Autonomous AI Agent-based Virtual Lab for Predicting Novel Ligand-Receptor Interactions Between Brain-derived Plasma Proteins and Immune Receptors

Anonymous Author(s)
Affiliation
Address
email

Abstract

The accelerating global population aging has intensified the urgent need to decipher the biological mechanisms underlying inflammaging and age-related pathologies. While the neuroimmune axis—a bidirectional communication network between the central nervous system and the immune system—has emerged as a critical regulator of longevity, the molecular effectors governing this dialogue remain poorly characterized. Prior research has predominantly focused on canonical cytokine pathways; however, emerging evidence suggests that non-cytokine proteins may dominate age-dependent inter-organ communication. To address this knowledge gap, we introduce a novel multi-agent AI framework, termed the Virtual Lab, designed to systematically identify uncharacterized interactions between brain-derived plasma proteins and immune receptors. Unlike conventional deep learning approaches hindered by dataset biases (e.g., overrepresentation of cytokines) and susceptibility to hallucination, our framework leverages a collaborative, cross-validated pipeline: specialized AI agents—including an immunologist, machine learning specialist, computational scientist, and scientific critic—iteratively refine search strategies through structured debate and rigorous peer review. We curated a high-confidence library of 528 cell-surface immune receptors and paired them with 112 brain-derived plasma proteins exhibiting stringent tissue specificity. Using ESM-2 embeddings for rapid sequence-based screening, we identified high-potential candidates, which were subsequently validated for structural compatibility via AlphaFold-Multimer. Key validation metrics included interface confidence scores (ipTM + pTM ≥ 0.6), predicted alignment error (PAE) plots confirming reliable domain positioning, and per-residue confidence scores (pLDDT > 50), alongside physical interaction interfaces visualized via PyMOL. Crucially, our approach overcomes the limitations of existing tools—whose performance often degrades when applied to non-canonical interactions—by integrating domain-adapted negative controls and mitigating evolutionary biases. We report a prioritized set of novel receptor-ligand pairs, several involving age-upregulated brain proteins, offering unprecedented opportunities to dissect the molecular dialogue between the aging brain and the immune system. These findings establish a new paradigm for the unbiased discovery of organ crosstalk effectors, with significant implications for targeting inflammaging and promoting healthspan extension.

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

1 Introduction

The rapid acceleration of global population aging has become a critical public health challenge. This demographic shift is a primary driver for the increasing prevalence of age-related pathologies, placing a significant burden on healthcare systems. Consequently, identifying the biological mechanisms underlying aging and associated diseases has become a priority in biomedical research. Aging is accompanied by a chronic inflammatory state, often termed "inflammaging," which contributes to the onset of degenerative diseases. Within this context, the interaction between the central nervous system (CNS) and the immune system appears to be critical. Recent proteomic analyses have demonstrated that the brain and the immune system exhibit the highest inter-organ correlation during aging. Furthermore, this study indicated that maintaining a youthful biological state specifically in the neuroimmune axis is associated with reduced mortality and extended longevity, unlike in other organs. This suggests that the interplay between the brain and immune system is a key determinant of organismal survival. Despite the importance of the neuroimmune axis, the specific molecular mechanisms mediating these interactions remain largely uncharacterized. Historically, research has focused on canonical cytokine-receptor interactions. However, recent profiling of the plasma proteome reveals that cytokines comprise a minor fraction of brain-derived proteins, particularly among those showing age-dependent changes. This indicates that non-cytokine proteins likely mediate a significant portion of brain-immune communication. Therefore, identifying these novel ligand-receptor interactions is essential for a comprehensive understanding of neuroimmune regulation. A major challenge in identifying novel protein-protein interactions (PPIs) is determining binding affinity between uncharacterized pairs. Traditional experimental screening is labor-intensive and limited in scale. However, the use of deep learning-based structural biology tools, such as AlphaFold and ESMFold, provides a solution. These models allow for the prediction of protein structures and binding probabilities in silico. This enables high-throughput screening of large datasets to identify potential interactions based on structural compatibility rather than existing literature. To implement these computational tools effectively, we utilized a multi-agent AI system, referred to as a "Virtual Lab." Unlike static models, this framework dynamically generates agents with specialized roles tailored to the specific research inquiry. By facilitating autonomous discussion and critical debate among these agents, the system refines experimental strategies and enhances the analytical depth of the proposed methods. Crucially, this collaborative cross-verification process significantly mitigates the risk of hallucination inherent in large language models, ensuring a higher level of methodological reliability compared to singular model utilization. In this study, we applied this Virtual Lab framework to systematically identify previously uncharacterized interactions between brain-derived plasma proteins and immune cell receptors. We report a set of candidate pairs validated by structural modeling, providing potential targets for investigating the synchronization between the aging brain and the immune system.

