StemCell-GPT: A Specialized AI Agent For Human Stem Cell Engineering

Jingwen Hui¹, Freja Kjellaug Amalia Ekman², Hana Yousef Ghanim², Sridhar Selvaraj² Yuanhao Qu³, Matthew Porteus², Le Cong⁴

¹Shu Chien – Gene Lay Department of Bioengineering, UC San Diego
²Department of Pediatrics, Stanford University
³Cancer Biology Program, Stanford University
⁴Department of Pathology, Stanford University

Abstract

CRISPR technology has revolutionized genetic medicine, enabling programmable genome modifications. Central to its success is high-performance guide RNA (gRNA) that directs Cas9 nuclease to desired genomic targets. Although the first CRISPR therapy was approved by FDA in 2023 to edit hematopoietic stem cells, it remains the only proven gene-editing treatment to date. While numerous gRNA design tools exist, they are primarily geared toward cancer cells. As a result, designing high-efficiency, context-aware gRNAs for precise stem cell editing remains a critical bottleneck. This work introduces StemCell-GPT, a specialized AI agent designed to automate and enhance stem cell editing through multi-objective SNV-aware CRISPR gRNA design, solving stem cell-specific engineering queries, and designing ssODN templates for homologous directed repair. The StemCell-GPT pipeline first generates candidate gRNAs in a target region and refines their on-target scores using DNA language model embeddings calibrated by a Random Forest regressor (achieving MAE = 4.94%, $R^2 = 0.685$). It then fuses calibrated scores with insertion-deletion mutation profiles, using a feed-forward network to optimally weight each component based on real-world stem cell editing data. The entire workflow is managed by a large language model-based agentic process, and we benchmarked its performance on four therapeutic genes (CCR5, HBB, STING1, CFTR), achieving an average Spearman correlation of 0.85 between the predicted and experimental gRNA design rankings. These results highlight the value of agentic exploration in navigating a vast bio-design space. By streamlining multi-objective optimization and providing context-aware engineering support, StemCell-GPT accelerates precision gene editing in stem cell research and paves the way for robust, high-throughput clinical applications.

1 Introduction

CRISPR–Cas9 systems have revolutionized genetic engineering by creating precise double-strand breaks at defined genomic loci, thereby enabling homology-directed repair (HDR) to insert exogenous DNA with high fidelity [Cong, 2013]. However, HDR's efficiency remains constrained by competition from faster, error-prone pathways such as non-homologous end joining (NHEJ). In mammalian cells, NHEJ is the dominant repair mechanism, but it leads to unpredictable DNA insertions and deletions [Guo, 2018]. On the other hand, microhomology-mediated end joining (MMEJ) can substantially elevate HDR frequencies [Fu, 2021]. Recent studies have shown that CRISPR guide RNAs (gRNAs) that bias towards MMEJ repair will yield a higher HDR efficiency in therapeutic scenarios [Tatiossian, 2021]. However, current gRNA design tools focus on generic on-target activity and off-target

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).



Figure 1: a) and b) LLM interprets the user query and identifies target genomic regions and retrieves the corresponding DNA sequence. c) SNV-aware guide-RNA design generates candidate guides by filtering out sequences overlapping population-level (gnomAD) or individual-specific (VCF) variants. FlashFry is then used solely to compute raw on- and off-target scores for these SNV-filtered candidates. FORECasT simulates spCas9 activity to derive the MMEJ/NHEJ repair ratio. d) Random forest model is used to fine-tune raw on-target scores with DNABERT-generated sequence embeddings. e) A small neural network is trained to learn sequence-dependent weights that balance on target score and the repair pathway ratio. f) Raw scores of each guide are recomputed with fine-tuned models. g) Guides are ranked by their predicted HDR% for output. h) RAG retriever fetches relevant stem-cell engineering passages from a pre-indexed literature corpus, lowering the learning curve by letting scientists pose natural-language questions and receive precise, domain-specific guidance. Optionally generate ssODN designs with homology arms, desired edits, and silent mutations—to accompany each selected guide.

minimization. In addition, current tools neither consider repair-pathway biases nor leverage large language models (LLMs), neural networks pretrained to understand plain English.

