceptor. Enriched models should be treated as *qualitative* hypotheses that can guide experiments or more detailed simulations, rather than as final, quantitatively precise holo structures [5].

1.8. Thesis Structure

Chapter 2 details the pipeline design and implementation. Chapter 3 reports evaluation on diverse targets and discusses quality indicators and failure modes. Chapter 4 concludes and outlines avenues for future work, including refinement protocols and multi-state/complex modelling.

Chapter 2. Methods

2.1. Overview

The framework implements a modular pipeline for systematic augmentation and assessment of predicted protein structures using experimentally resolved homologous complexes and associated ligands. The architecture consists of (i)

The framework implements a modular pipeline for systematic augmentation and assessment of predicted protein structures using experimentally resolved homologous complexes and associated ligands. Its architecture comprises (i) structured acquisition of target and donor structural data, (ii) detection and characterization of binding environments, (iii) sequence/structure alignment and geometric mapping, (iv) ligand and pocket feature transplantation with conflict resolution, (v) multi-criteria filtering and scoring, and (vi) standardized evaluation and visualization of resulting models and annotations. This layered decomposition promotes separation of concerns, controlled extensibility, and reproducible execution.

Inputs consist of one or more target protein identifiers (e.g., UniProt accessions) whose corresponding predicted three-dimensional models are retrieved (or validated if cached) and normalized. Donor candidates are gathered from curated structural repositories and homolog databases according to configurable similarity, coverage, and quality constraints. Auxiliary metadata (ligands, binding site residues, cavity descriptors, and experimental provenance) are harvested through dedicated tool adapters. Each external data source is encapsulated behind a dedicated integration component to isolate protocol specifics from core orchestration logic.

Alignment and transplantation proceed via staged refinement: coarse-grained sequence or domain-level correspondence, optional structural superposition, residue-level mapping, and coordinate transfer for ligand entities and binding site annotations. Conflict handling addresses steric clashes, incomplete residue definitions, alternate conformers, and chain indexing inconsistencies. A filtering and scoring layer then applies geometric, physicochemical, and provenance-derived criteria (e.g., clash metrics, distance thresholds, ligand completeness, resolution provenance proxies) to retain only high-confidence transplanted assemblies. Scoring outputs are structured to permit downstream comparative analyses across targets, donors, and parameter settings.

Evaluation utilities quantify improvement and reliability through metrics such as structural alignment quality, pocket conservation, ligand retention integrity, and scoring distribution profiles. Visualization components generate comparative plots and summary artifacts to support interpretability and methodological diagnosis. Configuration management and structured logging ensure parameter traceability, deterministic re-runs, and auditability of intermediate decisions. The overall design facilitates insertion of new data sources, scoring strategies, or analytical endpoints with minimal impact on existing workflow stages.

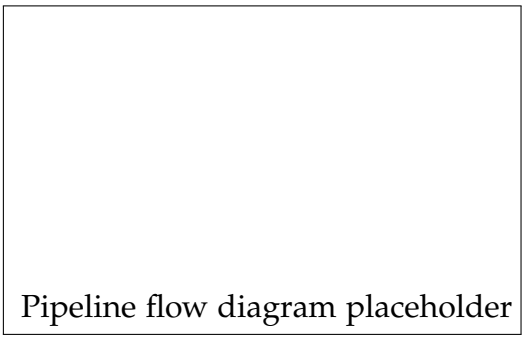

Figure 2.1.: High-level pipeline stages: data acquisition; binding site and ligand characterization; alignment and geometric mapping; transplantation; filtering and scoring; evaluation and visualization.

2.2. Data and Inputs

Todo: Describe the input protein models (AlphaFold/UniProt IDs), how they are fetched, and the benchmark dataset in evaluation/data (size, selection criteria). Include table summarizing proteins used.

Todo: Specify configuration via config.toml: key parameters (alignment thresholds, filters), seeds, and runtime knobs.

2.3. Components and Tools

2.3.1. AlphaFold Fetcher

Todo: Explain how predicted structures are obtained: AlphaFold DB URLs, version fallback v4→v3, file formats (PDB), and caching/normalization strategy.

Todo: Detail preprocessing: pLDDT-based trimming of unreliable segments ($\geq$5 consecutive residues with pLDDT $<$ 50) as implemented in the pipeline; justify threshold and segment-based approach versus per-residue masking.

Todo: Note chain indexing/altloc handling from AlphaFold PDBs, and any renaming or path conventions used in the output directory structure (`AlphaFold/`).

2.3.2. Binding Site Prediction (DoGSite3)

Todo: Summarize DoGSite3's purpose (pocket detection via Difference of Gaussian on protein surface; physicochemical descriptors). Include citation and tool version.

Todo: Describe invocation used in the pipeline: `--writeSiteResiduesEDF`, input PDB from AlphaFold, output directory structure (`Dogsite3/`).

Todo: State selection of the top-ranked pocket EDF (`output_P_1_res.edf`) and the REFERENCE fix-up to the absolute input PDB path performed post-run.

Todo: List key pocket features exported in EDF that downstream tools (SIENA) consume.

2.3.3. Donor Databases and Retrieval (PDB/UniProt/SIENA DB)

Todo: Detail sources of donor structures and ligands: PDB (version/date), UniProt mapping, and SIENA database preparation path. Include database versions and citations.

Todo: Explain how the SIENA database is generated/initialized once per run (input PDB directory, format), where it is stored, and when regeneration is skipped.

Todo: State any filtering on donors prior to alignment (resolution, polymer type, ligand presence).

2.3.4. Binding-Site Similarity Search and Alignment (SIENA)

Todo: Describe inputs (EDF pocket from DoGSite3) and the SIENA search/alignment workflow; include citation and tool version.

Todo: Document command parameters used: `--edf`, `--database`, `--output .`, identity cutoff (`--identity 0.85`); justify chosen thresholds.

Todo: Explain result parsing: reading `resultStatistic.csv` (semicolon separated), column cleanup, sorting primarily by Active site identity (desc), then Backbone and All-atom RMSD (asc).

Todo: Clarify use of `ensemble/.pdb` files for downstream ligand extraction/transplantation and how many top alignments are retained (configurable `siena_max_alignments`).

2.3.5. Ligand Optimization and Scoring (JAMDA)

Todo: Summarize JAMDA's role: energy-based pose optimization/scoring of transplanted ligands; include citation and tool version.

Todo: Describe invocation pattern per ligand: inputs (AlphaFold PDB, ligand SDF), outputs (optimized SDF), and flags used (`--optimize`); output organization (`JamdaScorer/<PDB>/ligand_id.sdf`).

Todo: State what scores/outputs are captured and how they inform filtering (e.g., clash reduction, score improvements). Cross-reference concrete thresholds in Filtering and Scoring.

2.3.6. Alignment

Todo: Describe global/local alignment methods used; define RMSD/score variants and parameters.

2.3.7. Transplantation

Todo: Explain coordinate mapping, handling of alternate locations, protonation states, and chain/residue mapping.

2.3.8. Filtering and Scoring

Todo: Define clash checks, distance cutoffs, and quality indicators; justify chosen thresholds. Reference how SIENA metrics (identity, RMSDs) and JAMDA outputs contribute to accept/reject decisions.

2.4. Implementation Details

Todo: Summarize code structure (packages in foldfusion/), logging, CLI entry points, and dependency management (pyproject/uv.lock).

Todo: Discuss error handling and retries for external resources; note offline/cached mode.

2.5. Reproducibility

Todo: Document exact environment (Python version, key packages), compute hardware, and commands to reproduce results. Reference Appendix for full config and commit hash.

2.6. Ethical and Licensing Considerations

Todo: Note database/tool licenses and usage restrictions; acknowledge limitations of homology-based inference.

3 Chapter 3. Results

3.1. Evaluation Setup

Todo: Briefly restate dataset and metrics. Point to Methods for details; specify commit hash and config used for final runs.

3.2. Overall Transplantation Performance

Todo: Report counts: number of targets, donors found, ligands considered, successful transplants. Include a summary table.

3.3. Quality Indicators

Todo: Present distributions of RMSD, clash scores, and confidence indicators. Add figure placeholders.

3.4. Alignment Quality Analysis

Todo: Insert analysis and figure from `scripts/evaluation_visualisations.py`. Explain trends and outliers.

3.5. Optimization and Filtering Effects

Todo: Show before/after metrics; explain trade-offs. Include relevant figures (`optimization_*`).

3.6. Case Studies

Todo: Pick 2–3 representative proteins; show visualizations (ligand placement), discuss plausibility and caveats.

3.7. Failure Modes

Todo: Describe common errors (misaligned donors, steric clashes, metal coordination errors) with examples and how filters mitigate them.

3.8. Ablations (optional)

Todo: If available, compare alternative alignment thresholds or scoring components. Summarize in a small table.

Chapter 4. Conclusion

4.1. Summary of Findings

Todo: Summarize key results and what they imply for ligand transplantation into predicted protein models. Reference main figures/tables.

4.2. Limitations

Todo: State major limitations: dependence on donor availability/quality, alignment sensitivity, metal coordination accuracy, protonation/tautomer issues.

4.3. Future Work

Todo: List concrete next steps: refinement with energy minimization, multi-state modelling, better scoring for metals, integration with docking/MD pipelines.

4.4. Concluding Remarks

Todo: Provide a concise take-home message and broader impact.

References

[1] John Jumper et al. "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 09/05/2025).

[2] Minkyung Baek et al. "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network". In: *Science* 373.6557 (Aug. 20, 2021), pp. 871–876. DOI: 10.1126/science.abj8754. URL: https://www.science.org/doi/10.1126/science.abj8754 (visited on 09/05/2025).

[3] Kathryn Tunyasuvunakool et al. "Highly Accurate Protein Structure Prediction for the Human Proteome". In: *Nature* 596.7873 (Aug. 2021), pp. 590–596. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03828-1. URL: https://www.nature.com/articles/s41586-021-03828-1 (visited on 09/05/2025).

[4] Stephen K Burley et al. "RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy". In: *Nucleic Acids Research* 47.D1 (Jan. 8, 2019), pp. D464–D474. ISSN: 0305-1048. DOI: 10.1093/nar/gky1004. URL: https://doi.org/10.1093/nar/gky1004 (visited on 09/05/2025).

[5] Maarten L. Hekkelman et al. "AlphaFill: Enriching AlphaFold Models with Ligands and Cofactors". In: *Nature Methods* 20.2 (Feb. 2023), pp. 205–213. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01685-y. URL: https://www.nature.com/articles/s41592-022-01685-y (visited on 09/05/2025).

[6] Joel Graef, Christiane Ehrt, and Matthias Rarey. "Binding Site Detection Remastered: Enabling Fast, Robust, and Reliable Binding Site Detection and Descriptor Calculation with DoGSite3". In: *Journal of Chemical Information and Modeling* 63.10 (May 22, 2023), pp. 3128–3137. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c00336. URL: https://doi.org/10.1021/acs.jcim.3c00336 (visited on 09/05/2025).

[7] Stefan Bietz and Matthias Rarey. "SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles". In: *Journal of Chemical Information and Modeling* 56.1 (Jan. 25, 2016), pp. 248–259. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00588. URL: https://doi.org/10.1021/acs.jcim.5b00588 (visited on 09/05/2025).

[8] Florian Flachsenberg et al. "A Consistent Scheme for Gradient-Based Optimization of Protein–Ligand Poses". In: *Journal of Chemical Information and Modeling* 60.12 (Dec. 28, 2020), pp. 6502–6522. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c01095. URL: https://doi.org/10.1021/acs.jcim.0c01095 (visited on 09/05/2025).

A Appendix A. Reproducibility Details

Todo: Insert environment details: OS, Python version, key packages with versions, hardware.

Todo: Insert exact commands (Makefile targets or CLI) used to generate results and figures.

Todo: Record git commit hash and config snapshot used for final evaluation.

Appendix B.

B Supplementary Tables and Figures

Todo: Add tables summarizing dataset proteins, donors, ligands, and transplant success by category.

Todo: Include any extended figures not in the main text.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich vorliegende Masterarbeit im Studiengang Bioinformatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den 9. September 2025

_____

Marius Rüve

Ich bin mit einer Einstellung der Masterarbeit in den Bestand der Bibliothek des Departments Informatik einverstanden.

Hamburg, den 9. September 2025

_____

Marius Rüve