2 Virtual lab

2.1 Design of the Autonomous Virtual Lab Framework

To overcome the limitations of singular large language models (LLMs) in complex scientific reasoning, we established an autonomous "Virtual Lab"—a hierarchical multi-agent system designed to mimic the collaborative structure of a human research institute. Unlike conventional prompting approaches, this framework operates through distinct, specialized roles: a human supervisor who defines the overarching research scope; an AI Principal Investigator (PI) who orchestrates the workflow and synthesizes decision-making; three AI Scientist Agents who act as domain experts (e.g., Immunologist, Machine Learning Specialist); and an AI Scientific Critic who enforces rigorous peer review. To mitigate the inherent stochasticity of LLMs and ensure the robustness of the proposed methodologies, the system executes five parallel discussion threads for each agenda. The AI PI then integrates these divergent outputs into a single, optimized research strategy, effectively filtering out hallucinations and logical inconsistencies through collective intelligence.

2.2 Team Composition and Domain Adaptation

The research initiated with a high-level directive from the human supervisor: "Predict novel ligand-receptor interactions between brain-derived plasma proteins and immune cell receptors based on structural compatibility and binding affinity." Upon receiving this agenda, the AI PI autonomously

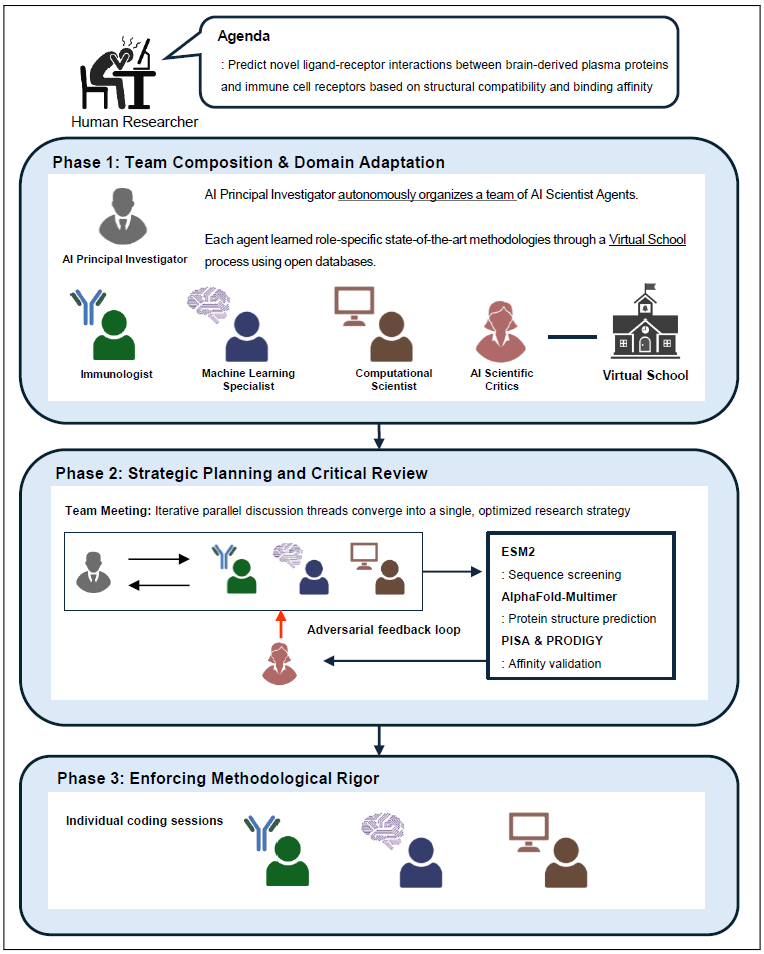

Figure 1: Virtual Lab Pipeline.

orchestrated the team selection process, determining that a multidisciplinary approach was essential. Consequently, the PI appointed three specialized agents: a machine learning specialist, a computational scientist, and an immunologist. To ensure domain proficiency, each agent underwent a "Virtual School" process—a retrieval-augmented learning phase where agents ingested relevant literature from Pubmed to internalize higher level of methodologies and contexts specific to their assigned roles.

2.3 Strategic Planning and Critical Review

Following team assembly, the Virtual Lab proceeded to the Project Specification and Tool Selection phases. The agents engaged in iterative team meetings to address key operational challenges, such as efficient receptor candidate reduction and computational resource management. Through this dialectic process, the agents formulated a "computational funnel strategy." The Immunologist proposed constructing an initial gene list using databases like ImmPort and UniProt, explicitly excluding intracellular proteins. To address the computational burden of high-throughput screening, the Machine Learning Specialist and Computational Scientist devised a cascaded filtration pipeline: utilizing pre-trained ESM2 embeddings for rapid sequence-based screening, followed by AlphaFold-Multimer for high-fidelity structural prediction, and finally validating affinities via PISA and PRODIGY.

2.4 Enforcing Methodological Rigor

A pivotal component of this framework was the role of the AI Scientific Critic. During the strategy formulation, the Critic challenged qualitative proposals, demanding quantitative justifications for all decision nodes. Specifically, the Critic required the definition of explicit thresholds for binding

energies, structural accuracy metrics (e.g., RMSD, TM-score), and estimates of computational costs per prediction. This adversarial feedback loop ensured that the final workflow was not merely a collection of tools, but a coherent, scientifically valid pipeline with clear exclusion criteria. Following this rigorous vetting, the agents transitioned to the Implementation phase, where they conducted individual coding sessions to generate the executable scripts required to operationalize the derived funnel strategy.

2.5 Advantages of Modular Coding

Following the strategic consensus, the Virtual Lab transitioned to the implementation phase, governed by a structured series of Individual Meetings. Unlike the collective brainstorming of team sessions, this phase employed a targeted one-on-one protocol between the human supervisor and each specialized agent. This segregation of duties was critical for maintaining high context fidelity within the large language model, ensuring that the generated code remained focused, modular, and free from cross-domain interference. By isolating the coding tasks, we achieved a level of granular control and debugging efficiency that is difficult to attain in monolithic code generation processes.

3 Preprocessing

To systematically identify immune targets capable of interacting with brain-derived factors, the AI Immunologist agent executed a hierarchical filtration protocol to curate a high-confidence receptor library. The agent first interrogated the UniProt database to retrieve proteins expressed in major immune cell types known to traffic to or interact with the CNS, including T-cells, B-cells, monocytes, dendritic cells, natural Killer cells, neutrophils, mast cells, and macrophages. To isolate biologically relevant targets for circulating ligands, the agent applied a subcellular localization filter, retaining only proteins anchored to the plasma membrane while strictly excluding intracellular or nuclear proteins. Subsequently, the agent utilized the CellphoneDB database to apply complex validity logic, ensuring that the selected targets function as components of viable signaling complexes. As a final purification step to remove downstream signaling moieties, the agent excluded intracellular kinases and GPCR-associated cytosolic proteins. This autonomous curation process yielded a final library of 528 immune cell receptors, optimized for extracellular interaction screening. For the brain-derived plasma protein list, we utilized a curated list of 112 brain-derived plasma proteins, identified in a recent study leveraging UK Biobank proteomics data. These candidates were selected based on a stringent tissue-specificity criterion, defined as exhibiting expression levels in the brain at least four-fold higher than in the next most highly enriched organ.

4 ESM

Following the dataset curation, the AI Machine Learning Specialist engineered a computational pipeline to prepare the biological data for analysis. The agent first retrieved the full-length amino acid sequences for the curated immune receptors and brain-derived plasma proteins from the result of the preprocessing. To focus on the biologically relevant physical interface and optimize computational efficiency, the agent implemented a sequence truncation protocol. This process computationally removed signal peptides, transmembrane helices, and intracellular domains, thereby isolating the extracellular domains (ECDs) where ligand-receptor binding events occur. For protein representation, the agent utilized ESM-2 (esm2_t33_650M_UR50D), a transformer-based protein language model developed by Meta AI. This model comprises 33 layers and 650 million parameters and is trained on evolutionary patterns from millions of protein sequences. By processing the truncated ECD sequences through this architecture, the agent transformed each protein into a 1,280-dimensional latent vector. This embedding strategy captures physicochemical and evolutionary features without the need for multiple sequence alignments (MSAs). To determine the most accurate method for predicting novel interactions using these embeddings, we conducted a benchmark analysis using a balanced dataset of 2,500 validated positive pairs and 2,500 randomly generated negative pairs. We evaluated five distinct interaction metrics, comparing a supervised approach against four zero-shot vector operations. The zero-shot baselines included cosine similarity, which assesses the directional alignment of vectors; dot product, which incorporates both signal magnitude and direction; Euclidean distance, which measures geometric proximity in the latent space; and ESM-2 attention maps, which infer residue-level contact probabilities directly from the model's internal cross-attention weights. The comparative analysis

Method	AUC	Max Accuracy
Cosine Similarity	0.7516	0.839
Euclidean Dist.	0.732	0.8502
Dot Product	0.5896	0.7955
ESM-2 Attention	0.5804	0.81
MLP (Supervised)	0.9934	0.9839

Figure 2: ESM accuracy.

indicated that the supervised Multi-Layer Perceptron (MLP) classifier, trained on the concatenated embeddings of the receptor-ligand pairs, outperformed the zero-shot baselines in both ROC-AUC scores and accuracy. While the zero-shot metrics provided geometric insights, the supervised model more effectively captured the non-linear complexities inherent in specific protein-protein binding. Consequently, the trained MLP model was selected to calculate and rank the interaction probabilities between the 528 curated immune receptors and the brain-derived plasma proteins.

This performance is exceptional, even when compared to existing studies such as ESMDNN-PPI (BioRxiv, 2024) or PLM-interact, which have reported high accuracy on general human PPI datasets. Notably, as highlighted in the recent TUnA study, conventional models often suffer from network bias toward well-studied protein families like cytokines or receptor tyrosine kinases. This bias frequently causes their generalization performance to plummet to the 0.6 range when attempting to identify truly novel interactions. In contrast, our approach proactively mitigated this bias through AI agent-driven domain refinement and rigorous negative control design. Consequently, we have achieved a level of reliability capable of sophisticatedly excavating the "dark matter" of the proteome—specifically, atypical brain-derived proteins that have been overlooked by existing databases.

5 Alphafold-Multimer

Next, we utilized AlphaFold-Multimer to predict and analyze the potential Protein-Protein Interactions (PPI) of the top 150 ligand-receptor pairs selected from approximately 60,000 ranked pairs. AlphaFold-Multimer generates a single prediction model for each input sequence pair, and we employed the following ranking metric to evaluate the reliability of each model:

Here, ipTM denotes the interface predicted TM-score, and pTM denotes the predicted TM-score. We initially aimed to select models with a Model Confidence score of 0.7 or higher as primary candidates; however, only one out of the 150 predicted models met this criterion. Consequently, we selected a total of five models with a Model Confidence score of 0.4 or higher for further analysis. Among these, one pair (Rank 16) was a well-known interaction, serving as a positive control that substantiated the reliability of the structural predictions by AlphaFold-Multimer. However, Model Confidence alone does not guarantee the existence of a binding interface between two proteins. Therefore, we performed visual verification using Predicted Aligned Error (PAE) plots to further assess the reliability of the predicted interactions. First, upon examining the PAE plot of the Rank 16 pair, blue blocks in the off-diagonal regions—indicating the presence of a strong interface between the receptor and ligand—were clearly observed (Fig 3). Similar blue blocks were also identified in the PAE plots of the top 4 selected models, excluding the positive control pair (Fig 4). In contrast, a comparative analysis of the bottom 4 candidates revealed no such blue blocks exhibiting low error in the off-diagonal regions (Fig 5). Subsequently, we visualized the tertiary structures of the candidate pairs utilizing PyMOL. To verify the physicochemical reliability of the predicted interfaces in the PAE plots, we visualized the pLDDT (predicted Local Distance Difference Test) scores of the residues constituting these regions. The visualized structures provide a concrete depiction of the spatial arrangement of the predicted binding interfaces. The receptor protein is represented in blue and the ligand protein in red,

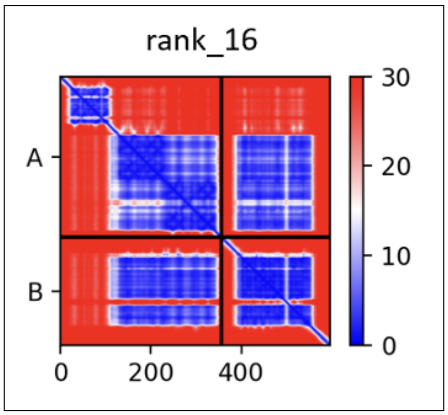

Figure 3: PAE plot of the well-known pair.

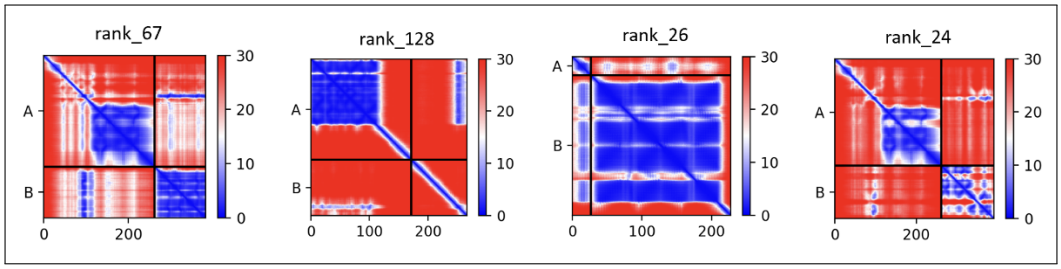

Figure 4: PAE plots of top 4 models ranked by model confidence.

with color intensity proportional to the pLDDT score. Analysis of whether residues corresponding to the predicted interaction regions (blue blocks) in the PAE plots possess high pLDDT values revealed that the Rank 16 pair exhibited high pLDDT levels across the entire protein, including the interface (Fig 6). Furthermore, the interfaces of the pairs ranked 67 and 24 were found to possess moderate pLDDT levels (Fig 7).

6 Discussion

In this work, we present an end-to-end pipeline for discovering novel receptor–ligand candidates based on the concept of a virtual lab. Unlike conventional approaches that use AI solely as an analytical tool, our framework positions AI as an active research agent involved in hypothesis formulation, validation strategy design, and result interpretation. In particular, the collaborative interaction of specialized AI agents enables iterative critique and refinement of research strategies, better reflecting the constraints and uncertainties of real-world biological research. The Virtual lab proposed fast sequence-based screening with structure-based validation. Virtual lab agent first trained an ESM2-based model using a balanced dataset comprising known receptor–ligand interactions and unknown pairs. The model demonstrated strong predictive performance when trained on 80%

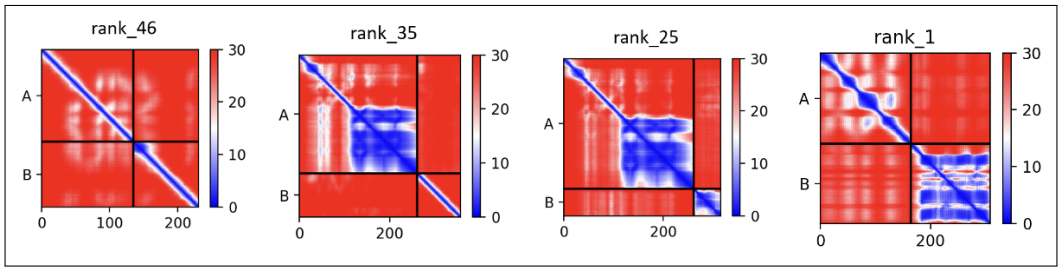

Figure 5: PAE plots of bottom 4 models ranked by model confidence.

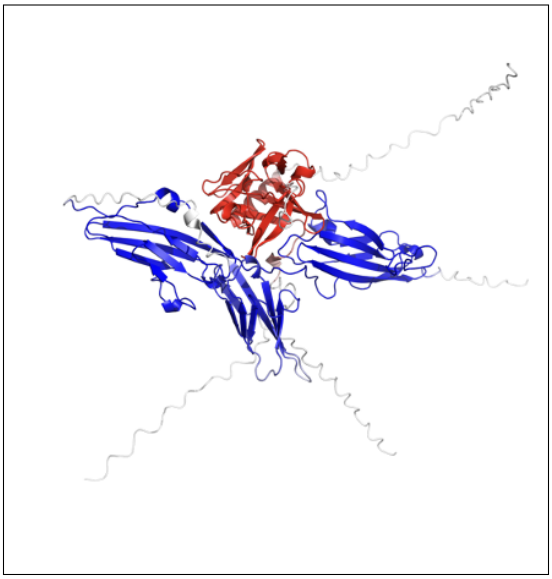

Figure 6: AlphaFold-Multimer prediction of the well-known pair(Rank 16).

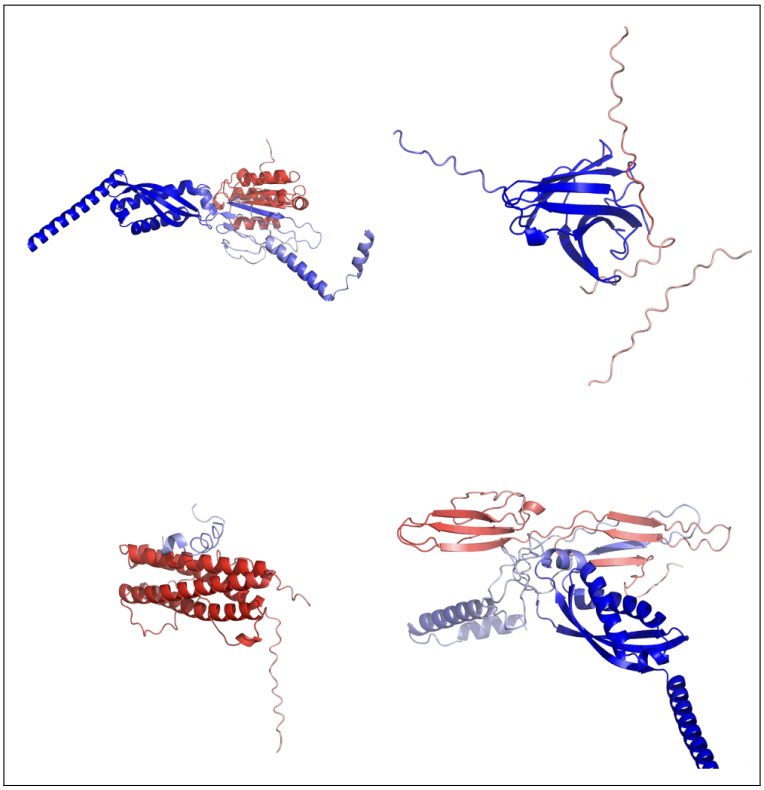

Figure 7: AlphaFold-Multimer prediction of the top 4 models(Rank 67, Rank 128, Rank 26, Rank 24).

of the data and evaluated on the remaining 20%. With this model, the agent screened interaction candidates between brain-derived plasma proteins from UK Biobank proteomics data and immune cell receptors obtained from CellPhoneDB. The top candidates were subsequently evaluated using AlphaFold, yielding a list of structurally plausible interaction candidates. The identified candidates are biologically meaningful, particularly in the context of aging. Brain-derived plasma proteins that increase with age are underrepresented among previously known immune ligands. The newly proposed receptor–ligand interactions therefore provide a starting point for exploring immune–brain communication mechanisms associated with aging.

7 Limitations and future work

Several limitations should be acknowledged. First, the predicted receptor–ligand interactions have not yet been experimentally validated. Future in vitro and in vivo studies will be required to confirm physical binding and functional relevance. Second, the incompleteness of the underlying datasets must be considered: current plasma proteomics does not capture all circulating proteins, and the repertoire of immune cell receptors remains partially characterized. These limitations constrain the interaction space that can be explored computationally. Despite these limitations, the proposed pipeline is highly extensible. By incorporating condition-specific ligand and receptor datasets, the framework can be readily applied to other biological contexts such as neurodegeneration, cancer, and autoimmune diseases. This flexibility highlights the potential of our work as a generalizable approach for systematic receptor–ligand discovery.

References

[1] Oh HS, Le Guen Y, Rappoport N, Urey DY, Farinas A, Rutledge J, Channappa D, Wagner AD, Mormino E, Brunet A, Greicius MD, Wyss-Coray T. Plasma proteomics links brain and immune system aging with healthspan and longevity. Nat Med. 2025 Aug;31(8):2703-2711. doi: 10.1038/s41591-025-03798-1. Epub 2025 Jul 9. PMID: 40634782; PMCID: PMC12353788.

[2] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

[3] Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J, Thomson E, Monteiro R, Guimaraes P, Walters B, Wiser J, Butte AJ. ImmPort: disseminating data to the public for the future of immunology. Immunol Res. 2014 May;58(2-3):234-9. doi: 10.1007/s12026-014-8516-1. PMID: 24791905.

[4] Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020 Apr;15(4):1484-1506. doi: 10.1038/s41596-020-0292-x. Epub 2020 Feb 26. PMID: 32103204.

[5] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123-1130. doi: 10.1126/science.ade2574. Epub 2023 Mar 16. PMID: 36927031.

[6] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021 Apr 13;118(15):e2016239118. doi: 10.1073/pnas.2016239118. PMID: 33876751; PMCID: PMC8053943.

[7] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.

[8] R. Evans, et al. (2021). Protein complex prediction with AlphaFold-Multimer. bioRxiv preprint.

[9] Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics. 2016 Dec 1;32(23):3676-3678. doi: 10.1093/bioinformatics/btw514. Epub 2016 Aug 8. PMID: 27503228.

[10] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007 Sep 21;372(3):774-97. doi: 10.1016/j.jmb.2007.05.022. Epub 2007 May 13. PMID: 17681537.

[11] Vangone A, Bonvin AMJJ. PRODIGY: A Contact-based Predictor of Binding Affinity in Protein-protein Complexes. Bio Protoc. 2017 Feb 5;7(3):e2124. doi: 10.21769/BioProtoc.2124. PMID: 34458447; PMCID: PMC8376549.

[12] Liu D, Young F, Lamb KD, Claudio Quiros A, Pancheva A, Miller CJ, Macdonald C, Robertson DL, Yuan K. PLM-interact: extending protein language models to predict protein-protein interactions. Nat Commun. 2025 Oct 27;16(1):9012. doi: 10.1038/s41467-025-64512-w. PMID: 41145424; PMCID: PMC12559430.

[13] Li Y, Wang C, Gu H, Long Z, Fan M, Li L. MPIDNN-GPPI: multi-protein language model with an improved deep neural network for generalized protein–protein interaction prediction. BMC Genomics. 2025 Nov 19;26(1):1059. doi: 10.1186/s12864-025-12228-y. PMID: 41257586; PMCID: PMC12628903.

[14] Franceschi C, Garagnani P, Vitale G, Capri M, Salvioli S. Inflammaging and 'Garb-aging'. Trends Endocrinol Metab. 2017 Mar;28(3):199-212. doi: 10.1016/j.tem.2016.09.005. Epub 2016 Oct 24. PMID: 27789101.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [N/A] .
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[N/A] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the main contribution is a multi-agent "Virtual Lab" framework for identifying novel brain-derived plasma protein–immune receptor interactions using sequence-based screening and structure-based validation. The scope is explicitly limited to in silico prediction and hypothesis generation, and these claims are consistent with the ESM-2 and AlphaFold-Multimer pipeline and the results discussed in the main text.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The "Limitations and future work" section explicitly acknowledges the lack of experimental validation and the incompleteness of available proteomics and receptor datasets. The paper also frames the findings as hypothesis-generating rather than biologically conclusive.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce new theoretical results, formal theorems, or mathematical proofs, and instead focuses on a computational and empirical pipeline for protein–protein interaction discovery.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the data sources (UniProt, ImmPort, CellPhoneDB, UK Biobank-derived protein list), the use of ESM-2 embeddings, the supervised MLP classifier, and the AlphaFold-Multimer confidence metrics (ipTM, pTM, pLDDT, and PAE). These descriptions provide a clear procedural outline for reproducing the main experimental workflow.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All data sources used in the study are publicly accessible databases (e.g., ImmPort, UniProt). While the full orchestration code for the AI agent system, detailed pseudocode, configuration descriptions, and execution logs are not released due to ongoing development.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [No]

Justification: While the paper reports the use of an 80/20 train–test split and compares supervised and zero-shot methods, it does not specify key training details such as optimizer type, learning rate, batch size, number of epochs, random seeds, or hyperparameter selection procedures, which limits full reproducibility.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports performance using threshold-independent metrics such as ROC-AUC and accuracy, which summarize model discrimination over the full range of decision thresholds rather than single-point estimates. However, these metrics are presented without confidence intervals, standard deviations, error bars, or statistical significance testing across multiple runs or data splits.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The manuscript does not disclose the type of hardware used (e.g., GPU/CPU models), memory requirements, runtime per experiment, or the total computational budget for training and AlphaFold-Multimer inference.

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research relies exclusively on publicly available biological databases and in silico modeling, does not involve personal or identifiable human data, and properly cites external resources, indicating alignment with the NeurIPS Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Although the work implies potential benefits for understanding aging and immune–brain communication, it does not explicitly discuss possible negative societal impacts, misuse risks, or ethical considerations related to downstream biomedical or AI applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk assets such as generative models, large-scale scraped datasets, or deployable systems that would require safeguards against misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: While the paper cites the sources of datasets and models (e.g., UniProt, ImmPort, CellPhoneDB, ESM-2, AlphaFold), it does not explicitly state the licenses or terms of use for these assets.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, codebases, or trained models for public use.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve human participants, surveys, or crowdsourced data collection, relying solely on existing public biological databases and computational modeling.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects or the collection of personal data, and therefore does not require IRB approval or equivalent ethical review.