1.1 CRISPR-Cas9 System and DNA Repair Pathways

CRISPR–Cas9 introduces a targeted double-strand break (DSB) by pairing a 20-nt guide RNA (gRNA) with its complementary genomic locus and recruiting the Cas9 endonuclease [Redman, 2016]. Once the DSB is cut by the endonuclease, eukaryotic cells employ one of several repair pathways. Non-homologous end joining (NHEJ) is the fastest and most prevalent in mammalian cells. It rejoins broken ends directly but often introduces small insertions or deletions (indels), producing undesired outcomes [Davis, 2013]. Homology-directed repair (HDR) utilizes a homologous DNA template to achieve precise sequence insertion or correction, resulting in high repair accuracy and making it the optimal choice for gene editing tasks. But, HDR is inherently slower and active only during S/G2 phases of the cell cycle [Rein, 2018]. Microhomologous sequences flanking the break to realign ends [Wang, 2017]. MMEJ yields more predictable indels and, in many contexts, enhances HDR when coupled with appropriate gRNA design [Yuan, 2024]. As a result, gRNA design remains as a crucial task for researchers.

1.2 Existing gRNA Design Tools

Several computational platforms assist researchers in selecting CRISPR–Cas9 guide RNAs. However, each tool addresses only a subset of the overall design problem. Web-based tools such as CHOPCHOP allow users to input gene names, genomic coordinates, or raw sequences and apply empirical models (e.g., Doench 2016 Rule Set 2) alongside Bowtie-based off-target scanning to rank candidates by cleavage efficiency and specificity McKenna [2018], Doench [2016]. The Broad Institute's CRISPick similarly uses gradient-boosted decision trees trained on large-scale knockout data and penalizes potential off-targets via the CFD score. However, it requires precise genomic inputs and does not model repair-pathway preferences Doench [2016], Horlbeck [2016]. Command-line softwares like FlashFry excel at high-throughput enumeration and annotation of all possible 20-bp spacers in a locus and reporting multiple on- and off-target metrics. But Flashfry leaves composite ranking and repair-outcome prediction to downstream processing McKenna [2018]. Deep learning frameworks

such as DeepCRISPR learn sequence and epigenetic features to predict both efficacy and off-target risks in a single model, yet they omit explicit predictions of NHEJ versus MMEJ repair pathways and offer no natural-language interface Chuai [2018]. More comprehensive servers like CRISPOR consolidate multiple scoring schemes (Doench 2016, CFD) and estimate frameshift frequencies across over 120 genomes, but still rely on structured inputs and do not integrate large-language models or HDR-focused repair modeling Haeussler [2016].

2 Our Contributions

We present StemCell-GPT, the first end-to-end, sequence-aware AI-agent pipeline that integrates natural-language-driven genomic targeting, SNV-aware sgRNA design (population and personal variant filtering), DNABERT-informed score calibration, automated repair pathway predictions, sequence-dependent sgRNA ranking, and retrieval-augmented generation (RAG). Additionally, we introduce automated ssODN design, streamlining experimental editing workflows.

3 Methods

StemCell-GPT is a specialized AI agent for CRISPR guide RNA (gRNA) design in stem cell engineering. StemCell-GPT integrates LLM-driven natural-language understanding of user-specified tasks, SNV-aware sgRNA design at both population-wide and personal levels, DNABERT-informed calibration of on-target score prediction, and agentic balance of on-target and DNA repair-pathway preferences for optimized ranking. Additionally, the system introduces basic ssODN design capabilities to facilitate experimental workflows. Below we detail the datasets and computational methods used

3.1 Datasets and Preprocessing

Wet-lab: HDR efficiencies for 58 CRISPR guides were measured as detailed in Supplementary Appendix A.1. In brief, CD34⁺ HSPCs were electroporated with Cas9 RNPs and donor templates; editing outcomes were quantified by Sanger sequencing followed by ICE analysis.

Feature Scores: Raw on-target activity scores were generated by FlashFry, sequence embeddings were extracted via DNABERT and reduced to three PCA components, and repair ratios were predicted with FORECasT.

Normalization and Cross-Validation: All features were scaled to [0, 1] using Min-Max normalization. We then employed 5-fold cross-validation (shuffle, random seed 42) for hyperparameter tuning and evaluation of both the Random Forest calibration and sequence-aware weight models.

