DATA REQUIREMENTS FOR AN AI-GUIDED STEM-CELL-ISLET PRODUCT MODEL IN TYPE 1 DIABETES

OpenAI 5.6 Sol
GloGlo Research Team

Abstract. Cellular foundation models can represent molecular state, but a predictive stem-cell-islet product model requires independent manufacturing campaigns joined from bank, edit and process history to post-thaw function, release, immune and graft outcomes. We audited 52 public or catalogued publication–repository objects and separately examined the closest public SC-islet process datasets at the campaign level. No object met the complete minimum for its proposed decision claim, and no accessible cohort supplied an end-to-end campaign-to-outcome join. Existing resources support narrower work on state representation, hypoxia, assay design and baseline physiology; they do not validate product-to-release or product-to-immune-safety prediction. We formalize the missing object as an all-initiated typed lineage graph and propose a minimum campaign-to-outcome data contract. The immediate priority is prospective data architecture, with existing models used as challengers on held-out campaigns.

1 Introduction

Stem-cell-derived islets have moved from a source-cell hypothesis to a product development problem. Stem-cell-derived islets can restore glucose-responsive function after transplantation [1], while immune-evasive strategies seek to remove the requirement for chronic systemic immunosuppression. The remaining questions concern a living product: whether manufacturing changes preserve post-thaw potency, whether an edited product survives relevant immune routes without losing host defense or beta-cell function, and whether release measurements predict graft behavior.

Large cell atlases and cellular foundation models provide rich molecular representations. Their scale can obscure a basic distinction. A model can learn regularities among cells while receiving almost no independent evidence about manufacturing campaigns, immune donors, graft recipients or failed lots. The number of sequenced cells is then unrelated to the number of units available to test a product decision. Likewise, datasets from different cell lines, process versions and assay systems cannot be joined by matching gene names alone: study identity may become a shortcut for product and protocol.

We therefore asked two linked questions. First, what claims are supported by the public and catalogued T1D data objects already identified by a broad publication-to-data audit? Second, does any public or demonstrably requestable object support a narrower SC-islet manufacturing-to-outcome model? This article reports the resulting resource map and translates the negative answer into a constructive data contract.

Date: August 30, 2026.
Key words and phrases. Type 1 diabetes, stem-cell-derived islets, cellular foundation models, data adequacy, manufacturing campaigns, potency, immune evasion, prospective validation.
OpenAI 5.6 Sol carried out the evidence synthesis and prepared the article. The GloGlo research team posed the research question, set the program direction, and guided the investigation.

Molecular foundation models can provide useful representations and generate hypotheses. A product model becomes testable when its target outcome, physical lineage and independent unit are observed together.

2 Audit design

2.1 Two evidence strata

The first stratum was a frozen matrix of 52 publication or repository objects covering five claim classes: beta-cell fate and function, immune–beta interaction, antigen-specific tolerance, disease-linked long-read HLA/KIR and joined multi-omics. Perturbation-screen reliability and assay-control objects were included where they could change whether an analysis was admissible.

The second stratum was a targeted search for SC-islet process data joined to product outcomes. It included public process-development, scale-up, hypoxia, maturation and multi-omic datasets; controlled-access catalogues; clinical islet-transplant analogues; public regulatory records; and public descriptions of sponsor or consortium programs. Evidence was frozen on August 30, 2026. Later releases and partner-held corpora can update the resource map.

The inferential unit in this audit is a *publication or repository object*. Multiple files belonging to one deposit do not become independent objects, and multiple cells, wells, aliquots or time points do not become independent manufacturing campaigns. A source was counted once at the most specific object level needed to decide access, lineage, endpoint and unit adequacy.

2.2 Claim-specific eligibility

An object was eligible only when it exposed the biological unit, denominator, pairing structure, complete tested or initiated universe, direct endpoint and sample map needed for its proposed claim. Access alone was insufficient. For example, an open transcriptomic matrix may support cell-state annotation while failing a beta-cell-protection claim because it has no absolute fate or dynamic function. An immunopeptidomic deposit may support presentation controls while failing an immune-protection claim because it has no paired effector, killing or post-challenge beta-cell function.

For the manufacturing question, the minimum was stricter. A qualifying object required:

1. independent initiated SC-islet campaigns, with aborted, failed, marginal and deviating campaigns retained;
2. immutable lineage from bank, edit and process version through campaign, product lot, final fill, shipment, thaw and assay specimen;
3. time-stamped process measurements and intermediate quality controls;
4. post-thaw viable beta-cell mass and direct dynamic potency;
5. a campaign-linked release, safety, graft or clinical outcome; and
6. a reconstructable machine-readable join and rights compatible with independent evaluation.

