Fast, accurate ranking of engineered proteins by receptor binding propensity using structural modeling

Xiaozhe Ding,* Xinhong Chen, Erin E. Sullivan, Timothy F. Shay, and Viviana Gradinaru*

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA

E-mail: xding@caltech.edu; viviana@caltech.edu

Abstract

Deep learning-based methods for protein structure prediction have achieved unprecedented accuracy. However, the power of these tools to guide the engineering of protein-based therapeutics remains limited due to a gap between the ability to predict the structures of candidate proteins and the ability to assess which of those proteins are most likely to bind to a target receptor. Here we bridge this gap by introducing Automated Pairwise Peptide-Receptor AnalysIs for Screening Engineered proteins (APPRAISE), a method for predicting the receptor binding propensity of engineered proteins. After generating models of engineered proteins competing for binding to a target using an established structure-prediction tool such as AlphaFold2-multimer or ESMFold, APPRAISE performs a rapid (under 1 CPU second per model) scoring analysis that takes into account biophysical and geometrical constraints. As a proof-of-concept, we demonstrate that APPRAISE can accurately classify receptor-dependent vs. receptor-independent engineered adeno-associated viral vectors, as well as diverse

1

classes of engineered proteins such as miniproteins targeting the SARS-CoV-2 spike protein, nanobodies targeting a G-protein-coupled receptor, and peptides that specifically bind to transferrin receptor and PD-L1. With its high accuracy, interpretability, and generalizability, APPRAISE has the potential to expand the utility of current structural prediction and accelerate protein engineering for biomedical applications.

Introduction

Many protein-based biologics rely on specific targeting. As a result, considerable effort has been devoted to engineering proteins that bind to a specific receptor, using strategies such as directed evolution[1–4] and rational design.[5–7] Currently, evaluating the receptor binding propensities of candidate molecules with *in vitro* and *in vivo* tests is a major bottleneck, which can be eased by computational prioritization of the candidate proteins.[8]

In order to predict protein function, two strategies have been employed: end-to-end sequence-function and two-step sequence-structure/structure-function. The sequence-function strategy, typically using end-to-end machine learning models,[9,10] is best suited for highly specialized functions, such as enzyme activities, which are challenging to predict from physical principles.[11] However, these models require specific training datasets in order to be applied to new contexts. In comparison, the two-step sequence-structure/structure-function strategy offers a more generalizable solution for functional prediction, particularly for functions with clear biophysical mechanisms, such as protein-protein binding.

The rapid development of deep learning-based methods has brought unprecedented accuracy to the first step of the sequence-structure/structure-function strategy. Since AlphaFold2 (AF2)'s outstanding performance in CASP14 in 2020,[12] a number of new deep learning-based structure-prediction tools have been released,[13–24] providing a powerful set of tools for generating protein models with atomic-level precision. Some of the most recent structure-prediction tools were trained to predict special targets such as antibody monomers,[18] and some,[14,17] notably AlphaFold-multimer,[14] can accurately model multi-chain complexes. Im-

portantly, these structure-prediction tools allow the generation of models in less than one GPU hour each, a level of throughput that experimental methods cannot match.

The second step, ranking target binding propensies based on modeled structures, has been less attended compared to the first step. Structure-prediction neural networks generate confidence scores for predicted complex models, such as pLDDT and pTM scores (used by AF2),[12] interface pTM scores (used by AF2-multimer),[14] which have been used as metrics to evaluate the probability that a protein binds to the target in some cases.[17,25] However, previous reports[26] and our experience have revealed that these scores alone are, in some cases, not reflective of binding propensities, particularly when the interaction is weak or transient. Utilizing biophysical principles to analyze the predicted structure may help to improve the ranking of binders by extracting relevant information stored in the 3D coordinates.

Ranking the binding probability of engineered proteins through modeled structures presents unique challenges. One challenge is imposed by the high sequence similarity between candidate molecules. Engineered protein variants are often constructed by modifying a short variable region to a common scaffold. Due to this similarity, the energy difference between the candidate binders can be very small, sometimes buried in the error of the energy function used for candidate ranking.[27,28] This problem is compounded by structure-prediction methods that rely heavily on co-evolutionary information or homology, causing them to generate similar binding poses for the candidate proteins. Another challenge is assessing a large number of predicted structure models efficiently. Direct quantification of protein-protein interface energy using interpretable, physics-based methods trades off accuracy and speed.[29] For instance, methods that use molecular dynamics simulations can cost more than 10^3 CPU hours per model. Faster, less rigorous methods with better-than-random ability to predict the impact of interface mutations still require 1 CPU minute to 1 CPU hour per non-antibody-antigen model.[29] In the post-AlphaFold era, an interpretable and efficient method of predicting the target binding of a large number of models would greatly accelerate protein engineering efforts.

Recently, Chang and Perez utilized competitive modeling with AF2-multimer to demonstrate a correlation between competition results and peptide binding affinities.[26] Nevertheless, the method that the study uses to assess the competition results necessitates a comparison of the modeled structures to an experimentally solved "native" structure, which is not available for many engineered proteins.

To bridge the remaining gap between structure prediction and protein engineering, here we present Automated Pairwise Peptide-Receptor AnalysIs for Screening Engineered proteins (APPRAISE), a readily interpretable and generalizable method for ranking the receptor binding propensity of engineered proteins based on competitive structure modeling and fast physics-informed structure analysis.

Results and discussion

The workflow of APPRAISE (Figure 1) comprises four main components. In the first step, pairs of peptides from N candidate protein molecules (N^2 pairs total) are modeled with a target receptor using a state-of-the-art structure modeling method such as AF2-multimer.[14] In the second stage, two simplified energetic binding scores are calculated for both the peptide of interest (POI) and its competitor in each individual model. In the third, optional, step, geometrical constraints for effective binding are applied to these scores. Finally, the score difference between the POI and the competitor is used to determine the result of each competition, and the peptides are ranked based on the matrix of competition results.
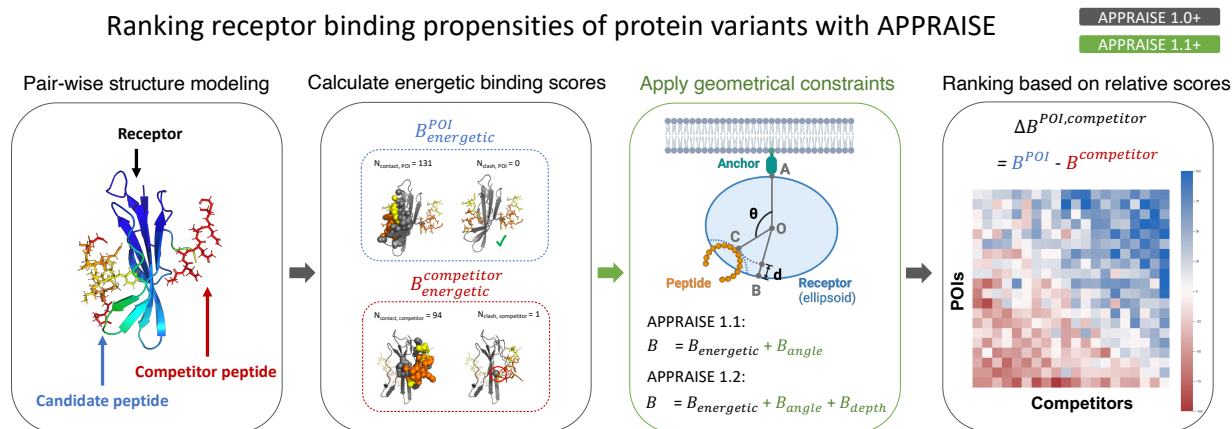
Figure 1: **Workflow of Automated Pairwise Peptide-Receptor AnalysIs for Screening Engineered proteins (APPRAISE).** First, peptides from the receptor-binding region of the engineered protein candidates are modeled in pairs in complex with the target receptor using tools like AF2-multimer or ESMFold. Second, a non-negative energetic binding score based on atom counting is calculated for each peptide in the pair. Third, in APPRAISE 1.1+, additional geometrical constraints that are important for peptide binding are considered. Finally, a relative score for each individual match is calculated by taking the difference of the score for the peptides, and a matrix of the averaged relative scores is used to determine the ranking of the peptides.

APPRAISE can accurately classify receptor-mediated brain transduction of viral vectors

We first developed APPRAISE to predict the binding propensities of engineered Adeno-Associated Viral (AAV) capsids for brain receptors. Recombinant AAVs are widely used as delivery vectors for gene therapy due to their relative safety as well as their broad and engineerable tropism. *In vivo* selections from libraries of randomized peptide-displaying AAV variants have yielded capsids that can transduce the animal brain,[1,2,30–34] an organ tightly protected by the blood-brain barrier (BBB). Widely-known examples among these capsids are AAV-PHP.B[1] and AAV-PHP.eB,[30] two AAV9-based[35] variants displaying short (7aa-9aa) surface peptides. The two variants can efficiently deliver genetic cargo to the brains of a subset of rodent strains. Genetic and biophysical studies have revealed that the BBB receptor for PHP.B/PHP.eB in these trains is LY6A, a GPI-anchored membrane protein.[36–38] A dataset comprising peptide-displaying AAV capsids that were engineered in a similar way

as PHP.B/eB was collected in order to train the APPRAISE method(Figure S1). Although binding between the AAV and the LY6A receptor is dynamic[39,40] and therefore challenging to be quantitatively measured, we could infer the binary LY6A-binding profiles of AAV capsids from their differential brain transduction profiles in mice strains with and without the receptor, producing a training set of peptide-displaying AAV capsids(Figure S1).

One challenge for modeling AAV capsids is that they are huge complexes made of 30,000+ amino acids (aa). In order to reduce computational costs for structure modeling and avoid complications arising from non-specific interactions, we modeled each AAV capsid variant using a single peptide spanning the engineered region (Figure 2a). This peptide (residues 587-594 in the VP1 sequence) includes 7 inserted residues and 8 contextual residues flanking the insertion. All of these residues are surface-exposed and may make direct contact with the receptor in the assembled capsid. Modeling this surface peptide (15 aa) is far less computationally intensive than modeling the entire capsid or even an asymmetric capsid subunit (500+ aa). In addition, compared to the latter, it may improve accuracy by eliminating competing interactions of residues normally buried in inter-subunit interfaces.

To discriminate relatively small differences in receptor binding propensities of candidate peptides, we modeled the peptides pairwise in competition for the target receptor.[26,41] To evaluate the competition results efficiently, we designed a score based on simple atom counting as a rough estimate of the interface free energy between the peptide of interest (POI) and the receptor in a structure model (Figure 2b). This score, which we term the energetic binding score ($B_{energetic}^{POI}$, simplified as B_0^{POI}), is a non-negative value calculated from the numbers of contacting and clashing atoms at the interface (Eq. 1). We describe the detailed rationale behind this score in the Methods.

$$B_0^{POI} = B_{energetic}^{POI} = \max(N_{contact}^{POI} - 10^3 \cdot N_{clash}^{POI}, 0) \tag{1}$$

To take full advantage of the information encoded in the competitive binding models, we further derived a "relative binding score" as inspired by the "specificity strategy" for protein-protein interface design.[42] The relative score by taking the difference between the absolute scores for the POI and competitor peptide (Eq. 2), rewarding POIs that destabilize the binding of competing peptides.

$$\Delta B_0^{POI,competitor} = B_0^{POI} - B_0^{competitor} \tag{2}$$

For an engineered protein to effectively bind to a membrane receptor, it must meet certain geometrical constraints (Figure 2c). In order to utilize this geometrical information, which is likely unused by structure-prediction tools, we incorporated two constraints that are essential for effective binding: the binding angle and the binding depth (Figure 2c-e).

The first constraint comes from the angle a binding protein can make (Figure 2c,d). In modeling a peptide-receptor complex using the extracellular domain of the membrane receptor (e.g., LY6A), most structure predictors (e.g., AF2-multimer) would consider the whole surface of the domain to be accessible by the peptide. However, in biological conditions, the membrane-facing side of the receptor is inaccessible to the engineered peptide. This polarity of accessibility is a general property of any receptor that is closely anchored to a larger complex. To account for the potentially huge energy cost of an engineered peptide binding these inaccessible locations, we used a steep polynomial term to penalize peptides that bind to the anchor-facing part of the receptor (Figure 2d, defined in the Methods by Eq. 5). B_0^{POI} is adjusted by this geometrical constraint term, rectified to be non-negative, and $\Delta B_0^{POI,competitor}$ is also re-calculated accordingly, yielding new scores B_1^{POI} and $\Delta B_1^{POI,competitor}$ (Eq. 3).

$$\Delta B_1^{POI,competitor} = B_1^{POI} - B_1^{competitor} \qquad (3)$$

$$= \max(B_{energetic}^{POI} + B_{angle}^{POI}, 0)$$

$$- \max(B_{energetic}^{competitor} + B_{angle}^{competitor}, 0)$$

The second constraint concerns the depth of the binding pocket (Figure 2c,e). Inspired by a previous report that the peptide-receptor distance predicted by AlphaFold2 is correlated with the IC50 of an inhibitor peptide,[26] we hypothesized that peptides binding to a deeper pocket on the receptor surface might benefit from longer receptor residence time, which is vital for the efficacy of many therapeutics.[43] Based on this hypothesis, we included a pocket depth consideration in APPRAISE. We used a relative pocket depth measurement, as opposed to an absolute peptide-receptor distance measurement, to avoid possible bias caused by the sizes of different receptors. We used an odd polynomial term to reward peptides that insert into deep pockets on the receptor while penalizing peptides that attach to surface humps (Figure 2e, defined in the Methods by Eq. 6). This gives us an adjusted score that considers both the angle constraint and the depth constraint, $\Delta B_2^{POI,competitor}$ (Eq. 4).

$$\Delta B_2^{POI,competitor} = B_2^{POI} - B_2^{competitor} \qquad (4)$$

$$= \max(B_{energetic}^{POI} + B_{angle}^{POI} + B_{depth}^{POI}, 0)$$

$$- \max(B_{energetic}^{competitor} + B_{angle}^{competitor} +$$

$$B_{depth}^{competitor}, 0)$$

We compared different versions of scoring methods based on competitive modeling results using AF2-multimer modeling (Figure 2g-i). Individual matching scores with statistical significance were used to determine wins and losses, and the total matching points in a tournament were used to rank all candidate engineered proteins (Methods). We found that simple atom-counting-based B_0^{POI} can already differentiate LY6A-binding peptides from non-binders(Figure 2g, j). Compared to B_0^{POI} alone, the relative score $\Delta B_0^{POI,competitor}$ showed improved prediction power, a ROC-AUC of 0.800 and an AUPRC of 0.756 for the training dataset (Figure 2h, j, k). Adding both geometrical terms, B_{angle} and B_{depth}, into consideration indeed improved the prediction accuracy of the binding score (Figure 2i-k), yielding a ROC-AUC of 0.838 and an AUPRC of 0.845 (Figure 2j, k). Importantly, the improvement in ROC-AUC mainly came from the low-false-positive-rate segment of the ROC curve, which is crucial for *in silico* screening of engineered proteins. We name the version of APPRAISE that considers only the angle constraint (through score ΔB_1) APPRAISE 1.1 (Figure S2a), and the version that considers both angle and depth constraints (through score ΔB_2) was named APPRAISE 1.2 (Figure 2i).

We then compared AF2-multimer-based APPRAISE 1.2 with other structure-based peptide affinity ranking methods on the AAV dataset (Figure 2k). With this particular dataset, the model confidence scores pLDDT, pTM, and interface pTM fail to differentiate whether an AAV variant is an LY6A binder, producing worse-than-random prediction (ROC AUC < 0.5). This is possibly due to the dynamic nature of the interaction between LY6A-binding AAV variants and the receptor,[39,40] which causes the confidence scores of the complex models to be generally low. APPRAISE 1.2 utilizing ESMFold as the structure prediction engine, however, performed at a comparable level to AF2-multimer-APPRAISE 1.2 (Figure S2b), with a ROC AUC of 0.895 and AUPRC 0f 0.818 (Figure 2k).

AF2-multimer-APPRAISE 1.2 ranking outperformed all other ranking methods at the

low-false-positive-rate end of the ROC curve, with a true positive rate of 0.714 and no false positive predictions. The performance with stringent cut-off values is particularly relevant for protein engineering applications, where the goal is typically to identify a few positive binders from a large number of negative, non-binding candidates. The superiority of AF2-multimer-APPRAISE 1.2 in dealing with this kind of imbalanced library is also evidenced by its highest AUPRC. Because of this, we chose to characterize AF2-multimer-APPRAISE 1.2 further. In the following text, 'APPRAISE' will be used to refer to AF2-multimer-APPRAISE 1.2 unless otherwise specified.
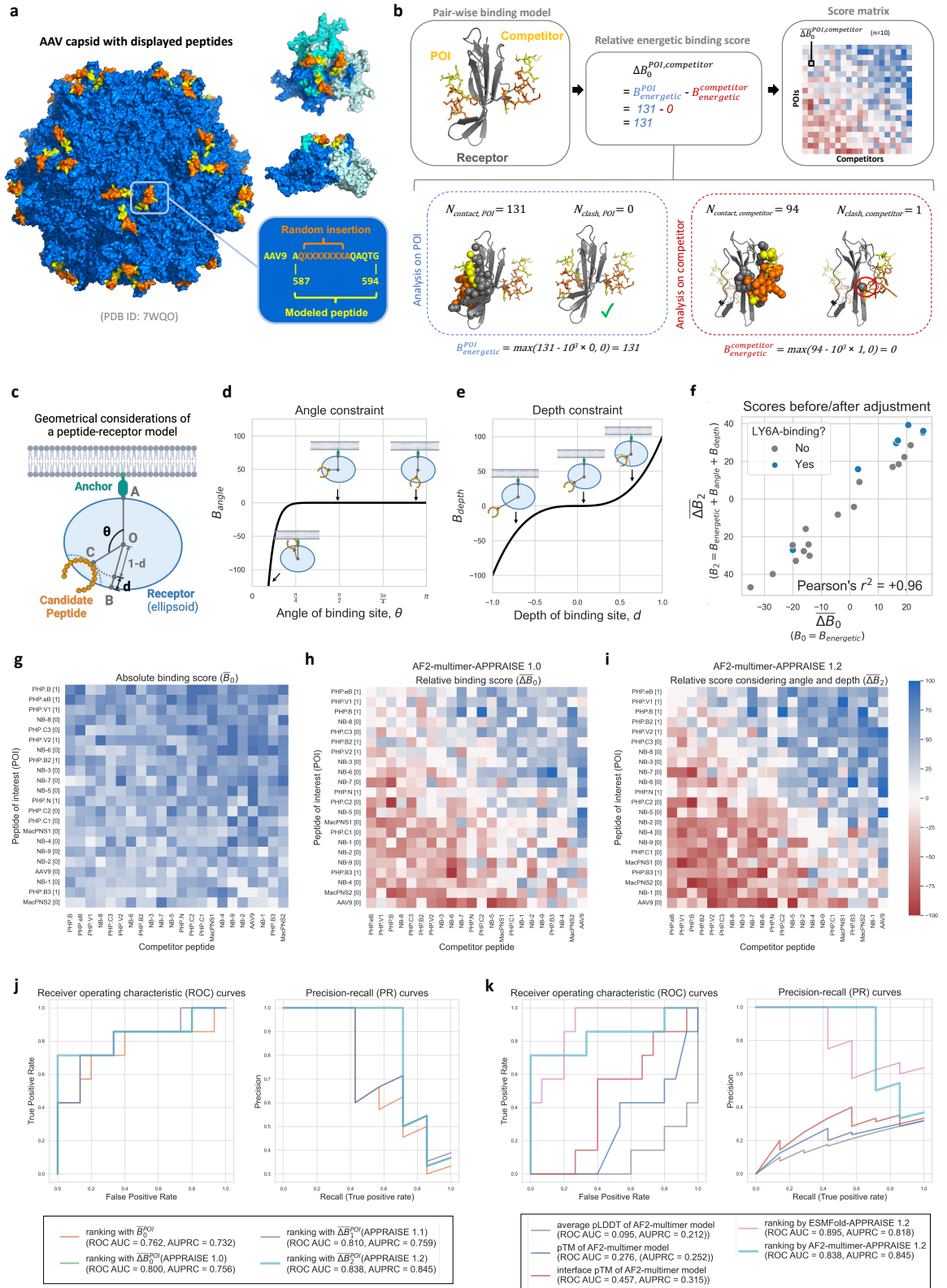
Figure 2: **Binary classification of receptor-binding AAV capsids using physical and geometrical principles.** See the next page for the caption.

