

DeepScan - an AI platform for discovery and design of potent cell surface display elements

Corresponding Author: Professor Sidi Chen

Parts of this Peer Review File have been redacted as indicated as we could not obtain permission to publish the reports from one of the peer reviewers.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This ambitious and timely study assembles an end-to-end AI pipeline—DeepScan → Genesis (Dawn/Dusk) → Omni (5-in-1)—to learn sequence–function rules of cell-surface display (CSD) elements and to generate de novo CSDs with superior cell-surface translocation (CST) strength. The scale and rigor are noteworthy: the authors design ~3,700 gCSDs, benchmark >540 models, and curate ~45 training datasets spanning ~170k sequences. The experimental part is also strong, with arrayed in vitro validation across multiple antigens and cell lines, appropriate held-out and independent test sets, and transparent statistical procedures. Impressively, seven gCSDs achieve CST strengths on similar or exceeding—the best natural CSDs, underscoring practical impact for vaccines, synthetic receptors, and cellular therapies.

Overall, the work offers a practical, generalizable tool to CSD optimization that could accelerate mRNA vaccine design, CAR/synthetic receptor engineering, and surface protein display across modalities. The demonstration that learned sequence rules translate into multi-antigen gains is especially compelling for translational pipelines. The combination of scale, methodological care, orthogonal validations, and interpretable biological signals makes this a strong contribution. My suggestions below are minor and aimed at further enhancing clarity and usability.

1. Figure 2f vs 2g discrepancy. The text suggests that the predicted CSD distributions mirror the screened CSD results. However, the separation between SPTM-non-CM and CST0–2 appears absent in the predicted CSDs while it is clearly present in the screened data. Please reconcile this or clarify whether this reflects a modeling limitation, class imbalance, or post-hoc normalization.
2. Self-training ablation. Please add a summary of self-training label counts vs. performance for Omni. A concise plot/table showing how performance scales with pseudo-label volume/quality would guide practitioners adopting this component.
3. Accessibility for non-computational users. In addition to releasing code, consider a user-friendly web portal to run the core inference/design workflows (with guardrails and example inputs). This would substantially broaden impact and facilitate adoption in biology labs.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The manuscript by Fang and Chen describes a pipeline for generation and discovery of new single-pass transmembrane anchor domains. If we could develop an AI/ML approach to predict transmembrane display anchors reliably and generally, it could enable applications in synthetic surface display and receptor engineering. They use flow-cytometry screening and a variety of machine learning models to predict and design sequences in order to achieve high levels of mammalian cell surface display. They screen ~100 natural modules fused to two antigens (A29 and E7) in two cell lines, and ultimately claim

27 generative potent cell surface display (CSD) modules that outperform the majority of those tested.

In my opinion the study falls short in several areas, and I do not expect that in the current form it would be among the top ranked papers in this field for the calendar year. I found the methodology to be somewhat lacking in terms of rigor. Some of the main methodological issues are related to small and biased training sets relative to the massive design space they are applied to, as well as limited independent validation across multiple (more than 2) cell types and antigens. At several points in the complicated ML pipeline, it appears they rely on pseudo-labels (they call it “supervised self-training”) which can inflate performance metrics. The validation of the various claims was therefore less than I hoped for. The communication style and heavy use of acronyms makes the work extremely difficult to follow and understand.

Major points:

1. Heavy use of acronyms: Key acronyms are introduced late, unclearly, or not at all. For example, “SPTM-nonCM” is used early and frequently, but they do not provide a clear definition or labeling rule for this negative class. But this class is an important one for classifier training. The definition must be inferred from this sentence: “Proteins in the Swiss-prot database are categorized into non-transmembrane proteins (nonTM) and transmembrane proteins (TM), which are further categorized into single-pass (SP) and multi-pass (MP) TM on cell membrane (CM) or endomembrane systems (nonCM).” From this, the reader should understand what “SPTM-nonCM” means, but this sentence is already ambiguous, and blends biological and protein topological characteristics. Much of the language in the paper lacks clarity and precision.
2. The training and validation of the DeepScan CST model relies on a small screen and needs more validation. The CST classifier was trained on ~110-constructs of A29 antigen. The text claims that 24/26 (92%) predicted CST3 were validated. The validation is very limited (i.e. same same two lines 293T/Lenti-X as used to produce the training labels). It does not establish the generality of these membrane anchors across cell types, antigens, or assay formats. Stronger evidence would include validation in other cell types or using alternative readouts and antigens.
3. Self-training (pseudo-labels): If I am understanding this text correctly, then the Omni models are trained with labels generated by averaging logits from the Genesis models. Even if the authors keep fixed test sets, those sets share similarities with sequences used for pseudo-labeling. This will tend to bias the performance metrics to be too optimistic. They should exclude pseudo-labeled sequences with high identity to examples in the test-set. They should also use a truly external test set that is assembled independently.
4. Generative design models (Genesis Dusk / ESM3) lack pre-specified selection criteria. The round-2 designs were selected for validation using a mixture of model scores, sequence heuristics, and “inferred biophysical propensities” (page 8). In this case, selection bias is hard to exclude. It was not clear how the synthetic (generative) sequences were assembled prior to model-based in silico screening. The validation was only performed in HEK293T/Lenti-X with two antigens (A29 and E7), which limits the claims of generality. They should provide the full ranked list (it was unclear to me if this was contained in the supporting files). The validation should be in 3 or more cell types and with 3 or more unrelated antigens. They should also report on functionality of the displayed antigens (not only antibody labeling). It was not clear if an expression tag was used for labeling or if the antigens themselves are labeled and detected.
5. There is a mismatch in terms of the labeled training data vs. very large query/design space. They apply models trained on ~100 or so labeled sequences (plus homologs) to tens to hundreds of thousands of sequences. This is not a problem per se, but should be done using out-of-distribution (OOD) analysis and uncertainty estimates. They should demonstrate that the query sets are similarly distributed as the training set (depending on representation this can take the form of PCA analysis). They can therefore consider adding OOD metrics, embedding-space analysis (e.g., train-test nearest-neighbor identity and ESM embedding distances).
6. The manuscript states they used standard ANOVAs with multiple-comparison tests but provides too few details on the exact n per construct. The text misspells Tukey’s test (spelled “Turkey’s”) and “Dunnett” tests; ensure consistency and justification.
7. Code and raw data should all be made available (partly this was done).
8. The clarity and quality of communication is poor. The manuscript is too difficult to read. There are too many dense acronyms. The various definitions (DeepScan, Genesis/Omni tasks, CST classes) are scattered. The layout of the manuscript maximizes confusion (figures are at the end; captions are separated from figures; etc...). I do not know if this is a journal requirement for submission, but it makes the task of the reviewer much more difficult. There are not sufficient plain language descriptions of the assay and methods. Furthermore, the claims about applications (mRNA vaccines, receptors, therapeutics) are not supported by data. A thorough rewrite is needed to clarify the work (especially the ML part), and limit the scope of claims to what is experimentally demonstrated.

Minor points

- They should provide early, formal definitions for the acronyms (or reduce their use) e.g., SPTM-CM vs SPTM-nonCM, CST0–5, “hTMC,” and each Genesis/Omni subtask (ID, Topo, CST, TM, Quant).
- The ML reporting and calibration should include ROC/PR curves, reliability diagrams, and class-wise confusion matrices.
- The supplementary information refers to mouse experimentation and mouse licenses, but as far as I can tell no mice were used in this study.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors present a systematic approach to designing cell surface display (CSD) elements for therapeutic applications, particularly mRNA vaccines. They begin by screening natural CSD sequences, develop multiple generations of deep

learning models (DeepSCan, Genesis, and Omni), and ultimately validate generative CSDs that match or exceed the performance of natural sequences.

This paper is a mixed bag. It starts strong but loses focus midway through. The problem is compelling and the first half of the paper is on a promising trajectory: collect and analyze natural sequences, build predictive models, recognize their limitations, then iterate. However, after Figure 3, the work diverges into an unfocused exploration of model architectures and training strategies that obscures rather than advances the core insights. The paper would have been considerably stronger if it had maintained its initial focus on understanding what makes CSDs effective and systematically exploiting those principles.

The systematic curation of SPTM-CM sequences in Figures 1-2 is well-executed, with clear datasets and thoughtful experimental design. The identification of polybasic residues at the ICD-TMD junction as a key feature is valuable. However, Figure 2F reveals a significant limitation that the authors don't adequately address: their models primarily distinguish extreme CST values (CST0 vs CST3) but show poor resolution for intermediate activities. This suggests the problem formulation needs reconsideration. A regression approach with appropriately transformed targets might better capture the continuous nature of CST strength, or alternatively, a binary classification at a meaningful threshold could provide clearer decision boundaries. The current multi-class approach may miss the ordinal nature of the categories.

Unfortunately, things go haywire from Fig 4 onwards. The dataset construction becomes progressively more opaque: we encounter "8 Genesis datasets" with vague descriptions of "augmented FL and ECD variants, CSD homologs...and proportional negative controls." We're presented with over 540 models across three "generations" without systematic ablation studies to determine which components actually matter. Basic baselines are missing - would simple CXCL+EREK combos work? The proliferation of naming schemes (Dawn, Dusk, Omni) adds unnecessary complexity. The multi-task formulation remains unclear-- what exactly are the five tasks in the Omni model and why were these chosen? Fig 5a was difficult for me to understand. The "self-training" approach using predicted labels from earlier models lacks theoretical grounding and risks amplifying model biases. There's no clean held-out validation set for hyperparameter selection, making it impossible to assess whether the chosen architectures are actually optimal or just overfitted to the particular data splits.

There also seem to be some troubling misunderstandings about ML models. In Fig 4b, the "linear" and "dense" models are conceptually equivalent: multiple linear layers without intervening nonlinearities are mathematically equivalent to a single linear transformation. This raises questions about what was actually implemented. What does "attention pooling" mean, exactly?

The generative modeling is substantially oversold. Genesis Dusk essentially performs off-the-shelf masked language modeling. The elaborate description of "forging new routes in the dark space of artificial CSD sequences" is indicative of the sometimes-overwrought writing. The constant switching between metrics (PE integral, normalized PE integral, CST categories, MSE, accuracy) without justification makes it difficult to track what's actually being optimized.

Model interpretability was a missed opportunity. There is insufficient discussion of the dominance of CXCL16 and EREG variants among successful gCSDs. What makes these sequences special? The limited attention weight analysis hints at the importance of the ICD-TMD interface but the authors don't pursue this systematically. A simpler approach, e.g., creating chimeras combining CXCL16 and EREG features or inserting their key motifs into other natural sequences, might achieve similar or better results.

I can't help but feel that this paper might actually have been stronger if it ended at Figure 3, presenting a focused empirical study with clear findings.

(Remarks on code availability)

Author Rebuttal letter:

Point by point response to reviewers' comments:

We thank reviewers for their constructive suggestions and valuable feedback. While complete details are described in the point-by-point response and the actual manuscript/figures, below we summarize the key changes in the revised manuscript, in three main categories (new experiments, new computational analyses, ms clarity).

I. New experiments of independent, orthogonal and functional validations.

1. Performed new validation experiments of DeepSCan-designed gCSDs using five diCerenet cell lines, three unrelated display proteins, and ~20 CSD chimeras, all of which validated the gCSDs' performance.
2. Performed new validation experiments of gCSDs using an orthogonal assay (confocal microscopy) as independent from the previous flow cytometry data.
3. Performed new functional validation experiments, establishing gCSD's biological impact on antigen surface expression and thereby CAR-T cell cytotoxicity in a co-culture

assay.

II. New computational analyses.

4. Built DeepSCan webserver that predicts cell surface translocation (CST) strength, topology, and membrane domain of protein sequences in FASTA format based on deep learning models. Reviewers can test this webserver here, DeepSCan Cell Surface Display Sequence Predictor (<https://deepscan.sidichenlab.org>).
5. Performed self-training ablation studies, observing model performance decay (MSE increase) as the number of self-training labels decreased.
6. For interpretability, we analyzed sequence features at the TMD–ICD junction in the labeled natural screen, which distilled the screens and DeepScan AI models into concrete design rules and shared motifs between strong natural and designed CSDs, strengthening the biological insight.
7. Ensured low similarity between pseudo-labeled training data and the independent test set using MMSeqs2 and pair-wise identity matrices.
8. Analyzed similarity between Genesis 8k training data and 170k query sequences using PCA analysis, OOD metrics and nearest neighbor distance.
9. Demonstrated CSD applications in a functional CAR-T cytotoxicity assay.
10. Evaluated DeepSCan models using ROC curves, reliability diagram, confusion matrices.
11. Reanalyzed Fig. 2F using binary classification and regression approaches

III. Manuscript clarity enhancements:

12. Further breaking down diYerent CST categories to enhance resolution in Fig. 2f–2g.
13. Simplified acronym use by separating common acronyms from custom acronyms and removing dispensable terms.
14. Streamlined original main Figs. 4–6 by consolidating them into new Fig. 4 to maintain initial focus on empirical study and sequence features that make CSD effective. Relocated the additional into extended and supplementary figures.

15. Discussed sequence features associated with the dominance of EREG- and CXCL16-derived CSDs.
16. Discussed linear, dense models and the definition of attention pooling.
17. Clarified details in manuscript and methods section, including Genesis datasets.
18. Added sample number per construct in ANOVA, corrected misspelling, shared code and raw data, removed mice-related description in methods.
19. Defined key model names and moved other model names to supplementary files

We believe these revisions substantially strengthen the manuscript, by validating gCSDs with new independent, orthogonal and functional approaches demonstrating that AI-designed CSDs support biologically productive immune recognition rather than surface expression alone; by a series of computational analyses that enhanced the rigor; and by sharpening its focus and improving methodological transparency.

The following point-to-point responses (in dark blue text) are provided below each original comment (in black) made by all reviewers:

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

This ambitious and timely study assembles an end-to-end AI pipeline—DeepSCan → Genesis (Dawn/Dusk) → Omni (5-in-1)—to learn sequence–function rules of cell-surface display (CSD) elements and to generate de novo CSDs with superior cell-surface translocation (CST) strength. The scale and rigor are noteworthy: the authors design ~3,700 gCSDs, benchmark >540 models, and curate ~45 training datasets spanning ~170k sequences. The experimental part is also strong, with arrayed in vitro validation across multiple antigens and cell lines, appropriate held-out and independent test sets, and transparent statistical procedures. Impressively, seven gCSDs achieve CST strengths on similar or exceeding—the best natural CSDs, underscoring practical impact for vaccines, synthetic receptors, and cellular therapies.

Overall, the work offers a practical, generalizable tool to CSD optimization that could

accelerate mRNA vaccine design, CAR/synthetic receptor engineering, and surface protein display across modalities. The demonstration that learned sequence rules translate into multi-antigen gains is especially compelling for translational pipelines. The combination of scale, methodological care, orthogonal validations, and interpretable biological signals makes this a strong contribution. My suggestions below are minor and aimed at further enhancing clarity and usability.

Response: We thank reviewer 1 for this positive feedback and valuable suggestions. We have resolved Fig. 2f-2g discrepancy, conducted self-training ablation analysis, and built a web portal for the DeepScan platform. Details of these changes can be found in the revised manuscript and point-to-point response below.

1. Figure 2f vs 2g discrepancy. The text suggests that the predicted CSD distributions mirror the screened CSD results. However, the separation between SPTM-non-CM and CST0–2 appears absent in the predicted CSDs while it is clearly present in the screened data. Please reconcile this or clarify whether this reflects a modeling limitation, class imbalance, or post-hoc normalization.

Response: We thank reviewer for bringing up this discrepancy between Fig. 2f-2g and suggesting potential factors behind this discrepancy. Our results showed that predicted CST3 elements mirrored conserved characteristics of screened CST3 modules. Although our results section did not directly discuss CST0-2 group, we notice that the original Fig. 2f-2g combined CST1 and CST2 categories into one group (CST0-2), of which CST1 or CST2 composition may introduce variations in ICD Arg + Lys percentage. To better illustrate ICD Arg + Lys percentage in each CST category, we separated the CST1 and CST2 and found there is a trend of increasing ICD Arg + Lys percent from CST1 to CST3 in both screened and predicted CSDs, reflecting DeepScan model's capability to capture features of CSD

modules with varying potency. Indeed, this is due to CST class imbalance, and we have updated Fig. 2f-2g to reconcile the discrepancy in Fig. 2f-2g.

2. Self-training ablation. Please add a summary of self-training label counts vs. performance for Omni. A concise plot/table showing how performance scales with pseudo-label volume/quality would guide practitioners adopting this component.

Response: We sincerely appreciate reviewer 1 for this insightful suggestion. First, suggested by other reviewers and editor, we streamlined the manuscript to maintain a focus on empirical CSD optimization, and have moved Omni model and self-training related data to Extended Data Figure 9 and Supplementary Fig. 17. We have updated the related Supplementary Fig. 17 and its data summary file to include additional parameters for self-training ablation study. Based on this additional information, we plotted the new Supplementary Fig. 17 showing the self-training label ablation eCets on model performance measured by MSE between predicted vs. experimental CST labels. In this figure containing ~900 Omni models, we observed a steady-state baseline of MSE at tick marks of 2000 to 4000 self-training labels and as the self-training labels decrease, MSE increases, which is more evident in the 68k group without CSD homolog labels. Notably, MSE reached a local minimum when 2000 self-training labels were used.

Supplementary Fig. 17

3. Accessibility for non-computational users. In addition to releasing code, consider a user-friendly web portal to run the core inference/design workflows (with guardrails and example inputs). This would substantially broaden impact and facilitate adoption in biology labs.

Response: Thank you for this great suggestion. We have established this user-friendly web portal that runs the latest DeepScan model inferences including Genesis Quant and Omni 5-in-1 models. Reviewers and readers can access this web portal through this link, <https://deepscan.sidichenlab.org/>. It is capable of batch-processing protein sequences in FASTA format and reliably predicts input sequence characteristics, including CST scores,

CST category, topology, and transmembrane domain boundaries. Moreover, its predictions are highly consistent with the model outputs in data summary files (examples in Fig. 4d-4e) or annotations in Uniprot. We have also included the source codes and models in GitHub (as a supplementary zip file to this manuscript), to be released upon publication for the community's broad adoption.

Below we show a snapshot of DeepSCan front end design (Supplementary Fig. S19) and reviewers can cross validate its outputs based on sequences in data summary file, Fig. 4d-4e.

Supplementary Fig. 19

In Supplementary Fig. 18, we evaluated and compared Genesis and Omni models' performance in CST category classification through confusion matrices. CST category of Genesis Quant and Omni Quant models were derived by dividing the continuous logits with five cutoC values shown in Fig. 4a, including normalized Log PE integral of 0.3, 0.55, 0.8, 1, 1.2. Since the separation of CST0 and non-zero categories are highly accurate ($\geq 98\%$ row-normalized), our primary focus is whether DeepSCan models can identify potent CSD modules (CST3-5) among non-zero categories. This question can be directly addressed by calculating area ratios of accurately predicted CST3-5 count over total CST1-5 count. This parameter showed that Genesis Quant 8k model achieved highest accuracy of potent module prediction (53%). For this reason, our DeepSCan web portal shows the Genesis Quant-derived CST category and compares logits of Genesis Quant and Omni Quant models.

Supplementary Fig. 18

On the web page, users can directly view the sequence information and prediction results. Additionally, they have the option to download results in CSV file, which also records predicted TMD sequences. The current web portal is deployed on AWS EC2 vCPU instances and can process ~1000 tokens or 6-8 protein sequences of 100-200 amino acids in one batch in 30 seconds. And its processing capability can be readily boosted using the uploaded scripts and on-cloud or in-house GPU units owned by users.

Reviewer #2 (Remarks to the Author):

The manuscript by Fang and Chen describes a pipeline for generations and discovery of new single-pass transmembrane anchor domains. If we could develop an AI/ML approach to predict transmembrane display anchors reliably and generally, it could enable applications in synthetic surface display and receptor engineering. They use flow-cytometry screening and a variety of machine learning models to predict and design sequences in order to achieve high levels of mammalian cell surface display. They screen ~100 natural modules fused to two antigens (A29 and E7) in two cell lines, and ultimately claim 27 generative potent cell surface display (CSD) modules that outperform the majority of those tested.

In my opinion study falls short in several areas, and I do not expect that in the current form it would be among the top ranked papers in this field for the calendar year. I found the methodology to be somewhat lacking in terms of rigor. Some of the main methodological issues are related to small and biased training sets relative to the massive design space they are applied to, as well as limited independent validation across multiple (more than 2) cell types and antigens. At several points in the complicated ML pipeline, it appears they rely on pseudo-labels (they call it "supervised self-training") which can inflate performance metrics. The validation of the various claims was therefore less than I hoped for. The communication style and heavy use of acronyms makes the work extremely difficult to follow and understand.

Response: We thank reviewer 2 for pointing out the limitations in our study and offering

constructive feedback to improve our manuscript. As outlined in “summary of revision” in the beginning of this rebuttal letter, we have since performed substantial new work to address major concerns and improve various, including the following areas pointed out by the reviewer: training set size, independent validations using various experimental approaches, pseudo-labels evaluation, rigor of the computational work, structure of the paper and communication styles. Below we provide detail point-by-point responses:

Major points:

1. Heavy use of acronyms: Key acronyms are introduced late, unclearly, or not at all. For example, “SPTM-nonCM” is used early and frequently, but they do not provide a clear definition or labeling rule for this negative class. But this class is an important one for classifier training. The definition must be inferred from this sentence: “Proteins in the Swiss-prot database are categorized into non-transmembrane proteins (nonTM) and transmembrane proteins (TM), which are further categorized into single-pass (SP) and multi-pass (MP) TM on cell membrane (CM) or endomembrane systems (nonCM).” From this, the reader should understand what “SPTM-nonCM” means, but this sentence is already ambiguous and blends biological and protein topological characteristics. Much of the language in the paper lacks clarity and precision.

Response: We thank reviewer for pointing out the lack of clarity in acronym-related languages.

We agree that SPTM-nonCM and SPTM-CM are important acronyms that should be better defined in the main text. To address this concern, we define the SPTM-CM and SPTM-nonCM early in the first paragraph of results section: “Based on UniProt transmembrane domain features and subcellular location annotations, proteins are categorized into non-transmembrane proteins (nonTM) and transmembrane proteins (TM). And TM are further separated into single-pass (SP) and multi-pass (MP) TMs on cell membrane (CM) or other endomembrane systems (nonCM). For example, SPTM-CM designates a group of SPTM that are annotated as on cell membrane, virion, virion membrane, extracellular space, extracellular matrix, cell outer membrane, cell surface, host cell membrane or apical cell membrane. And SPTM-nonCM is the complementary subset of SPTM-CM in the broader category of SPTM and it includes SPTM localized on other endomembrane systems and SPTM annotated with ambiguous subcellular localization, such as “proteins on membrane”.

To comprehensively evaluate the acronym use, we listed all acronyms in the main text in the last version. In total the main text contains ~25 acronyms, including 10 commonly used acronyms, 11 custom acronyms and 4 rarely used acronyms.

In the revised manuscript, we ensure that full terms of all acronyms are introduced at their first appearance in the main text or results section. Notably, we removed four acronyms that rarely appear in main text and replaced them with their full terms. The common acronyms frequently appear in other research articles and are widely accepted within scientific community. Therefore, they are kept in our manuscript. For the custom acronyms that are introduced by this study, they designate important concepts in our analysis and are repeatedly used in the manuscript, necessitating their use to ensure conciseness and avoid redundancy. While we used a number of custom acronyms, they can be easily grouped into three types including protein classification related acronyms, cell surface display related acronyms and other frequent acronyms. Most acronyms’ full terms are self-explanatory. For important acronyms such as SPTM-CM and SPTM-nonCM, we also provided more precise definitions as guidelines for curating these protein categories within UniProt database.

10 Common acronyms (including approximate counts):

1. Machine learning related: ML (Machine Learning, 10 counts), AI (Artificial Intelligence, 2 counts), genAI (Generative Artificial Intelligence, 4 counts), DL (Deep Learning, 15 counts)
2. Model names: ID (Identity, 18 counts), Topo (Topological, 16 counts), TM (Transmembrane, 20 counts).
3. SPTM domains: ICD (Intracellular Domain, 10 counts), ECD (Ectodomain, 13 counts), TMD (Transmembrane Domain, 28 counts).
4. Other frequent acronyms: MSE (Mean Squared Error, 8 counts), GO (Gene Ontology, 5 counts).

11 Custom acronyms:

1. Protein classification related acronyms: A set of five acronyms including SPTM-CM (single-pass transmembrane protein on cell membrane, 35 counts), SPTM-nonCM (3 counts), MPTM-CM (1 count), nonTM (1 count).
2. Cell surface display related acronyms: a set of four acronyms including CSD (Cell Surface Display, 34 counts), naturally occurring CSD (nCSD, 9 counts), generative CSD (gCSD, 17 counts), CST (Cell Surface Translocation, 39 counts). The CST0-5 categories are defined in Fig. 4a, which shows the five cut-offs separating CST0-5, namely normalized Log PE integral 0.3, 0.55, 0.8, 1, 1.2.
3. Other frequent acronyms: DL1 (Double Label 1, 7 counts), Quant (Quantitative, 21 counts).
4. Model names (codename rather than acronym):
 - a. DeepSCan: Deep learning model used to scan and identify surface protein sequences, 48 counts
 - b. Genesis Dawn: a new beginning of training strategies used to build a series of models that shed light on the surface protein classic characteristics, 2 counts, related figures were moved from original Fig. 5 to ED. Fig. 9
 - c. Genesis Dusk: training strategy update to improve models' ability to generate SPTM-CM like sequences and explore dark space of artificial CSD sequences.
 - d. Omni: it comes from Latin word "omnis", which implies unified deep learning models for TM, ID, CST, Quant and topology tasks.

4 Removed acronyms:

1. NBME (Nature Biomedical Engineering): Appeared three times. Replaced it with reference 9 (published, not in press anymore).
2. scFv (single-chain variable fragment): Appeared once. This acronym has been removed.
3. LNP (Lipid Nanoparticle): Appeared twice. Use of acronyms is not necessary.
4. MVP (Modular Vaccine Platform): Appeared three times. Replaced it with full term.

We hope the revised paper has improved clarity, including the use of well-defined acronyms only when necessary or helpful.

2. The training and validation of the DeepSCan CST model relies on a small screen and needs more validation. The CST classifier was trained on ~110-constructs of A29 antigen. The text claims that 24/26 (92%) predicted CST3 were validated. The validation is very limited (i.e. same two lines 293T/Lenti-X as used to produce the training labels). It does not establish the generality of these membrane anchors across cell types, antigens, or assay formats. Stronger evidence would include validation in other cell types or using alternative readouts and antigens.

Response: We thank reviewer 2 for raising model validation concern and offering constructive suggestions. Indeed, the overarching theme of our study is to iteratively train and validate DeepSCan models to improve its robustness and generality in CSD module prediction and generation. The DeepSCan CST model is a first-generation DeepSCan classification model evaluating input sequence's membrane sorting signal strength and it

was built based on DeepSCan ID and Topo framework with incorporation of ~110 type I CSD and ~40 type II CSD modules' experimental CST labels. The training dataset was further augmented with CSD modules' homologs and full-length variants, resulting in a total of ~1400 CST non-zero class labels. The experimental and augmented non-zero labels together constitute ~10% of training labels, which are within a reasonable range given the expense and effort required to obtain experimental labels.

As suggested by the reviewer, we have performed substantial new validation experiments for these membrane anchors, (1) across cell types, (2) with new antigens, and (3) with new orthogonal validation assay formats (new Fig. 5), in order to further strengthen their generality. Brief summary of our new validations are below:

Performed new validation experiments of DeepSCan-designed gCSDs across 5 different cell lines of distinct types

We first focused on the top-performing round 2 gCSDs to assess reproducibility across cellular contexts and to evaluate whether gCSD-mediated antigen surface display supports downstream immune activation. Quantitative evaluation across five distinct cell lines (HCT116, H1299, HeLa, K562, and Raji) demonstrated that the top three round 2 gCSDs fused to A29 exhibited consistent and reproducible surface expression across cellular contexts, as measured by CST-based flow cytometry (Fig. 5a).

Performed new validation experiments of gCSDs using an orthogonal assay (confocal microscopy)

To confirm that CST measurements reflected genuine cell-surface localization, we performed confocal microscopy under non-permeabilizing conditions in A29-transfected HEK293T cells. gCSD-fused constructs exhibited robust membrane-localized signal, whereas A29 expressed without a gCSD or mScarlet alone showed no detectable surface staining, confirming gCSD-dependent membrane display (Fig. 5b, c).

Performed new functional validation experiments (new type of cancer antigen surface expression and CAR-T killing)

We next tested whether gCSDs support modular surface presentation of biologically relevant antigens. As a clinically established immunotherapy target, we selected B-cell maturation antigen (BCMA), a tumor-associated surface protein targeted by CAR-T therapies. To enable modular display, BCMA was truncated to its extracellular domain and fused to candidate gCSDs. gCSD-mediated surface presentation of BCMA was observed in both HEK293T and Raji cells by flow cytometry and confocal imaging, whereas BCMA expressed without a gCSD failed to localize to the cell surface, indicating that membrane presentation was strictly gCSD-dependent (Fig. 5d–f).

To functionally evaluate gCSD-mediated antigen presentation, we generated BCMA-directed CAR-T cells via lentiviral transduction and co-cultured them with Raji-PL target cells expressing either a truncated, cytoplasmic BCMA construct lacking a gCSD (cytoBCMA) or BCMA fused to candidate gCSDs (mbBCMA) to enforce cell-surface localization (Fig. 5g). For downstream assays, we selected EREG_5_Ca9 as the candidate gCSD due to its relatively high transfection efficiency in Raji cells. Prior to functional assays, expression and antigen-binding capacity of the BCMA CAR, containing a novel scFv developed in-house, were validated by flow cytometry. Incubation of CAR-T cells with recombinant BCMA revealed robust and specific antigen binding, with BCMA binding intensity scaling linearly with CAR surface density, confirming functional CAR expression and antigen engagement (Fig. 5h). CAR-T cells were subsequently enriched by fluorescence-activated cell sorting (FACS) to obtain a uniform CAR-positive population (Fig. 5i).

Cytotoxicity assays performed across multiple effector-to-target (E:T) ratios demonstrated efficient, dose-dependent killing of Raji cells expressing EREG_5_Ca9-anchored BCMA by BCMA CAR-T cells (Fig. 5j, k). In contrast, Raji cells expressing cytoBCMA were not eliminated, consistent with the requirement for cell-surface antigen presentation to support CAR-mediated recognition. Importantly, no appreciable cytotoxicity was observed when targets were co-cultured with non-transduced T cells, confirming CAR specificity.

In quick summary, the new experimental data above provided additional validation for the CSD elements, in 5 distinct cell types, with alternative readouts (confocal and flow cytometry), with additional antigen (BCMA), and with functional assay (CAR-T killing).

In addition to the new experimental data above, we collected multiple screening datasets and performed new analysis to validate DeepScan models and designed CSDs in existing figures. Considering the limited validation of predicted naturally occurring CSD (nCSD) in Figure 2, we designed additional gCSD validation experiments in Figure 3 to further demonstrate the generality and robustness of DeepScan CST models. We first applied DeepScan CST model to predict 1100 gCSD sorting signal strength. Based on predicted CST class, we selected and tested 44 gCSD potency using two cell lines (293T and Lenti-X) in flow surface staining experiment. 37/44 (84%) generative CSDs (gCSD) that are predicted in CST3 category were validated by flow surface staining. The high prediction accuracy in both nCSD (92%) and gCSD (84%) validation experiments reflects reliability of DeepScan CST model predictions.

To further test the generality of potent membrane anchors across antigens and expand validation dataset, our A29/E7 dual antigen screens using 38 validated gCSD, 21 validated nCSD and 15 additional random nCSD. For A29-CSD screens, gCSD templates and validated nCSD/gCSD showed near 100% validation rate. Moreover, random nCSD's CST3 positive rate (60%, 9/15) was consistent with primary A29-CSD screen (60%, 62/104) in Fig. 2a and was much lower than that of DeepSCan-enriched nCSDs (92%, 24/26) in Fig. 2h (Fig. 3d-3e), highlighting superior performance of CSD sequences selected by DeepSCan CST model. Interestingly, compared to A29-CSD results, E7-CSD showed a trend of lower CST3 positive rate in random nCSD and validated gCSD groups, suggestive of a greater challenge posed by E7 surface translocation. Overall speaking, A29-CSD and E7-CSD showed similar CST3 validation rates, providing further evidence for the generality of DeepSCan-enriched nCSD and gCSD.

To further strengthen DeepSCan model's robustness, we developed the second-generation Genesis models and third-generation Omni models trained with expanded CST labels collected from five independent CSD screens. The training data was supplemented with varying number of CSD homologs, full-length variants, negative controls and self-training labels. The next-generation DeepSCan models explored the optimal homolog numbers, self-training label numbers, and output layer structures to maximize CST prediction performance. In addition to training approach optimization, we also experimentally validated CST prediction performance of Genesis Quant, Omni and Omni + self-training in A29/E7 dual antigen screen using over 80 round2 gCSD constructs. In total, we developed three generations of DeepSCan models, greatly increased CSD experimental label number and validated major models using 26 nCSD, 124 gCSD and 384 CSD-fused chimeric antigens.

All these data and analyses together provided stronger evidence that DeepSCan models exhibited robust prediction performance in single/dual antigen screens and its validated CSD modules demonstrated consistent ability to translocate distinct antigens to cell surface, as a common outcome across cell type, antigen and assays.

3. Self-training (pseudo-labels): If I am understanding this text correctly, then the Omni

models are trained with labels generated by averaging logits from the Genesis models. Even if the authors keep fixed test sets, those sets share similarities with sequences used for pseudo-labeling. This will tend to bias the performance metrics to be too optimistic. They should exclude pseudo-labeled sequences with high identity to examples in the test-set. They should also use a truly external test set that is assembled independently.

Response: We thank reviewer 2 for bringing up this concern regarding test set similarity with self-training labels and training dataset. We agree with reviewer 2 that independent test sets with low similarity to training data are critical for model evaluation. As illustrated in Extended Data Fig. 9d, the Omni training dataset combined eight Quant training datasets, ID 58k training dataset and proportional controls. So, Omni training datasets do not include the average logits from diCerent Genesis models.

Indeed, avoiding pseudo-labeled sequences with high identity to test set is a major consideration when we prepare the self-training dataset. When we prepare the CSD sequences for surface staining experiments, we clustered CSD sequences based on their identity and picked cluster-representative CSD sequences, which have less than 50% identity with each other (Extended Data Fig. 3).

From 235 experimentally tested CSD sequences, we isolated ~20% sequences as constant and independent test set as illustrated in Fig. 4a. Because of this clustering step, Quant 101 test set has low sequence identity with other CSD sequences in either training datasets or

untested 8952 query set. To further validate this result, we performed MMseqs2 clustering of Quant 101 test sequences and ~4000 Omni self-training sequences using a cluster cutoff at 70% identity (with longest sequence as the denominator and default 80% coverage). And we found there is no significant overlap between Quant 101 clusters and ~4k self-training CSD clusters, suggesting that Quant 101 test set is assembled independently and shares low sequence identity with pseudo-labeled CSD sequences.

To better illustrate the sequence identity between Quant 101 test sequences and 4k self-training labels, we showed their pair-wise sequence identity in the appendix figure below and found the overall average identity among the ~400k sequence pairs is ~11%. Among them, we identified 22 self-training sequences sharing over 50% identity with Quant 101 test set and therefore removed them from 4k pseudo labels. We trained the representative optimal Omni models with 4k pseudo labels with or without the 22 high-identity pseudo labels and showed their performance differences below (Appendix Figure 1). 22 high-identity pseudo labels led to an optimistic estimate of Quant and ID predictions, and showed relatively minor effects on CST classification, Topology and TMD boundary predictions as the pseudo labels

exclusively introduced distilled knowledge of CST score predictions rather than topology or TMD features. Notably, distillation, which trains next-generation deep learning models using distilled knowledge from existing model output, is becoming popular in large language model training¹⁻³ and sometimes in real-world scenarios it is inevitable to have overlap between distilled knowledge or pseudo labels and test set labels. It is a question worth exploring in our future studies how to properly use the pseudo labels in the DeepScan model self-training.

Appendix Figure 1 in this letter

Appendix Figure 1 | Pair-wise sequence identity analysis identified 22 pseudo labels that share >50% identity with Quant 101 test set and led to overestimates of self-trained Omni model performance, especially in Quant and ID evaluations. a, Pair-wise sequence identity analysis between Quant 101 test set and ~4k CSDs used as self-training labels. Sequences were aligned using python package pairwise2 alignment method and identity was calculated as the number of residue matches divided by the maximum length of target or query sequence. b, Comparing Quant, CST classification, ID, Topology and TMD boundary prediction performance of sampled representative Omni models trained with or without the 22 high-identity sequences in the 4k pseudo labels (n = 3 checkpoints). Data are presented as mean ± SEM. * p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001. Unpaired t-test was used to determine statistical significance.

To present clear results and maintain a focus on empirical study (how native and generated CSDs enhance cell surface localization / redirecting), we have consolidated the previous Figures 4-6 into one single figure (new Fig. 4), for which we moved computational data related to Omni models to Extended Data Fig. 9 and Supplementary Fig. 17. The revised main figures illustrate the main storyline of designing, screening, and optimizing CSD elements and develop a side storyline of establishing deep learning models to facilitate the design and optimization of CSD modules. We acknowledge the limitation of pseudo labeling, and it's sometimes inevitable in machine learning / AI training due to the limitation of available datasets. We agree that there are vast potentials in the future to improve the computational context and explore the optimal way to incorporate these self-training or pseudo labels.

Lastly, we want to explain how we assembled the fixed test sets. To assess Omni model performance, we use three sets of data including test set from train_test_split function(MSE1), independent Omni 1399 test set (MSE2 or MSE1399) and independent Quant 101 test set designed for Genesis Quant models (MSE101). Omni 1399 independent test set is a combination of Quant 101 test set, ID 1029 independent test set and their homologs from

Omni 128k training set (Extended Data Fig. 9d). Quant101 contains ~100 CSD sequences and ECD variant controls to evaluate deep learning model performance in CST score prediction. Notably, to avoid test set homologs appearing in Omni training datasets, we have

conducted MMSeqs2 cluster search against Omni 128k training datasets using the independent Quant101 and ID1029 test sets and moved these homologs from training data to Omni 1399 test set (Extended Data Fig. 9d). Therefore, the Omni training data share low identity with independent Omni 1399 test set. Similarly, we evaluate Genesis Quant model using dependent test set from train_test_split function(MSE1) and independent Quant 101 test set (MSE2 or MSE101) to enable robust model evaluation.

4. Generative design models (Genesis Dusk / ESM3) lack pre-specified selection criteria. The round-2 designs were selected for validation using a mixture of model scores, sequence heuristics, and “inferred biophysical propensities” (page 8). In this case, selection bias is hard to exclude. It was not clear how the synthetic (generative) sequences were assembled prior to model-based in silico screening. The validation was only performed in HEK293T/Lenti-X with two antigens (A29 and E7), which limits the claims of generality. They should provide the full ranked list (it was unclear to me if this was contained in the supporting files). The validation should be in 3 or more cell types and with 3 or more unrelated antigens. They should also report on functionality of the displayed antigens (not only antibody labeling). It was not clear if an expression tag was used for labeling or if the antigens themselves were labeled and detected.

Response: We thank reviewer for pointing out these concerns regarding synthetic sequences selection and validation. The synthetic CSD sequences were generated by three diCerent approaches, including 1) masking less essential residues and predicting masked sequences; 2) shortening terminal sequences; 3) CSD domain shuffling or recombination with oligomerization domains. Prior to model-based in silico screening, we selected best CSD candidates in round 1 dual antigen screens with an emphasis on shortening terminal sequences and refinement of less essential residues at N-term and C-term. In addition, we

recombined EREG-5 with diCerent oligomerization domains, hTM or ICD of Ca9 and ACE to seek optimal CSD chimera. We used 16 top-performing candidates in round 1 screens as templates of round 2 synthetic sequences, which include EREG, CXCL16, EREG-5, CXCL16-69, ACE, Ca9, NDV-F, EPCAM, EPCAM-78, Tacstd2, SARS2_Spike, SARS2_80, SDC3, sdc4-a, sdc4-b, and HA1B (Data summary file tab Fig. 4e_gCSD_templates).

For masked residue prediction, in total, we computationally designed 2600 round 2 gCSD (r2 gCSD) based on 6 nCSD, 3 gCSD templates and 13 masked prompts. The masked positions were chosen based on their homolog variation/conservation, model attention distribution and each residue's biochemical characteristics. The r2gCSD design was generated using Genesis Dusk and ESM3 models, with each model producing 100 gCSD per prompt (Data summary file tab Fig. 4e). The gCSDs were ranked based on Genesis dusk model loss function, Genesis CST model logits, basic residue percentage in ICD, and CST value predicted by Genesis Quant and Omni models, which can be found in data summary file, Fig. 4e_r2_gCSD_v2 tab. From this r2 gCSD pool, we selected 48 sequences for experimental validation based on multiple selection criteria including Genesis Quant logits, Omni model logits, Genesis Dusk loss functions, basic residue (R, K) usage frequency, evolutionally tolerated variations and inferred biophysical propensities. In addition, we also rationally designed 27 CSD length variants and 8 EREG chimera, with 4 EREG fused to nCSD truncations and 4 EREG fused to oligomerization domains. The CSD length variants were chosen based on homolog sequence alignment, conserved motif distribution and secondary structure predictions. For CSD fusion constructs, we selected EREG-5 as a template because of its great surface translocation potential, sequence plasticity and relatively short length. The EREG-5 complementary truncations were fused to either hinge-transmembrane domain (hTM) or intracellular domain (ICD) truncations of two most potent nCSDs (ACE or Ca9) in round 1 screen. In addition, we also fused EREG-5 full length to four classic oligomerization domains including trimerization, tetramerization and hexamerization domains.

The chimeric Flag-A29 antigens are labeled with Flag expression tag and detected with PE anti-Flag antibody (BioLegend, Cat. No. 637310), while E7-CSDs are directly detected by commercial PE anti-E7 antibody from Santa Cruz as described in methods (Santa Cruz Biotechnology, sc-6981 PE). The new protein BMCA was with His tag and detected by PE anti-His antibody.

To further validate CSD-fusion proteins display activity and their functionality, we focused

on the three top-performing round 2 gCSDs and performed substantial new validation experiments. As shown above, and summarized below, we have now provided experimental evidence in the following areas (although not all-by-all):

1. Across five distinct cell lines (HCT116 – colorectal origin, H1299 - lung, HeLa - cervical, K562 – hematological / leukemia, and Raji - hematological / lymphoma), all of which exhibited consistent and reproducible surface expression (Fig. 5a).
2. Across three distinct antigens (A29, E7, BCMA), all of which showed enhanced cell surface expression when linked to the defined CSDs or gCSDs. (various Figs)
3. Across two orthogonal assays (flow surface staining and confocal microscopy) expression (Fig. 5b-c for confocal; various Figs for flow).
4. With functionality of the displayed antigens, i.e. enhanced surface localization, leading to potent and augmented CAR-T cytotoxicity on target cells (not just surface expression alone). (Fig. 5d-k)

By validating gCSD performance across multiple cell types, antigens, and immune assays, we further generalized learned sequence–function rules beyond the original screening format. Importantly, the functional assays presented here move beyond antibody labeling or surface detection alone. gCSD-mediated display of BCMA is sufficient to drive antigen-specific CAR-T cell activation and cytotoxicity, demonstrating that gCSDs can be engineered to enable new biologically productive immune recognition rather than merely increasing apparent surface abundance. These results indicate that gCSDs can be leveraged to augment antigen accessibility necessary for effective immune engagement, rather than simply acting as membrane anchors.

5. There is a mismatch in terms of the labeled training data vs. very large query/design space. They apply models trained on ~100 or so labeled sequences (plus homologs) to tens to hundreds of thousands of sequences. This is not a problem per se but should be done using out-of-distribution (OOD) analysis and uncertainty estimates. They should demonstrate that the query sets are similarly distributed as the training set (depending on representation this can take the form of PCA analysis). They can therefore consider adding OOD metrics, embedding-space analysis (e.g., train-test nearest-neighbor identity and ESM embedding distances).

Response: We thank reviewer for providing these constructive suggestions. To better understand sequence similarity of training, query and test datasets, we have computed the embeddings of Genesis Quant 8k training sequences, Query 170k and Test 1399 protein sequences in the whole dataset using the Genesis 8k attention pooling model. The 480-dimension CLS embeddings were then projected to two-dimensional space to visualize the similarity and distribution of protein sequences in diCERENT CST categories (Principal component 1 and 2, PCA1 and PCA2, Extended Data Fig. 10c-e). In both training 8k and query 177k datasets, we notice that strong CST category sequences are mainly distributed on the right side of PCA 2-D plane, suggesting the major components in the model embeddings are strongly associated with protein sequences' cell surface translocation (CST) signals. The training 8k and query 170k datasets showed similar distribution on PCA1 and PCA2. Moreover, we estimated dataset's uncertainty using k nearest neighbor (kNN) out of distribution (OOD) distances. We indexed all training embeddings in the model's latent space and computed the kNN distance for each training test, or query sequence (k=10, 10

nearest neighbors' distance average). Larger distances indicate higher OOD likelihood or uncertainty.

Not surprisingly, the 170k query set and Omni 1399 test set showed overall higher OOD distance than training 8k set (Extended Data Fig. 10f). Based on the kNN distances, we calculated each query sequence's identity with its nearest neighbor sequence in 8k training set and found majority of them are below 50% identity (Extended Data Fig. 10g). According to Sun et al. study in ICML in 2022, the kNN distances of top 5% test set were considered as out of distribution⁴. Based on this assumption, we derived an OOD threshold of 0.286. Among all 177k query set, ~2.6% of sequences are labeled as out of distribution, and their CST predictions should be used with caution (OOD score and labels in Data summary file

ED Fig. 10).

Extended Data Fig. 10c-10g

6. The manuscript states they used standard ANOVAs with multiple-comparison tests but provides too few details on the exact n per construct. The text misspells Tukey's test (spelled "Turkey's") and "Dunnett" tests; ensure consistency and justification.

Response: Thank you for pointing out this oversight. We have added the sample number n in Fig. 1i, Fig. 2f-2g, and ED. Fig. 6. Fig. 1i has 3 samples per construct, while Fig. 2f-2g have one sample per construct. And we have corrected the misspelled Tukey's test and Dunnett test.

7. Code and raw data should all be made available (partly this was done).

Response: We have organized all major codes in one zipped folder and uploaded them through manuscript submission system. The major raw data have been provided in the data summary file.

8. The clarity and quality of communication is poor. The manuscript is too difficult to read. There are too many dense acronyms. The various definitions (DeepSCan, Genesis/Omni tasks, CST classes) are scattered. The layout of the manuscript maximizes confusion (figures are at the end; captions are separated from figures; etc.). I do not know if this is a journal requirement for submission, but it makes the task of the reviewer much more difficult. There are not sufficient plain language descriptions of the assay and methods. Furthermore, the claims about applications (mRNA vaccines, receptors, therapeutics) are not supported by data. A thorough rewrite is needed to clarify the work (especially the ML part) and limit the scope of claims to what is experimentally demonstrated.

Response: We apologize for the lack of clarity, and thank reviewer 2 for offering these critiques and valuable suggestions.

Clarity improvement:

We have addressed the heavy use of acronyms in response to your first major comment. In brief, we removed four acronyms that are rarely used and separated the remaining acronyms into two groups including 10 common acronyms that frequently appear in other studies and 11 custom acronyms that are defined by this study. In addition, we have expanded details of assays described in methods section, including newly added PCA and kNN OOD analysis.

We have put definitions of DeepSCan, Genesis and Omni models together in the introduction as shown below. "To address these questions, we harness high throughput screens to systematically profile CSD signal strength and develop a series of DeepSCan (Scanning and identifying surface protein sequences using deep learning models), Genesis (Emergence of enhanced model training datasets and strategies) and Omni (Originates from Latin word "Omnis", which means unified deep learning model that integrates multiple tasks of CSD predictions) models to accurately predict cell surface translocation strength, TMD boundary and topology of protein sequences, facilitating sophisticated design of novel potent CSD elements."

And CST0-5 categories are defined in Fig. 4a, which shows the five cut-offs separating CST0-5, namely normalized Log PE integral 0.3, 0.55, 0.8, 1, 1.2. This was also added to the main text: "The CST categories were defined by five cut-off values of normalized Log PE integral as shown in Fig. 4a."

Figure/legend formatting for review:

It is typical that NPG journals (Nature, NBT, etc) recommend separate text and figure submission and that all figures are appended at the end (so that text files are light and clean and easily editable). In order to facilitate readability and review, we have put together the figures and legends together at the end of manuscript to make it more reader-friendly for reviewers without having to go back-and-forth.

Interpretations:

Furthermore, the claims about applications (mRNA vaccines, receptors, therapeutics) are not supported by data. A thorough rewrite is needed to clarify the work (especially the ML part) and limit the scope of claims to what is experimentally demonstrated.

We also added a set of experiments to demonstrate CSDs' applications in cell therapy by displaying BCMA ectodomain on cell surface to activate CAR T cells (New Fig. 5). The functional assays presented here move beyond antibody labeling or surface detection alone. gCSD-mediated display of BCMA is sufficient to drive antigen-specific CAR-T cell activation and cytotoxicity, demonstrating that gCSDs enable biologically productive immune recognition rather than merely increasing apparent surface abundance. We have added this part to the revised manuscript (relevant to CAR-T therapeutics).

We agree with the reviewer that the applications in mRNA vaccines and receptors are not supported by our data, thus we removed these claims from the abstract and results. We only leave relevant languages in the introduction or discussion in which relevant context was mentioned (and if so proper reference of prior work was cited).

We also followed the reviewer's suggestion to have thoroughly rewritten the text, to clarify the ML part of the work, and limit the scope of claims to what is experimentally demonstrated.

Minor points

- They should provide early, formal definitions for the acronyms (or reduce their use) e.g., SPTM-CM vs SPTM-nonCM, CST0-5, "hTMC," and each Genesis/Omni subtask (ID, Topo, CST, TM, Quant).

Response: We have provided formal definitions for the acronyms mentioned by reviewer2. CST0-5 are six classes of cell surface translocation strength that are defined by five normalized Log PE integral threshold (0.3, 0.55, 0.8, 1, 1.2) as shown in Fig. 4a. The hTMC stands for hinge, transmembrane domain and cytosolic domain connected to each other in one construct. Genesis is a deep learning model name that marks a next generation learning model and a new beginning of model training. Omni is the third-generation deep learning model name that comes from the Latin word "omnis" and means universality, which covers all five DeepScan/Genesis model subtasks.

To address the acronym definition concerns, we define the SPTM-CM and SPTM-nonCM early in the first paragraph of results section: "Based on UniProt transmembrane domain features and subcellular location annotations, proteins are categorized into non-transmembrane proteins (nonTM) and transmembrane proteins (TM). And TM are further separated into single-pass (SP) and multi-pass (MP) TMs on cell membrane (CM) or other endomembrane systems (nonCM). For example, SPTM-CM designates a group of SPTM that are annotated as on cell membrane, virion, virion membrane, extracellular space, extracellular matrix, cell outer membrane, cell surface, host cell membrane or apical cell membrane. And SPTM-nonCM is the complementary subset of SPTM-CM in the broader category of SPTM and it include SPTM localized on other endomembrane systems and SPTM annotated with ambiguous subcellular localization, such as "proteins on membrane".

The five tasks of Genesis/Omni models are predictions of CSD identity (ID model, whether a sequence can display protein on cell surface), topology (Topo model, whether its extracellular domain is at N-term or C-term), transmembrane domain boundary (TM model, where its key membrane contact site is), CST strength classification (CST model, classification model) and CST strength value (Quant model, regression model). They were chosen because they are key features that define the CSD sequence functions and characteristics. When we design a cell surface display sequence, we want to know which protein terminus (N-term or C-term) it displays on cell surface, where its transmembrane domain or key contact site is and how efficiently it can display the protein of interest on cell surface. And the five tasks in Genesis/Omni model are designed to address these questions. The membrane domain boundary helps us refine the key membrane contact sites and better understand its interaction mechanisms with lipid bilayers. The CSD topology maximizes display efficiency and offers options to display different parts of protein on cell surface. The CST classification and strength value provide a straightforward way to evaluate CSD module activity. All five tasks share the same foundation model and differ in the final output layer. The 2-class identity, 3-class topology and 6-class CST category tasks rely on per-sequence classification model structures, while the CST strength task relies on per-sequence regression model structure. And the last 3-class transmembrane domain task relies on per-

token classification model architecture. Extended Data Fig. 9a shows that first-generation DeepSCan models and second-generation Genesis models can complete three single tasks (ID, Topo and CST classification) and five single tasks (ID, Topo, CST, Quant and TM) respectively. Their harmonized data and combined model architectures lay foundation for third-generation multi-task Omni models.

- The ML reporting and calibration should include ROC/PR curves, reliability diagrams, and class-wise confusion matrices.

Response: Based on your suggestions, we have prepared ROC curves, confusion matrices and calibration diagrams (reliability diagram variants for regression model) and put them together in Supplementary Fig. 18.

Supplementary Fig. 18

In this figure, we first compared row-normalized confusion matrices of Genesis Quant 8k model, Omni Quant model (no self-training), and Omni CST model (no self-training). For Genesis Quant and Omni Quant regression models, six CST classes were derived from logits bins separated by following normalized Log PE integral thresholds, -1, 0.3, 0.55, 0.8, 1, 1.2, 2, as indicated in Fig. 4a. Confusion matrices of all three models showed high accuracy in CST0 vs. non CST0 category (CSDs) prediction and indicated that identifying a CSD sequence is not a primary challenge here. The question of interest is whether a given sequence is a potent CSD or not. To address this question, we treat CST3-5 categories as potent categories and calculated accuracy of potent CSD predictions, which is accurately predicted CST3-5 sequence count over total CST1-5 sequence count. The result shows Genesis Quant model possesses higher accuracy in potent CSD module predictions, while Omni model exhibited better performance in CST0 vs. non-zero predictions (Supplementary Fig. 18a-18c). To evaluate the reliability of Genesis or Omni Quant model predictions, we showed their calibration plots (reliability diagram variant for regression model, as opposed to classification model). Interestingly, both plots showed systematic underprediction in the lower range (0, 0.4/0.5) and overpredictions at higher values (Supplementary Fig. 18d-18e). For Omni CST classification models, we analyzed its Receiver Operating Characteristic (ROC) curves and found that the class 0 showed higher curves, indicating that the model is more accurate when identifying CST0 sequences. Among

CST1-5 curves, CST2, CST4 and CST5 have area under curves over 0.9, highlighting models' better ability to distinguish these three classes than others (Supplementary Fig. 18f).

- The supplementary information refers to mouse experimentation and mouse licenses, but as far as I can tell no mice were used in this study.

Response: Thank you for pointing this out. We have removed supplementary information related to mouse experimentation and mouse licenses and changed it to this statement. "This study does not involve human subjects or animal experiments."

Reviewer #3 (Remarks to the Author):

The authors present a systematic approach to designing cell surface display (CSD) elements for therapeutic applications, particularly mRNA vaccines. They begin by screening natural CSD sequences, developing multiple generations of deep learning models (DeepSCan, Genesis, and Omni), and ultimately validate generative CSDs that match or exceed the performance of natural sequences.

This paper is a mixed bag. It starts strong but loses focus midway through. The problem is compelling and the first half of the paper is on a promising trajectory: collect and analyze natural sequences, build predictive models, recognize their limitations, then iterate. However, after Figure 3, the work diverges into an unfocused exploration of model architecture and training strategies that obscures rather than advances the core insights.

The paper would have been considerably stronger if it had maintained its initial focus on understanding what makes CSDs eCective and systematically exploiting those principles.

Response: We thank reviewer 3 for providing positive feedback and constructive suggestions. We have followed exactly that in the revision: (1) Streamline the paper, focus on the initial goal of understanding of what makes CSDs eCective; and (2) more systematically validate the findings using a number of new approaches. These revisions help us achieve “collect and analyze natural sequences, build predictive models, recognize their limitations,” generate new sequences that are potentially better, validate them, “then iterate”, in this study. Details are provided below:

(1) Streamlining the paper:

We have reorganized the manuscript to maintain its focus on what makes CSDs eCective and how to systematically exploit those principles, especially after Figure 3. To maintain the initial focus, we have designed new experiments to better understand key sequences regulating CSD module generality/activity and combined original Fig. 4-6 into single Fig. 4, which mainly shows the Genesis Quant model development and round 2 gCSD screens. The updated Fig. 1-4 presents a main storyline of empirically designing, screening and optimizing CSD elements and briefly touch on the side storyline of developing deep learning models to accelerate this empirical process.

Figure 4

(2) More systematically validate the findings:

To more systematically explore and validate CSD-fusion proteins display activity and their functionality, we have now provided experimental evidence in the following areas:

Performed new validation experiments of DeepSCan-designed gCSDs across 5 different cell lines of distinct types

We first focused on the top-performing round 2 gCSDs to assess reproducibility across cellular contexts and to evaluate whether gCSD-mediated antigen surface display supports downstream immune activation. Quantitative evaluation across five distinct cell lines (HCT116, H1299, HeLa, K562, and Raji) demonstrated that the top three round 2 gCSDs fused to A29 exhibited consistent and reproducible surface expression across cellular contexts, as measured by CST-based flow cytometry (Fig. 5a).

Performed new validation experiments of gCSDs using an orthogonal assay (confocal microscopy)

To confirm that CST measurements reflected genuine cell-surface localization, we performed confocal microscopy under non-permeabilizing conditions in A29-transfected HEK293T cells. gCSD-fused constructs exhibited robust membrane-localized signal, whereas A29 expressed without a gCSD or mScarlet alone showed no detectable surface staining, confirming gCSD-dependent membrane display (Fig. 5b, c).

Performed new functional validation experiments (new type of cancer antigen surface expression and CAR-T killing)

We next tested whether gCSDs support modular surface presentation of biologically relevant antigens. As a clinically established immunotherapy target, we selected B-cell maturation antigen (BCMA), a tumor-associated surface protein targeted by CAR-T therapies. To enable modular display, BCMA was truncated to its extracellular domain and fused to candidate gCSDs. gCSD-mediated surface presentation of BCMA was observed in both HEK293T and Raji cells by flow cytometry and confocal imaging, whereas BCMA expressed without a gCSD failed to localize to the cell surface, indicating that membrane presentation was strictly gCSD-dependent (Fig. 5d–f).

To functionally evaluate gCSD-mediated antigen presentation, we generated BCMA-directed CAR-T cells via lentiviral transduction and co-cultured them with Raji-PL target cells expressing either a truncated, cytoplasmic BCMA construct lacking a gCSD (cytoBCMA) or BCMA fused to candidate gCSDs (mbBCMA) to enforce cell-surface localization (Fig. 5g). For

downstream assays, we selected EREG_5_Ca9 as the candidate gCSD due to its relatively high transfection efficiency in Raji cells. Prior to functional assays, expression and antigen-binding capacity of the BCMA CAR, containing a novel scFv developed in-house, were validated by flow cytometry. Incubation of CAR-T cells with recombinant BCMA revealed robust and specific antigen binding, with BCMA binding intensity scaling linearly with CAR surface density, confirming functional CAR expression and antigen engagement (Fig. 5h). CAR-T cells were subsequently enriched by fluorescence-activated cell sorting (FACS) to obtain a uniform CAR-positive population (Fig. 5i).

Cytotoxicity assays performed across multiple effector-to-target (E:T) ratios demonstrated efficient, dose-dependent killing of Raji cells expressing EREG_5_Ca9-anchored BCMA by BCMA CAR-T cells (Fig. 5j, k). In contrast, Raji cells expressing cytoBCMA were not eliminated, consistent with the requirement for cell-surface antigen presentation to support CAR-mediated recognition. Importantly, no appreciable cytotoxicity was observed when targets were co-cultured with non-transduced T cells, confirming CAR specificity.

By validating gCSD performance across multiple cell types, antigens, and immune assays, we further generalized the learned sequence-function rules beyond the original screening format. Importantly, the functional assays presented here move beyond antibody labeling or surface detection alone. gCSD-mediated display of BCMA is sufficient to drive antigen-specific CAR-T cell activation and cytotoxicity, demonstrating that gCSDs can be engineered

to enable new biologically productive immune recognition rather than merely increasing apparent surface abundance. These results indicate that gCSDs can be leveraged to augment antigen accessibility necessary for effective immune engagement, rather than simply acting as membrane anchors.

The systematic curation of SPTM-CM sequences in Figures 1-2 is well-executed, with clear datasets and thoughtful experimental design. The identification of polybasic residues at the ICD-TMD junction as a key feature is valuable. However, Figure 2F reveals a significant limitation that the authors don't adequately address: their models primarily distinguish extreme CST values (CST0 vs CST3) but show poor resolution for intermediate activities. This suggests the problem formulation needs reconsideration. A regression approach with appropriately transformed targets might better capture the continuous nature of CST strength, or alternatively, a binary classification at a meaningful threshold could provide clearer decision boundaries. The current multi-class approach may miss the ordinal nature of the categories.

Response: We thank reviewer 3 for bringing up this consideration related to Fig. 2f. We have calculated CST strength of screened CSD sequences in Fig. 2f using normalized log PE integral and plotted the CST strength (x axis) vs. Lys + Arg percent in intracellular domain (ICD, y axis) in the figure below. Based on this result, we decided to conduct a binary or quaternary classification of all data points separating CST0-1 and CST2-3 two categories (Supplementary Fig. 5a-5b) or CST0-3 four categories (Fig. 2f-2g). We found in both situations CST2, CST3 or CST2-3 showed significantly higher Arg + Lys percent than CST0, CST1 or CST0-1 categories (Fig. 2f-2g or Supplementary Fig. 5a-5b), reflecting a positive correlation between RK percent in ICD and CST strength of screened sequences. The regression approach and quaternary classification approach improved the resolution for intermediate CST activity in Fig. 2f-2g, while the binary classification approach showed higher statistical power and sample number per group.

Fig. 2f-2g

Supplementary Fig. 5

Unfortunately, things go haywire from Fig 4 onwards. The dataset construction becomes progressively opaquer: we encounter "8 Genesis datasets" with vague descriptions of

"augmented FL and ECD variants, CSD homologs...and proportional negative controls." We're presented with over 540 models across three "generations" without systematic ablation studies to determine which components actually matter. Basic baselines are missing - would simple CXCL+EREK combos work? The proliferation of naming schemes (Dawn, Dusk, Omni) adds unnecessary complexity.

Response: We thank reviewer 3 for pointing out this concern regarding model description and ablation studies. To avoid egression and confusing concepts, as mentioned above, we

streamlined the paper, largely according to the reviewer's suggestion. We have combined original Fig. 4-6 to one Fig. 4 and moved Omni model data to Extended Data Figure 9 or Supplementary Figure S17. The corresponding results and figure descriptions were moved from main text to supplementary discussion, which removed model names such as Dawn and Dusk in the main text. In addition, we have removed four unnecessary acronyms that only appear in the manuscript 2-3 times to avoid proliferation of naming schemes. Based on this suggestion, we have revised the dataset description in method section to explicitly define the Genesis datasets and their composition. The revised text clarifies the role of datasets in model training and evaluation, improving transparency and reproducibility.

We defined the Genesis datasets in the main text and method section to improve clarity.

On top of this, we have conducted self-training ablation study and shown the results in Supplementary Fig. 17b to evaluate the importance of self-training labels. As the self-training label number increases, we observed a gradual and dependent improvement of model performance in CST strength prediction, as evidenced by the evaluation metrics, average of MSE101 and MSE1399. To test CSD chimeras in r2-gCSD experiment (Fig. 4e), we generated and characterized the chimeras of EREG-5, which recombined the best round 1 gCSD, EREG-5 with best round 1 nCSDs, ACE or Ca9. The result showed higher-than-WT potency of EREG-5 chimeras especially after fusing to ACE or Ca9 hTM (hinge-transmembrane domain).

Supplementary Fig. 17b

Fig. 4e

Moreover, we have provided definitions for the model names, including DeepSCan, Genesis and Omni in the introduction of manuscript and moved dispensable naming schemes such as Genesis Dawn and Genesis Dusk to supplementary discussions.

To comprehensively evaluate the acronym use, we listed all acronyms in the main text. In total the main text contains ~25 acronyms, including 10 commonly used acronyms, 11 custom acronyms and 4 rarely used acronyms. Notably, we removed four acronyms that rarely appear in main text and replaced them with their full terms. For the custom acronyms that are introduced by this study, they designate important concepts in our analysis and are repeatedly used in the manuscript, necessitating their use to ensure conciseness and avoid redundancy. While we used a number of custom acronyms, they can be easily grouped into three types including protein classification related acronyms, cell surface display related acronyms and other frequent acronyms.

11 Common acronyms (including approximate counts):

1. Machine learning related: ML (Machine Learning, 10 counts), AI (Artificial Intelligence, 2 counts), genAI (Generative Artificial Intelligence, 4 counts), DL (Deep Learning, 15 counts)
2. Model names: ID (Identity, 18 counts), Topo (Topological, 16 counts), TM (Transmembrane, 20 counts).
3. SPTM domains: ICD (Intracellular Domain, 10 counts), ECD (Ectodomain, 13 counts), TMD (Transmembrane Domain, 28 counts).
4. Other frequent acronyms: MSE (Mean Squared Error, 8 counts), GO (Gene Ontology, 5 counts).

11 Custom acronyms:

1. Protein classification related acronyms: A set of five acronyms including SPTM-CM (single-pass transmembrane protein on cell membrane, 35 counts), SPTM-nonCM (3 counts), MPTM-CM (1 count), nonTM (1 count).
2. Cell surface display related acronyms: a set of four acronyms including CSD (Cell Surface Display, 34 counts), naturally occurring CSD (nCSD, 9 counts), generative CSD

(gCSD, 17 counts), CST (Cell Surface Translocation, 39 counts). The CST0-5 categories are defined in Fig. 4a, which shows the five cut-offs separating CST0-5, namely normalized Log PE integral 0.3, 0.55, 0.8, 1, 1.2.

3. Other frequent acronyms: DL1 (Double Label 1, 7 counts), Quant (Quantitative, 21 counts).

4. Model names (codename rather than acronym):

- a. DeepScan: Deep learning model used to scan and identify surface protein sequences, 48 counts
- b. Genesis Dawn: a new beginning of training strategies used to build a series of models that shed light on the surface protein classic characteristics, 2 counts, related figures were moved from original Fig. 5 to ED. Fig. 9
- c. Genesis Dusk: training strategy update to improve models' ability to generate SPTM-CM like sequences and explore dark space of artificial CSD sequences.
- d. Omni: it comes from Latin word "omnis", which implies unified deep learning models for TM, ID, CST, Quant and topology tasks.

4 Removed acronyms:

1. NBME (Nature Biomedical Engineering): Appeared three times. Replaced it with reference 9 (published, not in press anymore).
2. scFv (single-chain variable fragment): Appeared once. This acronym has been removed.
3. LNP (Lipid Nanoparticle): Appeared twice. Use of acronyms is not necessary.
4. MVP (Modular Vaccine Platform): Appeared three times. Replaced it with full term.

The multi-task formulation remains unclear--what exactly are the five tasks in the Omni model and why were these chosen? Fig 5a was difficult for me to understand. The "self-training" approach using predicted labels from earlier models lacks theoretical grounding and risks amplifying model biases. There's no clean held-out validation set for hyperparameter selection, making it impossible to assess whether the chosen architectures are actually optimal or just overfitted to the particular data splits.

Response: We thank reviewers for raising concerns regarding multi-task formulation and self-training approach. The five tasks of Omni models are predictions of CSD identity (ID model, whether a sequence can display protein on cell surface), topology (Topo model, whether its extracellular domain is at N-term or C-term), transmembrane domain boundary (TM model), CST strength classification (CST model) and CST strength value (Quant model). They were chosen because they are key features that define the CSD sequence functions and characteristics. When we design a cell surface display sequence, we want to know which protein terminus (N-term or C-term) it displays on cell surface, where its transmembrane domain is and how efficiently it can display the protein of interest on cell surface. And the five tasks in Omni model are designed to address these questions. The membrane domain boundary helps us refine the key membrane contact sites and better understand its interaction mechanisms with lipid bilayers. The CSD topology maximizes display efficiency and offers options to display different parts of protein on cell surface. The CST classification and strength value provide a straightforward way to evaluate CSD module

activity. All five tasks share the same foundation model and differ in the final output layer. The 2-class identity, 3-class topology and 6-class CST category tasks rely on per-sequence classification model structures, while the CST strength task relies on per-sequence regression model structure. And the last 3-class transmembrane domain task relies on per-token classification model architecture. Extended Data Fig. 9 shows that first-generation DeepScan models and second-generation Genesis models can complete three single tasks (ID, Topo and CST classification) and five single tasks (ID, Topo, CST, Quant and TM) respectively. Their harmonized data and combined model architectures lay foundation for third-generation multi-task Omni models.

We noticed that the self-training approach raised concerns from both reviewer 2 and reviewer 3 and may distract readers from the primary goal of this study, which is designing optimal CSD element and deciphering the underlying sequence features/mechanisms. For this consideration, we have moved data related to self-training approach and Omni models

to Extended Data Fig. 9 and Supplementary Fig. 17. Several machine learning studies introduced and published the self-training approach across vision and natural language processing fields (NLP). For example, in an influential study by Lee et al. in 2013, model predictions are treated as training labels during subsequent training rounds³. And Xie et al. in 2020 proposed to use a teacher model to label unlabeled datasets, then use them to train a student model, which has become mainstream in image classification¹. In this distillation process, the trained model output becomes training data for the next iteration model training, which has been accepted in academia². This distillation or self-training approach has been repeatedly shown improved generalization and were used internally and externally across major tech companies including OpenAI, Google and Meta. Notably, the distilled knowledge in large language models sometime inevitably resembles the data in test set at some degree. It remains an interesting question to answer in our next study about how to incorporate these self-training labels to maximize generality and minimize biases.

We agree that the clean held-out validation set is important to identify optimal training stop point in the self-training approach, which may introduce potential model biases due to overfitting. To evaluate model's generalization ability, we want to ensure low similarity between Omni1399 test set and Omni training dataset. In the self-trained Omni models, we used the independent Omni1399 test set, which is a combination of Quant 101, ID1029 and their homologs from Omni 128k training set (Extended Data Fig. 9d). This is our independent held-out validation set. Because we moved Omni1399 homologs from training dataset to test set, the validation set shares low sequence identity with Omni training data. In addition, we performed the MMseqs2 cluster search using the Quant 101 and 4000 self-training CST labels and found no overlapped clusters between these two groups at an identity threshold of 70%. Because the self-training approach has become a secondary focus rather than a central focus, we keep the hyperparameter, model structure and train_test_split optimization as our future directions in the next study.

There also seem to be some troubling misunderstandings about ML models. In Fig. 4b, the "linear" and "dense" models are conceptually equivalent: multiple linear layers without

intervening nonlinearities are mathematically equivalent to a single linear transformation. This raises questions about what was actually implemented. What does "attention pooling" mean, exactly?

Response: We thank reviewer 3 for bringing up concerns regarding the final output layer structure and definition. In total, we tested six output layer structures, which are displayed in Extended Data Fig. 8a. And the linear and dense models share similar architectures and the dense layer stacks two linear layers together, which are indeed mathematically equivalent to a single linear transformation. But they still exhibited some extent of difference in layer structure and model performance when used to train and evaluate Genesis Quant models. The dense layer is a two-linear-layer head, consisting of a linear projection preserving the embedding dimension (480 dimension), followed by a contraction to a scalar output. Its mapping remains linear like single linear layer; however, the extra parameters can help optimizations such as conditioning and regularization effects as we did observe distinct model performance of these two output layers in Fig. 4b.

Extended Data Fig. 8a

Fig. 4b

The linear head uses a single projection from 480 dimensions to one dimension for Genesis Quant model, while the dense head is implemented as two stacked single linear layers with following dimensional transformation ($480 \Rightarrow 480$) and ($480 \Rightarrow 1$) in absence of nonlinear activation. Moreover, we have uploaded the Omni model script in this submission which showed what was implemented in PyTorch code. Here is the PyTorch code example in Genesis Quant 8k model:

```
Linear head: self.head = nn.Linear(hidden_size, 1)
Dense head: self.head = nn.Sequential(
    nn.Linear(hidden_size, hidden_size),
```

```
nn.Linear(hidden_size, 1)
)
Hidden_size = 480
```

The attention pooling is a learnable way to aggregate token-level embeddings into a sequence-length vector by assigning diCereent weights to each amino acid positions. The overall goal is to assign embedding-derived weights to each position and find optimal weights and important residues for the model output, which is, in our case, the scalar CST value. Core mechanism involves two steps including calculation of normalized weights (α_i) and analysis of pooled representation or weighted sum of token embeddings (z).

```
 $\alpha_i = \text{softmax}(w \cdot x_i)$ 
#
```

```
 $z = \sum \alpha_i x_i$ 
!$%
```

- $w \in \mathbb{R}$: learnable attention vector
- α_i : normalized importance weights
- Output stays in same dimension d

The generative modeling is substantially oversold. Genesis Dusk essentially performs oC-the-shelf masked language modeling. The elaborate description of "forging new routes in the dark space of artificial CSD sequences" is indicative of the sometimes-overwrought writing. The constant switching between metrics (PE integral, normalized PE integral, CST categories, MSE, accuracy) without justification makes it diCicult to track what's actually being optimized.

Response: We thank reviewer for this helpful comment. We agree that the generative component should not be overstated. We have substantially toned down the language in the manuscript and clarified that Genesis Dusk employed the masked language modeling rather than a generative framework. Our intention was to emphasize its role as a representative learning framework for downstream CSD discovery. This change helps us refocus the narrative on its practical contribution to learning informative sequence features and representations rather than the generative model methodology.

Updated description of Genesis Dusk model in main text is shown below:

"We established the second-generation Genesis Dusk model, a masked language model that sheds light on masked tokens in the dark space of artificial CSD sequences."

Thank you for pointing out these metric switching issues. To avoid metric switching issues and keep consistent evaluation metrics in main figures, we have updated the evaluation metrics and use the normalized log PE integral in all related figures including Fig. 2a, 2h, 3c, 3d, 4e, Extended Data Fig. 5d, Supplementary Fig. 5c.

It is worth mentioning that in parallel with development of three generations of DeepScan models, we iteratively refined associated evaluation metrics to better reflect objectives in each stage. These models and metrics concepts were developed and refined over two-year timeframe. Our original figures present the trajectory from a rudimentary idea to progressively refined models and evaluation metrics. The PE integral was developed and designed for single CSD screen experiments in Fig. 1-3, while the normalized log PE integral was designed to harmonize data across multiple CSD screens using CSDs with similar activity as the denominator, such as HLA or CLEC4F (Fig. 4a).

Model interpretability was a missed opportunity. There is insuCicient discussion of the dominance of CXCL16 and EREG variants among successful gCSDs. What makes these sequences special? The limited attention weight analysis hints at the importance of the ICD-TMD interface but the authors don't pursue this systematically. A simpler approach, e.g., creating chimeras combining CXCL16 and EREG features or inserting their key motifs into other natural sequences might achieve similar or better results.

Response: Thank you for pointing out this important direction in successful gCSD feature analysis. In our study, we performed two rounds of gCSD design, testing and optimization, which provided valuable insights regarding the key motifs involved in the successful gCSDs, although our main text does not discuss these features in depth. In the revised manuscript, we have added in-depth analysis and discussion of potent gCSD features.

In round1 gCSD optimization, we designed 44 gCSDs based on 11 nCSD templates and these gCSD can be categorized into three groups, including gCSDs maintained/enhanced surface translocation (EREG, CXCL16, EPCAM), gCSDs modestly reduced surface translocation, and gCSDs significantly ablated surface translocation (>5-fold reduction, IGHE, ACE, LCT etc) (Fig. 3c). Among round1 gCSDs, the two extreme categories, gCSDs enhanced or ablated surface translocation, are most interesting as they either preserved or destroyed the key motifs mediating surface translocation. EREG and CXCL16 derived gCSDs showed highest surface translocation potential, indicating their templates preserved the key motifs mediating surface translocation. It is worth noting that all four EREG gCSDs maintained or enhanced surface translocation, while CXCL16_30 completely abolished the surface translocation (>10 fold reduction), highlighting the absence of key motifs in the gCSDs. Based on the observations in r1 gCSD testing, we can get these conclusions:

1. EREG key motifs are preserved within the shared templates, which include two major fragments located at N term and C term of TMD (Supplementary Fig. 7, x means any residues, Φ designates hydrophobic residues):
 - a. SKE(Y/V)xALTV;
 - b. ITVVGSTY Φ CRWYRNRSKEPKKExERV
2. Compared to other EREG gCSDs or nCSD, EREG_5 inserted more charged residues in extracellular hinge and ICD. It also altered the polarity and space filling properties at the center of TMD. Notably, EREG_5 specific residues that received high ML model attention are hydrophobic residues VLI and LI in TMD (Data summary file, Extended Data Fig. 8b-8c). These features are likely associated with its improved function.
3. For CXCL16 key motifs, the shared sequence features among gCSDs include:
 - a. ART in extracellular hinge
 - b. LCLLA Φ xxLTAALSYxx in TMD
 - c. CKRRRGQSPxSSP in ICD
4. Compared to other CXCL16 nCSD and gCSDs, CXCL16_69 contains more charged residues in extracellular hinge and ICD. Moreover, its introduced LP and KA dipeptides at N term and C term of TMD have high attention weights (Data summary file, Extended Data Fig. 8b-8c), signifying their uniqueness and importance in surface translocation.
5. On the other hand, the CXCL16_30 high attention weights mainly fell on the N term and C term of TMD (Data summary file Extended Data Fig. 8b-8c). Among these key residues, the CXCL16_30 specific dipeptide YF at ICD-TMD interface constitutes a bulky three consecutive aromatic ring stretch (YYF) which likely contributed to its poor surface translocation performance.
6. Coincidentally, CXCL16_30 (VL => YF) and EREG_5 (YF => LI) as two opposite examples demonstrate that removal of aromatic residues at TMD and ICD interfaces enhances CSD translocation.

Supplementary Fig. 7

On top of the optimal r1 gCSD sequences, we designed and tested round2 gCSDs, including 8 EREG_5 mutation variants, 5 EREG_5 length variants, 8 EREG-5 fusion variants, 4 CXCL16_69 mutation variants and 5 CXCL16_69 length variants. In r2 gCSD design of EREG-5 and CXCL16, we kept the shared motifs of r1 gCSDs and focused on optimization of the ICD and hinge regions. Among EREG_5 and CXCL16_69 variants tested, we identified a set of variants that showed higher surface translocation potential. The representative variants include CXCL16_69_c95, EREG_5_a11, EREG_5_Ca9_hTM and EREG_5_Ca9_ICD. The

CXCL16_69_c95 sequence suggests that the C terminus of ICD is dispensable and key motifs exist within its first 39 residues (Supplementary Fig. 16) with high attention weights on variant-specific LP and KA dipeptides at N-term and C-term of TMD (Source Data of Extended Data Fig. 8b-8c). Interestingly, the CXCL16_69_c95 achieved highest per-residue efficiency at displaying protein of interest on cell surface among ~235 screened CSD modules (potent CSD with merely 43 residues).

For the EREG_5 fusion variants, domain shuClng with Ca9 or ACE maintained or enhanced EREG_5 surface translocation, suggesting potent CSD functional domains are

interchangeable. ACE hinge-transmembrane domain (hTM) and ICD fusion of EREG_5 behaved distinctly when displaying E7 antigens indicating that hTM domain inherits major characteristics of CSD sequences. For EREG_5_a11, it introduced more charged residues in ICD compared to EREG_5 template and its high attention weights share similar distribution as EREG_5 template indicating that key motifs at ICD-TMD interface in EREG_5_a11 variant remain unchanged.

Supplementary Fig. 16

To sum up, EREG_5_a11 and CXCL16_69_c95 represent the most advanced gCSD design to date in each category. Their key motifs include 1) conserved residues in r1 gCSD templates; 2) r1 gCSD specific residues within or adjacent to TMD; 3) r1 & r2 introduced charged residues at extracellular hinge and ICD.

In addition, to address reviewer requests for interpretability, we analyzed sequence features at the TMD–ICD junction in the labeled natural screen (Fig. 2h, 3c, 3d, and 4e). High CST sequences exhibit increased basic residue enrichment and polybasic motif patterns immediately downstream of the TMD, as shown by position-wise enrichment (Supplementary Fig. 6) and sequence logos analysis (Supplementary Fig. 8, Appendix Figure 2). These analyses distill the screens and deep learning models into concrete design

rules and shared motifs between strong natural and designed CSDs, strengthening the biological insight of the paper. The Appendix Figure 2 supports the idea that polybasic character near the TMD–ICD boundary is associated with higher CST activity.

Appendix Figure 2

Appendix Figure 2 | Sequence logo analysis of first 20 aa at N-term of natural ICD (ICD head) from CSD modules with low or high cell surface translocation (CST) activity. a-b, The sequence logo analysis revealed visibly higher enrichment of basic residues (R/K) in high CST (Appendix Figure 2a) versus Low CST (Appendix Figure 2b), concentrated near the immediate post-TMD region.

I can't help but feel that this paper might actually have been stronger if it ended at Figure 3, presenting a focused empirical study with clear findings.

Response: We truly appreciate this suggestion. As shown above, we have followed this instruction to have streamlined the paper, and strengthen it with new systematic validation data and new analysis, with the initial focus of “understand what makes CSD good and AI-engineer them for better”. We kept the original Figs. 1-3, combined original Fig. 4-6 into a new Fig. 4 and moved model architecture-related figures to Extended Data Fig. 8-10. We also provided new experimental validation data for the gCSDs in new Fig. 5. We hope the revised paper now presents a more focused empirical study with better clarity.

References cited for rebuttal letter:

- 1 al., X. e. Self-training with Noisy Student improves ImageNet classification. arXiv (2020).
- 2 Hinton G., e. a. Distilling the Knowledge in a Neural Network. arXiv arXiv:1503.02531 (2014).
- 3 Lee, D.-H. e. a. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. Challenges in Representation Learning (2013).
- 4 Sun, Y., et al. Out-of-Distribution Detection with Deep Nearest Neighbors. ICML (2022).

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have fully addressed my concern, including a web portal for easy use. I think it would be an impactful study.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

My compliments to the authors for the pretty extensive lift! A lot of my comments were well-addressed, esp around organization, and interpretability. My critiques on self-training or the difference between dense and linear were kind of "talked through", but I'll live with that.

This is solid work, and I have no additional feedback.

(Remarks on code availability)

Author Rebuttal letter:

Reviewer #1:

Remarks to the Author:

The authors have fully addressed my concern, including a web portal for easy use. I think it would be an impactful study.

Reply: We thank the reviewer for your valuable time and constructive feedback that help enhanced the work.

Reviewer #3:

Remarks to the Author:

My compliments to the authors for the pretty extensive lift! A lot of my comments were well-addressed, esp around organization, and interpretability. My critiques on self-training or the difference between dense and linear were kind of "talked through", but I'll live with that.

This is solid work, and I have no additional feedback.

Reply: We thank the reviewer for your valuable time and constructive feedback that help enhanced the work. We echo that the self-training part is responded textually and remain an important topic for future studies.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>