

De Novo Protein Binder Design for Advanced Glycation End Products: A Computational Approach to Targeting Age-Related Glycative Damage (Stage 1 Registered Report)

Kejun (Albert) Ying^{1,2,*}

Published: March 25, 2026

¹ Stanford University, Department of Neurology (Tony Wyss-Coray Lab).

² University of Washington, Institute for Protein Design (David Baker Lab).

*Corresponding author: albert.k.ying@gmail.com

ABSTRACT

Advanced glycation end products (AGEs) accumulate with age and contribute to cardiovascular disease, neurodegeneration, and diabetic complications through receptor for advanced glycation end products (RAGE) activation and direct protein cross-linking. Despite decades of research, no approved therapies specifically target AGEs. This project applies deep learning-based protein design to create de novo protein binders that sequester carboxymethyllysine (CML), the most abundant and well-characterized AGE species. Using RFdiffusion All-Atom, we will generate 10,000 protein scaffold designs conditioned on the CML molecular structure, followed by sequence optimization with ProteinMPNN and in silico filtering using AlphaFold2 and Rosetta. The top 100–200 computational designs will be validated experimentally through yeast surface display screening against CML-modified bovine serum albumin, and top candidates will undergo binding affinity characterization by biolayer interferometry. We hypothesize that this pipeline will yield at least one binder with measurable binding to CML-modified protein and sub-micromolar affinity. All computational designs, experimental protocols, and raw data will be made openly available. This work represents the first application of RFdiffusion All-Atom to AGE targeting and will establish whether computationally designed protein binders can serve as a new class of anti-glycation therapeutics.

INTRODUCTION

Research Question: Background, Importance, and Relevance

AGEs form through non-enzymatic glycation when reducing sugars react with amino groups on proteins, lipids, and nucleic acids. The Maillard reaction produces early glycation intermediates that undergo oxidation and rearrangement to form stable AGE adducts. The most abundant and best-characterized AGEs include carboxymethyllysine (CML), carboxyethyllysine (CEL), and pentosidine. These modifications accumulate in nearly all mammalian tissues with age and are accelerated by hyperglycemia, oxidative stress, and metabolic dysfunction^{1,2}.

AGEs cause pathology through two primary mechanisms. First, they directly modify proteins, causing cross-linking that impairs function. This is particularly damaging in long-lived extracellular matrix proteins like collagen, where AGE cross-links increase tissue stiffness and reduce turnover. Second, AGEs bind to the receptor for advanced glycation end products

(RAGE), activating NF- κ B signaling and triggering chronic inflammation, oxidative stress, and cellular dysfunction³. The RAGE-AGE axis has been implicated in diabetic complications, atherosclerosis, Alzheimer's disease, and accelerated aging⁴.

Current therapeutic approaches fall into three categories: AGE formation inhibitors (aminoguanidine, pyridoxamine), AGE cross-link breakers (alagebrium), and RAGE antagonists. None have achieved regulatory approval for AGE-related indications. Aminoguanidine reached Phase III trials but was discontinued due to adverse effects. Recent interest has shifted to natural polyphenols and dietary interventions, but these lack specificity and potency⁵.

This project proposes a novel strategy: designing de novo protein binders that recognize and sequester specific AGE adducts. Unlike small molecules, designed proteins can achieve high affinity and specificity through extensive complementary surfaces. The recent availability of RFdiffusion All-Atom enables design of proteins that bind small molecules by generating folded scaffolds around the target ligand^{6,7}. This approach has been validated for other small molecule targets including digoxigenin and heme. No group has yet applied this technology to AGE targeting.

A designed AGE binder could function therapeutically by intercepting circulating AGEs before they engage RAGE or cross-link tissue proteins. Beyond direct therapeutic potential, such binders would serve as research tools for detecting and quantifying specific AGE species, and as scaffolds for downstream enzyme design to actively degrade AGEs.

Hypotheses

This study will test three preregistered hypotheses:

H1: RFdiffusion All-Atom can generate protein scaffolds with designed binding pockets for CML (the most abundant AGE species).

H2: At least one computationally designed binder will show measurable binding to CML-modified bovine serum albumin (AGE-BSA) in yeast surface display screening, defined as a signal-to-background ratio greater than 3.

H3: Top binder candidates will exhibit sub-micromolar affinity for AGE-BSA as measured by biolayer interferometry (BLI).

H1 is the primary outcome as it establishes the foundational feasibility of the computational approach. H2 constitutes the primary experimental outcome, since yeast display screening is the initial gate for identifying active binders. H3 is a secondary outcome that characterizes the quality of candidates passing the primary screen. In the event no binding is detected in the initial yeast display screen, contingent analyses are pre-specified in the Methods section.

