AI-Driven *De Novo* Generation of Protein Binders

Adrian Jasiński, Michał Barej, Victor Murcia Pienkowski, Joanna Marczyńska-Grzelak, Piotr Skoczylas, Maciej Jasiński & Jan Kaczmarczyk

Ardigen, Krakow, Poland

BACKGROUND

This project aimed to leverage AI-driven methods, complemented by advanced computational techniques such as Molecular Modelling (MM) and Molecular Dynamics (MD) simulations, to generate novel binders for two prespecified protein targets: T1 and T2. The entire binder discovery process - from initial target analysis to final candidate selection - was performed *in silico*.

The core objective was to generate a ranked list of up to 100 binder candidates for each target through a comprehensive computational approach. This process encompasses the identification of promising binding sites, the design of novel binder structures and sequences using generative AI models, and rigorous *in silico* filtering and ranking based on predicted binding, stability, solubility, and other key biochemical properties.

METHODS

We implemented a three-phase computational pipeline to generate high-confidence *de novo* binders for two targets (Figure 1).

- **Target analysis (Phase I):** Binding sites on each target were identified and prioritized using MaSIF [1] and ARDock, our in-house peptide-docking module. The top 500 predicted complexes (ranked by docking score) were selected for further design (Figure 2).
- **Generative design (Phase II):** State-of-the-art models: RFdiffusion [2] and ProteinMPNN [3] were used to generate diverse binder scaffolds and amino acid sequences tailored to the binding sites identified during Phase I (Figure 3 & Table 1). Two strategies were used: *de novo* design from backbone structures and template-guided expansion of fragments. Reverse-folding methods further refined the candidate binders.
- **Filtering and ranking (Phase III):** Candidates were triaged by structural confidence, docking scores, isoelectric point, amino acid composition, and key biophysical properties, ensuring only the most promising molecules progressed (Figure 4 & Table 2). Candidates were ranked based on the values of four scores, PAE interaction, complex pLDDT, pDockQ, and binder RMSD [4-6]. Based on these, rankings of the top 100 binders for each binding site were created.

As a final verification step, candidate sequences were subjected to an anti-drug antibody (ADA) development risk assessment using Ardigen's ARDisplay II model, alongside screening against natural repertoires and patent databases via a tiered risk assessment of sequence similarity and coverage to ensure uniqueness (Figure 5).

RESULTS AND DISCUSSION

This project applied advanced AI-driven methods, complemented by molecular modeling (MM/MD) and *in silico* filtering techniques, to design novel protein binders for two targets. The primary aim was to identify promising binding sites and generate a highly refined list of *de novo* binder candidates for experimental validation.

To achieve this goal, we employed a binder design pipeline adaptable to a wide range of protein targets, with flexibility to account for differences in structural complexity, binding site accessibility, and availability of experimental data.

Key results:

- We executed an *in silico* workflow, encompassing detailed analysis of target protein surface, generative *de novo* design of binder structures and sequences using state-of-the-art AI models and rigorous multiparametric filtering, including patent-deposited sequence screening
- While target T1 benefited from high-resolution crystal structures that enabled confident, template-guided design, target T2 was more challenging because it lacked an unbound protein structure. Therefore, we used custom structural modeling and molecular dynamics simulations to find viable binding regions for *de novo* design.
- We generated a highly curated set of top-scored binder candidates for each binding site identified on the target proteins.
- Through a comprehensive computational evaluation, we modeled binder-target complexes using sequence-based structure prediction tools to assess interaction quality. This multi-faceted approach yielded **a ranklist of the top 100 binder candidates** for each of the 4 selected binding sites.

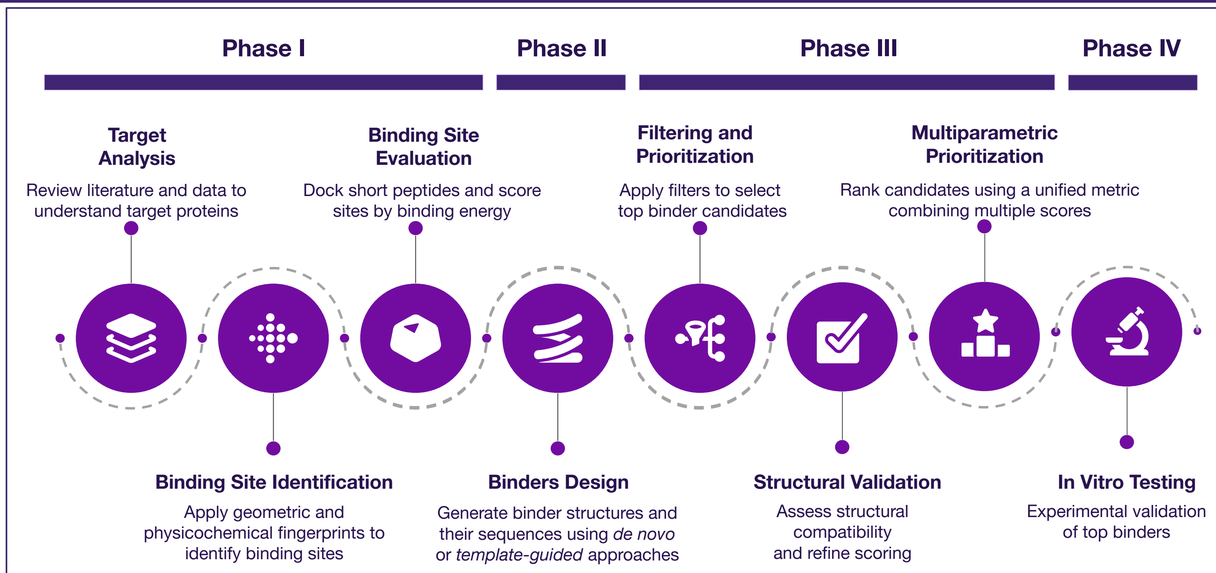


Figure 1: *Project Pipeline Overview. Diagram outlining the full project workflow, divided into four main phases: (I) target analysis, (II) generative design, (III) filtering and ranking, and (IV) in vitro validation.*

Figure 2: *Visual representation of Phase I: This phase involved using a geometric deep learning method, MaSIF, and an in-house peptide-docking platform, ARDock, to identify and prioritize binding sites on target proteins.*

