



An AI-Lab for small-molecule discovery at the IL-23R–p19 interface

AI-Lab: PlayMolecule AI

Supervised by: Gianni De Fabritiis

Research manuscript.

Abstract

The contact between IL-23 and its receptor offers a target for small molecules that could alter IL-23 signaling in psoriasis. We present an AI-Lab, a team of AI agents working autonomously under human supervision, and describe its search for such molecules at the IL-23R–p19 interface. A frontend agent coordinated communication, while independent agents chose their next steps and shared progress through a linked HTML wiki and a molecule leaderboard for human review. The AI-Lab analyzed protein structures, retrieved compounds, designed analogues and fragments, predicted binding poses by docking, ran molecular dynamics simulations and compared relative binding free energies within selected compound series. An initial library contained 2,429 molecular records, of which 2,426 were retained after curation. A separate comparison of 23 entries generated 230 three-dimensional conformations and 4,600 docking poses across cytokine-side and receptor-side models, with 138 structures selected for scoring. Reviews of molecular identity, analysis methods and proposed mechanisms guided the next calculations. Repeated simulations tested whether compounds remained in their predicted binding positions. Structural comparisons assessed whether those positions overlapped the space occupied by the receptor or cytokine in the natural complex. Three candidates, W5_004, PS2B_013 and W5_003, met their respective criteria for remaining in the pocket across three 40 ns runs each. These candidates and two exploratory receptor-side compounds formed a panel for experimental testing. The study shows how the AI-Lab used computational results to select compounds for tests of binding and interference with the receptor–cytokine interaction.

1 Introduction

1.1 Target biology and therapeutic rationale

IL-23 helps sustain inflammation in psoriasis, including by maintaining lymphocytes that produce IL-17 [1]. This secreted cytokine contains two subunits: p19, encoded by IL23A, and p40, encoded by IL12B. The subunits bind IL-23R and IL-12R β 1, respectively, to form a signaling complex outside the cell. This complex activates the intracellular JAK2/TYK2–STAT pathway, particularly STAT3 [2, 3]. Blocking the contact between p19 and IL-23R could therefore interrupt cytokine recognition before this intracellular signaling begins.

IL-12 shares p40 and IL-12R β 1 with IL-23. It uses p35 in place of p19 and IL-12R β 2 in place of IL-23R. We therefore focused on the p19–IL-23R contact, which is specific to IL-23, with the aim of preserving IL-12 function. The intended mechanism was to block cytokine binding to its receptor outside the cell [2, 3].

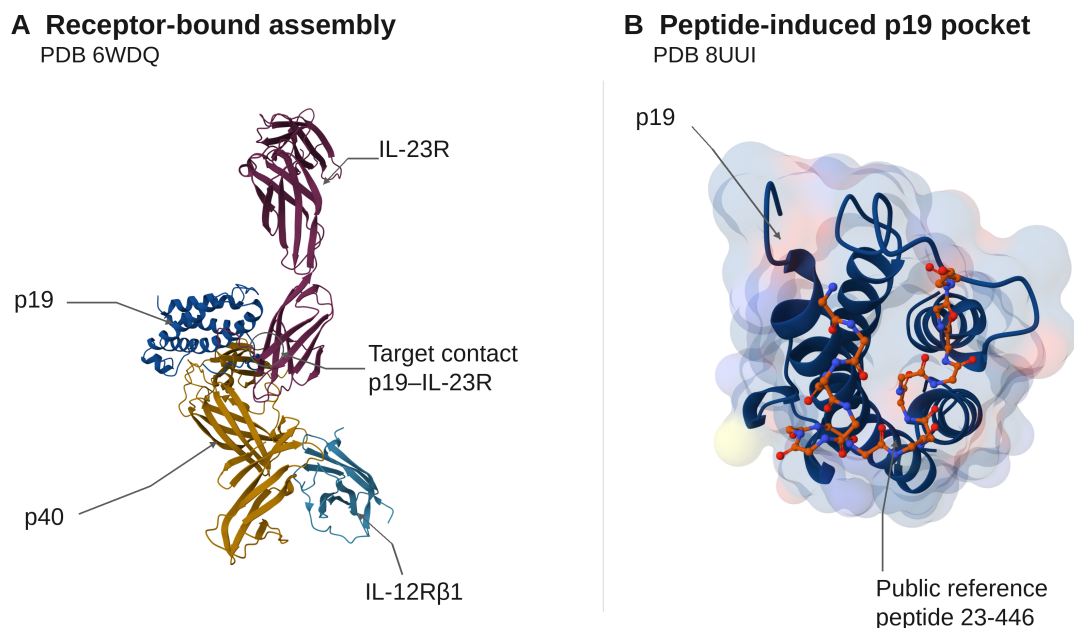


Figure 1. Structures used to guide small-molecule discovery at the IL-23R–p19 interface. **(A)** Receptor-bound complex, PDB 6WDQ, showing the p19–IL-23R contact targeted in this study and the p40–IL-12R β 1 components shared with IL-12 signaling. The panel shows the extracellular protein fragments resolved in the structure. IL-12R β 1 is particularly incomplete, and the deposited IL-23R construct contains the L310P mutation. Glycans, waters and missing residues are omitted. **(B)** Peptide-bound p19, PDB 8UUI, with the resolved backbone of reference peptide 23-446; p40 is omitted for clarity. The three disconnected p19 segments belong to one biological chain. The translucent surface outlines the molecule. Each panel uses its own orientation and scale to show the binding interface. Protein colors are consistent across panels; peptide carbons are orange and other peptide atoms use element colors [3, 6].

1.2 Structural starting points

To alter this interaction, a small molecule must bind a suitable protein conformation and change how the cytokine and receptor interact. Published structures show how the receptor binds IL-23 and identify the contact surface on p19 [3–5]. A structure containing the inhibitory peptide 23-446 shows a pocket induced in p19 near this surface. This pocket provided a starting point for designing non-peptide ligands [6]. These structures supported two approaches: binding in the pocket on the cytokine and binding at the contact surface on the receptor.

Figure 1 shows the structures used for these approaches. On the cytokine side, we asked whether a small molecule could occupy the peptide-induced pocket and affect receptor binding. On the receptor side, we asked whether a molecule could bind in the region normally approached by p19. We overlaid the modeled complexes with the natural receptor–cytokine complex to assess whether a ligand occupied space needed by the binding partner. We also simulated the contacts that held each ligand in place. Together, these analyses examined how a predicted binding position might interfere with the receptor–cytokine interaction. [F03][F18]

1.3 Target product profile and decision criteria

The target product profile (TPP) set the goals for the intended medicine. It called for a chemically defined, non-peptide compound that would directly, selectively and reversibly inhibit the p19–IL-23R interaction. The proposed product was an oral solid dose for adults with moderate-to-severe plaque psoriasis eligible for systemic treatment. Once-daily dosing was preferred, with twice-daily dosing acceptable if justified. The profile set criteria for binding affinity, blockade of the

receptor interaction, effects in cells, selectivity, drug exposure, safety and readiness for further development (Table 1). ^[F02]

Table 1. Target product profile guiding the campaign. The minimum and preferred columns set goals for selecting a preclinical candidate, including required evidence and preferred chemical properties. ^[F02]

Objective	Minimum objective	Preferred objective
Direct affinity and blockade	$K_d \leq 100$ nM and biochemical PPI $IC_{50} \leq 100$ nM; two independent assay principles; identified binding partner and supported site	$K_d \leq 10$ nM and PPI $IC_{50} \leq 30$ nM; ligand-bound structure
Primary human-cell function	IL-23 pSTAT3 $IC_{50} \leq 100$ nM; maximal inhibition $\geq 80\%$; meaningful effects across ≥ 6 donors	$IC_{50} \leq 30$ nM; maximal inhibition $\geq 90\%$; matching later cytokine suppression
Whole-blood potency	Nominal $IC_{50} \leq 300$ nM and directly measured IC_{90} within achievable exposure	Nominal $IC_{50} \leq 100$ nM; modest matrix shift; reproducible ex-vivo coverage
Functional selectivity	≥ 30 -fold IC_{50} separation from IL-12, IL-6, IFN α and viability; exclude relevant JAK/TYK2 activity as a mechanistic control	≥ 100 -fold separation; no material counterpathway suppression at projected free clinical C_{max} (working interpretation $<20\%$, preferably $<10\%$)
Exposure coverage	Unbound target-site trough $\geq 1 \times$ matrix-adjusted IC_{90} , or validated time-dependent PK/PD rationale	$\geq 3 \times$ IC_{90} trough coverage with acceptable margins
Oral pharmacokinetics	Bioavailability $F \geq 20\%$ in a suitable species, or lower F with a convincing dose/exposure case	$F \geq 40\%$; predicted human profile supports once-daily dosing
Physicochemistry	Prefer molecular weight ≤ 650 and calculated lipophilicity (cLogP) approximately 1–4; solubility supports assays and dose	Molecular weight ≤ 550 ; fewer exposed hydrogen-bond donors and measured absorption-compatible polarity
Efficacy package	Exposure-related activity in a target-relevant skin model plus a human ex-vivo system	$\geq 70\%$ suppression of the prespecified model disease increment, with concordant pathway and histology effects
Safety margin	Target ≥ 10 -fold unbound AUC and C_{max} separation at the relevant NOAEL versus projected efficacious human exposure	≥ 30 -fold separation; no concerning organ or immune findings at covered exposures
Early safety and drug interactions	Appropriate genotoxicity negative; acceptable safety pharmacology, CYP and transporter profile	Initial hERG IC_{50} /free $C_{max} \geq 30$, plus integrated cardiac assessment; low drug–drug interaction potential
Development readiness	Characterized active material, scalable route, validated bioanalysis, species and formulation plan	Reproducible batches, stable formulation, backup chemistry and designs for studies supporting clinical entry ready

The eleven objectives guided compound selection. Direct binding and blockade of the receptor interaction would establish the proposed mechanism. Tests in primary human cells and whole blood would measure its functional effects. Comparisons with IL-12, IL-6, IFN α and JAK/TYK2 activity would assess whether the effects were specific to the intended pathway. Solubility, unbound drug exposure, oral pharmacokinetics, efficacy and safety would determine whether the activity could support a practical dosing regimen. Molecular weight and lipophilicity were design

preferences. Larger non-peptide compounds would be assessed using measured absorption and solubility, together with the feasibility of the required dose. ^[F02]

Binding affinity is expressed as the dissociation constant, K_d . Functional potency is expressed as the concentration producing 50% or 90% inhibition in a specified assay (IC_{50} or IC_{90}). PPI denotes the protein–protein interaction, and phosphorylated STAT3 (pSTAT3) measures downstream signaling in cells. Exposure criteria refer to pharmacokinetics (PK), which describes drug concentrations over time, and pharmacodynamics (PD), which describes the drug’s effects. AUC is the area under the concentration–time curve, C_{max} is the peak concentration, and NOAEL is the no-observed-adverse-effect level. Exposure coverage compares unbound drug concentrations with potency adjusted for the biological medium used in the assay. Safety margins compare exposure at the no-observed-adverse-effect level with projected efficacious human exposure. ^[F02]

The planned experimental sequence began by confirming compound identity and measuring direct binding with independent methods and controls for assay artifacts. Later tests would assess receptor blockade, selective effects in primary cells, potency in the presence of serum and activity at achievable drug concentrations. Selecting a candidate and backup compounds would also require a scalable synthesis, methods for measuring drug concentrations in biological samples, acceptable safety margins and a development plan. These goals defined the questions that experimental tests of the computationally selected compounds should answer. ^{[F02][F41][F44]}

1.4 Study scope

Here we present an AI-Lab: a team of AI agents working autonomously under human supervision. We describe its operation through a computational search for small molecules targeting the IL-23R–p19 interface. The agents proposed compounds, predicted binding positions by docking and tested those positions with molecular dynamics simulations. They also compared relative binding free energies within selected compound series and examined how the predicted positions related to the natural receptor–cytokine complex. A frontend agent coordinated communication, while independent agents chose their own next steps and tested their ideas. The human followed progress mainly through a linked HTML wiki and a molecule leaderboard. Reviews of molecular identity and proposed mechanisms guided the next selections. After an initial exploration phase, the AI-Lab compared a separate set of compounds prepared by the same procedure and refined alternative chemical series. ^{[F45][F13][F16][F18][F26][F30][F31]}

The campaign selected compounds for experimental testing, including candidates that remained in a predicted pocket during simulations and exploratory compounds targeting the receptor side. We describe how the AI-Lab organized its work, revised it in response to results and chose the final panel. We also explain what experiments on each candidate could reveal and how those tests relate to the therapeutic goals above. ^{[F34][F41][F44]}

2 Operation and progression of the AI-Lab

The AI-Lab carried out the campaign through independent agents working autonomously under human supervision. Agents proposed molecules, ran calculations, reviewed results and chose their next steps. The human supplied the overall goal and resource limits, with very little guidance during the work. Progress was followed mainly through a shared wiki and a leaderboard of candidate molecules and their supporting evidence. ^[F45]

2.1 Working instructions and independent agents

The instructions were maintained in an `AGENTS.md` file and changed somewhat during the study. Table 2 gives a representative reconstruction provided by the human supervisor. The

prompt defined the research goal, a stopping rule, shared memory, team responsibilities and resource limits. It asked the team to keep working until stopped by the human, who would review the leaderboard to select 20 molecules to make. This was the intended selection goal; the later purchase review used the five-compound budget described in [Section 5](#). ^{[F45][F44]}

Table 2. Representative AGENTS .md instructions. The instructions evolved slightly during the work. ^[F45]

GOAL: Engineer the best small molecules based on the target product profile (TPP) of a small-molecule inhibitor therapeutic for the IL23R-p19 interaction for psoriasis at the preclinical level.

TERMINATION: I will evaluate a leaderboard of the best molecules to make 20. Work autonomously towards the GOAL till I stop you.

MEMORY: Maintain a wiki-like responsive HTML structure with internal links summarizing everything known and learned (the campaign memory) and state.

TEAM of INDEPENDENT AGENTS:

- **FRONTEND AGENT:** The frontend agent (PlayMoleculeAI) creates agents and communicates with them and with me. The lead does not do the team's work.
- **INDEPENDENT:** Agents autonomously try to solve the GOAL, define their own next steps, and test their own ideas.

RESOURCES: The team shares a budget of a maximum of 4 GPUs and 10 concurrent agents.

The frontend agent, PlayMoleculeAI, created agents and handled communication within the team and with the human. Scientific work was carried out by the independent agents, which chose their own next steps and tested their own ideas against the shared goal. The activities described below emerged from that work. The human retained the decision to stop the campaign and review the candidates for synthesis. ^[F45]

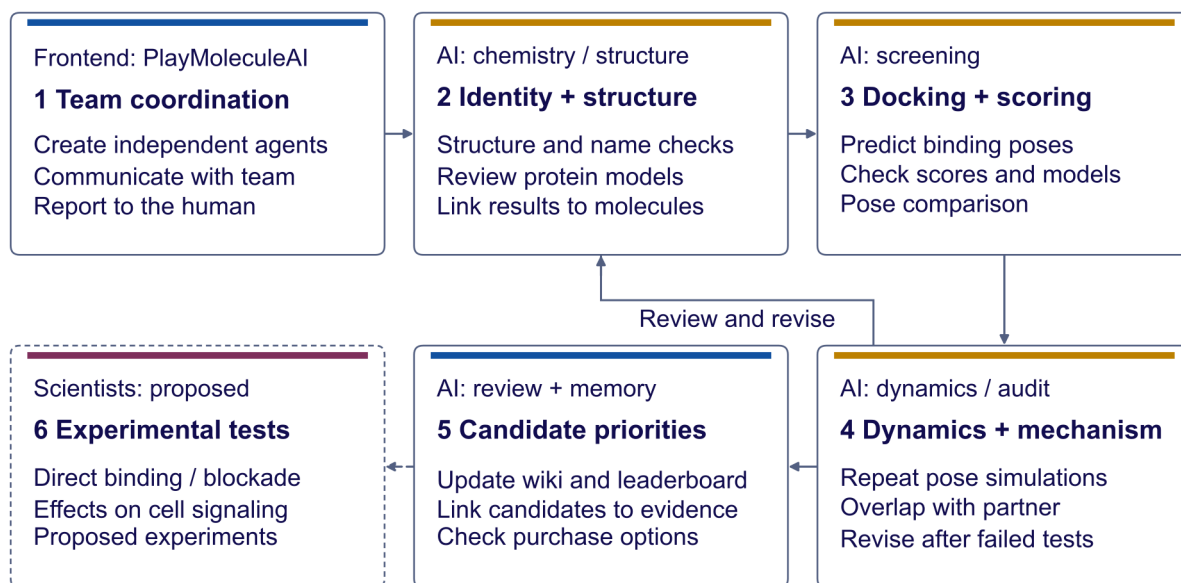
2.2 Shared memory and human supervision

The team maintained its campaign memory as a responsive HTML wiki with internal links. It summarized what was known, what had been learned and the current state of the work. This gave the human a common place to review progress across the agents' activities. Supervision consisted mainly of reading the wiki or asking the frontend to display leading molecules from the scoreboard, together with the evidence supporting them. These reviews focused on the emerging candidates and their evidence; agents chose the detailed computational steps with minimal further guidance. ^[F45]

2.3 Task coordination and scientific feedback

The independent agents pursued structural analysis, library design, protein and ligand preparation, docking, molecular dynamics and relative binding free-energy calculations. These categories describe the work performed across the team. Docking proposed positions and orientations of molecules in a binding site, called poses. Molecular dynamics then followed their movement over time. Each task produced molecular records, poses, simulation trajectories or comparison tables that other tasks could use. Identity checks linked each result to the intended molecule; structural checks assessed whether a proposed pose could interfere with the IL-23R-p19 interaction. [Figure 2](#) and [Table 3](#) summarize these roles and how they informed candidate selection. ^{[F03][F05][F13][F40]}

Results and errors both guided the next task. A failed database query prompted a different search procedure. Incorrect compound labels prompted a review and repair of the links between structures and results. Molecules that remained in a pocket during simulations were then checked



Outputs from separate studies

Phase 1: exploratory screening

2,429 library records → 2,426 curated records.
48,500 early pose-score rows across eight site tables.

Separate simulation groups

Top three candidates: 3 × 40 ns production runs each.
Wave 14: 16 completed production series of 100 ns each.

Phase 2: comparative screening

23 entries / 230 conformers; 4,600 poses across two protein models.
138 structures assessed with combined scoring methods.

Figure 2. AI-Lab workflow for proposing molecules, testing them computationally and selecting candidates for experiments. The frontend coordinated communication; independent agents chose their next steps under human supervision. Boxes summarize activities across the team. The wiki and molecule leaderboard supported human review. Phase 1 and Phase 2 studied separate sets of inputs; arrows show how tasks shared results and feedback.

for overlap with the space occupied by the natural binding partner. For a molecule placed on p19, that partner is IL-23R; for a molecule placed on IL-23R, it is p19. Reviews also revisited earlier results when problems with units, coordinates or reference structures were found. This feedback allowed the AI-Lab to revise both its candidate choices and its analysis procedures. [F04][F05][F07][F18][F22]

The review of completed work compared planned jobs, submitted inputs, simulation outputs and analysis results. When summaries disagreed with the underlying records, molecular input files, simulation tables and analysis code were used to resolve the discrepancy. This established which calculations had run and which results could inform candidate selection. [F22][F33][F40]

Table 3. AI-Lab tasks, outputs and uses in candidate selection. Counts are kept separate for each set of inputs and type of output.

Role / tools	AI-Lab work	Output	Use in selection
Library design and search	Propose and check structures; replace a failed database query with a search of a defined subset	Initial 2,429 records; 2,426 after checks	Establish the starting library. [F04]
Preparation and docking; AceDock/rDock	Prepare structures, dock molecules and collect scores	Eight early tables: 48,500 pose/score rows	Compare predicted poses across sites. [F05]
Identity checks; structure and stereochemistry	Compare input structures, output labels and simulated molecules	Repaired labels and separate records for stereoisomers	Link each result to the correct structure. [F05] [F14]
Dynamics; System-Builder/ACEMD	Build simulation systems, repeat runs and assess positions and contacts	Three main candidates: three 40 ns production runs each	Find poses that persisted across runs. [F13] [F42]
Relative binding free energies; QuantumBind-RBFE	Map corresponding atoms, compare molecule pairs and check agreement	Completed comparisons; rejected or conflicting results	Assess whether the comparisons were consistent. [F17] [F23]
Phase 2 preparation, docking and combined scoring	Use the same preparation procedure for all entries; dock to cytokine and receptor models	23 entries; 230 conformers; 4,600 raw poses; 138 structures scored by combined methods	Compare molecules and their prepared states at the two targets. [F26] [F27] [F28]
Mechanism and experiment planning	Compare poses with the natural binding partner; review structure identity and purchase records	Three primary and two exploratory priorities	Assess pocket retention, possible interference with the partner and readiness for purchase. [F18] [F41] [F43]

2.4 Campaign sequence

Phase 1 explored candidate structures and assessed their predicted poses. Later work expanded the search to related molecules and alternative binding arrangements. Phase 2 used a separate set of inputs prepared with a common procedure. [Table 4](#) summarizes the questions and decisions across these overlapping lines of work. Each set of inputs is counted separately. [\[F01\]](#)[\[F36\]](#)

2.5 Phase 1: building the library and checking compound identities

The initial structural survey examined the region where IL-23R contacts p19, the pocket opened by the reference peptide, a site farther from the interface, and regions that attracted small molecular probes. A search based on a VHH antibody binding site identified a receptor region that largely overlapped the natural p19 contact area. Docking at the more distant site generated 800 pose records from 16 entries, providing a comparison with the main interface sites. [\[F03\]](#)

The initial library combined 1,200 records from combinatorial design, 1,200 from commercial chemical collections and 29 starting compounds or related analogues. Three exclusions reduced the 2,429 records to 2,426. Some starting compounds were known for activity at other targets. When a database query failed, the search switched to local filtering and reproducible sampling of approximately 2% of the commercial collection. This search collected 5,108 entries before selecting the 1,200-record subset. Each library version was tracked separately, including a later version with 2,444 records. [\[F04\]](#)[\[F36\]](#)

Table 4. Campaign stages, completed work and the decisions it informed.

Stage	Work performed	Decision or revision
Phase 1	Survey binding sites, build libraries, dock molecules, check identities and examine early simulations	Correct links between structures and results; review analysis rules before extending comparisons. [F03][F04][F05][F07]
Expansion	Search for analogues, attach fragments, design variants, repeat simulations and compare energies and partner overlap	Identify persistent poses and test how they relate to the receptor interface. [F11][F13][F17][F18]
Later evaluation	Combine scores, rerun simulations after setup repairs and review affinity calculations	Revise score interpretations and update the record of completed simulations. [F21][F22][F23]
Phase 2	Prepare entries consistently, dock to both targets, run short simulations and test different starting arrangements	Reassess molecular states and poses; prepare a proposal for eight further simulations. [F26][F27][F31][F32][F33]
Source and purchase review	Check identities, units and evidence for each candidate	Select primary and exploratory candidates for experimental testing. [F01][F40][F41]

Two docking rounds generated 48,500 pose/score rows across eight site tables: four tables of 6,125 rows and four of 6,000. An incorrectly formatted first input record caused some early results to be assigned to the wrong compounds. A subsequent check compared structures and labels for 4,900 output-file/site combinations, representing 1,225 output-file records across four cavities. It found 4,876 matches under strict structural criteria and 4,900 under less restrictive criteria. Of 24 inspected simulation records, eight had incorrect compound assignments and sixteen matched their intended identities. Correcting the labels restored the links between structures and results used in later comparisons. Remaining stereochemistry and record-assignment issues are addressed in Section 6. [F05][F14]

Early molecular dynamics tested whether molecules remained near the proposed sites. A short receptor simulation, run in two 2 ns segments, was used afterward to compare movement thresholds across sites. Two longer simulations were each extended by 40 ns after repairing a restart setting that had initially added no simulation time. Their reported 100 ns comprised 10 ns of equilibration and 90 ns of production. Review of the analysis code then found a unit error: distances in nm were compared with cutoffs labeled as Å, making the intended 4/6 Å thresholds effectively 40/60 Å. The inclusion of hydrogen atoms and problems with the displacement reference and time axis also affected these two analyses. The review withdrew the earlier conclusion that these poses had been validated and called for corrected analysis. [F06][F07]

Checks of the scoring rules revealed further problems. In a five-compound docking comparison across sites, an overlap rule marked every case as selective and therefore could not distinguish them. A separate mismatch between an earlier score and the structure used to verify it remained after the label corrections. These findings prompted closer checks of how scores were calculated and linked to structures. Searches based on molecular shape and pocket fit then added more candidates: one pocket search returned 200 records, of which 49 passed filtering. Another docking calculation used 40 entries with ten conformers each and generated 2,000 poses. After correcting an indexing error between batches, the identities of all 2,000 poses were checked before the results were interpreted. [F08][F09]

CrypticScout simulations used mixtures of water and small molecular probes to identify favorable interaction regions around the prepared peptide-bound protein fragment. Of five proposed runs, the records showed two equilibration runs and one 40 ns production run. Two analyses of probe accumulation failed because the simulation-box dimensions were zero during periodic-boundary

processing. The available observations came from the benzene analysis and an alternative contact analysis using 80 saved frames. These results described the modeled pocket environment and identified a simulation-box processing problem for further review. ^[F10]

2.6 Expanding the search and testing possible binding mechanisms

During the expansion phase, the AI-Lab searched for related molecules, attached fragments to existing structures and designed small chemical changes. The search supplied 250 records and local design added 19, giving 269 docking inputs with checked atom connectivity. Attaching fragments at four positions produced 288 structures; 254 passed filtering and 150 were selected for docking. Preparation differed by route. Analogues were protonated for pH 7.4 and given up to ten conformers. Fragment variants kept the existing protonation treatment. The W5 series used ETKDG to generate one three-dimensional conformation from each molecular structure, followed by optimization with the MMFF force field. ^[F11]

Systematic combinations of chemical changes produced W5_003 and W5_004. Across the docking comparisons, the same chemical change did not always have the same effect. A later search for molecules with similar shapes checked 1,393 retrieved records and evaluated 30 candidates alongside five reference molecules. An earlier batch had been excluded because its results were linked to the wrong entries. For W10_LI_004, the best docking pose lacked a required anchoring contact, so the proposed free-energy calculation was stopped. This illustrates how a failed structural check changed the next computational step. ^{[F11][F12]}

Repeated simulations then tested whether molecules stayed in the pocket, while comparisons with the natural protein partner tested whether their positions could obstruct its binding. Four pocket candidates retained their poses across repeated runs under the rules used for each candidate; others failed the required contact or geometry checks. Section 3 explains where these outcomes depended on the reference structure or analysis rule. These results led to a combined test: a molecule had to keep its anchoring contact while also occupying space used by the natural binding partner. This tested whether a persistent pocket pose could also support a proposed mechanism of interference. ^{[F13][F14][F15][F18][F19]}

2.7 Later screening and review of simulation records

Later screening compared molecular shape, pocket fit and individual chemical groups, with control structures aligned in the same coordinate system. A corrected shape-score table contained 540 rows and replaced an invalid earlier table. Comparisons across target and control settings exposed calibration errors and incorrect score signs in the first round. In later rounds with valid calculations, no candidate passed all of the screening rules. Attaching fragments and constraining docking alignments improved some placement scores. However, tests requiring both contact retention and overlap with the partner failed their combined requirements. Later receptor-side tests, including W13R_014, also failed the specified contact checks even when overlap with p19 remained. These comparisons clarified which features of a proposed pose were supported by each test. ^{[F16][F19][F20][F21]}

A review of simulation records also resolved a mismatch between early setup failures and later completed runs. Initial receptor simulations with a coordinate-frame problem had been set aside, but an outdated summary still reported those failures. Checking the current production records recovered sixteen 100 ns series across eight cases, with two runs per case; twelve series involved receptor systems. Two further runs were canceled near 5.5 and 5.6 ns. The review updated the list of completed simulations. Interpreting the molecular motion in those trajectories remained a separate task. ^[F22]

2.8 Phase 2: consistent preparation and short simulations

Before further design, Phase 2 checked the structures associated with reported orders, duplicate entries and the origins of proposed molecules. It considered six designs and a diverse set of 16 retrieved entries. A search of 88,887 distinct molecular structures found no match for the six designs, either with stereochemistry included or when only atom connectivity was compared. Earlier compound names were linked to their existing molecular records. ^[F25]

Preparing 23 entries with a common pH 7.4 procedure produced 230 valid unbound conformers, ten per entry. Checks covered atom connectivity, names and stereochemistry after charge normalization. Unconstrained docking to models of p19+p40 and IL-23R generated 2,300 poses for each target. Combined scoring methods selected 138 structures, with checks of molecular state, coordinates and input identity. A correction to the receptor model's connectivity records accounted for its own glycans. All 4,600 raw poses and 138 scored poses were then rechecked, with no change to the flags for severe atomic clashes. Further checks of contact arrangements and MMFF potential energies calculated in vacuum changed the provisional candidate priorities. ^{[F26][F27][F28][F29]}

Short simulations tested how the results depended on the starting pose. One run, scheduled for 1.5 ns, tested an alternative contact arrangement and showed local movement near the protein; a related run failed before computation. Eight runs testing geometry and eight runs testing repeat starts were completed with nominal settings of 1.5 ns at 300 K and restraints on protein heavy atoms. Issues with molecular state, charge and timing limit their interpretation to local behavior near the chosen starting poses (Section 6). ^{[F30][F31]}

A later nine-entry preparation produced 90 conformers and 900 unconstrained docking poses. A CPU test of simulation setup passed checks of atom connectivity and protonation for the specified molecular state. A proposed batch of eight simulations, each scheduled for 5 ns, was prepared but still awaiting submission at the last input review. The reviews of source records and purchase options then brought the completed work together into the experimental priorities discussed below. ^{[F32][F33][F41]}

3 Results

The campaign identified three primary candidates for acquisition and testing: W5_004, PS2B_013 and W5_003. They remained in their proposed pocket positions during repeated simulations under the specified assessment rules. Two further candidates, C_0918 and CONT_CAT_R013, were proposed for exploratory tests at the receptor surface. Selection considered the identity of each compound, its position during simulation and whether it occupied space normally used by the other protein in the IL-23–receptor complex. ^{[F41][F42]}

3.1 Repeated simulations identified compounds that remained in the pocket

Repeated molecular dynamics simulations tested whether compounds remained near their proposed binding positions and maintained specified contacts with the protein. W5_004, PS2B_013 and W5_003 each completed three separate 40 ns production runs. Representative runs used 10 ns equilibration followed by unrestrained production at 310 K. Table 5 summarizes the simulations and how their results informed candidate selection. ^{[F13][F42]}

W5_004 showed the most consistent positioning across repeated simulations of the reviewed finalists, making it the first practical priority. PS2B_013 also remained in the pocket and maintained contacts across runs, making it a close alternative. W5_003 maintained a different arrangement of contacts involving a carbonyl group, providing a candidate for testing another position within the pocket. It passed the platform's specified assessment, although one run passed or failed

Table 5. Simulations used to select candidates for testing. Durations refer to production runs or short tests of the starting pose. Results were assessed using the rules and reference structures specified for each group of calculations.