Timeline

The study will be completed over 6 months following Stage 1 peer review approval:

Months 1–2: Computational design and in silico screening (backbone generation, sequence design, AlphaFold2/Rosetta filtering).

Months 3–4: Gene synthesis and yeast surface display library construction and expression.

Months 5–6: Binding validation by FACS screening, protein expression, purification, and affinity characterization by BLI.

A Stage 2 submission is anticipated approximately 8 months after Stage 1 acceptance, allowing time for data analysis and manuscript preparation.

METHODS

The Methods section provides sufficient detail to allow full replication of all computational and experimental procedures.

Experimental Design

Target Selection and Structure Preparation

Carboxymethyllysine (CML) will serve as the primary design target. CML is the most abundant AGE in human tissues, is chemically stable, has commercially available standards, and its structure is well characterized. The molecular structure of CML bound to a lysine backbone will be used as the design input. Coordinates will be generated from SMILES notation using RDKit (version 2023.09) and energy-minimized using the MMFF94 force field prior to diffusion design.

Computational Design Pipeline

Design will follow the RFdiffusion All-Atom workflow⁷ in three sequential stages:

Stage 1 — Backbone generation: RFdiffusion All-Atom will generate protein scaffolds of 70–100 residues conditioned on the CML structure, producing diverse binding pocket geometries. A total of 10,000 backbone designs will be generated using the publicly available RFdiffusion All-Atom repository (github.com/baker-laboratory/rf_diffusion_all_atom).

Stage 2 — Sequence design: ProteinMPNN will assign amino acid sequences to each backbone, optimizing for the predicted binding interface. Five independent sequences will be generated per backbone.

Stage 3 — In silico filtering: Designs will be scored using AlphaFold2 (v2.3) for structural confidence (pLDDT) and Rosetta (REF2015 score function) for binding energy estimation. The top 100–200 designs ranked by a combined metric of pLDDT > 85 and Rosetta binding energy will advance to experimental testing.

Yeast Surface Display Screening

Designed binders will be displayed on yeast cell surfaces as Aga2p fusions. A pooled library of top designs will be constructed through oligonucleotide synthesis and Gibson assembly. The library will be sorted against biotinylated AGE-BSA (commercially available from R&D Systems) using fluorescence-activated cell sorting (FACS). Binding cells will be enriched over 2–3 rounds

of selection. Individual clones from the enriched population will be sequenced by Sanger sequencing and characterized.

Binding Characterization

Top hits from yeast display will be expressed in *E. coli* (BL21(DE3) strain) and purified by nickel affinity chromatography. Binding kinetics (k_{on} , k_{off}) and equilibrium affinity (K_D) will be measured using biolayer interferometry (BLI) on an Octet RED96 instrument (Sartorius) with AGE-BSA immobilized on streptavidin biosensors. Specificity will be assessed by parallel testing against unmodified BSA at equivalent concentrations.

Blinding and Bias Minimization

Given the computational nature of the primary design stage, blinding is not applicable to backbone generation or sequence design. For experimental validation, FACS sorting will be performed by a laboratory member who is not the principal investigator, and BLI data will be collected and processed before unblinding of design identities. No exclusion criteria apply to the computational designs; all designs passing the pre-specified *in silico* filtering thresholds will advance to experimental testing.

Sample and Statistical Power

Based on published RFdiffusion All-Atom benchmarks for small molecule binders, the experimental hit rate for top-filtered designs is estimated at 5–15%⁷. Testing 100–200 designs is therefore expected to yield 5–30 binders with detectable binding signal. This estimate is conservative given that AGE-BSA presents multiple CML epitopes per molecule, increasing the effective target valency and concentration relative to single-epitope targets used in prior benchmarks.

For the primary outcome (yeast display hit rate), the study is designed as a proof-of-concept screen rather than a hypothesis test with a predetermined power target. The minimum detectable effect is defined operationally: a hit rate of $\geq 1\%$ (at least 1 binder out of 100 tested) would be considered evidence that the computational pipeline can identify functional binders, consistent with first-in-class platform development. BLI measurements will use standard curves with at least 5 analyte concentrations per binder, fit to a 1:1 Langmuir binding model, to accurately estimate binding constants. No covariates or regressors are anticipated for primary analyses.

Data Collection, Processing, and Planned Analysis

Data Sources and Instruments

Data will be collected from four sources: (i) RFdiffusion All-Atom output files (PDB format) for backbone designs; (ii) AlphaFold2 pLDDT scores and predicted aligned error (PAE) matrices; (iii) Rosetta binding energy scores; and (iv) BLI sensorgrams from the Octet RED96. FACS data will be collected on a BD FACSAria instrument and analyzed using FlowJo (v10). All raw data files will be retained in their original formats.

Pre-Processing

Computational designs will be filtered by the following sequential pre-processing criteria prior to experimental testing: (1) pLDDT > 85 across the full design; (2) predicted TM-score to design model > 0.8; (3) no steric clashes with CML ligand as assessed by Rosetta fa_rep term < 5.0. Designs failing any criterion are excluded from experimental testing. BLI sensorgrams will be double-referenced (reference channel and buffer-only reference) prior to kinetic fitting using the Octet Analysis Studio software (Sartorius, v12).

Primary Analysis

The primary outcome is the fraction of designs showing a binding signal above background in yeast surface display (hit rate). Designs with a signal-to-background ratio > 3 relative to no-ligand controls will be classified as hits. The hit rate and its 95% Wilson confidence interval will be reported.

Secondary Analyses

- (i) Correlation between in silico metrics (AlphaFold2 pLDDT, Rosetta binding energy) and experimental binding, assessed by Spearman rank correlation.
- (ii) Structural diversity of successful binders, assessed by pairwise backbone RMSD clustering.
- (iii) Affinity distribution of BLI-validated hits, reported as KD values with 95% confidence intervals from curve fitting.

Variations from Intended Sample Size

The primary anticipated challenge is a lower-than-expected number of designs passing in silico filtering thresholds, which could reduce the experimental sample below 100 designs. In this event, all designs passing filtering will still be tested and the hit rate reported with appropriate confidence intervals reflecting the smaller denominator. Gene synthesis failures (estimated < 5% of sequences) will be replaced with the next-ranked designs from the computational pipeline. No imputation will be applied to missing BLI measurements; any binder for which a reliable fit cannot be obtained will be reported as “fit failed” with the raw data provided.

Pilot Data

Preliminary computational experiments confirm feasibility. Test runs of RFdiffusion All-Atom with CML as the target ligand successfully generated diverse protein scaffolds with plausible binding pockets. AlphaFold2 predicts high confidence (pLDDT > 85) for 15–20% of designs after ProteinMPNN sequence assignment, consistent with published benchmarks for the method⁷. No wet laboratory pilot data has been generated, which is appropriate given the exploratory nature of this project and the absence of prior art applying RFdiffusion All-Atom to AGE targets.

Statistical Methods

Hit rates from yeast display will be reported as proportions with 95% Wilson score confidence intervals. Spearman rank correlation (scipy.stats, Python 3.11) will be used for secondary analysis (i) to avoid assumptions of normality given the expected right-skewed distribution of binding scores. BLI kinetic data will be fit to a 1:1 Langmuir model using nonlinear least squares

regression in Octet Analysis Studio. Affinity values (KD) will be reported as mean $\pm$ standard deviation from triplicate measurements.

Missing values: Designs for which BLI fitting fails ($R^2 < 0.95$) will be excluded from affinity analyses but reported descriptively. Outliers in BLI data will be defined as measurements deviating more than 3 standard deviations from the replicate mean and will be excluded after documentation. Multiple hypothesis testing: Secondary analyses (i)–(iii) are exploratory; no correction for multiple comparisons will be applied, and all p-values will be reported without adjustment with a note that these analyses are hypothesis-generating.

Software: RFdiffusion All-Atom (github.com/baker-laboratory/rf_diffusion_all_atom), ProteinMPNN (github.com/dauparas/ProteinMPNN), AlphaFold2 v2.3 (github.com/google-deepmind/alphafold), Rosetta 3.13 (rosettacommons.org), FlowJo v10 (BD), Octet Analysis Studio v12 (Sartorius), Python 3.11 with scipy 1.11 and numpy 1.24.

Contingent Analysis

IF no designs show a signal-to-background ratio > 3 in the initial yeast display screen, THEN the following pre-specified contingent analyses will be executed: (a) testing binding to free CML-lysine conjugate (rather than AGE-modified protein) to determine whether the absence of binding reflects a scaffold problem or a target accessibility problem; and (b) assessing whether designs adopt the computationally predicted fold using circular dichroism (CD) spectroscopy, to determine whether folding failures explain the absence of binding. These analyses are contingent on a null result in the primary screen and will be clearly labeled as such in the Stage 2 report.

Interpreting Results

If H2 and H3 are supported (binding detected with sub-micromolar affinity), this will establish proof-of-concept for a new class of AGE-targeting protein therapeutics and produce binder sequences that can be optimized through directed evolution. This would represent the first demonstration that computationally designed proteins can specifically engage a post-translational glycation modification, advancing both the therapeutic and research tool landscape for aging biology.

If binding is not achieved (H2 not supported), the negative results will inform the field about the challenges of designing binders for post-translationally modified amino acids and guide future computational approaches. Specifically, results from the contingent analyses will distinguish between scaffold folding failures and target accessibility as limiting factors, providing actionable direction for subsequent design iterations. Either outcome has direct implications for the feasibility of protein-based anti-AGE interventions.

These findings will contribute to the literature on de novo protein design benchmarks and to the aging intervention field. There are no direct immediate policy implications, though successful development of anti-AGE biologics could eventually inform therapeutic strategies for diabetes complications and age-related diseases.

Dissemination of Information

All computational designs, screening data, binding characterization results, and raw data files will be deposited on Zenodo under a Creative Commons Attribution (CC-BY 4.0) license within 3 months of project completion, regardless of outcome. A preregistration will be maintained on ResearchHub, and results will be shared as a preprint on bioRxiv prior to journal submission. The Stage 2 registered report will be submitted to ResearchHub.

Study Status

At the time of this Stage 1 submission, data collection has not started. Preliminary computational feasibility tests (test runs of RFdiffusion All-Atom with CML, as described in the Pilot Data section) have been completed, but no experimental data have been collected. The full design generation campaign and all experimental work described in this protocol will commence following Stage 1 peer review approval.

OTHER REQUIRED INFORMATION

Acknowledgements

The author thanks the Baker Lab at the University of Washington Institute for Protein Design for providing computational infrastructure and technical guidance on the RFdiffusion All-Atom pipeline, and the Wyss-Coray Lab at Stanford University Department of Neurology for support and resources.

Funding

This work will be conducted by the author as part of ongoing postdoctoral research. Computational infrastructure (beyond cloud GPU) is available through existing laboratory resources at the Baker Lab. Cloud GPU computing costs ($\approx$ 1,000 USD) will be funded from discretionary postdoctoral research funds. This work is also supported by a ResearchHub Foundation grant. The author declares that no external grants are involved in supporting this specific piece of research.

Author Contributions

K.A.Y. conceived and designed the study, will perform all computational and experimental work, analyze the data, and write the manuscript. All authors read and approved the final manuscript.

Corresponding author: Kejun (Albert) Ying, University of Washington, Institute for Protein Design, 1959 NE Pacific St, Seattle, WA 98195. Email: albert.k.ying@gmail.com

Ethical Approval

This project involves no human subjects, vertebrate animals, or biohazardous materials. All work uses commercially available reagents and standard molecular biology techniques (*E. coli* and *Saccharomyces cerevisiae*). Institutional biosafety approval for routine *E. coli* and yeast work is already in place at the University of Washington. No additional ethics committee approvals are required for this phase of the project.

Registration

This study is not a clinical trial and does not require trial registration. A preregistration of this protocol will be posted on ResearchHub prior to commencement of data collection.

Conflicts of Interest

The author declares that there are no conflicts of interest.

REFERENCES

1. Uceda AB, Mariño L, Casasnovas R, Adrover M. An overview on glycation: molecular mechanisms, impact on proteins, pathogenesis, and inhibition. *Biophys Rev.* 2024;16(2):189–218. doi:10.1007/s12551-024-01188-4.
2. Zhang Y, Zhang Z, Tu C, Chen X, He R. Advanced glycation end products in disease development and potential interventions. *Antioxidants (Basel).* 2025;14(4):492. doi:10.3390/antiox14040492.
3. Twarda-Clapa A, Olczak A, Białkowska AM, Koziołkiewicz M. Advanced Glycation End-Products (AGEs): Formation, Chemistry, Classification, Receptors, and Diseases Related to AGEs. *Cells.* 2022;11(8):1312. doi:10.3390/cells11081312.
4. Rowan S, Jiang S, Korem T, Szymanski J, Chang M. Mechanistic targeting of advanced glycation end-products in age-related diseases. *Biochim Biophys Acta Mol Basis Dis.* 2019;1865:165578.
5. Sarmah S, Roy AS. A review on prevention of glycation of proteins: potential therapeutic substances to mitigate the severity of diabetes complications. *Int J Biol Macromol.* 2022;195:565–588.
6. Watson JL, Wang J, Juergens D, et al. De novo design of protein structure and function with RFdiffusion. *Nature.* 2023;620:1089–1100. doi:10.1038/s41586-023-06415-8.
7. Krishna R, Wang J, Ahern W, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science.* 2024;384(6693):eadl2528. doi:10.1126/science.adl2528.