Figure 2: **Binary classification of receptor-binding AAV capsids using physical and geometrical principles. a**, a structure model of AAV-PHP.eB, highlighting the site for inserting the displayed peptide (orange) and the peptide used for APPRAISE modeling (yellow or orange). The left image shows the whole AAV capsid made of 60 structurally identical subunits. The two images on the top right show a top view and a side view around the 3-fold axis, respectively. The three subunits that make the trimer are colored blue, cyan, and white. The sequence corresponding to the peptides is shown in the bottom right. **b**, An example showing the calculation process of a relative energetic binding score. The number of contacting atoms ($< 5\,\text{Å}$) and the number of clashing atoms ($< 1\,\text{Å}$) for each peptide in the competition are counted, and an absolute energetic binding score is calculated based on the counts according to Eq. 1. A difference between the two numbers, or the relative energetic binding score, is then calculated. The competition result between two peptides is determined using the average of relative binding scores across 10 replicates. The matrix of the mean scores is then used to rank the peptides of interest (POIs). **c**, A simplified geometrical representation of a peptide-receptor model, where the hull of the receptor is represented by an ellipsoid (blue). Point O: the center of mass of the receptor. Point A: the terminus of the receptor that is attached to an anchor. Segment OB: the minor axis of the ellipsoid receptor hull. Point C: the deepest point on the candidate peptide (orange). θ: the binding angle of the peptide. d: the binding pocket depth of the peptide. **d**, The angle constraint function. Three representative scenarios with different binding angles are highlighted. **e**, The depth constraint function. Three representative scenarios with different binding depths are highlighted. **f**, Comparison of the averaged relative binding energy scores before geometry-based adjustments vs. after adjustments. **g-i** Heatmaps representing the matrix of mean scores 22 AAV9-based capsid variants, including **g** mean absolute binding scores, **h** mean relative binding scores, and **i** mean relative binding scores that have considered both angle and depth constraints. All heatmap matrices were sorted by point-based round-robin tournaments (Methods). Bracketed numbers in the row labels are LY6A-binding profiles of the capsids inferred from experimental evidence (Figure S1). Each block in the heatmap represents the mean score measured from 10 independent models. **j-k**, comparison of different ranking methods used as binary classifiers to predict the LY6A-binding profile of 22 AAV9-based capsid variants.**j**, comparison between rankings given by different versions of APPRAISE scores using AF2-multimer as the structure prediction tool. **k**, comparison between rankings given by confidence scores of AF2-multimer versus rankings given by APPRAISE 1.2 using either AF2-multimer or ESMFold as prediction engines. The sequence and shape parameters of LY6A used for the modeling and analyses are included in Table S1.
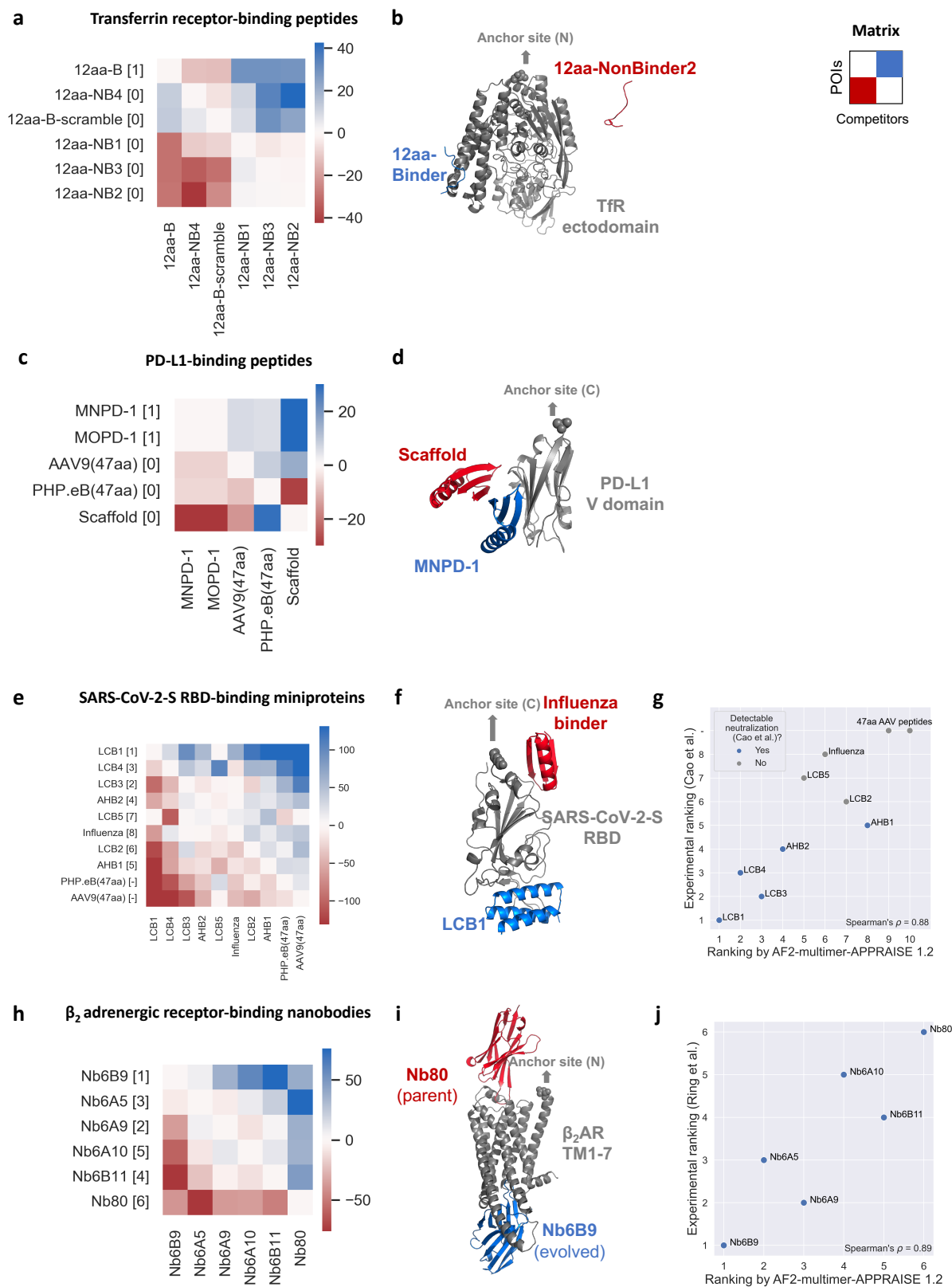
Figure 3: **AF2-multimer-APPRAISE 1.2 accurately ranks binding propensities of different classes of engineered proteins.** See the next page for the caption

Figure 3: **AF2-multimer-APPRAISE 1.2 accurately ranks binding propensities of different classes of engineered proteins. a-d,** APPRAISE 1.2 ranking of transferrin receptor-binding peptides and non-binding control peptides.[4] **a,** Pairwise score matrix and ranking of a panel of 12-aa peptides given by APPRAISE 1.2. Bracketed numbers in the row labels are experimentally determined transferrin receptor-binding profiles of each peptide.[4] **b,** A representative AF2-multimer model result of a binding peptide (blue) competing against a non-binding peptide (red) for binding to transferrin receptor. **c-d,** APPRAISE 1.2 ranking of PD-L1-binding peptides and non-binding control peptides.[7] **c,** Pairwise score matrix and ranking of a panel of 47-aa peptides given by APPRAISE 1.2. Bracketed numbers in the row labels show the PD-L1-binding profile of each peptide determined either experimentally (for MNPD-1, MOPD-1, and scaffold protein) or by expectation (for AAV9 and PHP.eB)..[7] **d,** A representative AF2-multimer model result of MNPD-1 (blue), a designed binding peptide, competing against a non-binding scaffold peptide (red) for binding to PD-L1. **e-g,** APPRAISE 1.2 ranking of SARS-CoV-2-S RBD-binding miniproteins.[5] **e,** Pairwise score matrix and ranking given by APPRAISE 1.2. Bracketed rankings in the row labels are determined based on experimentally measured IC50 of each miniprotein to neutralize live SARS-CoV-2.[5] **f,** A representative AF2-multimer model result of LCB1 (blue), a SARS-CoV-2-S RBD-binding miniprotein, competing against an influenza virus-binding miniprotein[6] (red). **g,** A scatter plot showing the correlation between APPRAISE-predicted ranking and experimentally-measured IC50 ranking of all miniproteins tested. Blue points highlight binders that showed the capability of complete neutralization of the SARS-CoV-2 virus in the tested range of concentration *in vitro*. **h-j,** APPRAISE 1.2 ranking of β_2 adrenergic receptor-binding nanobodies.[3] **h,** Pairwise score matrix and ranking given by APPRAISE 1.2. Bracketed numbers in the row labels are rankings of experimentally measured binding of each nanobody.[3]**i,** A representative AF2-multimer model result of Nb6B9 (blue), the strongest evolved binder to active β_2AR, competing against Nb80 (red), the nanobody used for the evolution. **j,** A scatter plot showing the correlation between APPRAISE-predicted ranking and experimentally-measured ranking by β_2AR binding of all nanobodies tested. Sequences and shape parameters of all receptors are included in Table S1. Each block in the heatmap represents the mean score measured from 10 independent models. For comparison, rankings given by AF2-multimer-APPRAISE 1.0, ESMFold-APPRAISE 1.2, and interface pTM of SARS-Cov2-S RBD binding miniproteins and β_2 adrenergic receptor-binding nanobodies are shown in Figure S3.

APPRAISE is generally applicable to diverse classes of engineered proteins

To determine the applicability of APPRAISE to different classes of engineered proteins, we applied the method to four classes of engineered protein binders targeting four representative receptors for therapeutics.

We first applied APPRAISE 1.2 to other short peptide binders (Figure 3a-d). In the first trial, the method successfully ranked a peptide selected by phage display to bind human transferrin receptor,[4] a well-characterized BBB receptor, over non-binding counterparts of the same length (Figure 3a). In the second trial, evaluating two 47aa-long, rationally-designed PD-L1 binding peptides[7]) against the scaffold and length-matched AAV variable region segments, both designed PD-L1 binding peptides were clear winners (3c).

We next tested whether APPRAISE 1.2 can be used to evaluate larger proteins, for example, computationally designed miniproteins (50-90 aa) that bind to the receptor-binding domain (RBD) of SARS-CoV-2 spike protein[5] (Figure 3e-g). Among the designed miniproteins, 5 can neutralize live SARS-CoV-2 virus *in vitro* with IC50 from $20\,\mathrm{pM}$ to $40\,\mathrm{nM}$.[5] The APPRAISE 1.2 rankings of the 5 neutralizing miniproteins matched well with their IC50 rankings (Spearman's $\rho = 0.90$, $p = 0.037$, Figure 3g). The predictive accuracy of AP-PRAISE decreased when non-neutralizing miniproteins were included (Spearman's $\rho = 0.88$, $p < 0.001$, Figure 3g); nevertheless, the top 4 binders still remained on the top.

Finally, we used APPRAISE to rank 6 nanobodies (120 aa) that were evolved experimentally[3] to bind to an activated conformation of β_2 adrenergic receptor (β_2AR), a G-protein-coupled receptor (GPCR) (Figure 3h-j). APPRAISE 1.2 correctly found the strongest evolved binder and placed the parent (the weakest binder among all candidates) at the bottom (Figure 3h). The overall predicted ranking correlated well with the ranking from experimentally determined binding readouts[3] (Spearman's $\rho = 0.89$, $p = 0.02$, Figure 3j). The ability of APPRAISE to predict the binding affinity of nanobodies, a widely-used therapeutic modality, expands its value for drug design and development.

We also tested the performance of alternative ranking methods on the miniprotein dataset and the nanobody dataset. AF2-multimer-APPRAISE1.2 again yielded the most accurate predictions when compared to AF2-multimer-APPRAISE 1.0, ESMFold-APPRAISE 1.2, or interface pTM scores given by AF2-multimer (Figure S3), reflected by higher Spearman's correlation to experimental rankings. It is worth noting that ESMFold-APPRAISE 1.2 failed completely with the miniprotein dataset (Figure S3b). Upon further inspection, we found that the failure can be explained by the unfolded SARS-Cov-2-S RBD in ESMFold-generated complex models.

Without any fine-tuning, AF2-multimer-APPRAISE 1.2 demonstrated consistent prediction ability for ranking all four classes of proteins, including experimentally-selected and rationally-designed peptides, computationally-designed miniproteins, and nanobodies.

HT-APPRAISE screening can identify novel receptor-dependent capsid variants

We next attempted to adapt APPRAISE 1.2 for *in silico* screening. The computational cost of the original pairwise competition grows quadratically with the number of input variants, which is unsuitable for high-throughput screening. To address this scalability issue, we designed a two-stage screening strategy we named high-throughput (HT)-APPRAISE (Figure 4a). The first stage aims to shrink the size of the variant library using a less accurate yet simpler modeling method. Variants are randomly pooled into groups of 4 and compete within each group. To reduce grouping bias, at least two sets of competitions with different groupings are performed in parallel. The variants are then ranked by their $\overline{B}_2^{POI}$. The number of competitions required for the first stage of HT-APPRAISE grows linearly with the number of variants in the starting library, making this stage highly scalable. In the second stage, the top variants selected from the first stage are subjected to standard pairwise competition with APPRAISE 1.2, yielding a matrix of $\overline{\Delta B}_2^{POI,competitor}$ and a more accurate ranking.

We tested our HT-APPRAISE *in silico* screening method for LY6A binding with a library

of 100 capsid variants (Figure 4b-c). This library is composed of 97 capsid variants randomly chosen from a list of 9,000 variants showing superior brain enrichment in C57BL/6J mice compared to the wild-type AAV9 capsid[2] as well as three spiked-in capsids: the variants PHP.B and PHP.D, and the wild-type AAV9 (Table S2). PHP.D is a brain-transducing capsid identified in a recent directed evolution campaign in our lab, the relevant receptor for which was unknown.

We completed the HT-APPRAISE screening within 24 hours using 3 parallel Google Colaboratory GPU sessions and a laptop computer. In both stages of the screening process, the most time-consuming step was the structural prediction, which took approximately 0.1-1 GPU minute per peptide-LY6A model (a complex made of 114 aa total). The time cost for structural analysis was negligible in comparison, taking less than 1 second per model on a CPU. After the first stage of screening, both PHP.D and PHP.B appeared in the top 15% of the library (ranked by $\overline{B}_2^{POI}$) (Figure 4b).). In the second stage, the top 18 capsids were ranked using pairwise APPRAISE 1.2 (Figure 4c). PHP.D and PHP.B were the 1st and the 4th in the final ranking.

The most intriguing aspect of PHP.D's result is that its variable region bears little sequence similarity to any of the LY6A-dependent variants used to develop the APPRAISE method (Figure 4d-e). To confirm this prediction result, we experimentally tested PHP.D's LY6A dependency. An *in vitro* viral infection assay showed that PHP.D indeed exhibits LY6A-enhanced transduction of HEK293T cells (Figure 4f). In addition, *in vivo* systemic delivery of PHP.D packaging a ubiquitously-expressed fluorescent protein revealed that the brain transduction capability of this capsid variant is restricted to LY6A-expressing mouse strains (Figure 4g). The ability of HT-APPRAISE to identify binders with distinct sequences highlights the generalizability of the physics-informed, sequence-structure/structure-function strategy.
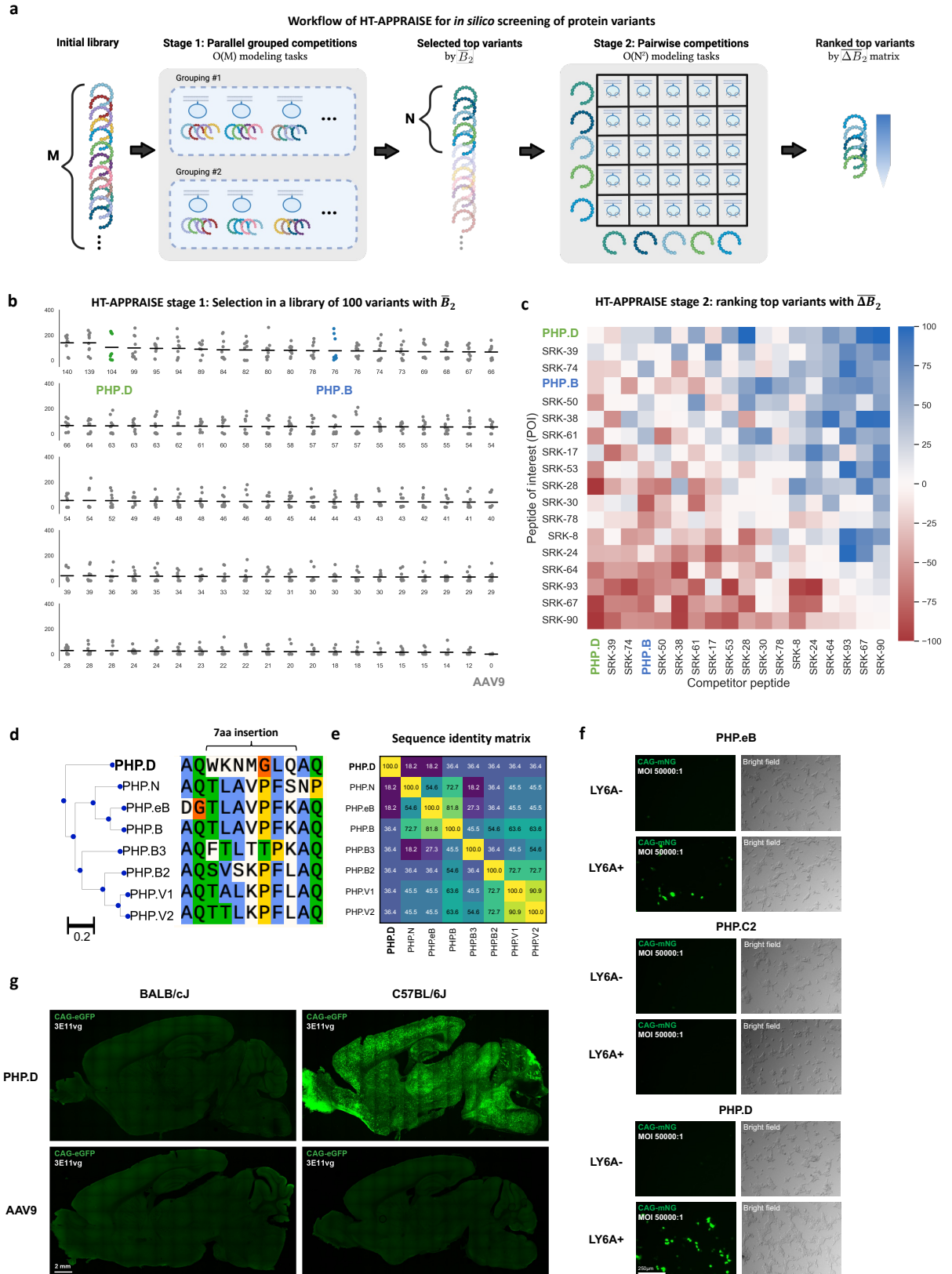
Figure 4: **An *in silico* HT-APPRAISE screening of a medium-sized AAV library identifies a LY6A-dependent variant with a distinct sequence.** See the next page for the caption.

Figure 4: **An *in silico* HT-APPRAISE screening of a medium-sized AAV library identifies a LY6A-dependent variant with a distinct sequence. a,** A schematic showing the two-stage strategy for efficient *in silico* screening of a variant library. In the first stage, M variants of interest are grouped into groups of 4 and compete against one another for receptor binding. At least two different groupings are used in parallel to reduce the grouping bias. Each peptide's mean absolute binding score is used for selecting the top N variants. In the second stage, the N variants compete pairwise using standard APPRAISE to get a more accurate ranking. **b-c,** Results from a proof-principle screening with 100 AAV9-based variants, including the wild-type control and variants with 7aa insertions. 97 variants were randomly chosen from a list of 9000 variants[2] that showed higher brain enrichment than the wild-type AAV9 after one round of screening in C57BL/6J mice. PHP.B and PHP.D, two known strong brain transducers, as well as wild-type AAV9, are spiked into the library. Peptide sequences used for the screening can be found in Table S2. **b,** Results from stage 1. Each variant's absolute binding scores from 10 separate models are shown as individual dots. The mean score is printed below the dots and shown as a horizontal bar. Scores of PHP.D (ranked 3rd), PHP.B (ranked 13th), and AAV9 (ranked 100th) are highlighted. **c,** Results from stage 2. Rows corresponding to Scores of PHP.D (ranked 1st), PHP.B (ranked 4th) are highlighted. Each block in the heatmap represents the mean score measured from 10 independent models. **d-g,** Characterization of PHP.D, a variant that tops the *in silico* screening. **d,** Sequence alignment and phylogenetic tree of known LY6A-dependent brain transducing variants. The sequence of PHP.D is very different from all other variants. The alignment and sequence distances were generated with Clustal Omega.[44] The colored alignment is plotted with Snapgene software. Blue: conserved hydrophobic residues; green: conserved hydrophilic residues; orange or yellow: conserved unique residues (glycine or proline). **e,** Sequence identity matrix of the LY6A-dependent variants. **f,** *In vitro* infectivity assay in HEK293T cells. PHP.D and PHP.eB showed LY6A-enhanced transduction, while the negative control PHP.C2 did not show LY6A-enhanced transduction. AAV capsids carrying a fluorescent protein expression cassette were applied to HEK293T cells either transfected with Ly6a or not at 5×10^8 vg per well in a 96-well plate. Images were taken 24hr after transduction. n=3 per condition. Scale bar, 250 μm **g,** *In vivo* brain transduction of PHP.D vs. AAV9 in two different mice strains. PHP.D showed strain-restricted transduction in only the LY6A+ strain, C57BL/6J. Viruses carrying a fluorescent protein expression cassette were injected intravenously at 3×10^{11} vg per animal, and the tissues were harvested and imaged 3 weeks after injection. n=3 per condition. Scale bar, 2 mm.

Discussion

Here we describe APPRAISE, a structure-based, physics-informed method that accurately ranks receptor binding propensities of engineered proteins. APPRAISE uses a competitive, pairwise modeling strategy to capture affinity differences between even proteins with similar sequences and takes into account biophysical and geometrical principles. Using state-of-the-art structure prediction tools like AF2-multimer, we show that APPRAISE's predictions are accurate and generalizable across five different classes of engineered proteins.

The competition-based structure modeling strategy addresses the challenge of assessing small differences in binding affinity with high accuracy. This challenge was highlighted by a recent benchmarking study using AF2 models for molecular docking of small-molecule antibiotic candidates. The authors reported that the prediction power of the direct physics-based scoring is no better than a random model.[45] By contrast, a competition-based modeling strategy might have helped cancel the shared noise and amplify the signal arising from the small affinity differences. The competition setup forces the structure-prediction neural network to put only the more probable binder close to the receptor (Figure 3b, d, f, i), converting a small probabilistic difference into a binary output.

A key feature of APPRAISE is that its analysis module uses only information stored in the 3D coordinates, making it potentially compatible with other computational tools for protein engineering. For example, the structure-prediction tool used in APPRAISE can, in theory, be replaced by any current or future structure-prediction tools. In addition, the scores generated by APPRAISE are complementary to confidence scores generated by structure-prediction neural networks that also predict binding.[12,14,17,46] Moreover, APPRAISE can be an orthogonal validation tool for structure-based protein design methods,[47–49] particularly those that rely on optimization of predicted confidence scores.[50]

The scalable, two-stage HT-APPRAISE strategy we designed allows *in silico* screening of protein candidates for receptor binding (Figure 4). Such screening can help prioritize leading candidates during drug discovery, reducing the huge time, financial, and environmental costs

of experimental validation. For example, the computational tasks needed for the screening of 100 AAV variants that we presented here (Figure 4) could be completed within 24hr with research-grade computational resources. Similar scaled *in vivo* characterization of the capsids would have taken several months. In addition, the scale of a similar screening could be dramatically expanded in an industry setting with high-performance computing resources.

APPRAISE has several limitations that are a result of competition-based ranking. One such limitation is that APPRAISE only outputs the relative, not the absolute, probability of binding. Therefore, a variant's position at the top of the final ranking does not, in and of itself, indicate that the variant has experimentally detectable binding affinity unless there are positive controls with known binding to compare against. Another limitation is that APPRAISE assumes that the binding of candidate proteins is mutually exclusive, though this can be proven false in some instances when the proteins bind to separate, far-apart epitopes.

Other limitations of APPRAISE may arise from the protein structure prediction engine that it relies on. For example, ESMFold-APPRAISE fails when the protein-language based prediction tool cannot fold the protein in a complex properly (Figure S3b), while AF2-multimer-APPRAISE results can possibly be biased by the specific selection of multiple-sequence alignments due to the dependence on co-evolutionary information by AF2-multimer. Moreover, the accuracy and speed of APPRAISE may be compromised when the modeled proteins contain long disordered regions or large domains that are unnecessary for binding. As a result, pre-screening of several truncated protein constructs for minimal folding domains with the particular structure prediction tool (similar to the common practice in structural biology) may sometimes be needed. Additionally, the APPRAISE method is not as effective in ranking very weak binders in a pool (e.g., Figure 3g), perhaps because the predicted structures do not offer many opportunities for meaningful interaction with the receptor in a competitive setting, resulting in near-zero B_2^{POI} values. Fortunately, this should not be a practical concern for most protein engineering applications since the most valuable candidates

are usually those that bind with higher affinities.

While APPRAISE has proven to be successful in ranking the binding propensities of different classes of protein variants, its accuracy and speed could be further improved. For example, the parameters of APPRAISE 1.2's scoring function have only been minimally tuned to proper orders of magnitudes to decrease the risk of over-fitting (Methods). With further fine-tuning of parameters and the ever-growing power of protein structure prediction, the APPRAISE method promises to streamline the process of engineering protein-based therapeutics.

Methods

APPRAISE Methodology

The code and demos for APPRAISE are available through a public Github repository: https://github.com/GradinaruLab/APPRAISE.

Modeling of peptide-receptor complexes using AF2-multimer

Peptide-receptor models are modeled using Colabfold (Python package index: alphafold-colabfold 2.1.14), an implementation of integrated multiple-sequence alignment generation with MMseqs2 and structure modeling with AF2-multimer-v2.[14,15,51–54] First, batches of *.fasta files containing combined receptor sequences (Table S1) and peptide sequences for the pairwise competition or pooled competition, where the protein chains are separated by the ":" symbol, are prepared using a customized Python function (appraise.input_fasta_prep.get_complex_fasta). Second, the *.fasta files are used as input files for the "batch" Jupyter notebook in the Co-labfold package, and the notebook is run on Google Colaboratory using a V100 SXM2 16GB GPU or an A100 SXM4 40GB GPU. The settings used for the modeling are listed below:

```
msa_mode = "MMseqs2 (UniRef+Environmental)"
```

```
num_models = 5

num_recycles = 3

stop_at_score = 100

use_custom_msa = False

use_amber = False

use_templates = True

model_type = "auto" #or "AF2-multimer-v2"
```

Modeling of peptide-receptor complexes using ESMFold

To model the peptide-receptor complexes using ESMFold, a process analogous to the one employed for AF2-multimer modeling is implemented. First, batches of *.fasta files containing combined receptor sequences (Table S1) and peptide sequences for the pairwise competition or pooled competition, where the protein chains are separated by a poly-glycine linker (30 glycine residues), are prepared using the same customized Python function (appraise.input_fasta_prep.get_complex_fasta()). Second, the *.fasta files are used as input files for a custom Jupyter notebook with codes adapted from the Colabfold package for batch modeling using ESMFold, and the notebook is run on Google Colaboratory using an A100 SXM4 40GB GPU. The custom Colab notebook is included in the APPRAISE package: `ColabFold_ESMFold_batch_run.ipynb`.

Physics-informed analysis of individual structure models

The output folder containing *.pdb files generated by alphafold-colabfold is downloaded to a local computer for processing. Key parameters in a predicted structure model are measured, and the measurements are used to generate binding scores for each peptide in a model.

Automated quantification of the peptide-receptor models

The structure models are analyzed using PyMOL 2.3.3 using a custom PyMOL script. Briefly, the script loads all *.pdb models in a directory, extracts metadata from the file names, and measures the relevant contact atom numbers, angles, and distances. The measurements are saved as a *.csv file. The custom PyMOL script is included in the APPRAISE package: `pymol_quantify_peptide_binding.py`.

Measurement of the R_{minor} of the receptor hull

The receptor shape parameter R_{minor}, which is necessary for APPRAISE 1.2, is obtained by measuring the shape parameters of an AlphaFold-modeled receptor structure. Briefly, the monomeric receptor (Table S1) is modeled using Colabfold (Python package index: alphafold-colabfold 2.1.14). The top model is then analyzed using HullRad v8.1[55] to obtain its major axis diameter D_{max} and aspect ratio P. R_{minor} is then calculated using the formula $R_{minor} = D_{max}/P/2$. Before the analysis, R_{minor} measurement is manually added as a column to pandas dataframe storing PyMOL measurements with a column title "R_minor".

Construction and calculation of $B_{energetic}$

We defined a contact atom to be a non-hydrogen atom of either the receptor or the peptide that is within 5Å of the binding partner in the peptide-receptor model since atoms within this distance cutoff are responsible for most protein-protein interactions.[56] We defined a clashing term to be the number of non-hydrogen atoms in the peptide that are within 1Å of the receptor since this distance is smaller than the typical diameter of an atom and can cause a huge Van der Waals strain. To find the suitable weight for the clashing term, we estimated the relative energy scales using Lennard-Jones potential and concluded that an order of magnitude of 10^3 should be proper (Eq. 1). Since most interfaces between the engineered peptide and the receptor have up to a few hundred of non-hydrogen atoms (tens of residues) in the interface, this heavy weight for the clashing atom practically sets the

24

B_0^{POI} of any peptide with steric clashing against the receptor to 0. In other words, Eq. 1 is practically equivalent to:

$$B_{energetic}^{peptide} = \begin{cases} N_{contact}^{peptide}, & \text{if } N_{clash}^{peptide} = 0 \\ 0, & \text{if } N_{clash}^{peptide} >= 1 \end{cases}$$

Construction and calculation of geometrical scores

The binding angle θ is defined as the angle between the vector from receptor center of mass to receptor anchor $\overrightarrow{OA}$ and the vector from receptor center of mass to peptide center of mass $\overrightarrow{OC'}$ (Figure 2f. Note that the peptide center of mass C' is usually very close to the deepest point C, and therefore point C and point C' are undifferentiated in this schematic). A steep function is used to penalize inaccessible binding angles that are close to the anchor point:

$$B_{angle}^{peptide} = \begin{cases} -10^3 \cdot (1 - \frac{\theta}{\frac{\pi}{2}})^{10}, & \text{if } \theta < \frac{\pi}{2} \\ 0, & \text{if } \frac{\pi}{2} <= \theta <= \pi \end{cases} \tag{5}$$

The definition of binding depth d is a simplification of previously defined travel depth:[57] we first calculate the hydrodynamic radius of the hull of the receptor at the minor axis (R_{minor}) using HullRad,[55] and then take the difference of the distance between the "closest point on the peptide" to the receptor center and R_{minor}. The ratio between the difference and R_{minor} is defined as the depth. In other words, binding depth $d = \frac{\|OB\| - \|OC\|}{\|OB\|}$ where $\|OB\|$ is the minor axis radius (in Å) of the receptor hull when considering it as an ellipsoid (Figure 2f), and $\|OC\|$ is the distance (in Å) between the center of mass of the receptor and the closest point on the peptide (Figure 2f). An odd polynomial function is used to construct the score to reflect both the positive effect of a deep binding pocket and the negative effect of a convex binding site:

25

$$B_{depth}^{peptide} = 10^2 \cdot d^3 \tag{6}$$

Calculation of scores for each peptide in a model

The total binding scores for each peptide in a model are calculated using Eqs. 1-4 in the main text.

Generation of the score matrix and a ranking

The total binding scores of a POI vs. a competitor across 10 replicate models are averaged to get $\overline{\Delta B}^{POI,competitor}$ ($\overline{\Delta B}_0^{POI,competitor}$ for APPRAISE 1.0, $\overline{\Delta B}_1^{POI,competitor}$ for APPRAISE 1.1, or $\overline{\Delta B}_2^{POI,competitor}$ for APPRAISE 1.2). These individual scores are then used to create a matrix and are plotted as a heatmap.

In the final score matrix, the POIs are ranked using a point-based round-robin tournament system[58] to avoid the bias caused by individual competitions with unusually high scores. Briefly, each $\overline{\Delta B}^{POI,competitor}$ in the matrix is considered as the match result between a POI and a competitor. A POI gains 1 point for winning over each match and loses 1 point for losing each match. (In the cases when $|\overline{\Delta B}^{POI,competitor}|$ does not reach the threshold of $p < 0.05$ using a one-sample, two-sided, Student's t test (degree of freedom=9), the match is called a tie, and the POI gets 0 points from the match.)

Demo

A Jupyter notebook demonstrating the procedures of APPRAISE, using the HT-APPRAISE screening (Figure 4) as an example, is included in the APPRAISE package:

`demo/HT-APPRAISE_demo.ipynb`

Additional demo notebooks for other applications can be found under the `additional_misc_demos` folder.

Experimental Validations

In vitro infectivity assay

HEK293T cells were seeded in 6-well plates at 80% confluency and maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 5% fetal bovine serum (FBS), 1% non-essential amino acids (NEAA), and 100 U/mL penicillin-streptomycin at 37°C in 5% CO2. Cells were transiently transfected with 2.53 µg plasmid DNA encoding an expression cassette for the LY6A receptor. The following day, receptor-expressing cells were transferred to black, clear bottom 96-well plates at 20% confluency and maintained in FluoroBrite™ DMEM supplemented with 0.5% FBS, 1% NEAA, 100 U/mL penicillin-streptomycin, 1x GlutaMAX, and 15 µM HEPES at 37°C in 5% CO2. Engineered AAV variants packaging a CAG-mNeonGreen transgene were dosed in triplicate at 5E8 vg per well once the cells were attached. Plates were imaged 24 hours after AAV was introduced to cells with a Keyence BZ-X700 using a 4x objective and NucBlue™ Live ReadyProbes™ Reagent (Hoechst 33342) to autofocus each well.

In vivo mouse experiment

For all the experiments performed in this study, the animals were randomly assigned, and the experimenters were not blinded while performing the experiments unless mentioned otherwise. All animal procedures in mice were approved by the California Institute of Technology Institutional Animal Care and Use Committee (IACUC), Caltech Office of Laboratory Animal Resources (OLAR) and were carried out in accordance with guidelines and regulations.

For the profiling of the novel AAVs in mice (C57BL/6J and BALB/cJ), the AAV vectors were injected intravenously via the retro-orbital route to 6-8 week old adult mice at a dose of 3×10^{11} vg per mouse. Retro-orbital injections were performed as described previously.[59] To harvest the tissues of interest after 3 weeks of expression, the mice were anesthetized with Euthasol (pentobarbital sodium and phenytoin sodium solution, Virbac AH) and tran-

scardially perfused using 50 mL of 0.1 M phosphate buffered saline (PBS) (pH 7.4), followed by 50 mL of 4% paraformaldehyde (PFA) in 0.1 M PBS. The organs were collected and post-fixed 24 h in 4% PFA at 4°C. Following this, the tissues were washed with 0.1 M PBS and stored in fresh PBS-azide (0.1 M PBS containing 0.05% sodium azide) at 4°C. Before imaging, the 100 µm slices of tissue were cut using a Leica VT1000S. Brain images were acquired with a Zeiss LSM 880 confocal microscope using a Plan-Apochromat 10× 0.45 M27 (working distance 2.0 mm) objective. Zen Black 2.3 SP1 was used to process the images.

Ethical approval

All animal procedures in mice were approved by the California Institute of Technology Institutional Animal Care and Use Committee (IACUC), Caltech Office of Laboratory Animal Resources (OLAR) and were carried out in accordance with guidelines and regulations.

Acknowledgement

Supporting Information Available

Supplementary Figures and Supplementary Tables

References

(1) Deverman, B. E.; Pravdo, P. L.; Simpson, B. P.; Kumar, S. R.; Chan, K. Y.; Banerjee, A.; Wu, W.-L.; Yang, B.; Huber, N.; Pasca, S. P.; Gradinaru, V. Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nature biotechnology* **2016**, *34*, 204–209.

(2) Ravindra Kumar, S.; Miles, T. F.; Chen, X.; Brown, D.; Dobreva, T.; Huang, Q.; Ding, X.; Luo, Y.; Einarsson, P. H.; Greenbaum, A.; Jang, M. J.; Deverman, B. E.; Gradinaru, V. Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. *Nature Methods* **2020**, *17*, 541–550.

(3) Ring, A. M.; Manglik, A.; Kruse, A. C.; Enos, M. D.; Weis, W. I.; Garcia, K. C.; Kobilka, B. K. Adrenaline-activated structure of the 2-adrenoceptor stabilized by an engineered nanobody. *Nature* **2013**, *502*, 575–579.

(4) Lee, J. H.; Engler, J. A.; Collawn, J. F.; Moore, B. A. Receptor mediated uptake of peptides that bind the human transferrin receptor. *European Journal of Biochemistry* **2001**, *268*, 2004–2012.

(5) Cao, L.; Goreshnik, I.; Coventry, B.; Case, J. B.; Miller, L.; Kozodoy, L.; Chen, R. E.; Carter, L.; Walls, A. C.; Park, Y.-J.; Strauch, E.-M.; Stewart, L.; Diamond, M. S.; Veesler, D.; Baker, D. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **2020**, *370*, 426–431, Publisher: American Association for the Advancement of Science.

(6) Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* **2017**, *550*, 74–79, Number: 7674 Publisher: Nature Publishing Group.

(7) Yin, H.; Zhou, X.; Huang, Y.-H.; King, G. J.; Collins, B. M.; Gao, Y.; Craik, D. J.; Wang, C. K. Rational Design of Potent Peptide Inhibitors of the PD-1:PD-L1 Interac-

tion for Cancer Immunotherapy. *Journal of the American Chemical Society* **2021**, *143*, 18536–18547, Publisher: American Chemical Society.

(8) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacological Reviews* **2014**, *66*, 334–395, Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Review Article.

(9) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature Methods* **2019**, 1.

(10) Bedbrook, C. N.; Yang, K. K.; Rice, A. J.; Gradinaru, V.; Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Computational Biology* **2017**, 1–21.

(11) Bolon, D. N.; Voigt, C. A.; Mayo, S. L. De novo design of biocatalysts. *Current Opinion in Chemical Biology* **2002**, *6*, 125–129.

(12) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**,

(13) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876, Publisher: American Association for the Advancement of Science.

(14) Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2021**,

(15) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein folding accessible to all. *Nature Methods* **2022**,

(16) Aderinwale, T.; Christoffer, C.; Kihara, D. RL-MLZerD: Multimeric protein docking using reinforcement learning . 2022; `https://www.frontiersin.org/articles/10.3389/fmolb.2022.969394`.

(17) Motmaen, A.; Dauparas, J.; Baek, M.; Abedi, M. H.; Baker, D.; Bradley, P. Peptide binding specificity prediction using fine-tuned protein structure prediction networks. 2022; `https://www.biorxiv.org/content/10.1101/2022.07.12.499365v1`, Pages: 2022.07.12.499365 Section: New Results.

(18) Ruffolo, J. A.; Chu, L.-S.; Mahajan, S. P.; Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. 2022; `https://www.biorxiv.org/content/10.1101/2022.04.20.488972v1`, Pages: 2022.04.20.488972 Section: New Results.

(19) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* **2022**, 2022.07.20.500902.

(20) Li, Z.; Liu, X.; Chen, W.; Shen, F.; Bi, H.; Ke, G.; Zhang, L. Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv* **2022**, 2022.08.04.502811.

(21) Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; Peng, J. High-resolution ¡em¿de novo¡/em¿ structure prediction from primary sequence. *bioRxiv* **2022**, 2022.07.21.500999.

(22) Wang, G.; Fang, X.; Wu, Z.; Liu, Y.; Xue, Y.; Xiang, Y.; Yu, D.; Wang, F.; Ma, Y. HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle. 2022; `https://arxiv.org/abs/2207.05477`.

(23) Cheng, S.; Wu, R.; Yu, Z.; Li, B.; Zhang, X.; Peng, J.; You, Y. FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. 2022; `https://arxiv.org/abs/2203.00854`.

(24) Liu, S. et al. PSP: Million-level Protein Sequence Dataset for Protein Structure Prediction. 2022; `https://arxiv.org/abs/2206.12240`.

(25) Bennett, N.; Coventry, B.; Goreshnik, I.; Huang, B.; Allen, A.; Vafeados, D.; Peng, Y. P.; Dauparas, J.; Baek, M.; Stewart, L.; DiMaio, F.; De Munck, S.; Savvides, S. N.; Baker, D. Improving ¡em¿de novo¡/em¿ Protein Binder Design with Deep Learning. *bioRxiv* **2022**, 2022.06.15.495993.

(26) Chang, L.; Perez, A. Ranking Peptide Binders by Affinity with AlphaFold. *Angewandte Chemie* **2022**,

(27) Baker, D. What has de novo protein design taught us about protein folding and biophysics? *Protein Science* **2019**, *28*, 678–683.

(28) Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017**, *13*, 3031–3048.

(29) Gonzalez, T. R.; Martin, K. P.; Barnes, J. E.; Patel, J. S.; Ytreberg, F. M. Assessment of software methods for estimating protein-protein relative binding affinities. *PLOS ONE* **2020**, *15*, e0240573.

(30) Chan, K. Y.; Jang, M. J.; Yoo, B. B.; Greenbaum, A.; Ravi, N.; Wu, W. L.; Sánchez-Guardado, L.; Lois, C.; Mazmanian, S. K.; Deverman, B. E.; Gradinaru, V. Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nature Neuroscience* **2017**, *20*, 1172–1179.

(31) Nonnenmacher, M.; Wang, W.; Child, M. A.; Ren, X.-Q.; Huang, C.; Ren, A. Z.; Tocci, J.; Chen, Q.; Bittner, K.; Tyson, K.; Pande, N.; Chung, C. H.-Y.; Paul, S. M.; Hou, J. Rapid evolution of blood-brain-barrier-penetrating AAV capsids by RNA-driven biopanning. *Molecular Therapy - Methods Clinical Development* **2021**, *20*, 366–378.
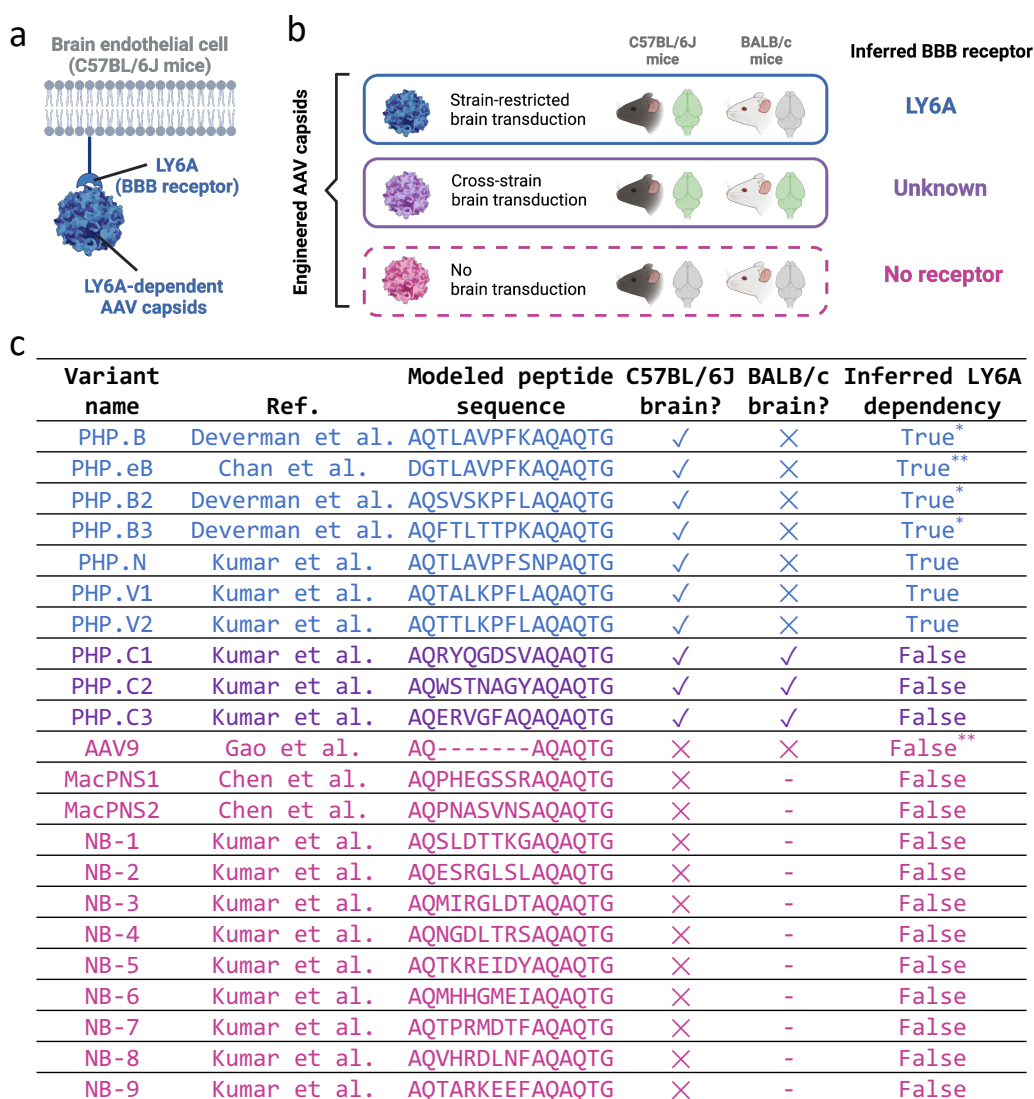
(32) Goertsen, D. et al. AAV capsid variants with brain-wide transgene expression and decreased liver targeting after intravenous delivery in mouse and marmoset. *Nature Neuroscience* **2022**, *25*, 106–115.

(33) Chuapoco, M. R. et al. Intravenous gene transfer throughout the brain of infant Old World primates using AAV. *bioRxiv* **2022**, 2022.01.08.475342.

(34) Chen, X. et al. Engineered AAVs for non-invasive gene delivery to rodent and non-human primate nervous systems. *Neuron* **2022**, *110*, 2242–2257.e6.

(35) Gao, G.; Vandenberghe, L. H.; Alvira, M. R.; Lu, Y.; Calcedo, R.; Zhou, X.; Wilson, J. M. Clades of Adeno-Associated Viruses Are Widely Disseminated in Human Tissues. *Journal of Virology* **2004**, *78*, 6381–6388.

(36) Huang, Q.; Chan, K. Y.; Tobey, I. G.; Chan, Y. A.; Poterba, T.; Boutros, C. L.; Balazs, A. B.; Daneman, R.; Bloom, J. M.; Seed, C.; Deverman, B. E. Delivering genes across the blood-brain barrier: LY6A, a novel cellular receptor for AAV-PHP.B capsids. *PLOS ONE* **2019**, *14*, e0225206.

(37) Hordeaux, J.; Yuan, Y.; Clark, P. M.; Wang, Q.; Martino, R. A.; Sims, J. J.; Bell, P.; Raymond, A.; Stanford, W. L.; Wilson, J. M. The GPI-Linked Protein LY6A Drives AAV-PHP.B Transport across the Blood-Brain Barrier. *Molecular Therapy* **2019**, *27*, 912–921.

(38) Batista, A. R.; King, O. D.; Reardon, C. P.; Davis, C.; Shankaracharya,; Philip, V.; Gray-Edwards, H.; Aronin, N.; Lutz, C.; Landers, J.; Sena-Esteves, M. Ly6a Differential Expression in Blood–Brain Barrier Is Responsible for Strain Specific Central Nervous System Transduction Profile of AAV-PHP.B. *Human Gene Therapy* **2019**, *31*, 90–102.

(39) Xu, G.; Zhang, R.; Li, H.; Yin, K.; Ma, X.; Lou, Z. Structural basis for the neurotropic AAV9 and the engineered AAVPHP.eB recognition with cellular receptors. *Molecular Therapy - Methods Clinical Development* **2022**, *26*, 52–60.

(40) Jang, S.; Shen, H. K.; Ding, X.; Miles, T. F.; Gradinaru, V. Structural basis of receptor usage by the engineered capsid AAV-PHP.eB. *Molecular Therapy - Methods Clinical Development* **2022**, *26*, 343–354.

(41) Morrone, J. A.; Perez, A.; MacCallum, J.; Dill, K. A. Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. *Journal of Chemical Theory and Computation* **2017**, *13*, 870–876.

(42) Bolon, D. N.; Grant, R. A.; Baker, T. A.; Sauer, R. T. Specificity versus stability in computational protein design. *Proceedings of the National Academy of Sciences* **2005**, *102*, 12724–12729.

(43) Copeland, R. A. The drug–target residence time model: a 10-year retrospective. *Nature Reviews Drug Discovery* **2016**, *15*, 87–95.

(44) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **2011**, *7*, 539.

(45) Wong, F.; Krishnan, A.; Zheng, E. J.; Stärk, H.; Manson, A. L.; Earl, A. M.; Jaakkola, T.; Collins, J. J. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology* **2022**, *18*, e11081.

(46) Bryant, P.; Pozzati, G.; Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications* **2022**, *13*, 1265.

(47) Ovchinnikov, S.; Huang, P.-S. Structure-based protein design with deep learning. *Mechanistic Biology * Machine Learning in Chemical Biology* **2021**, *65*, 136–144.

(48) Anand, N.; Eguchi, R.; Mathews, I. I.; Perez, C. P.; Derry, A.; Altman, R. B.; Huang, P.-

S. Protein sequence design with a learned potential. *Nature Communications* **2022**, *13*, 746.

(49) Dauparas, J. et al. Robust deep learning based protein sequence design using Protein-MPNN. *bioRxiv* **2022**, 2022.06.03.494563.

(50) Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **2022**, *377*, 387–394.

(51) Steinegger, M.; Meier, M.; Mirdita, M.; V"ohringer, H.; Haunsberger, S. J.; S"oding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **2019**, *20*, 473.

(52) Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **2019**,

(53) Mirdita, M.; Steinegger, M.; S"oding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **2019**, *35*, 2856–2858.

(54) Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; S"oding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **2017**, *45*, D170–D176.

(55) Fleming, P. J.; Fleming, K. G. HullRad: Fast Calculations of Folded and Disordered Protein and Nucleic Acid Hydrodynamic Properties. *Biophysical Journal* **2018**, *114*, 856–869.

(56) Salamanca Viloria, J.; Allega, M. F.; Lambrughi, M.; Papaleo, E. An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. *Scientific Reports* **2017**, *7*, 2838.

(57) Coleman, R. G.; Sharp, K. A. Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. *Journal of Molecular Biology* **2006**, *362*, 441–458.

(58) McCarthy, C. A.; Benjamin, A. T. Determinants of the Tournaments. *Mathematics Magazine* **1996**, *69*, 133–135.

(59) Challis, R. C.; Ravindra Kumar, S.; Chan, K. Y.; Challis, C.; Beadle, K.; Jang, M. J.; Kim, H. M.; Rajendran, P. S.; Tompkins, J. D.; Shivkumar, K.; Deverman, B. E.; Gradinaru, V. Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nature Protocols* **2019**, *14*, 1–1.

Supplementary Figures

a, Brain endothelial cell (C57BL/6J mice)
LY6A (BBB receptor)
LY6A-dependent AAV capsids

b, Engineered AAV capsids
C57BL/6J mice | BALB/c mice | Inferred BBB receptor
Strain-restricted brain transduction — LY6A
Cross-strain brain transduction — Unknown
No brain transduction — No receptor

c,

Variant name	Ref.	Modeled peptide sequence	C57BL/6J brain?	BALB/c brain?	Inferred LY6A dependency
PHP.B	Deverman et al.	AQTLAVPFKAQAQTG	✓	✗	True[*]
PHP.eB	Chan et al.	DGTLAVPFKAQAQTG	✓	✗	True[**]
PHP.B2	Deverman et al.	AQSVSKPFLAQAQTG	✓	✗	True[*]
PHP.B3	Deverman et al.	AQFTLTTPKAQAQTG	✓	✗	True[*]
PHP.N	Kumar et al.	AQTLAVPFSNPAQTG	✓	✗	True
PHP.V1	Kumar et al.	AQTALKPFLAQAQTG	✓	✗	True
PHP.V2	Kumar et al.	AQTTLKPFLAQAQTG	✓	✗	True
PHP.C1	Kumar et al.	AQRYQGDSVAQAQTG	✓	✓	False
PHP.C2	Kumar et al.	AQWSTNAGYAQAQTG	✓	✓	False
PHP.C3	Kumar et al.	AQERVGFAQAQAQTG	✓	✓	False
AAV9	Gao et al.	AQ-------AQAQTG	✗	✗	False[**]
MacPNS1	Chen et al.	AQPHEGSSRAQAQTG	✗	-	False
MacPNS2	Chen et al.	AQPNASVNSAQAQTG	✗	-	False
NB-1	Kumar et al.	AQSLDTTKGAQAQTG	✗	-	False
NB-2	Kumar et al.	AQESRGLSLAQAQTG	✗	-	False
NB-3	Kumar et al.	AQMIRGLDTAQAQTG	✗	-	False
NB-4	Kumar et al.	AQNGDLTRSAQAQTG	✗	-	False
NB-5	Kumar et al.	AQTKREIDYAQAQTG	✗	-	False
NB-6	Kumar et al.	AQMHHGMEIAQAQTG	✗	-	False
NB-7	Kumar et al.	AQTPRMDTFAQAQTG	✗	-	False
NB-8	Kumar et al.	AQVHRDLNFAQAQTG	✗	-	False
NB-9	Kumar et al.	AQTARKEEFAQAQTG	✗	-	False

* LY6A binding had been confirmed by cell binding assay
** LY6A dependency had been biochemical and genetic assays

Figure S1: **Prior experimental studies revealing the receptor dependency of some brain-transducing AAV variants. a,** A schematic showing an AAV capsid binding to a blood-brain barrier receptor (BBB) receptor that is only expressed at a high level on the endothelial cells of certain mouse strains.[36–38] **b,** A schematic showing how we can infer whether an AAV capsid can use LY6A, a mouse BBB receptor, by characterizing its brain transduction across different strains. A capsid that has strain-restricted brain transduction in C57BL/6J mice is likely LY6A-dependent, while a capsid that has cross-strain brain transduction or does not transduce the brain is not LY6A-dependent. **c,** A table summarizing all 22 capsids used in Figure 2 with their source literature, sequence, brain transduction profile, and inferred LY6A dependency.
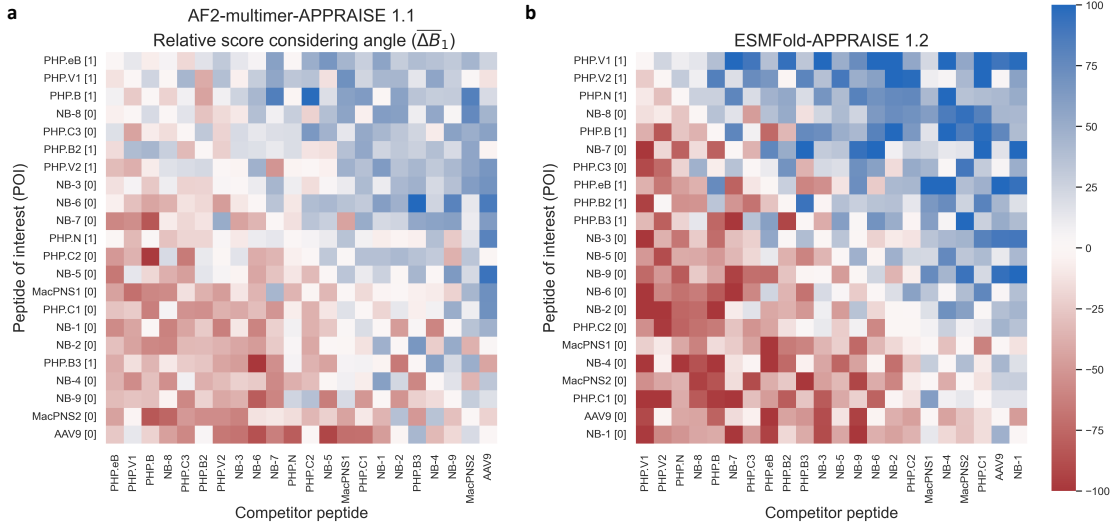
Figure S2: **Heatmaps representing score matrices of AF2-multimer-APPRAISE 1.1 and ESMFold-APPRAISE 1.2.**
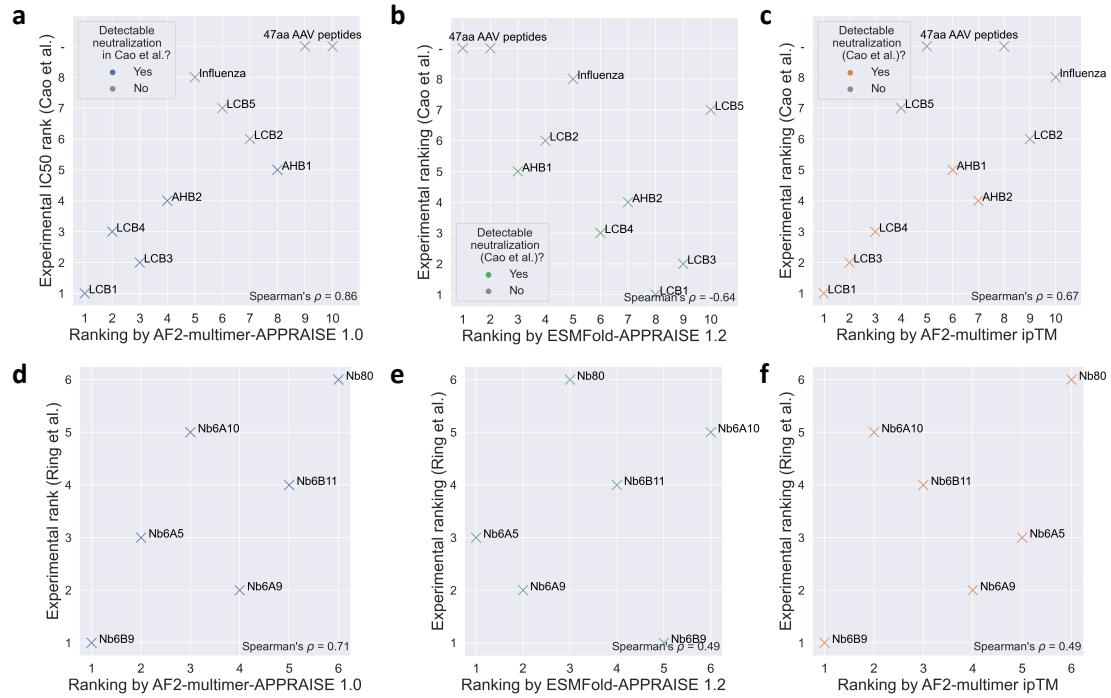

Figure S3: **Ranking protein binders using alternative methods.** Rankings of two groups of peptides analyzed in Figure 3g, j based on **a, d)** AF2-multimer-APPRAISE 1.0, **b, e)** ESMFold-APPRAISE 1.2, and **c, f)** interface pTM given by AlphaFold-multimer.

2

Supplementary Tables

Table S1: Receptor sequences and parameters used for APPRAISE analysis

Receptor name	Organism	Uniprot accession ID	Domain used for modeling	Residue indices (start-end)	D_{max} (Å) *	Axial ratio *	R_{minor} (Å) *	Anchor site	Sequence used for modeling
Ly6a	Mus musculus (Mouse)	P05533	Mature protein	27–110	46.68	1.74	13.4	C-term	LECYQCYGVPFETSCPSITCPYPDGV CVTQEAAVIVDSQTRKVKNNLCLPIC PPNIESMEILGTKVNVKTSCCQEDLC NVAVP
PD-L1	Homo sapiens (Human)	Q9NZQ7	V domain	18–132	46.50	1.51	15.4	C-term	AFTVTVPKDLYVVEYGSNMTIECKFP VEKQLDLAALIVYWEMEDKNIIQFVH GEEDLKVQHSSYRQRARLLKDQLSLG NAALQITDVKLQDAGVYRCMISYGGA DYKRITVKVNA
Beta-2 adrenergic receptor	Homo sapiens (Human)	P07550	TM1-TM7	29–342	93.8	2.23	21.0	N-term	DEVWVVGMGIVMSLIVLAIVFGNVLV ITAIAKFERLQTVTNYFITSLACADL VMGLAVVPFGAAHILMKMWTFGNFWC EFWTSIDVLCVTASIETLCVIAVDRY FAITSPFKYQSLLTKNKARVIILMVW IVSGLTSFLPIQMHWYRATHQEAINC YANETCCDFFTNQAYAIASSIVSFYV PLVIMVFVYSRVFQEAKRQLQKIDKS EGRFHVQNLSQVEQDGRTGHGLRRSS KFCLKEHKALKTLGIIMGTFTLCWLP FFIVNIVHVIQDNLIRKEVYILLNWI GYVNSGFNPLIYCRSPDFRIAFQELL CL
Transferrin receptor 1	Homo sapiens (Human)	P02786	Ectodomain	122–760	86.50	1.47	29.4	N-term	LYWDDLKRKLSEKLDSTDFTGTIKLL NENSYVPREAGSQKDENLALYVENQF REFKLSKVWRDQHFVKIQVKDSAQNS VIIVDKNGRLVYLVENPGGYVAYSKA ATVTGKLVHANFGTKKDFEDLYTPVN GSIVIVRAGKITFAEKVANAESLNAI GVLIYMDQTKFPIVNAELSFFGHAHL GTGDPYTPGFPSFNHTQFPPSRSSGL PNIPVQTISRAAAEKLFGNMEGDCPS DWKTDSTCRMVTSESKNVKLTVSNVL KEIKILNIFGVIKGFVEPDHYVVVGA QRDAWGPGAAKSGVGTALLLKLAQMF SDMVLKDGFQPSRSIIFASWSAGDFG SVGATEWLEGYLSSLHLKAFTYINLD KAVLGTSNFKVSASPLLYTLIEKTMQ NVKHPVTGQFLYQDSNWASKVEKLTL DNAAFPFLAYSGIPAVSFCFCEDTDY PYLGTTMDTYKELIERIPELNKVARA AAEVAGQFVIKLTHDVELNLDYERYN SQLLSFVRDLNQYRADIKEMGLSLQW LYSARGDFFRATSRLTTDFGNAEKTD RFVMKKLNDRVMRVEYHFLSPYVSPK ESPFRHVFWGSGSHTLPALLENLKLR KQNNGAFNETLFRNQLALATWTIQGA ANALSGDVWDIDNEF
Spike	SARS-CoV-2	P0DTC2	RBD	331–529	67.93	1.87	18.2	C-term	NITNLCPFGEVFNATRFASVYAWNRK RISNCVADYSVLYNSASFSTFKCYGV SPTKLNDLCFTNVYADSFVIRGDEVR QIAPGQTGKIADYNYKLPDDFTGCVI AWNSNNLDSKVGGNYNYLYRLFRKSN LKPFERDISTEIYQAGSTPCNGVEGF NCYFPLQSYGFQPTNGVGYQPYRVVV LSFELLHAPATVCGPKK

Table S2: Peptides used for *in silico* screening

Peptide name	Peptide sequence (residues 587-594 in VP1)	Original source
AAV9	AQ-------AQAQTG	Gao et al., 2002
PHP.B	AQTLAVPFKAQAQTG	Deverman et al. 2016
PHP.D	AQWKNMGLQAQAQTG	Unpublished in vivo selection
SRK-1	AQLYHGGSTAQAQTG	Kumar et al. 2020
SRK-2	AQNNSVRQLAQAQTG	Kumar et al. 2020
SRK-3	AQVNSTRNVAQAQTG	Kumar et al. 2020
SRK-4	AQGNMTKFTAQAQTG	Kumar et al. 2020
SRK-5	AQTAIQPPKAQAQTG	Kumar et al. 2020
SRK-6	AQITTDQPFAQAQTG	Kumar et al. 2020
SRK-7	AQDTANTARAQAQTG	Kumar et al. 2020
SRK-8	AQTHDAQAWAQAQTG	Kumar et al. 2020
SRK-9	AQQPLAEEAAQAQTG	Kumar et al. 2020
SRK-10	AQTALANQKAQAQTG	Kumar et al. 2020
SRK-11	AQTGTERLSAQAQTG	Kumar et al. 2020
SRK-12	AQNGVTQSKAQAQTG	Kumar et al. 2020
SRK-13	AQWTEQRLVAQAQTG	Kumar et al. 2020
SRK-14	AQDTGLNNRAQAQTG	Kumar et al. 2020
SRK-15	AQPLPPTSIAQAQTG	Kumar et al. 2020
SRK-16	AQSDPGKFMAQAQTG	Kumar et al. 2020
SRK-17	AQTTMGTMLAQAQTG	Kumar et al. 2020
SRK-18	AQKQTQDSSAQAQTG	Kumar et al. 2020
SRK-19	AQLAHNSALAQAQTG	Kumar et al. 2020
SRK-20	AQVVPSTYRAQAQTG	Kumar et al. 2020
SRK-21	AQFRHLTGAAQAQTG	Kumar et al. 2020
SRK-22	AQSANLLSSAQAQTG	Kumar et al. 2020
SRK-23	AQFSNTHALAQAQTG	Kumar et al. 2020
SRK-24	AQFNSKLQLAQAQTG	Kumar et al. 2020
SRK-25	AQFKTNISAAQAQTG	Kumar et al. 2020
SRK-26	AQYPVPLKQAQAQTG	Kumar et al. 2020
SRK-27	AQHVNHMAPAQAQTG	Kumar et al. 2020
SRK-28	AQIVSNQMSAQAQTG	Kumar et al. 2020
SRK-29	AQPRPERMYAQAQTG	Kumar et al. 2020
SRK-30	AQNMKIQHVAQAQTG	Kumar et al. 2020
SRK-31	AQNTNVPAMAQAQTG	Kumar et al. 2020
SRK-32	AQSAQLRSSAQAQTG	Kumar et al. 2020
SRK-33	AQSHHEQVSAQAQTG	Kumar et al. 2020
SRK-34	AQGATGHLTAQAQTG	Kumar et al. 2020
SRK-35	AQHNLRDSIAQAQTG	Kumar et al. 2020
SRK-36	AQGPGTSFKAQAQTG	Kumar et al. 2020
SRK-37	AQSPPVQGLAQAQTG	Kumar et al. 2020
SRK-38	AQTLYNAIHAQAQTG	Kumar et al. 2020
SRK-39	AQLGDITGFAQAQTG	Kumar et al. 2020
SRK-40	AQGFNSMKPAQAQTG	Kumar et al. 2020
SRK-41	AQSNGLNGLAQAQTG	Kumar et al. 2020
SRK-42	AQVRIPGALAQAQTG	Kumar et al. 2020
SRK-43	AQDMGTDNLAQAQTG	Kumar et al. 2020
SRK-44	AQNYATKSQAQAQTG	Kumar et al. 2020
SRK-45	AQSVTTSHVAQAQTG	Kumar et al. 2020
SRK-46	AQTSGTDGIAQAQTG	Kumar et al. 2020
SRK-47	AQARTAHGYAQAQTG	Kumar et al. 2020
SRK-48	AQHSANMSKAQAQTG	Kumar et al. 2020
SRK-49	AQHDERANMAQAQTG	Kumar et al. 2020
SRK-50	AQNNFNASLAQAQTG	Kumar et al. 2020
SRK-51	AQSASLVSHAQAQTG	Kumar et al. 2020
SRK-52	AQAPRIDNAAQAQTG	Kumar et al. 2020
SRK-53	AQLTSSNALAQAQTG	Kumar et al. 2020
SRK-54	AQTLNSIRAAQAQTG	Kumar et al. 2020
SRK-55	AQSGTGRQQAQAQTG	Kumar et al. 2020
SRK-56	AQKTTLASGAQAQTG	Kumar et al. 2020
SRK-57	AQMRVNTEEAQAQTG	Kumar et al. 2020
SRK-58	AQFETLHKTAQAQTG	Kumar et al. 2020
SRK-59	AQTQHRFEMAQAQTG	Kumar et al. 2020
SRK-60	AQHTAEKAPAQAQTG	Kumar et al. 2020
SRK-61	AQNHMVRELAQAQTG	Kumar et al. 2020
SRK-62	AQRFQPSSAAQAQTG	Kumar et al. 2020
SRK-63	AQRSVANVPAQAQTG	Kumar et al. 2020
SRK-64	AQVFQATRTAQAQTG	Kumar et al. 2020
SRK-65	AQEQRTPSPAQAQTG	Kumar et al. 2020
SRK-66	AQGSSTASLAQAQTG	Kumar et al. 2020
SRK-67	AQQVPHLHSAQAQTG	Kumar et al. 2020
SRK-68	AQPSQPYTKAQAQTG	Kumar et al. 2020
SRK-69	AQTHTRDQGAQAQTG	Kumar et al. 2020
SRK-70	AQINPGITLAQAQTG	Kumar et al. 2020
SRK-71	AQLQPTKSSAQAQTG	Kumar et al. 2020
SRK-72	AQQDAKVTTAQAQTG	Kumar et al. 2020
SRK-73	AQGASTHNAAQAQTG	Kumar et al. 2020
SRK-74	AQIPVSIQAAQAQTG	Kumar et al. 2020
SRK-75	AQVTSAHPVAQAQTG	Kumar et al. 2020
SRK-76	AQTASLIASAQAQTG	Kumar et al. 2020
SRK-77	AQDRGTRTVAQAQTG	Kumar et al. 2020
SRK-78	AQTAYLEVKAQAQTG	Kumar et al. 2020
SRK-79	AQATTQMSSAQAQTG	Kumar et al. 2020
SRK-80	AQKYDASQSAQAQTG	Kumar et al. 2020
SRK-81	AQTGTSHLHAQAQTG	Kumar et al. 2020
SRK-82	AQTMTPSGIAQAQTG	Kumar et al. 2020
SRK-83	AQTPSSSGNAQAQTG	Kumar et al. 2020
SRK-84	AQKDVVNSNAQAQTG	Kumar et al. 2020
SRK-85	AQRSPATMLAQAQTG	Kumar et al. 2020
SRK-86	AQYDQKSLAAQAQTG	Kumar et al. 2020
SRK-87	AQMGARNLPAQAQTG	Kumar et al. 2020
SRK-88	AQLPISATEAQAQTG	Kumar et al. 2020
SRK-89	AQTRHTSLTAQAQTG	Kumar et al. 2020
SRK-90	AQNKLTANGAQAQTG	Kumar et al. 2020
SRK-91	AQNGDSHSHAQAQTG	Kumar et al. 2020
SRK-92	AQVRTDMDMAQAQTG	Kumar et al. 2020
SRK-93	AQSVSTPRGAQAQTG	Kumar et al. 2020
SRK-94	AQVSRQFEPAQAQTG	Kumar et al. 2020
SRK-95	AQSANNVRGAQAQTG	Kumar et al. 2020
SRK-96	AQIGTKSTNAQAQTG	Kumar et al. 2020
SRK-97	AQGSELRTGAQAQTG	Kumar et al. 2020