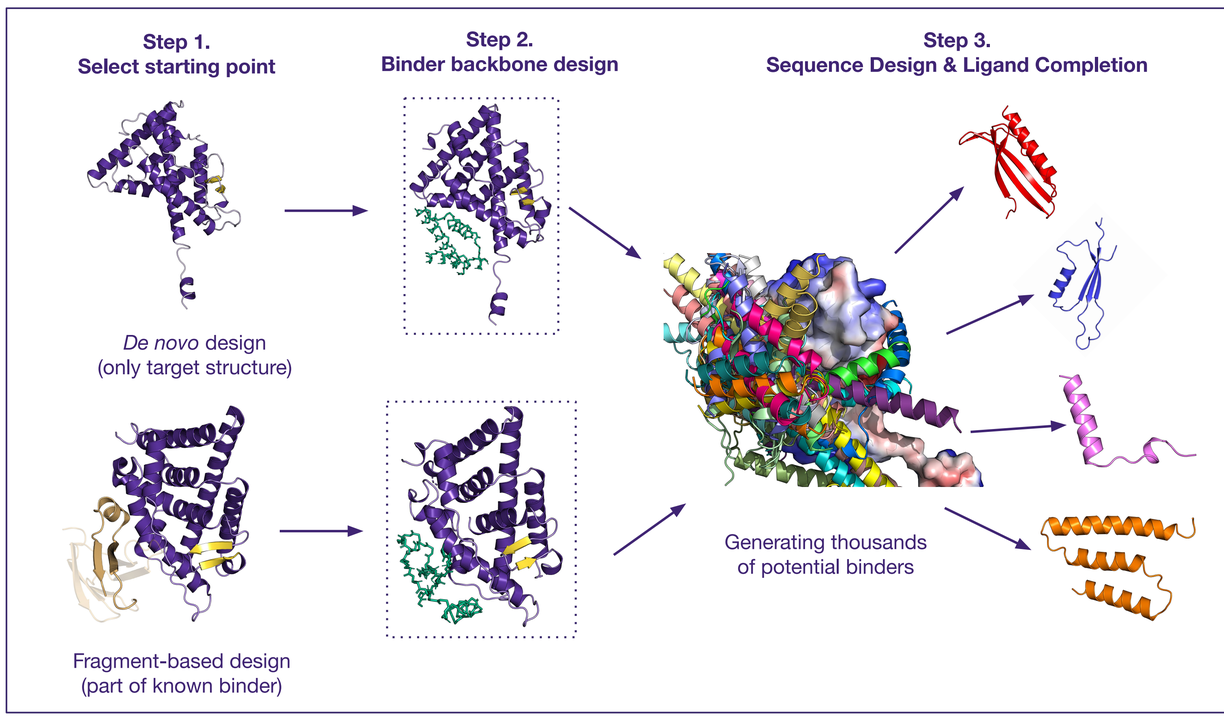


Figure 3: *Visual representation of Phase II: A de novo approach to generate new backbone structures and a template-guided approach to diversify known binder fragments as two complementary strategies for generative binder design.*

Target	Binding site	*De novo*	Using known binder fragments	Total
T1	P1	24,600	42,500	**67,100**
	P2	59,300	0	**59,300**
T2	P1	15,700	0	**15,700**
	P2	8,500	0	**8,500**

Table 1: *Number of generated binding candidates as a result of Phase II. To create potential binders, two deep learning models were used. First, RFdiffusion created three-dimensional protein backbones, either from scratch (de novo) or by incorporating fragments of known binders (template-guided). Next, ProteinMPNN generated multiple different amino acid sequences for each backbone to increase the likelihood of finding high-affinity binders.*

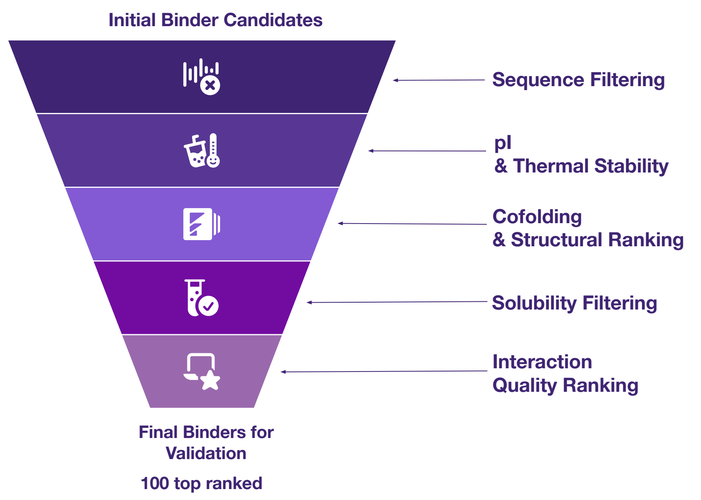


Figure 4: *Visual representation of Phase III: Filtering cascade overview.*

	T1		T2	
	P1	P2	P1	P2
Initial binder candidates	67,100	59,300	15,700	8,500
After sequence-based filters	16,843	19,321	5,912	1,091
After pI and thermal stability filters	8,986	15,573	2,861	962
After selection based on structure confidence score	1,800	930	900	470
After solubility filtering	1,120	677	519	313
Interaction Quality Ranking	100	100	100	100

Table 2: *Number of binder candidates generated and filtered for proteins T1 and T2. The generated binder candidates were evaluated and prioritized based on a comprehensive set of structural, physicochemical, and functional criteria, to select the most promising designs for experimental validation.*

REFERENCES

[1] Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020 Feb;17(2):184–92, DOI: 10.1038/s41592-019-0666-6

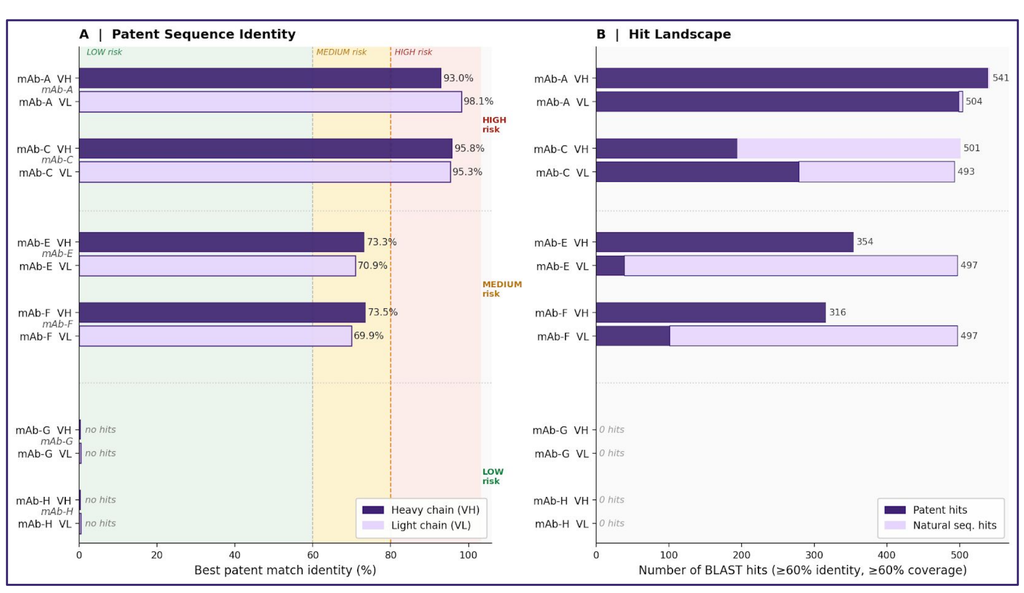
[2] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023 Aug 31;620(7976):1089–100, DOI: 10.1038/s41586-023-06415-8

[3] Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science. 2022 Oct 7;378(6615):49–56, DOI: 10.1126/science.add2187

[4] Wohlwend J, Corso G, Passaro S, Getz N, Reveiz M, Leidal K, et al. Boltz-1 Democratizing Biomolecular Interaction Modeling [Internet]. 2024. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.11.19.624167

[5] Genz LR, Mulvaney T, Nair S, Topf M. PICKLUSTER: a protein-interface clustering and analysis plug-in for UCSF ChimeraX. Bioinformatics. 2023 Nov 1;39(11):btad629. DOI: 10.1093/bioinformatics/btad629

[6] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 2022 Mar 10;13(1):1265, DOI: 10.1038/s41467-022-28865-w



The figure shows exemplary results (here for antibodies) selected to illustrate the applied solution, which is also applicable to other patent infringement verification contexts.

Figure 5. *Patent risk and sequence uniqueness assessment as the final pipeline verification step. As a last screening stage, candidate sequences are queried against a unified reference proteome comprising natural antibody repertoires (ABSD, ~5 species) and broad patent-deposited sequences (PLAbDab, NCBI pataa; >8.4 M entries total). Hits are retained at ≥60% sequence identity and ≥60% query coverage. Patent risk is tiered as HIGH (best patent match ≥80%), MEDIUM (patent hit present but <80%), or LOW (no qualifying patent hits). (A) Best patent match identity per chain, with risk-zone shading. (B) Composition of the hit landscape - patent-derived versus natural repertoire hits - revealing both the degree of IP overlap and the biological precedent of each sequence.*

IMPACT

Our **Biologics Discovery Platform** seamlessly integrates physics-based approaches with state-of-the-art AI methods to deliver high-confidence, ranked binder candidates for experimental validation.

- We successfully applied this pipeline to evaluate complex protein targets, generating *de novo* binders with optimal physicochemical properties.
- All candidates underwent rigorous screening against patent databases and natural repertoires to ensure sequence uniqueness and freedom to operate.

You can download the poster here: