FoldFusion-PoC: A Ligand Transplantation & Optimization Pipeline

Marius Rueve

May 19, 2025

Talk Outline

- Background: AlphaFill algorithm
- FoldFusion implementation overview
- Key differences between AlphaFill and FoldFusion
- AlphaFill's scientific evaluation
- FoldFusion's current evaluation approach & future improvements

What the AlphaFill Paper Did

Goal: “Transplant” small molecules & ions from experimentally determined PDB structures into AlphaFold models to enrich them with biologically relevant ligands.

1. BLAST each AlphaFold sequence against PDB-REDO database
2. Filter hits by $\geq$25% identity over $\geq$85 residues
3. Perform global (Cα) and local (backbone within 6 Å) structural alignments
4. Transplant compounds if no duplicate within 3.5 Å of centroid
5. Record global & local RMSD and calculate TCS (clash score)

AlphaFill's Scientific Evaluation

- **Validation Set:** 28,619 transplants from 100%-identity donors.
- **Quality Metrics:**
  - **LEV score:** All-atom RMSD of ligand + nearby protein atoms within 6 Å
  - **Local RMSD:** Proxy for LEV; calculated for every transplant
  - **TCS:** $\sqrt{\sum(\text{overlap})^2/N}$, quantifying van der Waals clashes
- **Refinement:** Energy-minimize selected complexes in YASARA; evaluate before/after TCS & LEV
- **Confidence Annotation:** Statistical cutoffs (IQR + 1.5×IQR) on local RMSD & TCS to label high/medium/low confidence

Abbreviations: RMSD = Root-Mean-Square Deviation; LEV = Local Environment Validation; TCS = Transplant Clash Score; IQR = Interquartile Range

What I Did in FoldFusion

Pipeline Components:

1. **AlphaFoldFetcher:** Download AF-DB models
2. **DogSite3:** Detect pockets & generate EDF files
3. **Siena:** Build ensemble of homolog structures around predicted pocket
4. **LigandExtractor:** Identify & extract ligands (HETATM) from SIENA PDBs
5. **JamdaScorer:** Optimize & score each ligand-protein complex

Implementation: Modular Python package (`foldfusion/`), TOML config, structured logging

FoldFusion's Current Scientific Evaluation

- **Core Metric:** Transplant Clash Score (TCS) via `clash_scorer.calculate_tcs`
  - Parses protein (PDB) & ligand (SDF) atoms
  - Computes van der Waals overlap squared sums within cutoff (4 Å)
  - Returns $\sqrt{\text{mean overlap}^2}$
- **What's Missing:**
  - Local RMSD / LEV-style metric for transplant geometry
  - Global RMSD to donor structures
  - Statistical confidence annotations (IQR cutoffs)
  - On-the-fly refinement & re-scoring akin to YASARA minimization

Key Differences

Aspect	**AlphaFill**	**FoldFusion**
Homology Search	Sequence-based BLAST vs. PDB-REDO	Pocket-centric: DogSite3 + SIENA
Transplant Criteria	Seq. identity ≥25%, local/global RMSD	Pocket detection & ensemble extraction
Evaluation	LEV, local RMSD, TCS + refinement	?
Refinement	YASARA energy minimization	JamdaScorer ”optimize” flag
Confidence	Statistical cutoffs on metrics	Not yet implemented

FoldFusion - What Can Be Improved

1. **Add Local RMSD Calculation:** Use structural alignment of pocket residues (backbone within 6 Å) to compute local RMSD, mirroring LEV
2. **Record Global RMSD:** After SIENA alignment, compute Cα-based global RMSD to donor PDB
3. **Implement Confidence Tiers:** Analyze distributions of local RMSD & TCS; define IQR-based cutoffs for high/med/low confidence
4. **Integrate On-Demand Refinement:** Hook into YASARA or another minimizer; recalculate TCS post-minimization
5. **Visual & Statistical Reporting:** Summarize per-model transplant quality (histograms, boxplots) for pipeline validation

Next Steps & Roadmap

- **Metric Expansion:** Add LEV-style and global RMSD metrics
- **Dashboard & Reports:** Generate per-project quality summaries
- **Benchmarking:** Compare against AlphaFill on a shared test set

Summary & Conclusions

- **AlphaFill** set the stage with homology-driven ligand transplantation, multi-metric validation, and confidence annotation
- **FoldFusion** reimagines the workflow around pocket detection (DogSite3) and ensemble extraction (SIENA), with a modular Python framework
- **Opportunities:** Enrich evaluation with additional metrics, statistical confidence levels, and refinement to match and surpass AlphaFill's rigor

Thank you!