Failure of the manufacturing-to-release minimum also precludes the broader product world model, which would additionally join immune, host-defense, abnormal-lineage, graft-site and clinical outcomes.

2.3 Decision rule

Each object received a terminal result for a named claim: usable now for a narrow estimand, metadata or access gated, useful as an assay precedent, or ineligible for the proposed model. Missing outcomes were not imputed from omics, static markers or related studies. An absent row from a selected-hit table was treated as missing rather than negative. Cells and technical replicates remained nested within their donors, campaigns or recipients.

3 Results

3.1 The 52-object audit

No object passed the complete minimum for a human intervention, immune–beta interaction, antigen-specific tolerance, disease-linked long-read HLA/KIR or joined outcome-linked multi-omics claim. The public landscape instead supports narrower uses separated by structural gaps (Table 1).

Claim class	What available objects can support	Missing decision evidence
Beta-cell fate and function	Baseline physiology, static secretion, cell-state and selected-perturbation descriptions	Starting and final viable beta-cell mass, early death, dynamic secretion, content/processing and identity in the same independent units
Immune–beta interaction	HLA-presentation pipelines, selected peptide and reporter controls	Typed target–effector pairing, absolute killing, matched beta-cell function and complete assignment controls
Antigen-specific tolerance	Paired immune transcriptional response and assay precedents	Exact payload, carrier and dose joined to lineage-resolved challenge, washout, recall and direct beta-cell outcome
HLA and long-read variation	Technical caller references and controlled-access leads	Phased HLA/KIR joined in the same people to beta-cell or immune response and molecular measurements
Joined multi-omics	State maps, stress responses and mechanism constraints	Authoritative donor/sample/modality/condition/batch crosswalks and direct outcomes at adequate independent-unit scale

Table 1: Valid uses and missing evidence in the frozen 52-object audit. Each row separates what the objects measure from the stronger decision claim for which they were considered.

Several objects are immediately useful within these limits. Complete guide tables can support screen-reliability analyses. Participant-paired RNA data can support a narrow immune transcriptional contrast. Public immunopeptidomics can qualify presentation controls. Human-islet and pancreas atlases can describe baseline or stress states after donor and modality crosswalks are resolved. None of these analyses creates an intervention outcome that was not measured.

3.2 No public campaign-to-outcome SC-islet cohort

The targeted manufacturing audit reached the same conclusion at a more specific level. No public or catalogued controlled-access object joined an adequate set of independent SC-islet campaigns from process history to post-thaw dynamic potency and a product outcome.

The nearest public objects are scientifically useful fragments (Table 2). GSE264176 describes process development and a 500-mL reactor setting but deposits RNA rather than a row-level reactor-

to-endpoint cohort [2]. GSE266878 reports vertical-wheel scale-up and dynamic function, while the deposited single-cell data comprise two final-stage libraries and do not expose a complete per-campaign join [3]. GSE263130/PXD062589 provides a valuable controlled hypoxia perturbation with multi-omic measurements, not uncontrolled campaign variation or a release decision [4]. GSE301598 is an acute hypoxia time course from one differentiation batch and one adult-islet donor, not a manufacturing cohort [5]. Developmental and maturation series similarly map biological state without a process-to-release join [6, 7].

Object	Narrow use	Why it is not a product-model cohort
GSE264176	Process measurement schema and stage comparison	Deposited RNA is not joined row by row to reactor telemetry, post-thaw dynamic potency and release outcome
GSE266878	Scale-up and assay precedent	Two deposited final-stage libraries; no complete campaign-level functional and release table
GSE263130 / PXD062589	Hypoxia-response benchmark	Programmed oxygen exposure is a treatment contrast, without campaign history or release outcome
GSE301598	Acute hypoxia biology	One SC-islet differentiation batch; time points are repeated observations, not independent campaigns
GSE167880 and GSE270220	Maturation and cross-line state references	No verified process-to-post-thaw-potency-to-release join
CIT-07/08 [8]	Clinical-islet manufacturing analogue	Deceased-donor primary islets rather than SC-islets; the complete machine-readable lot-to-recipient join was not verified

Table 2: Closest real data resources and their supported scope. The independent units and endpoints do not match the proposed product-model decision.

Public records make it reasonable to infer that sponsors maintain CMC, batch, release, deviation and traceability records. They do not demonstrate that those records form one analyzable corpus, retain all initiated failures, join to direct post-thaw function and later outcomes, or permit independent model development. Nor can a public molecular atlas be merged with a separate clinical report to manufacture the missing physical lineage.

3.3 Why scale does not repair the join

Let C denote an independent parental manufacturing-campaign component, X the measurements available at a declared prediction time, and Y the later product outcome. A decision-relevant benchmark requires observations of (C, X, Y) with lineages partitioned so that no connected parental component crosses development and test sets. Public fragments often provide abundant observations of X conditional on a small number of campaigns, or provide aggregate Y without a join to X. Increasing the number of cells within a campaign does not increase the number of independent (C, X, Y) units.

Three additional shortcuts fail for the same reason. First, restricting to released lots conditions on a downstream selection event and removes the failures needed to learn release risk. Second, combining products and protocols without explicit version fields allows the source study to predict the label. Third, synthetic outcomes or pretrained embeddings may improve regularization but cannot identify the unobserved campaign-to-outcome mapping.

4 A minimum campaign-to-outcome data contract

The missing object can be represented as a closed typed lineage graph. Its minimum semantic content consists of nine families:

U—Universe. One record for every initiated campaign, including aborted, failed, marginal, reworked and discarded paths.

S—Manufacturing state. Bank, cell line, edit, process, manufacturing-site version and planned and actual scale bound to each campaign and physical product.

X—Disposition. Time-stamped termination, deviation, rework, discard and release history.

P—Product lineage. Typed campaign-to-lot-to-final-fill relations with declared cardinality.

R—Release lineage. Final fill to post-thaw specimen or event to release decision, including timing and decision version.

D—Dose. A versioned full-dose-equivalent definition attached to the final fill or release record and explicit about pre- or post-thaw basis.

Q—Dynamic potency. Raw dynamic secretion on the relevant final product specimen, joined to viable beta-cell mass, content/processing, identity, assay precision and release decision.

C—Cold chain. Final fill to shipment, receipt, destination, thaw and post-thaw release lineage, with excursions retained.

K—Comparability. Named predecessor/successor process, scale, site or cryopreservation states joined to affected campaigns, prespecified critical quality attributes, pairing/blocking keys, acceptance rule and decision.

For immune-evasive products, the contract extends rather than compresses this graph. Split-lot descendants retain the same parental campaign identity while recording independent immune donor, challenge class and assay run. Direct allorecognition, recurrent autoimmunity, indirect or semidirect presentation, NK, macrophage, humoral, antibody-dependent and complement-dependent routes remain distinct endpoints. Host-defense, abnormal-lineage control and graft fate are separate non-compensatory modules. A single aggregate "safety score" would discard precisely the failure modes that the product design must resolve.

4.1 A two-stage access path

The contract supports a low-risk partner workflow. Stage 1 contains no biological values. A holder returns field availability, definitions, cardinalities, join structure, missingness counts, initiated-campaign status counts, grouping keys and whether a future outcome can remain concealed. A positive Stage-1 result means only that a row-level audit is worth requesting.

Stage 2 is separately permissioned and holder-local. It provides the smallest value-bearing bundle for one named decision: all initiated campaigns, physical lineage, time-zero inputs, a direct later outcome, failure and missingness records, and an untouched holdout. The model and simple baseline can execute inside the holder's environment; only frozen aggregate evaluation results need leave it.

4.2 Eligibility floors are not power calculations

For discovery conversations, we propose an eligibility floor of at least 60 development campaigns, including at least 20 observations on each side of a frozen decision boundary, plus at least 20 additional chronological or site-held-out campaigns. These numbers are a screening rule for whether a data-room analysis is plausible. They are *not* a power calculation and do not establish adequacy for any particular model.

The final sample requirement depends on outcome prevalence, effect size, parental-batch clustering, repeated process versions, missingness, assay reliability, the number of candidate predictors, the desired false-positive rate and the intended deployment shift. It must be calculated after a narrow estimand, decision boundary and variance structure are frozen. A smaller prospective set can support a preregistered challenger benchmark, but cannot be relabelled as a trained product world model.

5 Machine-readable resource package

The proposed supplement is designed for independent updating rather than a static narrative review. It contains four versioned objects:

1. `decision_matrix.tsv`: one row per publication or repository object, with accession, claim class, access class, independent unit, supported estimand, denominator and pairing gate, endpoint gate, terminal result, reversal condition and source artifact;

2. `scislet_object_manifest.tsv`: the targeted manufacturing sources, date last checked, campaign denominator, deposited modalities, process-to-endpoint join status, terms and exact scope;

3. `campaign_outcome_contract.json`: a closed schema for the U–S–X–P–R–D–Q–C–K lineage families, temporal roles, grouping keys, all-initiated status counts and holdout declarations; and

4. `provenance.tsv`: primary landing pages, persistent identifiers, retrieval dates, local checksums where files were acquired, and the audit decision each source supports.

Every release receives a date and schema version. New objects append rows; they do not silently replace prior negative terminals. A changed terminal must cite the newly available field, join, outcome or right that reverses the earlier decision. The supplement thereby separates a changing data landscape from a fixed scientific acceptance rule.

6 Implications for an AI-for-T1D research program

The audit supports three immediate activities. First, use public molecular data for bounded state, stress and assay benchmarks with donor- or campaign-correct evaluation. Second, test available foundation models only as challengers against frozen simple baselines; pretrained scale grants no product authority. Third, make prospective data design a primary research output: define the product lineage, preserve failures, conceal a future outcome and hold out parental campaign components before model selection.

It does not support pooling public fragments into a synthetic T1D product corpus or training a large product-specific model before the campaign ledger is shown. The first useful partner question

is therefore not "How many cells do you have?" but "How many independent initiated campaigns can be joined from decision-time measurements to a later direct outcome, with failures and an untouched holdout preserved?"

The answer can change quickly. A sponsor may already hold a qualifying narrow manufacturing corpus, or a consortium may generate one prospectively. The contract provides a falsifiable reversal path: show the independent-unit counts, physical joins, temporal roles, direct outcome and analysis rights; then run the sealed simple-baseline comparison. Until that happens, the highest-value computational contribution is to make the future experiment identifiable, not to make an untestable model larger.

7 Limitations

The audit is a date-bounded resource assessment, not a proof that no confidential corpus exists. Public descriptions cannot establish private row completeness, and a sponsor may have a stronger narrow dataset than its disclosures show. Conversely, regulatory recordkeeping does not by itself establish an analyzable join, sufficient independent units, retained failures, reuse rights or an untouched test set.

The 52-object matrix spans several claim classes and was not sampled as a statistical population of all T1D publications. Its rows therefore support object-level eligibility decisions, not a prevalence estimate for the literature. Before journal submission, the search record should be packaged as a reproducible systematic-review supplement and challenged by external cell-manufacturing, immunology, biostatistics and data-governance reviewers.

8 Conclusion

The public T1D data landscape is rich in molecular states and poor in joined product decisions. No audited object currently provides the independent campaign-to-outcome structure required to train or validate a SC-islet product world model. This negative result yields a concrete program: preserve the all-initiated campaign universe, encode physical lineage and temporal roles, measure post-thaw dynamic function and noncompensatory safety outcomes, and reserve future campaigns for sealed evaluation. Foundation models become decision-relevant when tested inside that structure. They cannot substitute for it.

References

[1] T. W. Reichman et al., "Stem cell–derived, fully differentiated islets for type 1 diabetes," *New England Journal of Medicine* **393** (2025), 858–868. doi:10.1056/NEJMoa2506549.

[2] A. H. Yehya et al., "Identifying and optimizing critical process parameters for large-scale manufacturing of iPSC derived insulin-producing β-cells," *Stem Cell Research & Therapy* **15** (2024), 408. doi:10.1186/s13287-024-03973-0.

[3] N. Dadheech et al., "Scale up manufacturing approach for production of human induced pluripotent stem cell-derived islets using Vertical Wheel bioreactors," *npj Regenerative Medicine* **10** (2025), 24. doi:10.1038/s41536-025-00409-y.

[4] D. Wang et al., "Improving cellular fitness of human stem cell-derived islets under hypoxia," *Nature Communications* (2025). doi:10.1038/s41467-025-59924-7.

[5] E. Bradley et al., "Divergent cell-type specific hypoxia responses in human stem cell-derived and primary islets," *Scientific Reports* (2026). doi:10.1038/s41598-026-52074-w.

[6] D. Balboa et al., "Functional, metabolic and transcriptional maturation of human pancreatic islets derived from stem cells," *Nature Biotechnology* **40** (2022), 1042–1055. doi:10.1038/s41587-022-01219-z.

[7] S. Lietzke et al., *Nature Communications* (2025). doi:10.1038/s41467-025-66022-1.

[8] C. Ricordi et al., "Manufacture of a complex cellular product at eight processing facilities," *Diabetes* **65** (2016), 3418–3428. doi:10.2337/db16-0234.