Compound label	Simulations	Outcome	Role in selection
W5_004	3 × 40 ns	Maintained specified contacts and showed consistent positioning across runs	First priority among the reviewed finalists. ^[F42]
PS2B_013	3 × 40 ns	Remained in the pocket and maintained contacts; third run documented separately	Close alternative and comparison within the chemical series. ^{[F13][F42]}
W5_003	3 × 40 ns	Maintained a different contact arrangement; one displacement test depended on the reference used	Candidate for testing a different pocket position. ^{[F13][F42]}
PS_010	3 × 40 ns	Result depended on the assessment rule; uncertainty in how structures were matched remained	Reserve candidate from the pocket series. ^{[F13][F42]}
PS3B_006; PS2B_018	Each 3 × 40 ns	Failed the final pose assessment under both sets of rules	Previously selected compounds with less favorable pose results. ^{[F34][F35]}
C_0918	2 × 40 ns	Met the revised surface-association rule but did not maintain the proposed p19-blocking position	Exploratory receptor-surface candidate. ^{[F15][F42]}
CONT_CAT_R013	Short tests with repeated starts	Behavior depended on the starting pose under protein restraints	Optional candidate with a different chemical scaffold. ^{[F31][F42]}

a movement test depending on the reference structure used. The assessment of PS_010 also depended on whether the old or revised rule was applied, and it was retained as a reserve. These differences determined the order of testing and which alternatives were included in the panel. ^{[F13][F41][F42]}

The broader simulation program included a later group, Wave 14, with 16 production series completed to 100 ns each. These totaled 1.6 μ s across eight candidate systems. A review of the execution records separated these completed series from archived outputs with coordinate-frame errors and canceled partial runs. Separately, PS3B_006 and PS2B_018, which had previously been selected for experimental testing, each had three 40 ns runs. Both compounds failed the final pose assessment under both sets of evaluation rules. Their simulation results contrasted with those of the candidates that remained in the pocket. ^{[F22][F34][F35]}

3.2 Free-energy comparisons and overlap with the partner protein informed selection

Relative binding free-energy calculations compared selected pairs of compounds. The review checked agreement between repeated calculations, consistency of the modeled binding positions and whether valid calculations connected different groups of compounds. Repeated calculations sometimes disagreed, compounds occupied different positions, and some calculations intended to connect groups failed. These problems limited which compounds could be compared on the same relative energy scale. Candidate selection therefore also drew on repeated simulations of each compound and included candidates that could test different binding mechanisms. Corrections

to compound labels, reference choices and units changed the interpretation of some earlier results. The campaign sequence and comparison procedures are described in [Sections 2 and 4](#). [\[F07\]\[F14\]\[F17\]\[F23\]](#)

Structural comparisons tested whether a compound occupied space normally used by the other protein in the IL-23–receptor complex. These comparisons used peptide-bound and receptor-bound structures as references [\[3–6\]](#). A compound could sit in the induced p19 pocket without extending into the receptor’s position. For C_0918, the revised trajectory analysis supported association with the receptor surface, but the compound did not maintain a position that would obstruct p19. Later simulations tested whether compounds could both maintain their specified protein contacts and occupy space used by the partner protein. The tested compounds failed to meet both requirements across the required runs. These results motivated paired experimental measurements of compound binding and IL-23–receptor complex formation, so that binding to the protein and interference with its partner could be assessed separately. [\[F15\]\[F18\]\[F19\]\[F20\]\[F42\]](#)

4 Computational methods

4.1 Structural models and ligand preparation

Two possible binding sites were investigated: the induced p19 pocket in [8UUI](#), and the receptor surface in [6WDQ](#). Additional protein-complex structures, [5MZV](#) and [8OE4](#), were used to test whether a proposed compound position occupied space normally used by the partner protein [\[3–6\]](#). Later cytokine models included p40 and repairs to missing structural regions. The models represented selected protein fragments and their local binding environments. [\[F03\]\[F18\]\[F27\]](#)

RDKit was used to check molecular connectivity, stereochemistry, calculated properties and conformers, which are different three-dimensional arrangements of the same molecule. MoleculeKit handled coordinates and analysis of the molecular systems, and ProteinPrepare prepared the proteins. Earlier docking preparations used pH 7.2; later ligand preparations and representative simulation setups used pH 7.4. AcePka predicted protonation states. Conformer preparation varied between groups of calculations: analogues used up to ten conformers, while some local designs used one conformer generated by distance geometry and optimized with molecular mechanics. Each input was checked separately for molecular connectivity, encoded stereochemistry and agreement between the molecular definition and its coordinates. [\[F05\]\[F11\]\[F14\]\[F26\]](#)

4.2 Compound searches, docking and scoring

AceSearch retrieved similar compounds. AceGraft generated candidates by adding molecular fragments to existing structures, while ClipSearch and AceShape searched for alternatives with shapes suited to the modeled pocket. AcePockets mapped regions where compounds might interact with the protein. AceDock/rDock generated possible binding positions, or poses. Verified early docking screens used a 6 Å search radius, five output poses and four CPUs per job. A reference structure defined the search region, while the compound was free to change its orientation. Separate calculations constrained parts of the compound to a reference position to test the effect of that assumption. These protocols differed from the later Phase 2 comparison, which used a shared ligand-preparation procedure and two protein models. [\[F09\]\[F11\]\[F12\]](#)

In Phase 2, the same preparation procedure was applied to a 23-entry batch before docking into cytokine and receptor models ([Table 6](#)). The table distinguishes input entries, conformers and docking poses. Selected poses were scored with AcePockets and AceMoiety, with additional scores from rDock. Repeating docking with identical inputs tested score variability. Comparisons with control models tested how scores changed with the protein model. [\[F16\]\[F26\]\[F27\]\[F28\]](#)

Stage	Inputs and tools	Completed work
Ligand preparation	AcePka and RDKit checks	23 entries; 230 valid unbound conformers.
Docking into two models	AceDock/rDock; cytokine and receptor models	4,600 raw poses: 2,300 per model.
Scoring selected poses	Combined AcePockets/AceMoiety assessment	138 selected structures: 69 per model.
Later molecular dynamics	ACEMD; eight candidate systems	16 completed 100 ns production series; 1.6 μ s for this group.

Table 6. Completed calculations in the Phase 2 comparison and the separate Wave 14 simulation group. Counts refer to entries, conformers, poses or production series as specified in each row.

4.3 Molecular dynamics and overlap with the partner protein

Representative systems were prepared with SystemBuilder, AMBER ff14SB protein parameters, GAFF2/AM1-BCC ligand parameters, TIP3P water and 0.15 M salt. Repeated simulations of proposed binding poses used ACEMD at 310 K, with 10 ns equilibration followed by 40 ns production. Positional restraints were gradually removed during equilibration. Production was unrestrained and used an effective 4 fs time step in the verified protocol. The later group in Table 6 used 100 ns production runs. These completed simulations were counted separately from archived runs with coordinate-frame errors and canceled partial runs. [F13][F22]

To assess each pose, the protein structures were first aligned so that compound movement could be measured relative to the protein. Analyses then measured compound displacement, tracked contacts between specified atoms and compared results across runs. Overlap with the partner protein was assessed by aligning the relevant protein-complex structures and checking whether the compound occupied any of the partner's space. This overlap test and contact tracking during simulation together assessed whether a compound could stay attached in a position that might obstruct the partner. The analysis also checked which molecule each atom belonged to, whether coordinates shared the same reference frame and how the simulation's periodic boundaries affected distances. Contact frequencies reported the fraction of saved frames containing each specified contact. [F13][F14][F18][F19]

Phase 2 also included shorter tests of the starting poses. Eight geometry tests and eight tests with repeated starts used a nominal duration of 1.5 ns at 300 K, with restraints on protein heavy atoms. A separate short calculation refined an alternative arrangement of ligand–protein contacts. These calculations tested how local movement depended on the starting arrangement while protein motion was restrained. [F30][F31]

4.4 Relative binding free energies and computational scope

QuantumBind-RBFE calculations compared the binding free energies of selected compound pairs. Preparation checked compound identity, stereochemistry, atom matching between the compounds, structural alignment and restraints. One verified configuration used 300 K, a 2 fs time step and a nominal 3.2 ns for each simulation window; settings varied in other calculations. Analysis assessed whether outputs were finite, repeated calculations agreed and assumptions about binding poses were consistent. Relative energies were interpreted using the reference assigned to each group of connected comparisons. Internal consistency was tested by adding energy differences around a closed loop of compound pairs: a consistent loop should sum to approximately zero. [F17][F23]

The representative instructions set a shared maximum of four GPUs and ten concurrent agents. Reported token consumption for the campaign was 7 billion tokens. ^[F45]

4.5 Sources of evidence

F01–F44 identify findings from the campaign records and subsequent reviews. F45 records the human supervisor’s retrospective account of the evolving prompt, team organization, shared memory, human supervision and total token consumption. The representative prompt in [Section 2](#) comes from that account. All findings are summarized in the finding register ([Appendix A](#)). Numbered literature citations identify the published studies used for biological and structural context.

5 Discussion

5.1 Selecting persistent poses and different mechanisms to test

Candidate selection considered two questions: did repeated simulations support the proposed binding position, and what could an experiment reveal about how the molecule might interfere with IL-23 signaling? The repeated simulations favored W5_004, PS2B_013 and W5_003. Their different contact patterns offered a way to compare related binding arrangements within the pocket. Candidates placed on the receptor offered a second way to investigate interference with the IL-23R–p19 interaction. ^{[F41][F42]}

The candidate review placed W5_004 first for purchase and testing, with PS2B_013 as a close alternative. This order reflected the consistency of the simulations, the practical requirements for obtaining the correct chemical structures and what each compound would add to the experiments. W5_003 provided a different arrangement of contacts involving a carbonyl group. C_0918 added a receptor-side binding proposal, while the optional CONT_CAT_R013 added a different core chemical structure for comparison on the receptor side. [Table 7](#) links each priority to its proposed experimental question. ^{[F41][F42]}

Table 7. Candidate priorities and proposed experiments. Priorities 1–3 form the primary set; entries 4–5 add exploratory tests.

Priority / compound	Computational basis	Experimental question
1. W5_004	Consistent pose and retention of the specified contacts across three runs.	Does the molecule bind in the proposed pocket and affect receptor binding?
2. PS2B_013	Pocket pose supported by repeated simulations; close alternative to the first candidate.	How does this related compound compare with the first priority and previously selected compounds?
3. W5_003	Repeated support for a different contact arrangement, followed by identity checks.	Does the different carbonyl-contact arrangement change binding or inhibition of the protein interaction?
4. C_0918	Association with the receptor surface under a revised analysis rule.	Can binding to the receptor reduce p19 binding?
5. CONT_CAT_R013	Different core structure placed on the receptor by docking and examined in short restrained simulations.	Can a different core structure provide another way to bind the receptor?

PS_010 was kept as a reserve that could replace the optional fifth candidate. It would add another pocket candidate in place of a different receptor-side structure. The review also considered the

previously reported orders of PS3B_006 and PS2B_018. If those two compounds count toward a five-compound budget, the three primary candidates would fill the remaining slots. The receptor-side candidates would then require substitutions or a larger test set. [F34][F41][F43][F44]

5.2 From computational predictions to experimental tests

The structural results suggest measuring binding and interference with the protein interaction separately. Direct binding measurements with intact IL-23 and IL-23R would identify which protein each compound binds. Tests using a different measurement method would then assess whether the compound prevents complex formation or displaces a bound partner. For pocket candidates, these experiments would connect occupation of the p19 pocket to its effect on receptor binding. For receptor-side candidates, they would test whether association with the receptor surface affects p19 binding. [F02][F18][F44]

The proposed experiments begin by checking the purchased material's identity, stereochemistry, purity and concentration, including how much remains dissolved under assay conditions. Independent binding measurements and controls for aggregation or interference with detection would help interpret the results. Cell experiments would then test the effects on IL-23 signaling and on related pathways. The target product profile in Table 1 sets the activity, selectivity and drug-exposure goals for this progression. [F02][F37][F44]

5.3 Implications for operating an AI-Lab

The campaign shows how a broad research goal and shared memory supported autonomous work by independent agents. The frontend handled communication, while the human followed the wiki and requested leading molecules and their evidence, giving very little guidance on individual calculations. Agents generated candidate molecules, checked which structures had been modeled and assessed their poses and possible effects on receptor binding. The sequence in Section 2 shows how this process expanded the search and revised earlier interpretations under human supervision. [F45][F04][F05][F11][F18]

Three practices were central to this process. First, every result needed a clear link to the molecule and prepared state used to produce it. Incorrect docking labels and later stereochemistry checks showed why those links mattered. Second, the analysis rules needed review alongside the calculations. The contact-distance unit error and a selectivity rule that passed every compound both changed which conclusions the results supported. Third, progress needed to be checked against input files and actual simulation outputs. This distinguished completed simulations from outdated failure reports and from a later batch that had been prepared but not submitted. These practices made the computational work traceable and helped define the next experiments. [F05][F07][F08][F14][F22][F33]

6 Limitations

This study used computation to propose and prioritize compounds. The reviewed records contain no qualifying measurements of direct binding, inhibition of the IL-23R-p19 interaction or cellular activity for the prioritized compounds. Purchase recommendations and reported orders also preceded verification of the physical samples. Supplier confirmation of availability and prices for the finished compounds remains outstanding. The selected compounds therefore represent proposals for experimental testing, ordered by computational support and the questions they could answer rather than measured potency. [F34][F37][F41][F43][F44]

The simulations depend on the structures and starting poses used. The peptide-bound and receptor-bound templates show different protein conformations and contain only parts of the

full proteins. The prepared models also omit parts of the complete glycosylated systems. The calculations do not establish how often the induced p19 pocket opens or the energy needed to open it. Conventional simulations follow specified starting poses for a limited time. Their contact frequencies count contacts in saved frames, which are correlated in time; they do not give equilibrium binding probabilities or residence times. Overlap with the natural binding partner tests whether a proposed pose could physically obstruct it. Effects transmitted through changes elsewhere in the protein, known as allosteric effects, require separate evidence. [\[F03\]\[F13\]\[F18\]\[F27\]](#)

The short Phase 2 simulations have further limits: protein motion was restrained, some runs used shared or related starting poses, and questions remained about molecular states and simulation timing. These runs describe local behavior near the starting arrangement. In the longer simulations, some pass/fail outcomes depended on the reference structure or analysis rule, as seen for W5_003 and PS_010. Unresolved issues in earlier work include trajectory processing, results still assigned to the wrong isomer and an uncertain contact-atom assignment for one selected compound. These issues limit comparisons between candidates; [Sections 2 and 3](#) describes the main corrections and the outcomes affected by analysis choices. [\[F07\]\[F14\]\[F30\]\[F31\]\[F35\]\[F42\]](#)

Docking and free-energy results also depend on the comparison being made. Repeated docking produced variable scores. Relative binding free-energy comparisons were affected by disagreement between runs, differences in sampled binding positions, choices of reference and failed calculations linking compound pairs. The resulting comparisons apply to the modeled systems under those assumptions. They do not provide experimentally calibrated absolute affinities or a ranking across different chemical series. [\[F16\]\[F17\]\[F23\]](#)

Finally, this paper describes one campaign through a review of its recorded work. It lacked a controlled comparison with a manual workflow and complete records of GPU and CPU hours. The study therefore describes the AI-Lab through the tasks it completed and the decisions those tasks informed. Measuring its effects on time, cost and discovery success will require prospective comparisons. [\[F01\]\[F40\]](#)

7 Conclusions

The AI-Lab combined searches for candidate molecules, structure preparation, docking, repeated molecular dynamics and relative binding free-energy calculations at the IL-23R–p19 interface. Checks of compound identity and comparisons with the natural binding partner helped select a focused set of candidates for experiments. Three candidates—W5_004, PS2B_013 and W5_003—retained their proposed pocket positions in repeated simulations under the criteria used for each candidate. Two exploratory receptor-side candidates added different possible binding mechanisms to test. [\[F05\]\[F13\]\[F18\]\[F41\]\[F42\]](#)

The main outcome is a set of compounds and proposed binding mechanisms for experimental evaluation, with W5_004 as the first priority. Binding and protein-interaction measurements on verified samples would guide the next stage of development. The campaign also shows how an AI-Lab can connect molecular design to computational tests, revise interpretations after checking identities and analysis methods, and select informative experiments through autonomous teamwork under human supervision. [\[F41\]\[F43\]\[F44\]](#)

References

- [1] MA Lowes, M Suárez-Fariñas, and JG Krueger. Immunology of psoriasis. *Annu Rev Immunol*, 32:227–255, 2014. URL <https://doi.org/10.1146/annurev-immunol-032713-120225>.
- [2] C Parham, M Chirica, J Timans, E Vaisberg, M Travis, J Cheung, S Pflanz, R Zhang, KP Singh, F Vega, W To, J Wagner, AM O’Farrell, T McClanahan, S Zurawski, C Hannum, D Gorman, DM Rennick, RA Kastelein, R de Waal Malefyt, and KW Moore. A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R. *J Immunol*, 168:5699–5708, 2002. URL <https://doi.org/10.4049/jimmunol.168.11.5699>.
- [3] C.R. Glassman, Y.K. Mathiharan, K.M. Jude, L. Su, O. Panova, P.J. Lupardus, J.B. Spangler, L.K. Ely, C. Thomas, G. Skiniotis, and K.C. Garcia. Structural basis for IL-12 and IL-23 receptor sharing reveals a gateway for shaping actions on T versus NK cells. *Cell*, 184:983–999.e24, 2021. URL <https://doi.org/10.1016/j.cell.2021.01.018>.
- [4] Y. Bloch, L. Bouchareychas, R. Merceron, K. Skladanowska, L. Van den Bossche, S. Detry, S. Govindarajan, D. Elewaut, F. Haerynck, M. Dullaers, I.E. Adamopoulos, and S.N. Savvides. Structural Activation of Pro-inflammatory Human Cytokine IL-23 by Cognate IL-23 Receptor Enables Recruitment of the Shared Receptor IL-12R beta 1. *Immunity*, 48:45–58.e6, 2018. URL <https://doi.org/10.1016/j.immuni.2017.12.008>.
- [5] Y. Bloch, J. Felix, R. Merceron, M. Provost, R.A. Symakani, R. De Backer, E. Lambert, A.R. Mehdipour, and S.N. Savvides. Structures of complete extracellular receptor assemblies mediated by IL-12 and IL-23. *Nat Struct Mol Biol*, 31:591–597, 2024. URL <https://doi.org/10.1038/s41594-023-01190-6>.
- [6] F.C. Lecomte, J.S. Joseph, J. Stalewski, Q. Shen, E. Arnoult, V. Sridhar, M. Liu, Y. Hu, J.G. Gasendo, H. Ben Arie, N. Keinan, L. Keidar, I. Aviv, E. Ruvinov, J. Grandjean, P.R. Does-Silva, A. Mak, B. Santoso, S. Kim, V. Shende, W.J. Wever, T. Mirzadegan, Z. Zhu, B. Fuchs, P. Pinton, and R. Szabady. Identification of an Induced Orthosteric Pocket in IL-23: A New Avenue for Non-biological Therapeutic Targeting. *ACS Chem Biol*, 20:1609–1618, 2025. URL <https://doi.org/10.1021/acscchembio.5c00181>.

Appendix A Finding register

ID	Finding and limits of evidence
F01	Project coverage and different date records. A broad inventory and targeted reviews covered recovered project records; not every file or event was scientifically validated. <i>Limits:</i> The inventory was collected over time. File dates, event labels and reporting dates describe different events.
F02	Preclinical product goals. The profile sets eleven minimum and preferred goals, rather than reporting achieved potency, exposure, safety or candidate selection. <i>Limits:</i> Earlier low-pM, fixed-half-life and universal 1,000-fold criteria were retired.
F03	Binding strategies and alternative sites. The study explored receptor sites, the peptide-induced cytokine pocket, more distant sites and regions favorable to small probes. Public structures do not show binding by campaign compounds. <i>Limits:</i> Incomplete protein structures, coordinate alignment and the distinction between geometry and function limit interpretation.
F04	Initial library and recovered searches. The initial 2,429-record library retained 2,426 after curation. Retrieved commercial compounds and starting compounds from unrelated targets were not positive controls for this target. <i>Limits:</i> Counts depend on the version and search route; coverage of chemical space is incomplete.
F05	Early docking and compound assignments. Eight early tables contain 48,500 score rows. Checks of 4,900 output-file/site combinations and 24 MD records found incorrect compound assignments that required correction. <i>Limits:</i> Pose rows, file assignments, molecular structures and MD records count different things.
F06	Preliminary test and revised screening rule. A short receptor-system model informed a revised cytokine-side screening rule; it was not an experimentally known p19 binder. <i>Limits:</i> Behavior modeled at another site does not validate a binding classifier.
F07	Early extensions ran, but their analyses had defects. Two 40 ns extensions ran. Unit, atom-selection, reference, timing and RMSD errors undermine conclusions about ligand retention and site validation. <i>Limits:</i> This concerns two early extension analyses; no trajectories were reanalyzed, and later MD is not generally invalidated.
F08	Early checks did not establish functional selectivity. A five-entry cross-docking rule marked every entry as selective. Unresolved links between structures and metrics limited independent checks. <i>Limits:</i> A failed computational test is not evidence of biological selectivity or cross-reactivity.
F09	Early compound searches and free-energy preparation. Shape/pocket searches and searches for alternative core structures produced traceable sets, including 2,000 checked poses. Relative binding free-energy preparation preceded later calculations. <i>Limits:</i> Counts from overlapping sets cannot be added to give a single screening total; preparation is not simulation.
F10	Completed tests of a proposed hidden pocket. Records showed one 40 ns production run with partial/fallback analysis, versus five planned runs; two hotspot analyses failed. <i>Limits:</i> Probe contacts in a prepared protein-fragment pocket do not establish pocket opening in the unbound protein, allostery or inhibition.
F11	Analogue design, fragment grafting and further variants. There were 269 analogue/design docking inputs and a fragment-grafting route from 288 to 254 to 150 entries, followed by tests of local changes and combinations. <i>Limits:</i> Molecular connectivity and docking effects do not establish stereochemistry or experimental structure–activity relationships.
F12	Later shape search and a failed follow-up criterion. The 1,393 retrieval records yielded 30 candidates plus 5 references. Failure of the required anchoring contact prevented the proposed relative binding free-energy calculation. <i>Limits:</i> Earlier misindexed outputs were set aside; screening does not measure activity.

ID	Finding and limits of evidence
F13	Four compounds repeatedly stayed in predicted pockets. PS_010, PS2B_013, W5_003 and W5_004 each showed pose persistence across three runs under the rules for their respective routes. <i>Limits:</i> There was no common numerical threshold; these results do not establish binding, residence time, target occupancy or inhibition.
F14	Corrected stereochemical identity and remaining record errors. PS011M_rekey differs from the historical PS_011 record. A failed third-run contact and retained old-isomer metrics prevent pooling results. <i>Limits:</i> A corrected record is not a new, experimentally validated sample.
F15	Surface association did not establish blockade. C_0918 supported a revised rule for staying near the protein surface, but support for the original anchoring contact and lasting exclusion of p19 was absent. <i>Limits:</i> The rule was revised partly after seeing results. Receptor-bound poses must be compared with the space needed by p19.
F16	Target/control docking comparisons were unstable. Score differences changed and reversed sign across repeat calculations. Correcting the sign did not make affinity or selectivity interpretable. <i>Limits:</i> The increase from 5.18 to 7.36 concerns score changes relative to a reference, not absolute selectivity differences or energies.
F17	Relative binding free energies and consistency checks. Selected comparisons within compound series ran; invalid results were excluded. Agreement around comparison loops and assigned reference energies did not establish absolute affinity. <i>Limits:</i> Different poses, disagreement between runs, sign/reference conventions and incomplete calculation-engine records limit interpretation.
F18	Structural overlap is an indirect test. The induced pocket was compared with the natural binding interface. Limited overlap with the receptor in static structures did not demonstrate blockade. <i>Limits:</i> Jaccard similarity between sets of contacted residues does not measure blocked surface area; lack of overlap does not disprove allostery.
F19	Combined contact and partner-overlap test failed. W12B_001e1 did not meet both registered requirements in both runs: the runs differed in which requirement was lost. <i>Limits:</i> This does not prove an unavoidable chemical tradeoff or establish inactivity.
F20	Later receptor-side compounds failed their evaluation rules. W13R_014 and related compounds failed the combined criteria for contacts and polar-atom assignments, despite geometric overlap. <i>Limits:</i> Later results update the earlier pending status; no biochemical outcome follows.
F21	Wave 14 screening and score corrections. A corrected 540-row shape table and several analyses of scores and controls extended evaluation beyond docking rank. <i>Limits:</i> Table rows are not unique active compounds; the requested independent types of scoring were not always established.
F22	Current simulation records and outdated conclusions. Tables linked to source records distinguish 16 current 100 ns series, including 12 receptor series, from archived coordinate-frame errors and two canceled partial runs. <i>Limits:</i> Completion and analysis values alone do not establish binding, dissociation or independent runs.
F23	Affinity, inhibition and residence-time definitions. Later reviews corrected reference-state, thermal-factor and response-definition assumptions. Hypothetical affinity, residence-time and dose estimates remain uncalibrated. <i>Limits:</i> Response-relative IC ₉₀ , absolute 90% inhibition and adequate clinical exposure are not interchangeable.
F24	Monitoring updates and modeled exposure scenarios. Later monitoring repeatedly recorded zero “edge detections” and 97 cumulative “gates.” Solubility and exposure sets described modeled scenarios. <i>Limits:</i> Repeated state records are not new gates, measured solubility, compound pharmacokinetics or computing time consumed.
F25	Phase 2 designs, searches and identity checks. New designs and a diverse 16-entry set were assembled. A limited scan of 88,887 molecular graphs did not establish patent novelty. <i>Limits:</i> CSV/SMILES searches omit some SDF-only records, tautomers and external identities; revisited compounds are not new molecules.

ID Finding and limits of evidence

- F26 **Compounds prepared by a common procedure.** A 23-entry batch yielded 230 valid unbound conformations, with checks of molecular structure, chemical state, names and stereochemistry. *Limits:* Entries/conformations are not distinct compound connectivities or binding poses.
- F27 **Docking at both sites completed.** Docking without fixed ligand positions in natural-context cytokine and receptor models yielded 4,600 raw poses, with chemistry, state and coordinate checks. Structural-context flags were recorded separately. *Limits:* Targets were partial, repaired models; chemical checks do not establish physical binding or inhibition.
- F28 **Combined scoring and corrected structural context.** The 138 scored structures passed checks linking output states and coordinate frames to source records. A recorded correction to the natural-complex context did not change severe-clash flags. *Limits:* Search-region overlap and no modeled glycan clashes do not independently prove blockade or capture full physiology.
- F29 **Additional model checks changed provisional priorities.** Checks of binding positions, contacts and vacuum potential energy revised a three-probe panel, without recommending synthesis/purchase or claiming activity. *Limits:* Vacuum potential energy is not binding strain or free energy; it cannot be added to docking scores or converted to affinity.
- F30 **One short refinement of an alternative binding position ran.** One nominal 1.5 ns run showed local ligand rearrangement near the protein; its paired run failed before computation. *Limits:* Staying near the protein, with incomplete model and timing records, does not validate the pose, affinity, residence time or blockade.
- F31 **Short tests of geometry and repeated starting structures.** Eight geometry and eight repeated-start jobs completed at a nominal 1.5 ns duration with restrained protein. Model-state and timing records had limitations. *Limits:* These were not long unrestrained MD runs or established statistically independent repeats.
- F32 **A later nine-entry preparation and docking set completed.** Fresh preparation produced 90 conformations and 900 poses from docking without fixed ligand positions, with separate quality checks. *Limits:* This later docking work is not a completed long simulation or experiment.
- F33 **Prepared inputs had not yet run.** A CPU build with strict state requirements passed topology and protonation checks. Eight requested 5 ns analogue runs remained prepared but unsubmitted. *Limits:* Passing input checks does not establish execution, force-model or timing details, binding, or 40 ns of completed work.
- F34 **Experimental selections and reported orders.** PS3B_006 and PS2B_018 were selected for validation. Later documents report orders for these IDs, but do not verify receipt, material identity or order dates. *Limits:* Reviewed confirmation and shortlist records establish neither the selection date nor the comparative rationale; this historical selection is not a new ranking.
- F35 **The selected pair had unfavorable simulation results.** Both compounds have their own three-run 40 ns records and unfavorable final model conclusions. The review found only one historical relative comparison. *Limits:* Failing a model rule does not establish inactivity. Unresolved oxygen-atom assignments and an unusable comparison between series prevent borrowing results or combining scales.
- F36 **Compound identities and changing list versions.** Labels, molecular graphs, connectivity and stereochemical flags differ between versions; a “top 20” list need not contain 20 unique named selections. *Limits:* Counts from one version do not describe the full screening sequence or an experimental ranking.
- F37 **Material and assay records.** Three assay registries contain zero data rows. All 23 material entries remain on hold: two have reported orders and 21 have no verified physical material. *Limits:* No qualifying campaign measurement was found in the reviewed records; this does not establish its absence elsewhere.
- F38 **Predicted drug properties and a failed ADMET tool.** Offline molecular descriptors and ESOL estimates are model outputs. The ADMET application failed before processing, leaving no endpoint predictions or measured product-goal compliance. *Limits:* A container startup failure does not establish faulty input chemistry or unfavorable pharmacology.

ID	Finding and limits of evidence
F39	Published experimental results provide external context. Published weak-fragment binding results do not validate campaign compounds, natural-receptor blockade or the identity of new material. <i>Limits:</i> Uncertainties in information transcribed from primary and secondary sources, and in material stereochemistry, remain; no unverified individual affinity is reported.
F40	Configured runs and incomplete resource records. The 3,129 configuration copies contain 3,035 run labels across 29 application/version groups. Failed prediction attempts and status markers do not establish productive jobs or spending. <i>Limits:</i> Total CPU/GPU hours, peak use, resource-limit compliance and human-intervention time remain incompletely documented.
F41	Five compounds prioritized for experimental testing. The reported order is W5_004, PS2B_013, W5_003, C_0918 and CONT_CAT_R013. The first three test pocket-binding hypotheses; the last two are less certain probes of mechanism or chemical scaffold. <i>Limits:</i> This is a new, conditional purchase/testing priority, not the historical selection, an affinity ranking, purchase authorization or candidate nomination. The first two are nearly tied in practice; PS_010 is the conservative reserve for the optional fifth slot.
F42	Compound-specific model support and uncertain mechanisms. The first three priorities have repeated evidence of remaining in a pocket, including an alternative binding position for W5_003. C_0918 and CONT_CAT_R013 offer exploratory receptor-side hypotheses with weaker model support. <i>Limits:</i> Staying in a pocket does not establish blockade of natural receptor recruitment. Relative binding free energies establish neither a W5 affinity advantage nor a common calibrated scale. Short restrained tests that depend on starting structures cannot be compared with longer production runs. Selectivity and experimental activity remain unknown.
F43	Purchase evidence and compound-identity corrections. No finalist has verified current stock of the complete compound, a supplier-confirmed quote or authenticated material in the reviewed records. Building-block listings and proposed routes support only conditional quote requests for the exact compound. <i>Limits:</i> Related products must not be substituted. Positional isomers, cage orientation and linker assignments require exact-identity checks; historical route descriptions need correction. Catalog/database entries, component stock and plausible routes do not establish a complete product, finished synthesis, material quality or availability.
F44	Purchase budget and sequence of assay checks. Confirm reported PS3B_006 and PS2B_018 orders before repeat purchases. If they use two of five total slots, only the first three priorities are new purchases; exploratory probes are optional if more slots remain. <i>Limits:</i> A reported order does not verify receipt or material identity. Confirm identity, purity and usable solubility, then test direct binding and natural receptor recruitment/blockade with independent controls for artifacts. Follow with cellular function, viability and tests of other pathways. No assay, PK, safety or product-goal achievement is established.
F45	First-hand account of AI-Lab operation. The human supervisor described an evolving goal-based prompt, a coordinating frontend, independent agents, a linked HTML wiki and review of a molecule leaderboard with supporting evidence. The representative instructions set a maximum of four GPUs and ten concurrent agents, with human review to select 20 molecules to make. The supervisor reported providing very little guidance and total campaign consumption of 7 billion tokens. <i>Limits:</i> This retrospective account reconstructs the instructions and describes human supervision and token consumption; it does not measure total compute hours, exact intervention time or verified peak resource use.
