Student Project Proposal

From Dimers to Drugs

PEARL - Peptide Extraction and AI-guided Refinement for Ligand design

Master area	AI for Drug Discovery; protein-protein interactions; structural bioinformatics; computational chemistry
Project type	Computational project with Python implementation, MD simulation and AI-guided molecule design
Duration	6–8 weeks
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Basic Python and machine learning; introductory biochemistry; familiarity with protein structure is useful
Main output	Documented Jupyter notebooks (one per module), a technical report (15–20 pages), and a short final presentation
Possible targets	EGFR kinase domain dimer (PDB: 3NJP); BRAF dimer (PDB: 4MNE); CDK2/Cyclin A (PDB: 1FIN)

The project asks the student to implement an end-to-end computational pipeline that reverse-engineers a kinase dimer interface to design peptide and small-molecule inhibitors. Starting from a crystallographic dimer structure, students identify binding hotspots by computational alanine scanning, simulate interface dynamics by molecular dynamics, extract a minimal seed peptide, and apply protein language models and generative AI tools to optimise and miniaturise the binder toward drug-like leads. The project provides hands-on experience with the computational methods — from structural bioinformatics to diffusion-based docking — that are reshaping modern pharmaceutical research.

1. Background and Rationale

Protein kinases are pivotal regulators of cellular signalling and among the most validated drug targets in oncology and inflammatory disease. Many kinases exert their biological function as dimers: the EGFR kinase, for example, is activated through an asymmetric homodimer in which one protomer acts as activator and the other as receiver. Oncogenic mutations (L858R, exon 19 deletions) stabilise this active conformation constitutively, driving uncontrolled proliferation in non-small-cell lung cancer. While ATP-competitive inhibitors have achieved clinical success, resistance invariably emerges, motivating orthogonal therapeutic strategies.

Targeting the protein-protein interface (PPI) that forms the dimer represents a mechanistically distinct approach. PPI interfaces are typically large (1000–3000 Å²) and featureless, resisting conventional small-molecule drug discovery. Nevertheless, computational and structural studies consistently show that a small subset of residues — the hotspots — contribute disproportionately to binding free energy, providing a tractable pharmacophoric target. Recent advances in protein language models (ESM-2), structure prediction (AlphaFold2, ESMFold), inverse folding (ProteinMPNN), and diffusion-based docking (DiffDock) have collectively transformed the feasibility of computational PPI drug discovery, enabling a student project that would have been impractical only five years ago.

This project places these AI tools in a physically grounded framework anchored by molecular dynamics simulation, ensuring that predictions are validated against the thermodynamic reality of the binding interface. The native dimer partner is itself the most optimised binder imaginable: the project extracts its binding essence, then uses AI to compress and refine it toward a drug-like molecule.

2. Project Objective

The objective of the project is to implement and validate a four-stage computational pipeline — interface analysis, molecular dynamics, AI-guided sequence optimisation, and lead generation — that transforms a kinase dimer crystal structure into a ranked set of peptide and peptidomimetic inhibitor candidates. The project should remain realistic for a 6–8 week student activity. It is not necessary to run long molecular dynamics simulations or to develop production-quality software. The student should focus on a well-defined prototype applied to one kinase dimer target, with clear validation at each stage.

3. Specific Aims

1. Retrieve a kinase dimer structure from the PDB, prepare the system, and identify interface residues by inter-chain contact analysis (≤ 4.5 Å).

2. Rank interface residues by binding energy contribution using computational alanine scanning (FoldX or Rosetta), defining the hotspot set.
3. Set up and run molecular dynamics simulations of the full dimer and of the isolated interface peptide using OpenMM, analysing contact persistence with MDAnalysis.
4. Extract a minimal seed peptide (15–25 residues) covering the hotspot region and quantify its binding affinity to the kinase monomer by MM-PBSA.
5. Apply ESM-2 masked language modelling and ProteinMPNN backbone-fixed design to generate optimised candidate sequences, filtered by predicted structure and solubility.
6. Extract a pharmacophore model from the MD-derived peptide-kinase ensemble and use REINVENT 4 and DiffDock to identify peptidomimetic and small-molecule leads.
7. Evaluate and rank candidates by docking score, MM-GBSA binding energy, and drug-likeness filters (Lipinski, Veber), and present the top lead with a structural rationale.

4. Proposed Methodology

4.1 Interface Analysis and Hotspot Identification

The student will retrieve the target dimer structure from the PDB and prepare it using standard protocols: removal of water and heteroatoms, addition of missing residues, and protonation at physiological pH. Interface residues will be identified as those with any heavy atom within 4.5 Å of the partner chain. Computational alanine scanning will be performed using FoldX (PositionScan command) or Rosetta (ddg_monomer protocol) to compute ΔΔG for each interface residue, defining hotspots as those with ΔΔG > 1.5 kcal/mol. Cross-family conservation will be assessed by aligning kinase domain sequences using BioPython and annotating conservation onto the structure in PyMOL.

4.2 Molecular Dynamics Simulation

The dimer will be parameterised using the AMBER14SB force field with TIP3P water in OpenMM. Following energy minimisation and NVT equilibration (2 ns), an NPT production run of 50–100 ns will be performed at 300 K and 1 atm. Interface contact persistence will be computed with MDAnalysis across the trajectory. The interface peptide (seed peptide) will be extracted, its termini capped (acetyl-N, amide-C), re-solvated, and subjected to a 20 ns MD run in isolation. MM-PBSA binding free energy will be estimated using MDAnalysis or gmx_MMPBSA on snapshots extracted from the dimer trajectory.

4.3 AI-Guided Sequence Optimisation

The seed peptide sequence will be processed by ESM-2 (650M parameter model via the fair-esm library) with hotspot positions masked, generating a ranked list of substitutions by log-likelihood. ProteinMPNN will be applied to the MD-averaged backbone geometry of the interface fragment to redesign the sequence while preserving the binding conformation. A simple Bayesian optimisation loop (BoTorch) will be used to guide iterative sequence evaluation, using ESMFold-predicted structure and a docking surrogate as objectives. Candidate sequences will be filtered by predicted secondary structure agreement, solubility score (CamSol), and absence of known immunogenic motifs.

4.4 Pharmacophore Extraction and Lead Generation

The pharmacophore model will be derived from the MD ensemble of the optimised peptide bound to the kinase monomer, using RDKit Pharm3D to map H-bond donors/acceptors, hydrophobic centres, and charged features at hotspot sidechain positions. REINVENT 4 will be used in a pharmacophore-constrained generation mode to produce small-molecule candidates. DiffDock will be applied for blind docking of all candidates onto the kinase monomer, followed by MM-GBSA rescoring of the top 50 poses. Final candidates will be filtered by Lipinski's Rule of Five and Veber's oral bioavailability criteria.

5. Indicative Work Plan

Period	Main Task	Expected Output
Weeks 1–2	PDB retrieval, structure preparation, interface contact analysis, alanine scanning	Hotspot residue list with ΔΔG values; PyMOL visualisation
Weeks 2–3	OpenMM MD setup, equilibration and NPT production run (dimer)	50–100 ns trajectory; contact persistence heat map
Week 3	Seed peptide extraction, isolated MD, MM-PBSA binding affinity	Validated seed peptide with quantified ΔG°bind
Weeks 4–5	ESM-2 substitution profiling; ProteinMPNN redesign; Bayesian optimisation loop	Top 5 optimised sequences with ESMFold structures and predicted scores
Week 5	RDKit pharmacophore model from MD ensemble	Pharmacophore map with feature positions and tolerances
Weeks 6–7	REINVENT 4 generation; DiffDock docking; MM-GBSA rescoring; drug-likeness filtering	Ranked peptidomimetic lead with binding pose and pharmacophore overlay

Week 7–8	Report writing, notebook finalisation, preparation of oral presentation	Final report (15–20 pp.), four annotated notebooks, 15-min presentation

6. Assessment and Deliverables

Assessment is distributed across three deliverables:

- Jupyter Notebooks (40%) — four reproducible notebooks, one per module, with clear annotations, inline figures, and interpretation of results at each stage
- Scientific Report (40%) — 15–20 page written report covering scientific background, methods, results, and critical discussion; should include at least one structural figure and one quantitative comparison of candidates
- Oral Presentation (20%) — 15-minute presentation covering the full pipeline from target to lead, with 5 minutes Q&A; assessed on scientific clarity and ability to justify methodological choices

7. Key References and Resources

Structural Targets

- PDB 3NJP — EGFR kinase domain asymmetric dimer
- PDB 4MNE — BRAF dimer with vemurafenib
- PDB 1FIN — CDK2/Cyclin A complex

Software and Tools (details can vary)

- OpenMM 8.x — molecular dynamics engine (openmm.org)
- MDAnalysis — trajectory analysis (mdanalysis.org)
- ESM-2 / ESMFold — protein language model and structure prediction (github.com/facebookresearch/esm)
- ProteinMPNN — inverse folding / sequence design (github.com/dauparas/ProteinMPNN)
- RDKit — cheminformatics and pharmacophore modelling (rdkit.org)
- REINVENT 4 — generative molecule design (github.com/MolecularAI/REINVENT4)
- DiffDock — diffusion model for molecular docking (github.com/gcorso/DiffDock)

Key Literature

- Arkhipov et al. (2013) Architecture and membrane interactions of the EGF receptor. Cell 152, 557–569
- Wells & McClendon (2007) Reaching for high-hanging fruit at protein-protein interfaces. Nature 450, 1001–1009
- Lin et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130
- Dauparas et al. (2022) Robust deep learning-based protein sequence design using ProteinMPNN. Science 378, 49–56
- Corso et al. (2023) DiffDock: Diffusion steps, twists, and turns for molecular docking. ICLR 2023
- Bagal et al. (2022) MolGPT: Molecular generation using a transformer-decoder model. J. Chem. Inf. Model. 62, 2064–2076

8. Recommended Student Background and Skills Acquired

Preferred Background

- Familiarity with Python programming (NumPy, pandas, Matplotlib)
- Basic biochemistry and molecular biology: protein structure, enzyme function, receptor-ligand interactions
- Introductory understanding of machine learning concepts
- Prior experience with MD simulation or structural bioinformatics is beneficial but not required

Skills Acquired

- Structural bioinformatics: PDB retrieval, interface characterisation, computational alanine scanning
- Molecular dynamics: simulation setup and parameterisation, trajectory analysis, MM-PBSA free energy estimation
- Protein language models: ESM-2 embeddings, masked language modelling for sequence design
- Generative AI for molecules: ProteinMPNN inverse folding, REINVENT 4 pharmacophore-constrained generation
- AI-guided docking: DiffDock blind docking, CNN-based scoring, pose evaluation and ranking
- Scientific communication: reproducible Jupyter notebooks, structured technical report, oral presentation