3.2 Computational Pipeline

3.2.1 Layer 1: LLM-Driven Query Interpretation and Sequence Retrieval

LLM is used as a "front end" that allows users to specify their genomic target in English (e.g. "I want to target the STING1 gene in human cells at 139480867–139480891"). StemCell-GPT passes this text into GPT-40 via LangChain and prompts it to return a small JSON output that contains four fields: the gene name, the species, a flag indicating whether the user explicitly supplied coordinates, and the precise genomic region to target. This structured output is then parsed to drive our downstream pipeline steps.

With the gene name and genomic location, StemCell-GPT automatically retrieves the corresponding DNA sequence using NCBI's Entrez API. We first look up the gene in NCBI's Entrez to retrieve its GRCh38 location. Then, we download the corresponding genomic DNA sequence (optionally, can include a flanking window) to produce the exact nucleotide context for guide-RNA design.

3.2.2 Layer 2: SNV-Aware Guide Design and Raw Scoring

Candidate sgRNAs in StemCell-GPT are generated through a two-level (population and personal) SNV-aware filtering strategy. First, at the population level, we pre-index common SNVs from gno-mAD (Minor Allele Frequency $\geq 10\%$) using sorted lists with binary-search lookups to exclude any guide overlapping these variants. Next, at the individual level, personal VCFs (e.g., NA12878/HG001)

are parsed using cyvcf2, and variants are stored in chromosome-indexed dictionaries to rapidly filter out sgRNAs that intersect personal SNVs. This ensures that all candidate guides avoid both common and personal variant sites, increasing specificity and reducing off-target risk.

After SNV-based filtering, we compute raw on-target scores using FlashFry's Doench-2014 model, along with standard off-target efficiency metrics. These raw scores feed into Layer 3 (Random Forest calibration), followed by DNABERT embeddings, PCA reduction, MMEJ/NHEJ prediction via FORECasT, and final sequence-aware weighting, as described in the subsequent layers.

3.2.3 Layer 3: Random Forest–Based On-Target Score Refinement

To further enhance the accuracy of Flashfry's raw on-target score predictions, we train a Random Forest Regressor that uses both the initial Flashfry score and sequence embeddings to produce a calibrated on-target score.

Sequence Learning We have 58 guides that are 23 base pairs long (20bp spacer + 3bp PAM). Each sequence was passed into a pretrained DNABert transformer model [Ji, 2021] to get a 768-dimensional vector embedding. Then, we reduce each embedding to just three principal components via PCA to reduce overfitting.

Calibration Steps We take each guide's raw FlashFry score and its three PCA-reduced DNABERT embeddings and use them to train a Random Forest model that maps those four inputs onto our experimentally measured on-target score. Specifically, we split the 58 guides into an 80/20 training/validation set. Then, we run a grid search with 5-fold cross-validation to choose the best tree count, maximum depth, and feature-sampling strategy. Once the optimal hyperparameters are identified, we retrain the Random Forest on the full training set, confirm its performance on the held-out validation set, and freeze the resulting model weights. These frozen parameters are then used in all downstream scoring steps. Therefore, we ensure that our calibrated on-target predictor remains consistent and reproducible across new guide evaluations.

3.2.4 Layer 4: Repair-Ratio Prediction

StemCell-GPT uses the FORECasT software to derive the MMEJ/NHEJ ratio. Specifically, FORE-CasT predicts the spectrum of indel outcomes for a given target sequence by simulating spCas9 cutting and subsequent repair, yielding counts of microhomology-mediated end joining (MMEJ) versus classical non-homologous end joining (NHEJ) events. We then compute the repair-ratio as

$$y = \frac{\text{total MMEJ events}}{\text{total NHEJ events}}.$$

This scalar y captures the ratio between the two DNA repair mechanisms for each guide and is used alongside our calibrated on-target score in the final ranking model.

3.2.5 Layer 5: Sequence-Aware Weight Model

We train a small feed-forward network that captures the guide-specific biases in how the on-target score and repair-pathway score combine. During training, each datapoint consists of the guide's DNABERT sequence embedding, its normalized and RF-calibrated on-target score x, and its normalized MMEJ/NHEJ ratio y, and the normalized HDR percentage s. Here, E denotes the DNABERT embedding of the full 23-nt guide plus its adjacent PAM sequence, capturing sequence motifs and structural features that influence HDR efficiency. The network uses E alone to predict a confidence weight

$$a_{\rm raw} \in [0, 1],$$

which then modulates the calibrated on-target score x and the MMEJ/NHEJ ratio y via ...

$$s_{\text{pred}} = 0.3 \, a_{\text{raw}} \, x + 0.7 \, a_{\text{raw}} \, y.$$

We optimize all network parameters by minimizing the mean-squared error

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (s_{\text{pred},i} - s_i)^2$$

using the Adam optimizer (learning rate 5×10^{-4} , weight decay 10^{-4}). Training is conducted for up to 1,000 epochs with a batch size of 16, applying 20% dropout after each hidden layer. A ReduceLROnPlateau scheduler monitors the validation MAE to adapt the learning rate. We employ a 5-fold cross-validation regime to ensure robust generalization, and upon completion we save the final, fold-averaged network weights for predicting $a_{\rm raw}$ in the downstream inference pipeline.

3.3 StemCell-GPT Workflow

Given a user query specifying the genomic target, StemCell-GPT proceeds as follows. First, it retrieves the relevant DNA sequence via LLM-based reasoning and NCBI Entrez lookup. Next, candidate sgRNAs are generated using a two-level SNV-aware filtering process: population-level filtering against gnomAD common SNPs (MAF $\geq 10\%$) and individual-level filtering using personal VCF data (e.g., NA12878). Only SNV-free guides pass through to FlashFry, which computes raw on-target scores with the Doench-2014 model and off-target metrics. These raw scores, denoted x, are then fed into the Random Forest calibration model, yielding calibrated on-target scores x_{cal} . FORECasT simulations produce the MMEJ/NHEJ ratio y. Each guide's DNABERT embedding is processed by a sequence-aware weight neural network to generate a confidence weight a_{raw} . The final efficiency score s is computed as:

$$s = 0.3 \, a_{\rm raw} \, x_{\rm cal} + 0.7 \, a_{\rm raw} \, y$$

This score is then rescaled to reflect predicted HDR% values, and guides are ranked accordingly.

Optionally, if ssODN design is requested, StemCell-GPT generates a candidate oligonucleotide following experimental guidelines. StemCell-GPT constructs a 90 bp sequences with 45 bp homology arms, the intended mutation, and silent mutations across the sgRNA site to prevent re-cutting on the same strand as the guide. Each ranked guide is returned alongside its corresponding ssODN.

3.4 Retrieval-Augmented Generation (RAG) for Stem Cell Queries

To enable StemCell-GPT to answer specialized stem cell engineering questions, we developed a retrieval-augmented generation (RAG) pipeline. This pipeline consists of three main components:

Paper Acquisition and Preprocessing We first curated a corpus by scraping relevant literature from CrossRef and Unpaywall APIs, downloading open-access PDFs, and extracting their textual content using PyMuPDF. Each paper was segmented into manageable chunks (1,000 tokens with 200-token overlaps) using the LangChain CharacterTextSplitter.

Embedding and Indexing Text chunks were embedded using the HuggingFace sentence-transformer model (all-MiniLM-L6-v2) and stored in a FAISS vector index for efficient retrieval.

Querying and Response Generation For query answering, we implemented a retrieval-enhanced pipeline using OpenAI's GPT-40 language model, integrated via LangChain. Given a natural language query, the model retrieves the most relevant text chunks from the FAISS index and synthesizes precise, domain-specific responses. Implementation details are described in Supplementary Appendix A.2.

3.5 Basic ssODN Design

StemCell-GPT v2 automates single-stranded oligodeoxynucleotide (ssODN) design, supporting insertions, deletions, and single-base substitutions with standard experimental guidelines:

- 45 bp homology arms flanking sgRNA cut sites (90 bp total).
- Introduction of specified mutations.
- Silent mutations near sgRNA cut sites extending to mutation points to prevent re-cutting.

• ssODN alignment matching sgRNA strand.

Further enhancements like GC content and thermodynamic stability are potential future developments.

4 Results

We present our evaluation in three parts. (1) On-Target score calibration (layer 3). (2) Sequence-aware weight training (layer 5). (3) End-to-end ranking performance on guides across multiple genes. (4) RAG System Query Demonstration.

4.1 On-Target Score Calibration (Layer 3)

We first optimized a Random Forest regressor to map FlashFry's raw on-target scores and DNABERTderived embeddings onto experimentally measured INDEL percentages. We tuned the model via a randomized search with 5-fold cross-validation. The best hyperparameters (50 trees, max depth = 3, and 1 sample per leaf) yielded a cross-validation MAE of 8.31%. When retrained on the full 58-guide dataset and followed by a linear calibration step, the model achieved an MAE of 4.94% and $R^2 = 0.685$ between observed and predicted INDEL rates (Figure 2).



Figure 2: True (experimentally measured) vs. calibrated editing efficiency, quantified as INDEL percentages, for the full stem cell editing dataset (scatter), with MAE and R^2 annotated.

These results, especially the high $R^2 = 0.685$ and tight alignment along the diagonal in Figure 2, show a strong correlation between true and calibrated INDEL percentages. It highlights the potential of this calibration method for reliable on-target score prediction.

4.2 Sequence-Aware Weight Model Performance (Layer 5)

We then evaluate our feed-forward "weight" network. This model learns the guide sequence-specific confidence $a \in [0, 1]$. We used a 5-fold cross-validation on the normalized on-target and repair0ratio inputs. Across all folds, validation MSE ranged from 0.015 to 0.069 and MAE from 0.095 to 0.198 on the normalized HDR% scale. These per-fold results are summarized in Table 1.

Table 1: Cross-validation MSE and MAE per fold for the sequence-aware weight model.

Fold	MSE	MAE		
1	0.065869	0.197804		
2	0.069732	0.194242		
3	0.043704	0.135646		
4	0.045504	0.170371		
5	0.015404	0.095251		

These results (particularly the fifth fold achieving the lowest MSE of 0.0154 and MAE of 0.0953) are impressive given our small 58-guide dataset and highlight the model's strong potential for sequence-dependent weight training in HDR prediction.

4.3 End-to-End Pipeline Validation Across Genes

Lastly, we applied the complete StemCell-GPT pipeline to four independent test genes (CCR5, HBB, STING1, and CFTR). We have 3-5 guides for each of those genes whose editing efficiencies have been determined experimentally. Table 2 reports the experimentally measured rank order versus our predicted ranking by HDR% for each gene.

Table 2: Comparison of the ground truth (from wet-lab stem cell engineering experiments) vs. predicted (using StemCell-GPT) gRNA rankings for 4 clinically relevant target genes.

CCR5		HBB		STING1		CFTR	
Real Rank	Predicted Rank						
1	3	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	1	3	4	3	3	3	3
4	4	4	3	_	_	_	_
5	5	_	_	_	-	-	-

To quantify the agreement between real vs predicted ranks, we calculate Spearman's rank-correlation coefficient for each gene as follows:

Spearman's rank-correlation coefficient is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \quad d_i = (\text{RealRank}_i - \text{PredRank}_i),$$

where n is the number of guides for that gene. Applying this definition to each gene yields:

$$\rho_{\rm CCR5} = 0.60, \quad \rho_{\rm HBB} = 0.80, \quad \rho_{\rm STING1} = 1.00, \quad \rho_{\rm CFTR} = 1.00.$$

These values demonstrate strong alignment between our predicted guide rankings and the experimental ground truth across all four genes.

4.4 RAG System Query Demonstration

We demonstrate StemCell-GPT's retrieval-augmented generation (RAG) capability with a representative example queryies found in Supplementary Appendix A.3. The responses are far more specific and literature-grounded than what we observe from a general-purpose LLMs like ChatGPT.

4.5 ssODN Design

StemCell-GPT also supports automated ssODN design. Upon user requests, it generates full ssODN templates—including homology arms, intended edits, and silent mutations—for each selected guide. A representative example is provided in Supplementary Appendix A.3.

5 Conclusion

The results show that integrating sequence embeddings and empirical calibration enhances CRISPR-Cas9 gRNA design for stem cell engineering. Particularly, in cases where HDR precision is critical. By using DNABERT embeddings with experimentally-informed Random Forest fine-tuning and FORECasT-derived DNA repair ratios, StemCell-GPT shows high potential in predicting HDRfocused editing efficiency. Nevertheless, future validation on more genomic targets and experimental validation is necessary to establish its generalizability. In particular, expanding our wet-lab guide pool beyond the current 58 gRNAs will likely further boost model performance and predictive power. In addition, the refinement of the model through in-house wet-lab data shows the importance of extending such an approach to more collaborative datasets to enhance robustness. Furthermore, the inclusion of a retrieval-augmented generation (RAG) system significantly expands StemCell-GPT's practical utility by enabling rapid, precise answers to complex, domain-specific queries based on the latest scientific literature. Future work will explore expanding the corpus to further enhance domain coverage and integrating user feedback to iteratively refine the retrieval system. Future system-level enhancements will refine the ssODN design module by incorporating additional factors such as GC-content and thermodynamics, further optimizing guide RNA and ssODN designs for complex genetic engineering tasks. We hope to expand StemCell-GPT's application to other genetic editing scenarios beyond wild-type CRISPR systems.

Acknowledgements

The research group hopes to acknowledge the Nebius team for their assistance with GPU credits.

References

- L. et al. Cong. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, Feb 2013. doi: 10.1126/science.1231143.
- T. et al. Guo. Harnessing accurate non-homologous end joining for efficient precise deletion in crispr/cas9mediated genome editing. *Genome Biol.*, 19:170, 2018. doi: 10.1186/s13059-018-1518-x.
- Y. W. et al. Fu. Dynamics and competition of crispr–cas9 ribonucleoproteins and aav donor–mediated nhej, mmej and hdr editing. *Nucleic Acids Res.*, 49(2):969–985, Jan 2021. doi: 10.1093/nar/gkaa1251.
- K. J. et al. Tatiossian. Rational selection of crispr–cas9 guide rnas for homology-directed genome editing. *Mol. Ther.*, 29(3):1057–1069, Mar 2021. doi: 10.1016/j.ymthe.2020.10.006.
- M. et al. Redman. What is crispr/cas9? Arch. Dis. Child. Educ. Pract. Ed., 101(4):213–215, Aug 2016. doi: 10.1136/archdischild-2016-310459.
- A. J. et al. Davis. Dna double strand break repair via non-homologous end-joining. *Transl. Cancer Res.*, 2(3): 130–143, Jun 2013. doi: 10.3978/j.issn.2218-676X.2013.04.02.
- L. A. M. et al. Rein. Applications of gene editing technologies to cellular therapies. *Biol. Blood Marrow Transplant.*, 24:1537–1545, 2018. doi: 10.1016/j.bbmt.2018.03.017.
- H. et al. Wang. Microhomology-mediated end joining: new players join the team. *Cell Biosci.*, 7:6, 2017. doi: 10.1186/s13578-017-0136-8.
- B. et al. Yuan. Modulation of the microhomology-mediated end joining pathway suppresses large deletions and enhances homology-directed repair following crispr-cas9-induced dna breaks. *BMC Biol.*, 22:101, 2024. doi: 10.1186/s12915-024-01896-z.
- A. et al. McKenna. Flashfry: a fast and flexible tool for large-scale crispr target design. *BMC Biol.*, 16:74, 2018. doi: 10.1186/s12915-018-0545-0.
- J. G. et al. Doench. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nat. Biotechnol.*, 34(2):184–191, 2016. doi: 10.1038/nbt.3437.
- M. A. et al. Horlbeck. Compact and highly active next-generation libraries for crispr-mediated gene repression and activation. *Elife*, 5:e19760, 2016. doi: 10.7554/eLife.19760.
- G. et al. Chuai. Deepcrispr: optimized crispr guide rna design by deep learning. *Genome Biol.*, 19:80, 2018. doi: 10.1186/s13059-018-1407-4.
- M. et al. Haeussler. Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor. *Genome Biol.*, 17:148, 2016. doi: 10.1186/s13059-016-1012-2.
- Y. et al. Ji. DNABERT: pre-trained bidirectional encoder representations from transformers model for dnalanguage in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

A Technical Appendices and Supplementary Material

A.1 Wet-Lab Procedures

A.1.1 Culture of iPSCs and CD34⁺ HSPCs

Human iPSCs were maintained under feeder-free conditions on Matrigel-coated plates (Corning, 354277) in mTeSR1 complete medium (STEMCELL Technologies, 85850) at 37 °C with 5% CO₂ and ambient oxygen. Human HSPCs were derived from fresh cord blood obtained through the Binns Family Program for Cord Blood Research and cultured in StemSpan SFEM II medium (STEMCELL Technologies, 09655) supplemented with 35 nM UM171, 100 ng/mL thrombopoietin (TPO), 100 ng/mL stem cell factor (SCF), 100 ng/mL FLT3 ligand, 100 ng/mL interleukin-6 (IL-6), 20 mg/mL streptomycin, and 20 U/mL penicillin at 37 °C with 5% CO₂ and 5% oxygen.

A.1.2 Genome Editing

Chemically modified single guide RNAs (sgRNAs) were obtained from Synthego or Integrated DNA Technologies, containing 2'-O-methyl-3'-phosphorothioate modifications at the three terminal nucleotides on both 5 and 3 ends. Ribonucleoprotein (RNP) complexes were formed by incubating $6 \mu g$ HiFi SpCas9 nuclease (Aldevron, 9214) with $3.2 \mu g$ sgRNA for 15 minutes at room temperature.

iPSCs were edited at ~80% confluency. Cells were dissociated using Accutase (STEMCELL Technologies) and resuspended in 20 μ L P3 Primary Cell Nucleofector Solution (Lonza) per reaction. Electroporation was performed in a 16-well nucleocuvette strip using the 4D-Nucleofector system (Lonza) with program CA-137. Cells were plated at a density of 100 000 cells per well in 48-well plates containing pre-warmed culture medium supplemented with 10 μ M Y-27632 ROCK inhibitor (Cayman Chemical, 10005583).

HSPCs were edited after 3 days in culture. Cells were resuspended in P3 buffer and electroporated using program DZ-100. Cells were plated at a density of 250 000 cells/mL in HSPC expansion medium.

For homology-directed repair (HDR), either $2.5 \,\mu$ M single-stranded oligonucleotide (ssODN) was added to the electroporation mixture, or AAV6 donor vectors were added directly to the culture medium immediately post-electroporation at an estimated multiplicity of infection (MOI) of 2500.

A.1.3 Indel Frequency Analysis by ICE

Genomic DNA was extracted 3 days post-electroporation using the QuickExtract DNA Extraction Solution (Lucigen, QE09050). Target loci were amplified by PCR using primers flanking the cut site (~ 600 bp amplicon), and PCR products were purified using the GeneJET Gel Extraction Kit (Thermo Fisher Scientific). Purified amplicons were submitted for Sanger sequencing (Genewiz or MCLabs), and INDEL frequencies were analyzed using the Synthego ICE (Inference of CRISPR Edits) online tool by comparing chromatograms from edited and unedited control samples.

A.1.4 ssODN Design Workflow (Wet-Lab)

StemCell-GPT automates ssODN design consistent with standard laboratory practices, following these steps:

- 1. Design 45 bp homology arms flanking the sgRNA-induced double-strand break site.
- 2. Precisely introduce specified mutations (insertions, deletions, substitutions).
- 3. Create silent mutations across the sgRNA recognition site extending to the desired mutation site to prevent repeated Cas9 cleavage.
- 4. Ensure ssODN design aligns with the sgRNA DNA strand.

A.2 Implementation of RAG Pipeline

The RAG system was implemented in Python, leveraging the HuggingFace sentence-transformers/all-MiniLM-L6-v2 model for embedding generation and FAISS for indexing. The query-answering pipeline utilized GPT-40 via the LangChain library.

A.2.1 Additional Query-Response Examples

Below are examples demonstrating the retrieval-enhanced capability of StemCell-GPT:

Query: What is significant about U2AF1 S34F mutations in lung adenocarcinomas?

StemCell-GPT Response:

- 1. They co-occur with ROS1 translocations, suggesting a functional link between these mutations in LUADs.
- 2. They alter RNA binding specificity, affecting introns with CAG trinucleotides at their 3' splice sites, leading to alternative splicing events.
- 3. They upregulate genes involved in epithelial-mesenchymal transition (EMT), thereby increasing tumor invasiveness.
- 4. They promote splicing of the longer SLC34A2–ROS1 isoform, associated with enhanced invasive characteristics, demonstrating mechanistic synergy between mutant U2AF1 and ROS1.

Query: *How to Improving the safety of human pluripotent stem cell therapies?*

Response: Improving the safety of human pluripotent stem cell (hPSC) therapies can be achieved through several strategies:

- **Genome-Edited Safeguards:** Implementing genome-edited orthogonal safeguards, such as the NANOGiCaspase9 system, which eliminates undifferentiated hPSCs to prevent teratoma formation, and the ACTBTK and ACTBOiCaspase9 systems, which can eliminate the entire hPSC-derived cell product if adverse events occur.
- **Genetic Stability:** Ensuring the genetic stability of stem cells during culture to prevent mutations that could lead to tumorigenesis.
- **Purity of Cell Populations:** Developing methods to ensure that differentiated cell populations are free from undifferentiated pluripotent stem cells, which could form teratomas.
- **Controlled Differentiation:** Refining differentiation protocols to produce homogeneous and functionally mature cell types that are less likely to cause adverse effects.
- **Immunogenicity:** Addressing potential immune rejection by using autologous cells or engineering cells to be less immunogenic.
- Monitoring and Quality Control: Implementing rigorous monitoring and quality control measures during cell production and before transplantation.
- **Regulatory Compliance:** Adhering to regulatory guidelines and conducting thorough preclinical and clinical testing to ensure safety and efficacy.

These strategies, along with consulting recent scientific literature and guidelines from regulatory bodies, can help improve the safety of hPSC therapies.

Query: How to perform gene correlation in HSC cells with X-linked sSevere cCombined ilmmunodeficiency (SCID-X1)?

Response: The process of gene correction in HSC cells with X-linked Severe Combined Immunodeficiency (SCID-X1) involves using a CRISPR-Cas9/AAV6 based strategy to achieve targeted integration of a cDNA into the endogenous start codon. This method functionally corrects diseasecausing mutations throughout the gene. The approach has been shown to achieve high targeting frequencies in CD34⁺ HSPCs from SCID-X1 patients and can rescue lymphopoietic defects both *in vitro* and *in vivo*. This technique is described as a clinically relevant, selection-free "universal" CRISPR-Cas9-rAAV6 gene editing methodology, capable of potentially correcting over 97% of known IL2RG pathogenic mutations. This method is termed "functional gene correction" because it corrects mutations by integrating cDNA to functionally address downstream mutations rather than directly correcting the mutation itself.

Query: How to use a mouse model to to study Cas9-AAV6-mediated HBB-correction in functional HSC?

Response: To study Cas9-AAV6-mediated HBB-correction in functional hematopoietic stem cells (HSCs) using a mouse model, you can follow these steps:

- Mouse Model Selection: Use the Townes-SCD mouse model, which is a humanized model carrying the human HBA and the sickle cell disease (SCD) mutant form of HBB (HbS). This model exhibits symptoms of SCD, such as short RBC half-life and RBC sickling.
- HSC Isolation and Culture: Isolate CD150⁺CD34⁻Kit⁺Sca1⁺Lineage⁻ cells from mouse bone marrow. Culture these HSCs in media that supports their ex vivo expansion.
- **Gene Editing Protocol:** Use Cas9-AAV6 technology for gene editing. Electroporate HSCs with HiFi Cas9 protein pre-complexed with sgRNA and an AAV6-repair template containing the desired genetic correction.
- Autologous Transplantation: Transplant the corrected HSCs back into irradiated Townes-SCD recipient mice. Use a radiation dose to clear endogenous erythropoiesis and support the mice with blood transfusions for recovery.
- Assessment of Gene Correction: Track engraftment and function of gene-corrected HSCs by analyzing peripheral blood for hemoglobin tetramers using HPLC, and assess hematological parameters like reticulocyte counts and RBC morphology.
- **Long-term Evaluation:** Monitor the long-term stability of hemoglobin A production and reduction of sickle hemoglobin in the transplanted mice. Evaluate the correlation between myeloid chimerism and erythropoietic correction.

A.3 ssODN Design Example

Below is a ssODN design examples. Desired mutations are shown in red, silent mutations in blue, and sgRNA sequences are <u>underlined</u>:

• HBB E06V (mutation offset from cut site) sgRNA: CTTGCCCCACAGGGCAGTAA

Engineered ssODN:

 $\label{eq:caacticacc} CAACTTCATCCACGGTTCACCCCCACAGGGCAGTGACTGCGGGATTTTTCCACAGGAGTCAGGATGCACCATGGTGTCTGTTTGAGGTT$

Wild Type Sequence: