Steering into competitive whitespace: a public-data, computation-driven campaign nominates GUCY2C and DKK1 and delivers de novo binder leads across the esophageal adenocarcinoma trajectory

Ruth-Anne Pai, PhD[1,2]

[1] Citizen Scientist, EoE patient-researcher; Built with Claude: Life Sciences Hackathon (June–July 2026)

[2] Pai Advisory, LLC

* Correspondence: Ruth-Anne Pai (Pai Advisory, LLC) — advising@ruthannepai.com

Preprint — not peer reviewed. This manuscript was generated with substantial AI assistance (Anthropic Claude Science platform) under human direction. All results are computational (in-silico); no experimental validation has been performed. This document is research and informational content only and is not medical advice.

Abstract

Esophageal adenocarcinoma (EAC) is a somatic, chromosomally-unstable cancer whose incidence continues to rise in Western populations, yet whose targeted-therapy landscape is concentrated on a small number of validated axes — HER2, PD-1/PD-L1, VEGF, and, most recently, Claudin-18.2. We asked whether a reproducible, competition-aware computational campaign built entirely on public data could identify differentiated therapeutic entry points and deliver credible molecular leads against them. Mining TCGA genomics (cBioPortal *esca_tcga_pan_can_atlas_2018*, n=182) confirmed a copy-number-driven disease — *TP53* altered in 87% of tumors, *CDKN2A* deleted in 39%, *CCND1* amplified in 35% — in which the drug-tractable drivers are dominated by amplified surface receptors and angiogenic factors. A structured competitive-landscape analysis (351 interventional EAC trials, 106 active; Open Targets known-drug sets) mapped where clinical effort is concentrated and, by contrast, where it is not. We deliberately steered target nomination away from the crowded axes and applied an explicit druggability filter that down-weighted lost tumor suppressors and secreted biomarkers not addressable by an engineered biologic. Two leads emerged spanning the Barrett's→dysplasia→carcinoma trajectory: **GUCY2C**, a gastrointestinal-luminal-restricted surface antigen selected for a T-cell-engager treatment arm in advanced EAC, and **DKK1**, a secreted Wnt-pathway modulator with existing clinical precedent selected for a neutralizing-trap interception arm in high-risk Barrett's. For both targets we designed de novo protein binders using an RFdiffusion→ProteinMPNN→Boltz-2 pipeline. In a pilot-scale campaign (eight backbones per target, single generation run), fifteen of sixteen designs cleared an interface-confidence threshold (ipTM > 0.5); the top GUCY2C binder reached ipTM 0.915 (complex pLDDT 0.896) and the top DKK1 binder ipTM 0.858 (complex pLDDT 0.908), each folding at a contiguous target interface. A sequence-composition liability screen tempers this: several high-ipTM designs — including the top GUCY2C binder — carry high alanine content and long homopolymer runs characteristic of a generic-scaffold failure mode, whereas the DKK1 lead is well-composed on both interface confidence and

developability. ipTM is a single-model docking-confidence proxy, not a measured affinity; these are pilot-scale, in-silico leads, not experimentally tested therapeutics. The program's two decisive questions — whether GUCY2C's luminal restriction provides a genuine therapeutic window for a T-cell engager, and whether a dysplasia-regression surrogate endpoint is acceptable for DKK1 interception — remain open and lie downstream of the work reported here.

Introduction

Esophageal adenocarcinoma is the predominant subtype of esophageal cancer in Western populations and the only one whose incidence has continued to rise over recent decades (Coleman et al. 2018). It develops through a well-characterized histologic sequence — Barrett's metaplasia, low- and high-grade dysplasia, intramucosal carcinoma, and finally invasive, locally advanced or metastatic disease (Spechler & Souza 2014). Despite this uniquely legible natural history, most patients present with advanced disease, and outcomes remain poor.

The molecular basis of EAC is now well described. It is a somatic, chromosomally-unstable adenocarcinoma initiated by early loss of *TP53* and *CDKN2A* and driven forward by whole-genome doubling and focal amplification of cell-cycle and receptor-tyrosine-kinase genes, rather than by a small set of recurrently mutated oncogenic drivers (TCGA Research Network 2017). This genomic architecture has shaped its therapeutic development: the drug-tractable drivers are amplified surface receptors (HER2, EGFR) and secreted angiogenic factors (VEGFA), and the approved and late-stage agents cluster tightly around them and around immune checkpoints. Trastuzumab and trastuzumab-deruxtecan (HER2) (Bang et al. 2010), pembrolizumab and nivolumab (PD-1, including the adjuvant setting) (Kelly et al. 2021), ramucirumab (VEGFR2) (Fuchs et al. 2014), and zolbetuximab (Claudin-18.2, approved 2024) (Shitara et al. 2023) define a competitive but narrow target space.

Two consequences follow. First, a differentiated program requires deliberate movement into competitive whitespace — biology that is credible but under-exploited. Second, the disease's legible trajectory creates an opportunity that a treatment-only strategy ignores: the chance to intercept malignant progression in the large, identifiable, high-risk Barrett's population before invasive cancer develops, a setting in which no molecular therapeutic is approved and endoscopic surveillance plus ablation remains the standard of care.

Here we report a computational discovery-to-design campaign designed around both of these considerations. Working entirely from public genomics, target-association, and clinical-landscape data, we nominated candidate targets under an explicit novelty steer and an honest druggability filter, selected two leads that together span the disease trajectory, and designed de novo protein binders for each using contemporary generative structure tools. We present the campaign as a reproducible, honesty-gated template: every quantitative claim traces to a named public source or a generated computational result, and every unvalidated inference is flagged as such.

Results

EAC is a copy-number-driven disease whose tractable drivers are clinically crowded

We first characterized the somatic landscape of EAC using the TCGA PanCancer Atlas cohort (cBioPortal *esca_tcga_pan_can_atlas_2018*, n=182 after restriction to the adenocarcinoma subtype). The alteration profile is dominated by tumor-suppressor loss and copy-number change rather than recurrent activating point mutation: *TP53* is altered in 87% of tumors, *CDKN2A* is deep-deleted in 39%, and *CCND1* is amplified in 35%, followed by *WWOX* deletion (34%) and amplification of *MYC* (22%), *ERBB2* (15%), *VEGFA* (12%), *EGFR* (12%), and *GATA6* (11%) (**Figure 1a**). This confirms a chromosomally-unstable, amplification-driven biology in which the readily drug-tractable events are amplified cell-surface receptors and secreted factors, while the earliest and most penetrant events — *TP53* and *CDKN2A* loss — are intracellular and hard to drug directly.

We then mapped the clinical landscape. Across ClinicalTrials.gov we identified 351 interventional EAC trials (106 active) and, separately, 203 interventional trials in the Barrett's/dysplasia precursor setting; Open Targets known-drug sets returned 106 agents associated with the indication. Ranking targets by active-trial intervention mentions against their development maturity (**Figure 1b**) shows clinical effort concentrated on PD-1/PD-L1 (the most-mentioned axis by a wide margin), HER2, and VEGF, with Claudin-18.2 and FGFR2b emerging. By contrast, several biologically credible axes are nearly untouched in EAC: the surface antigens B7-H3, TROP2, and GUCY2C; the secreted Wnt modulator DKK1; the immunometabolic target CD73; and, as a category, molecular interception of Barrett's progression.

A trajectory-spanning, two-arm strategy

Rather than compete on a crowded axis or pursue a single point on the disease course, we defined a two-arm strategy that attacks EAC at two under-exploited positions on its trajectory (**Figure 1c**): an **interception arm** acting across Barrett's metaplasia and dysplasia to slow or prevent progression, and a **treatment arm** acting on established early-through-advanced disease. This framing was reinforced by a patient-priorities synthesis emphasizing preservation of swallowing function, quality-of-life-weighted survival, minimization of treatment toxicity, and — in the precursor setting — a well-tolerated, less-invasive way to reduce progression risk and defer or avoid esophagectomy.

Honest druggability triage nominates GUCY2C and DKK1

We nominated sixteen candidate targets spanning both lanes and scored them on a transparent additive composite, composite = tumor_selectivity + novelty + tractability_score + 5 × disease_association (component definitions, ordinal anchors, and weights in Table S0). A weight-perturbation analysis (Monte Carlo over ±50% multiplicative variation in all four weights, 20,000 draws; Table S0) tested rank robustness: among biologic-addressable treatment-lane targets, GUCY2C is the top-ranked target in 80.7% of draws and top-three in 94.2%, with CDH17 the nearest competitor. This rank-1 dominance is, however, contingent on GUCY2C's optimistic tumor-selectivity score: recalibrating it from 5 to 4 to reflect that the safety window is unresolved (O-2) drops GUCY2C's composite to 12.00, lets CDH17 (12.38) overtake it, and reduces its Monte-Carlo rank-1 frequency to ~0% (top-three ~48%; Table S0) — so O-2 is decisive for lead selection, not only safety, and CDH17 is carried as a credible backup; in the interception lane, DKK1 is the only biologic-addressable candidate after the druggability filter, so its selection is driven by the filter, not by out-competing a field — an honest reflection of how thin the

addressable interception space is. Critically, we then applied an explicit biologic-druggability filter, because a high composite score does not by itself make a target reachable by an engineered protein (**Figure 2**). The single highest raw-composite candidate, *GPX7*, is a tumor suppressor lost in Barrett's — a restoration or synthetic-lethal problem, not an antibody-addressable one — and was down-weighted accordingly. Three further high-scoring interception candidates (*REG4*, *OLFM4*, *TFF3*) are secreted progression biomarkers whose therapeutic value is unproven, and the crowded validated receptors (*ERBB2*, *EGFR*, *CLDN18*) were deprioritized under the novelty steer.

After this filter, the highest-scoring biologic-addressable target for the treatment lane was **GUCY2C** (guanylyl cyclase C), a gastrointestinal-epithelial surface antigen whose apical, luminal-facing restriction in normal gut is the basis for a candidate therapeutic window, and which has antibody, ADC, bispecific, and CAR precedent in colorectal cancer. Normal-tissue RNA profiling (GTEx v8; **Figure S1**, Table S4) supports the selectivity premise: 85% of GUCY2C's normal expression is in GI tissues and its highest non-GI tissue reaches only ~1.7 TPM, a profile matched among comparators only by CDH17 and clearly distinct from the pan-epithelial TROP2 (TACSTD2). Two caveats are essential and unresolved. First, GTEx measures *normal* tissue only; it establishes tissue selectivity but does **not** establish that GUCY2C is expressed on EAC tumor cells at therapeutically actionable levels — the clinical precedent is colorectal, and EAC tumor-cell expression remains to be confirmed from tumor RNA/protein data (a gap we flag rather than paper over — GUCY2C tumor-mRNA quantification in the esca_tcga_pan_can_atlas_2018 cohort is the single highest-priority follow-up query, and the treatment-arm rationale should be considered provisional until it is reported). Second, GUCY2C is genuinely expressed in normal small intestine and colon (~28 TPM median), so "luminal restriction" is a polarity argument, not an absence-of-antigen argument, and the normal-GI safety window for a T-cell engager is precisely open question O-2. For the interception lane we selected **DKK1** (Dickkopf-1), a secreted Wnt-pathway modulator addressable by a neutralizing antibody or ligand trap. The DKN-01 program (Klempner et al. 2024) establishes that pharmacologic DKK1 neutralization is clinically tractable — but that precedent is in *advanced* gastric/gastro-esophageal adenocarcinoma, a late-stage treatment population with different safety tolerances and endpoints. It is **not** evidence for DKK1 interception in a largely pre-malignant Barrett's surveillance population; direct precursor-setting therapeutic evidence for DKK1 is essentially absent beyond progression-biomarker associations. We therefore carry DKK1 into the interception arm on mechanistic and druggability grounds, with the population mismatch stated explicitly as a risk (open question O-4). GUCY2C was routed to an anti-GUCY2C × anti-CD3 T-cell-engager format for advanced/peri-operative EAC; DKK1 to a neutralizing-trap format for high-risk Barrett's interception.

De novo binders were designed for both arms

For each target we prepared a design-ready structure from the AlphaFold database and isolated the relevant domain: the GUCY2C extracellular domain (UniProt P25092, residues 24–430; the intracellular kinase/cyclase domain was excluded as irrelevant to a surface binder) and the DKK1 CRD2 domain (UniProt O94907, residues 178–256; the LRP6-binding functional region and the high-confidence structured segment). Surface-exposed, high-confidence residue patches were selected as hotspots by a neighbor-count and pLDDT heuristic. We generated eight binder backbones per target with RFdiffusion, designed sequences with ProteinMPNN, and folded each best-per-backbone binder in complex with its target using Boltz-2 (**Figure 3**).

In-silico validation and lead complexes

In this pilot-scale run (eight backbones per target from a single generation seed), fifteen of sixteen designs cleared the interface-confidence threshold of ipTM > 0.5 (**Figure 3**). We report this as a pilot observation, not a generalizable success rate: n=8/target is too small to estimate a rate, ipTM > 0.5 is a permissive bar, and the values are single-best-of-five diffusion samples. The top treatment-arm binder by ipTM, gucy2c_bb2 (71 aa), reached ipTM 0.915 with a complex pLDDT of 0.896 (seven of eight GUCY2C designs passed); the top interception-arm binder, dkk1_bb1 (63 aa), reached ipTM 0.858 with a complex pLDDT of 0.908 (all eight DKK1 designs passed). A sequence-composition liability screen (Table S5) materially qualifies the treatment-arm lead: gucy2c_bb2 is 44% alanine with an eight-residue homopolymer run, and several other high-ipTM designs (dkk1_bb2/bb3/bb5) exceed 50% alanine — the signature of a generic low-complexity scaffold that can fold with spurious confidence. By this combined criterion, **dkk1_bb1 is the most robust lead** (ipTM 0.858, 17.5% alanine, maximum run 2), whereas gucy2c_bb2's headline ipTM must be read against its composition liability and treated as provisional pending redesign toward a lower-alanine backbone. In the predicted complexes (**Figure 4**), both binders dock against a contiguous, well-localized surface on their target: the GUCY2C binder contacts a patch of 13 target residues (Cα < 8 Å) on the extracellular domain (UniProt P25092 positions 153–402), and the DKK1 binder contacts an eight-residue patch mapping to UniProt O94907 positions 187–208 — confirming engagement squarely within CRD2 (178–256). A pairwise sequence-identity check across the design set (length-penalized) found maximum identities of 21.4% (GUCY2C set) and 32.9% (DKK1 set; the most-similar pair, dkk1_bb2/bb3, are both alanine-rich), well below any high-similarity threshold — the designs are diverse solutions, not a single repeated motif (Table S5). These are pilot-scale in-silico leads generated in a single design cycle. All structural confidence rests on single-model Boltz-2 ipTM/pLDDT without orthogonal consensus (AlphaFold-Multimer or Chai-1 fold-back, and monomer/decoy specificity controls, are the immediate next computational step, O-1); ipTM is a docking-confidence proxy, not a measurement of affinity; and no experimental characterization has been performed.

Figures

Figure 1

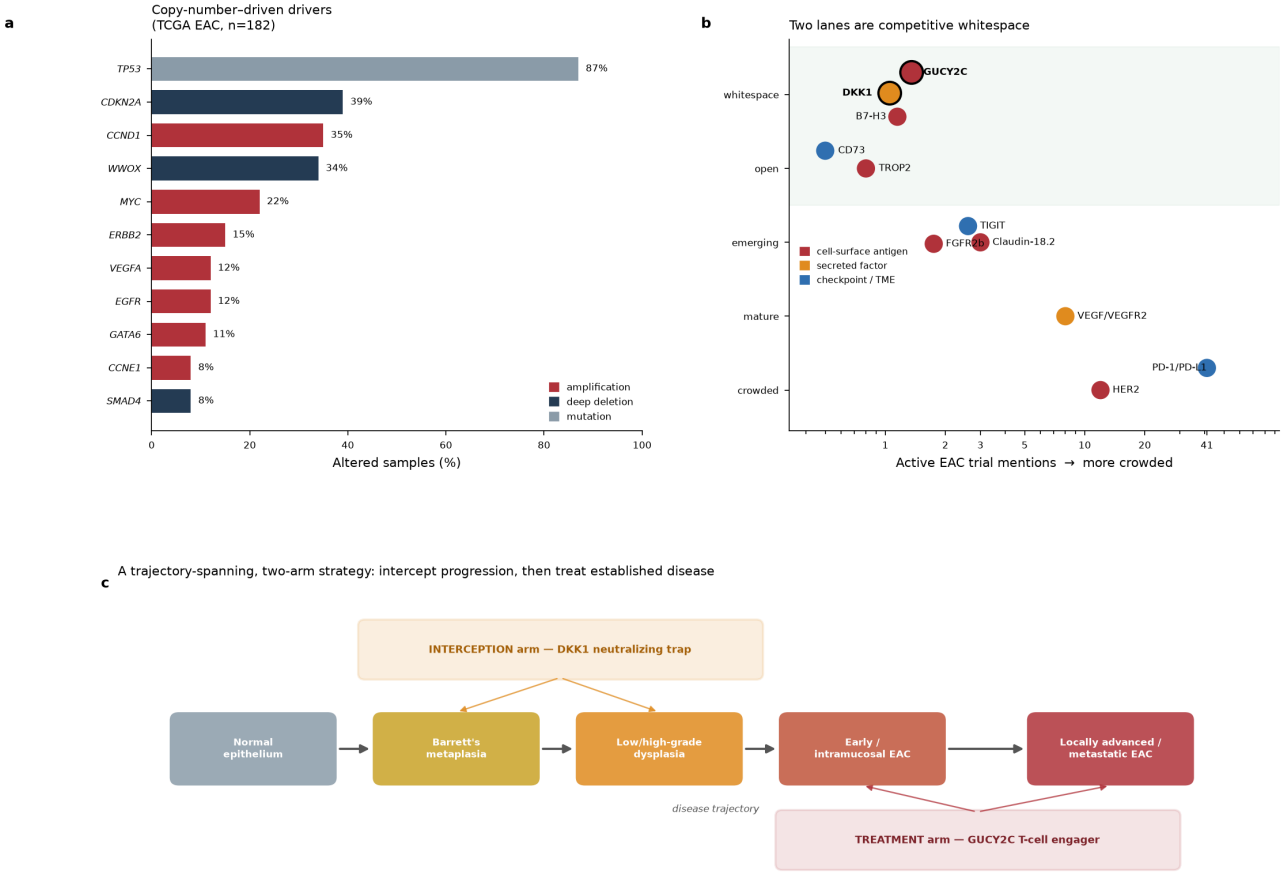

Figure 2

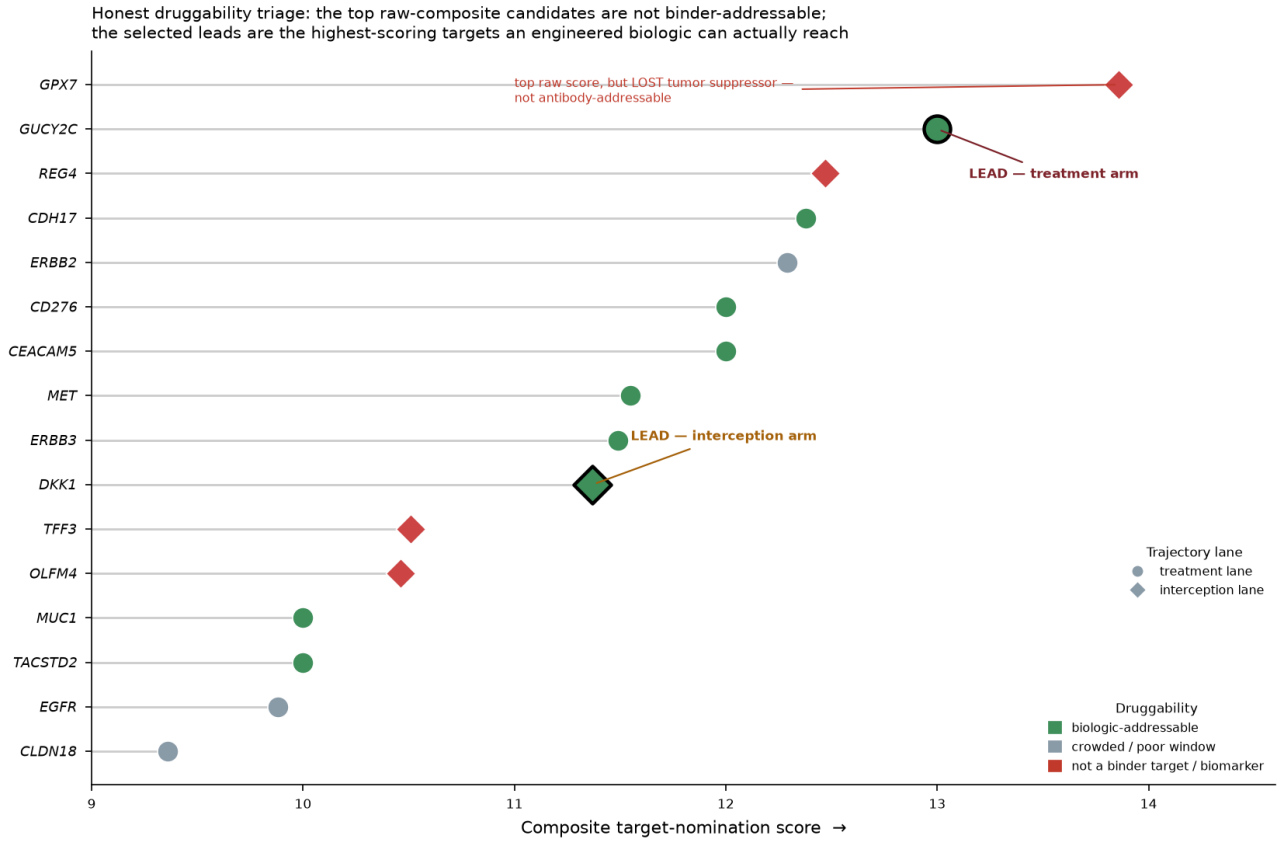

Figure 3

Stage 6 — de novo binder design: RFdiffusion → ProteinMPNN → Boltz-2 fold-back

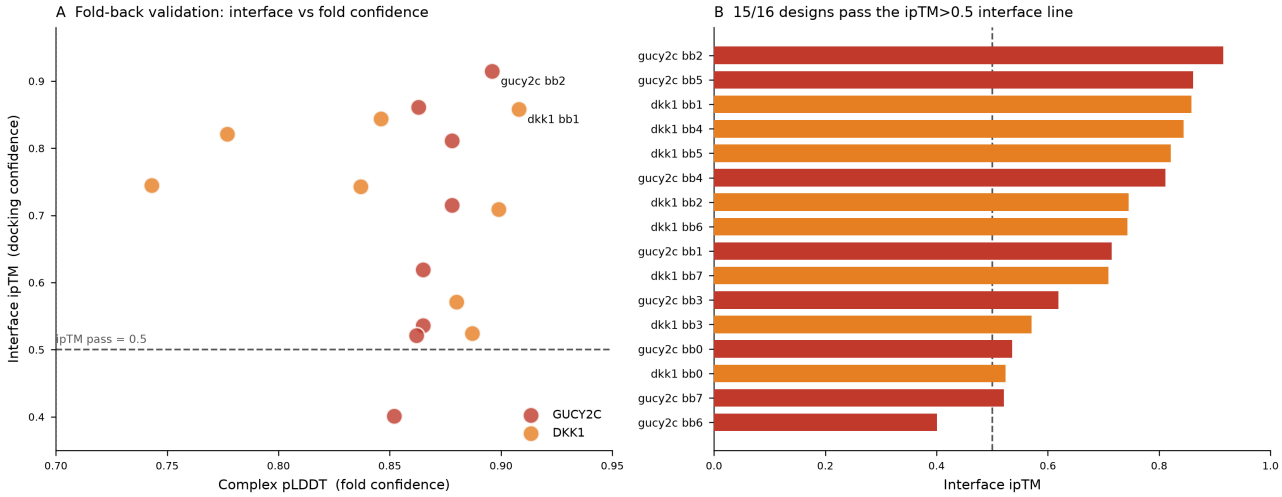

Figure 4

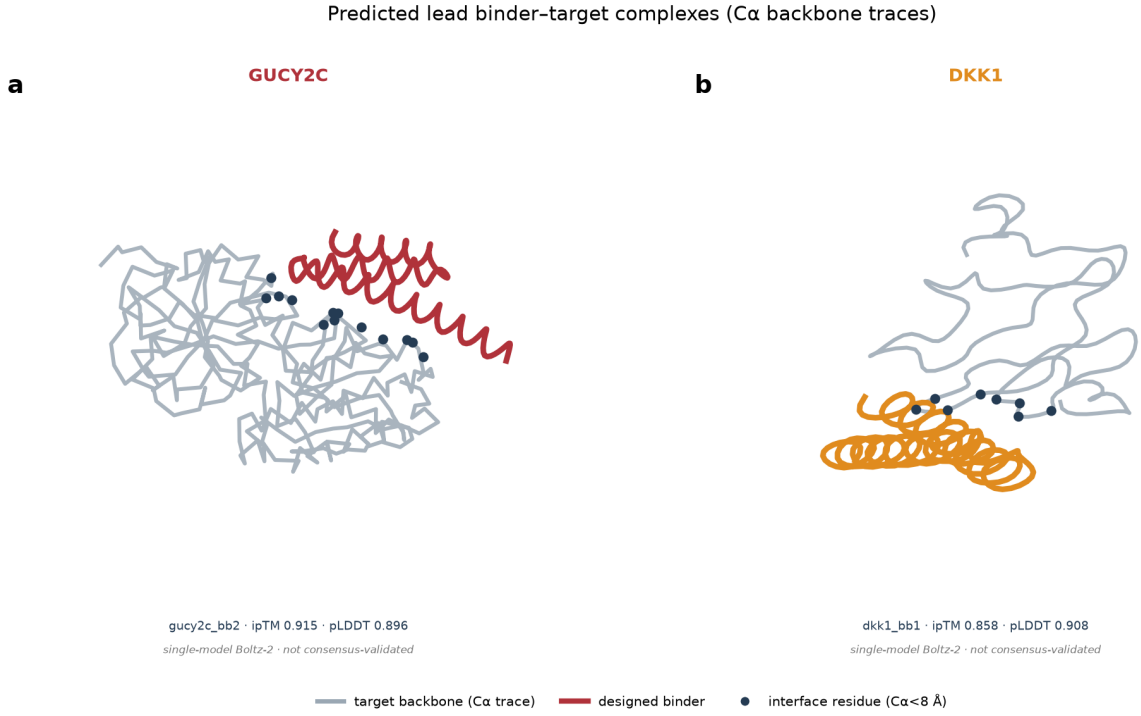
Predicted lead binder–target complexes (Cα backbone traces)
a
GUCY2C
b
DKK1
gucy2c_bb2 · ipTM 0.915 · pLDDT 0.896
single-model Boltz-2 · not consensus-validated
dkk1_bb1 · ipTM 0.858 · pLDDT 0.908
single-model Boltz-2 · not consensus-validated
target backbone (Cα trace)
designed binder
interface residue (Cα<8 Å)

Figure S1 (Supplementary)

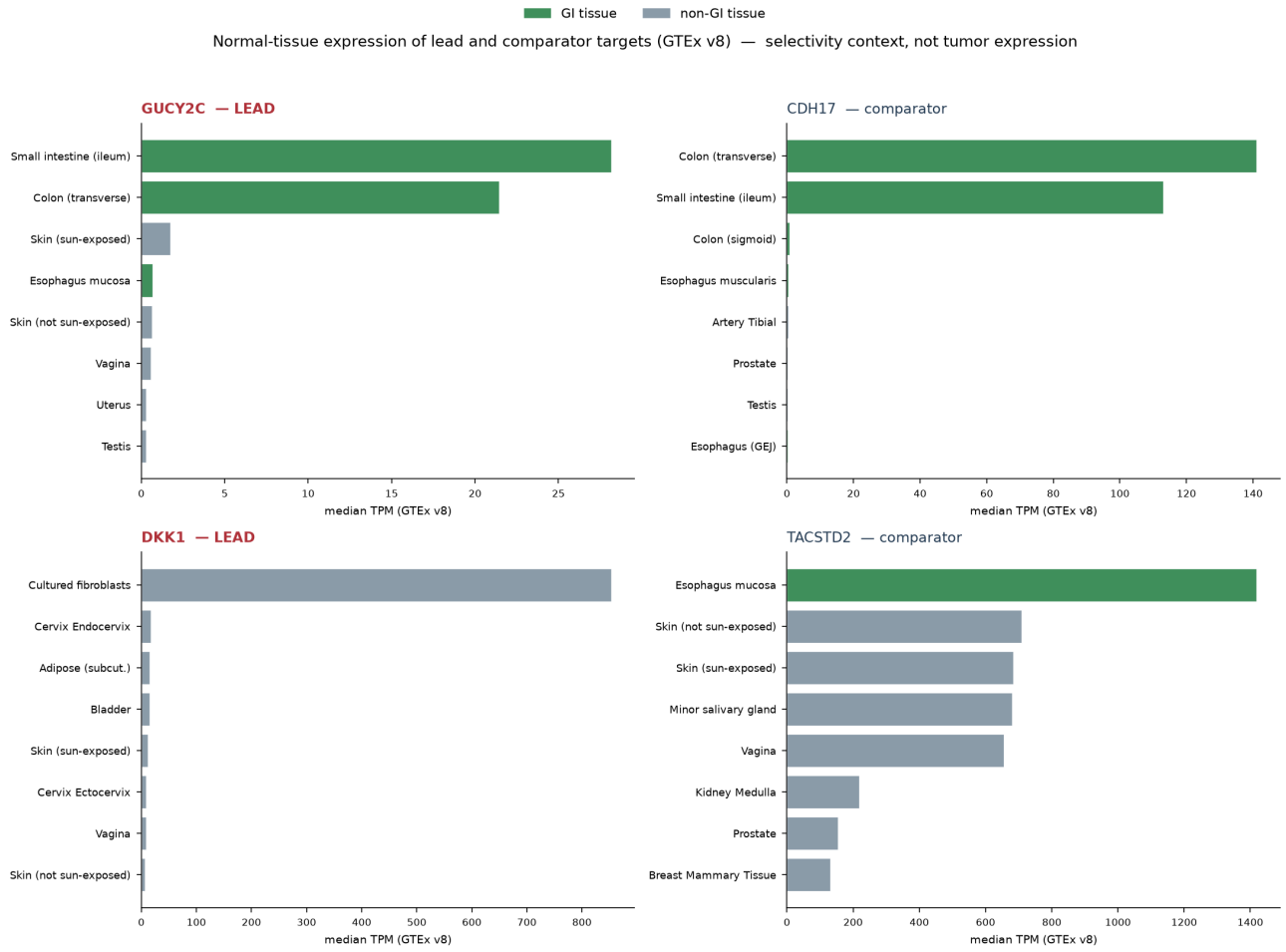

Figure Legends

Figure 1. EAC is a copy-number-driven cancer whose tractable drivers are clinically crowded, motivating a two-arm, trajectory-spanning strategy. (a) Somatic alteration frequencies in the TCGA PanCancer Atlas esophageal adenocarcinoma cohort (cBioPortal *esca_tcga_pan_can_atlas_2018*, n=182), colored by alteration class. The earliest and most penetrant events (*TP53* mutation 87%, *CDKN2A* deletion 39%) are intracellular and hard to drug; the tractable drivers are amplified surface receptors and secreted factors. (b) Competitive landscape of EAC target axes, positioning each axis by active-EAC-trial intervention mentions (x, log scale; higher = more crowded) against development maturity (y), colored by target biology. Clinical effort concentrates on PD-1/PD-L1, HER2, and VEGF; the selected leads GUCY2C and DKK1 (black outline) sit in competitive whitespace. x-axis values are non-de-duplicated active-EAC-trial intervention mentions and should be read as relative crowding indicators, not exact disjoint counts (see Methods). (c) The two-arm strategy on the disease trajectory: a DKK1 neutralizing-trap interception arm acting across Barrett's metaplasia and dysplasia, and a GUCY2C T-cell-engager treatment arm acting on early-through-advanced disease.

Figure 2. Honest druggability triage nominates GUCY2C and DKK1. Sixteen candidate targets scored on a composite of disease association, tumor selectivity, antibody tractability, and novelty (x-

axis), then filtered by whether an engineered biologic can actually reach them (color: green = biologic-addressable; grey = crowded/poor-window and deprioritized under the novelty steer; red = not a binder target or an unproven secreted biomarker). Marker shape encodes trajectory lane (circle = treatment, diamond = interception). The single highest raw-composite candidate, *GPX7*, is a lost tumor suppressor and is not antibody-addressable; the selected leads are the highest-scoring targets an engineered biologic can reach in each lane. Source: target_nomination.csv.

Figure 3. De novo binder design and in-silico validation. Binder campaign for both arms: RFdiffusion backbone generation -> ProteinMPNN sequence design -> Boltz-2 fold-back of each binder in complex with its target. Interface confidence (ipTM) and fold confidence (complex pLDDT) for all designs; 15 of 16 clear the ipTM>0.5 interface threshold. (Reproduced from the design campaign; source design_results.png / design_leads.csv.)

Figure 4. Predicted lead complexes fold at contiguous target interfaces. Boltz-2 predicted complexes for the two lead binders, shown as Calpha backbone traces (target grey, binder colored). Target residues within 8 A (Calpha-Calpha) of the binder are marked (navy). (a) gucy2c_bb2 (71 aa) on the GUCY2C extracellular domain (ipTM 0.915, complex pLDDT 0.896), contacting a 13-residue patch. (b) dkk1_bb1 (63 aa) on the DKK1 CRD2 / LRP6-binding region (ipTM 0.858, complex pLDDT 0.908), contacting an 8-residue patch. Each panel is annotated “single-model Boltz-2 · not consensus-validated”: the depicted structure is the highest-confidence of five Boltz-2 diffusion seeds, not an ensemble or an orthogonal (AF-Multimer/Chai-1) consensus. Interface residues map to UniProt P25092 153–402 (GUCY2C, within ECD) and O94907 187–208 (DKK1, within CRD2). Backbone traces depict topology and interface localization, not experimentally-determined secondary structure; ipTM is a docking-confidence proxy, not a measured affinity. Source: gucy2c_bb2_complex.pdb, dkk1_bb1_complex.pdb.

Figure S1. Normal-tissue expression of lead and comparator targets (GTEx v8). Top-8 tissues by median TPM for the two leads (GUCY2C, DKK1) and comparators (CDH17, TACSTD2/TROP2). GI tissues green, non-GI grey. GUCY2C and CDH17 are sharply GI-restricted; TROP2 is pan-epithelial. GTEx measures normal tissue only and does not establish tumor-cell expression; DKK1’s top row (cultured fibroblasts) is a cell-culture artifact (Table S4 footnote). Source: S4_gtex_selectivity.csv.

Discussion

This campaign shows that a public-data, competition-aware computational workflow can move from open genomics to differentiated therapeutic leads while deliberately routing around crowded target space. Its principal output is not a single molecule but a coherent, trajectory-spanning hypothesis backed by concrete designed leads: intercept EAC progression in Barrett’s by neutralizing secreted DKK1, and treat established disease by redirecting T cells against the GI-restricted surface antigen GUCY2C — two axes chosen precisely because the HER2/PD-1/VEGF field is saturated.

Several features distinguish the work from a purely aspirational target list. The druggability triage was applied honestly: the highest raw-scoring candidate was rejected because it is not addressable by the chosen modality, and secreted biomarkers with no therapeutic evidence were down-weighted rather than promoted for narrative convenience. The design step produced real, structurally-validated binders for both arms rather than a specification, with high interface and fold confidence.

The program's strengths are matched by two decisive, unresolved questions, both of which lie downstream of anything demonstrated here. The first is the **GUCY2C therapeutic window**: the luminal-restriction argument is a polarity- and species-dependent hypothesis, and whether it constitutes a genuine safety margin for a T-cell engager against an antigen also expressed on normal enterocytes can only be answered with human normal-tissue immunohistochemistry and redirected-cytotoxicity selectivity data. The second is the **DKK1 interception surrogate**: intervening in a largely pre-malignant Barrett's population sets a very high safety bar and requires regulatory acceptance of histologic dysplasia regression (or a comparable molecular surrogate) as a registrational endpoint — acceptance that is not established. Either question could halt its arm.

The limitations of the computational work should be stated plainly. (i) All structural validation rests on single-model Boltz-2 confidence without orthogonal consensus; AlphaFold-Multimer / Chai-1 fold-back and monomer/decoy specificity controls (O-1) are the immediate next step and have not yet been run, so ipTM values are provisional and cannot be distinguished from generic-scaffold artifacts without them. (ii) The sequence-composition screen (Table S5) already shows that the top GUCY2C design by ipTM carries a high-alanine, low-complexity liability, so interface-confidence rank and developability rank disagree and the treatment-arm lead should be regarded as a starting point for redesign rather than a finished binder. (iii) GTEx establishes GUCY2C's *normal-tissue* selectivity but not its EAC *tumor-cell* expression, which must be confirmed from tumor data before the treatment-arm rationale is secure. For DKK1, the interception arm targets an otherwise-well Barrett's surveillance population, so the safety bar is fundamentally higher than for the advanced-disease DKN-01 precedent: DKK1 has canonical roles in Wnt-dependent bone and skeletal homeostasis, its normal-tissue RNA (Table S4) is dominated by a cultured-fibroblast signal (a cell-culture artifact) with low solid-tissue levels consistent with a stromal/paracrine source, and systemic neutralization in well patients carries skeletal and developmental-signaling risks that late-stage oncology dosing does not have to clear. We therefore list skeletal/systemic toxicity explicitly, alongside GI on-target effects, as a threshold risk for the interception arm (gate G5, O-4). (iv) The epitope hotspots were chosen by a structural heuristic rather than experimental functional-site mapping, so it remains to be confirmed that the DKK1 binder blocks the LRP6 interface and that the GUCY2C binder engages a tumor-exposed epitope. (v) The GUCY2C binder is only one arm of a bispecific whose CD3 arm and format engineering are not done. (vi) The triage operated over a manually assembled sixteen-target shortlist rather than a fully enumerated candidate universe, and the trial-count extraction lacks recorded query strings and snapshot dates. No biophysical, cellular, or in-vivo data exist. The immediate next steps are correspondingly computational and experimental: an ESM-based developability and liability screen with orthogonal fold-back for confidence consensus, followed by expression and biophysical characterization, functional and normal-tissue-selectivity assays, and — for the interception arm — an early regulatory interaction to settle the surrogate-endpoint question before any long prevention trial is contemplated.

Framed appropriately, the campaign is a reproducible template for early, differentiated, honesty-gated target and lead generation, and a concrete dual-arm starting point for EAC that a wet-lab program could take forward.

Methods

Genomic characterization

Somatic alteration frequencies were obtained from cBioPortal for the TCGA PanCancer Atlas esophageal cohort (*esca_tcga_pan_can_atlas_2018*), restricted to the adenocarcinoma subtype (n=182). Alteration classes (amplification, deep deletion, mutation) were taken as reported by cBioPortal; the drivers shown in Figure 1a are the most frequently altered genes in the cohort.

Target-association and competitive-landscape analysis

Target–disease associations and tractability were drawn from Open Targets (EAC MONDO_0005028; Barrett's esophagus MONDO_0013662) and known-drug sets. The competitive landscape was assembled from ClinicalTrials.gov interventional-trial counts (351 EAC, 106 active; 203 Barrett's/dysplasia) reported in the upstream program artifacts, and from approved-agent records confirmed against Drugs@FDA. The per-axis "active EAC trial intervention mentions" (Figure 1b, x-axis; Table S3) are axis-level annotations of active interventional trials referencing each target/axis; they were not de-duplicated across combination trials and therefore do not sum to the 106 active-trial total, and should be read as relative crowding indicators rather than exact disjoint counts. The exact extraction query strings and snapshot dates were not recorded in the upstream artifacts and should be captured with a raw trial-ID list before external submission — we flag this as a reproducibility gap in the Limitations.

Target triage

Candidates were scored on the additive composite composite = tumor_selectivity + novelty + tractability_score + 5 × disease_association, where tumor_selectivity and novelty are 1–5 ordinal analyst scores (anchors in Table S0), tractability_score maps antibody-tractability tier to an integer (surface-confirmed = 3, clinical-stage = 4, clinical-validated = 5), and disease_association is the Open Targets association score (EAC for treatment-lane targets, Barrett's for interception-lane targets; 0 when unavailable). This rubric exactly reproduces all sixteen published composite values (maximum residual 0.005). A separate biologic-druggability category (STRONG/GOOD/MODERATE = addressable by an engineered biologic; CROWDED = deprioritized under the novelty steer; WEAK-BIOMARKER and NOT-A-BINDER-TARGET = down-weighted) was assigned from target biology and applied as an explicit filter over the raw composite (Figure 2). Rank robustness was assessed by Monte Carlo (20,000 draws, each weight scaled by an independent U(0.5,1.5) factor and the association weight by U(2.5,7.5)), restricted to biologic-addressable targets within each lane (Table S0). The candidate universe was the manually assembled sixteen-target shortlist; a fully enumerated association-ranked universe is noted as a limitation.

Target-structure preparation

Design-ready structures were taken from the AlphaFold database: GUCY2C (UniProt P25092) and DKK1 (UniProt O94907). The GUCY2C extracellular domain (residues 24–430) and the DKK1 CRD2 domain (residues 178–256) were isolated; disordered and functionally-irrelevant regions were excluded. Hotspot residue patches were selected by a combined neighbor-count and pLDDT heuristic (GUCY2C ECD patch

[239,240,267,422,424,425]; DKK1 CRD2 patch [181,182,183,194,195], in the isolated-domain numbering).

De novo binder design

Binder backbones were generated with RFdiffusion (Complex_base checkpoint, noise 0), eight backbones per target, binder length 70–100 aa (GUCY2C) and 60–90 aa (DKK1). Sequences were designed with ProteinMPNN (model v_48_020, binder chain only, eight sequences per backbone at sampling temperature 0.1). Each best-per-backbone binder was folded in complex with its target using Boltz-2 (target chain with --use_msa_server, binder as single sequence, five diffusion samples). Designs were assessed by interface ipTM (pass threshold > 0.5) and complex pLDDT (fold threshold > 0.7). Interface residues reported here were computed directly from the predicted complex coordinates as target residues with a Cα within 8 Å of any binder Cα, then mapped to canonical UniProt positions (GUCY2C → P25092 153–402; DKK1 → O94907 187–208, within CRD2; full offset mapping in Table S6). All GPU steps were run on cloud A100 hardware.

Reproducibility and honesty gating

Every quantitative claim in this manuscript traces to one of: cBioPortal, Open Targets, ClinicalTrials.gov, Drugs@FDA, or a generated computational result reported in the design tables. In-silico results are labeled as such throughout and are not represented as experimental validation; clinical, regulatory, and commercial inferences appear only in the referenced strategy documents (Supplementary Information) and are flagged there as assumptions requiring expert confirmation.

Funding

This work received no external funding. It was conducted during the Built with Claude: Life Sciences Hackathon (2026).

Data and code availability

The upstream program artifacts — machine-readable dossier, target-nomination and competitive-landscape tables, the full binder-design table (all sixteen designs with sequences and confidence metrics), the two lead complex structures (PDB), and the scientific/regulatory/commercial strategy briefs — accompany this manuscript as Supplementary Information. Public data sources (cBioPortal *esca_tcga_pan_can_atlas_2018*; Open Targets MONDO_0005028 and MONDO_0013662; ClinicalTrials.gov; AlphaFold DB entries P25092 and O94907) are openly accessible. Design tools (RFdiffusion, ProteinMPNN, Boltz-2) are publicly available.

Author contributions and AI-assistance disclosure

This campaign was executed as a human-supervised, AI-assisted computational workflow, and we distinguish three categories of contribution. (1) Deterministic bioinformatics — retrieval of TCGA/cBioPortal alteration frequencies, Open Targets associations, ClinicalTrials.gov and Drugs@FDA

records, GTEx expression, and AlphaFold structures; the additive triage composite and its Monte-Carlo robustness analysis; interface-residue and sequence-composition computations — was reproducible code, not model judgment. (2) Published ML methods as objects of study — RFdiffusion (backbone generation), ProteinMPNN (sequence design), and Boltz-2 (complex fold-back) — generated the binder designs and their confidence metrics. (3) AI-assisted synthesis and drafting — an LLM-based agent proposed the novelty steer, the two-arm framing, candidate shortlisting, the ordinal selectivity/novelty annotations in Table S2, and the manuscript prose; every such judgment was reviewed by a human supervisor, and every quantitative claim was regenerated from, or traced to, a deterministic source (see claims_to_sources.csv). The ordinal annotations and the "whitespace verdict" column reflect analyst judgment informed by the cited data, not an independent measurement. The author (Ruth-Anne Pai) directed the scientific strategy, made all decisions, and is responsible for the content.

Competing interests

The author is a person living with eosinophilic esophagitis and is the sole member of Pai Advisory, LLC. No specific grant funding was received. The author declares no other competing interests.

Acknowledgements

This work was conducted as a citizen-science effort during the Built with Claude: Life Sciences Hackathon (2026), organized by Anthropic, Gladstone Institutes, and Cerebral Valley. The author thanks the maintainers of the open data resources and structure-prediction tools used here, and acknowledges the eosinophilic esophagitis community whose lived experience motivates this work.

References

1. Coleman et al. (2018). *The Epidemiology of Esophageal Adenocarcinoma*. 10.1053/j.gastro.2017.07.046
2. Spechler & Souza (2014). *Barrett's Esophagus*. 10.1056/nejmra1314704
3. TCGA Research Network (2017). *Integrated genomic characterization of oesophageal carcinoma*. 10.1038/nature20805
4. Bang et al. (2010). *Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA)*. 10.1016/s0140-6736(10)61121-x
5. Kelly et al. (2021). *Adjuvant Nivolumab in Resected Esophageal or Gastroesophageal Junction Cancer (CheckMate 577)*. 10.1056/nejmoa2032125
6. Fuchs et al. (2014). *Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD)*. 10.1016/s0140-6736(13)61719-5
7. Shitara et al. (2023). *Zolbetuximab plus mFOLFOX6 in patients with CLDN18.2-positive, HER2-negative, untreated, locally advanced or metastatic gastric or gastro-oesophageal junction adenocarcinoma (SPOTLIGHT)*. 10.1016/s0140-6736(23)00620-7

8. Klempner et al. (2024). *DKN-01 in Combination With Tislelizumab and Chemotherapy as First-Line Therapy in Advanced Gastric or Gastroesophageal Junction Adenocarcinoma (DisTinGuish)*. 10.1200/jco.24.00410

Supplementary Information

Manuscript: Steering into competitive whitespace — a public-data, computation-driven campaign nominates GUCY2C and DKK1 and delivers de novo binder leads across the esophageal adenocarcinoma trajectory

All results are computational (in-silico). No experimental validation has been performed. Every value below is reproduced from the upstream program artifacts named in each section.

Table S0 — Triage scoring rubric, all component scores, and rank-robustness

Rubric (exact): composite = tumor_selectivity + novelty + tractability_score + 5 × disease_association - tumor_selectivity — analyst ordinal 1–5: 5 = tissue-restricted with a strong therapeutic window; 3 = moderate window / some normal expression; 1–2 = broad normal expression / poor window. - novelty — analyst ordinal 1–5: 5 = essentially unexploited in EAC; 3–4 = clinical-stage elsewhere / emerging; 1 = approved/crowded in EAC. - tractability_score — antibody-tractability tier mapped to integer: surface-confirmed = 3, clinical-stage = 4, clinical-validated = 5. - disease_association — Open Targets association score (EAC for treatment lane, Barrett's for interception lane; 0 if unavailable), weighted ×5.

This rubric reproduces all 16 published composite values (max residual 0.005).

Component scores:

target	lane	tumor_selectivity	novelty	ab_tractability	eac_assoc	barretts_assoc	composite	biologic_dru
GPX7	I	4	5	surface-confirmed	nan	0.372	13.86	NOT-A-BIND TARGET
GUCY2C	T	5	5	surface-confirmed	nan	nan	13.0	STRONG
REG4	I	4	5	surface-confirmed	nan	0.094	12.47	WEAK-BIOM
CDH17	T	4	5	surface-confirmed	nan	0.076	12.38	STRONG
ERBB2	T	4	1	clinical-validated	0.457	nan	12.29	CROWDED
CEACAM5	T	3	4	clinical-validated	nan	nan	12.0	STRONG
CD276	T	3	5	clinical-stage	nan	nan	12.0	STRONG

target	lane	tumor_selectivity	novelty	ab_tractability	eac_assoc	barretts_assoc	composite	biologic_dru
MET	T	3	2	clinical-validated	0.31	nan	11.55	GOOD
ERBB3	T	3	3	clinical-stage	0.298	nan	11.49	GOOD
DKK1	I	3	4	clinical-stage	nan	0.074	11.37	GOOD
TFF3	I	3	4	surface-confirmed	nan	0.102	10.51	WEAK-BIOM
OLFM4	I	3	4	surface-confirmed	nan	0.091	10.46	WEAK-BIOM
TACSTD2	T	3	4	surface-confirmed	nan	nan	10.0	GOOD
MUC1	T	3	3	clinical-stage	nan	nan	10.0	MODERATE
EGFR	T	2	1	clinical-validated	0.377	nan	9.88	CROWDED/P WINDOW
CLDN18	T	4	1	clinical-stage	nan	0.072	9.36	CROWDED

Rank robustness (Monte Carlo, 20,000 draws; each of the four weights independently scaled — selectivity/novelty/tractability by U(0.5,1.5), association by U(2.5,7.5) — restricted to biologic-addressable targets per lane):

Lane	Lead	# addressable targets	% draws lead ranks #1	% draws lead in top-3	Nearest competitor
Treatment (T)	GUCY2C	8	80.7%	94.2%	CDH17
Interception (I)	DKK1	1	100%*	100%*	(none addressable)

*DKK1 is the only biologic-addressable interception target after the druggability filter; its rank-1 result is therefore trivial and reflects the thinness of the addressable interception space rather than out-competing a field.

Selectivity recalibration (important sensitivity). GUCY2C's tumor_selectivity was recorded as 5 ("strong therapeutic window"); because that window is in fact unresolved (O-2), a more honest anchor is 4 ("tissue-restricted, window pending"). Re-running the triage with GUCY2C selectivity = 4 lowers its composite from 13.00 to 12.00 and **does change the lane ranking**: GUCY2C is overtaken by CDH17 (12.38) and falls into a tie with CEACAM5 and CD276 (all 12.00), and its Monte-Carlo rank-1 frequency drops from 80.7% to ~0% (top-three ~48%). In other words, GUCY2C's position as the single top treatment-lane target is contingent on the optimistic selectivity score, which is precisely what open

question O-2 tests. This does not disqualify GUCY2C — it remains a strong, biologic-addressable, novelty-consistent candidate for which we have a designed binder — but it means (i) resolving O-2 is decisive for lead choice as well as for safety, and (ii) **CDH17 is a credible backup treatment-lane target** that should be carried forward in parallel.

Table S1 — Complete de novo binder design set (all 16 designs)

Source: design_leads.csv. Pipeline: RFdiffusion (Complex_base, noise 0) -> ProteinMPNN (v_48_020, binder chain, 8 seq/backbone, T=0.1) -> Boltz-2 fold-back (target --use_msa_server, binder single-sequence, 5 diffusion samples). iptm = interface predicted TM (pass > 0.5); complex_plddt = fold confidence (>0.7). Sequences are the ProteinMPNN binder-chain designs.

design	target	binder_len	mpnn_score	iptm	complex_plddt	confidence	pass	seq
dkk1_bb1	DKK1	63	1.232	0.858	0.908	0.898	Y	SSEEKEKLAKEYEEKAKK
dkk1_bb4	DKK1	63	1.248	0.844	0.846	0.846	Y	STAEGLRKGREELLKKAE
dkk1_bb5	DKK1	60	1.275	0.821	0.777	0.785	Y	AAAAAAAAAAERAAAAAN
dkk1_bb2	DKK1	62	1.044	0.745	0.743	0.743	Y	MTREELIAAAGAAGLAYG
dkk1_bb6	DKK1	89	1.315	0.743	0.837	0.818	Y	ALAVLLLALLAAALTALLI
dkk1_bb7	DKK1	85	1.172	0.709	0.899	0.861	Y	MEKKKKAEELLKLAEEYI
dkk1_bb3	DKK1	76	1.008	0.571	0.88	0.818	Y	SAALAAAAAAAGLAALAV
dkk1_bb0	DKK1	89	1.189	0.524	0.887	0.815	Y	EELEKKKKEAEEKLKKAF
gucy2c_bb2	GUCY2C	71	0.95	0.915	0.896	0.9	Y	TPEDIFAAGKAAADAAAG
gucy2c_bb5	GUCY2C	80	0.834	0.861	0.863	0.863	Y	AELAALAAEAEALAAAVL
gucy2c_bb4	GUCY2C	90	0.872	0.811	0.878	0.865	Y	MKEQAKKALEAAKKAAE
gucy2c_bb1	GUCY2C	73	1.065	0.715	0.878	0.845	Y	AEEELKKKLEEEAKKKEE
gucy2c_bb3	GUCY2C	84	0.842	0.619	0.865	0.816	Y	AEEKAKEKIEKALELAKK
gucy2c_bb0	GUCY2C	87	0.909	0.536	0.865	0.8	Y	ELEEAGKKAEELMEKIKK
gucy2c_bb7	GUCY2C	81	0.881	0.521	0.862	0.793	Y	REAVAKRIEEALKKAEAG
gucy2c_bb6	GUCY2C	72	1.005	0.401	0.852	0.762	n	SEEVEKKKEELLKEADEY

Leads: gucy2c_bb2 (treatment arm; ipTM 0.915) and dkk1_bb1 (interception arm; ipTM 0.858). 15/16 designs pass the ipTM>0.5 interface line (only gucy2c_bb6 fails at 0.401).

Table S2 — Target nomination and honest druggability triage (16 candidates)

Source: target_nomination.csv. Lane: T = treatment (surface antigens), I = interception (Barrett's->dysplasia). composite = raw nomination score (association + selectivity + tractability + novelty). biologic_druggability is the explicit filter applied over the raw score: STRONG/GOOD/MODERATE = addressable by an engineered biologic; CROWDED = deprioritized under the novelty steer; WEAK-BIOMARKER / NOT-A-BINDER-TARGET = down-weighted. The highest raw composite (GPX7, 13.86) is a lost tumor suppressor and is *not* carried forward.

target	lane	composite	tumor_selectivity	novelty	biologic_druggability	trajectory	druggability_note
GPX7	I	13.86	4	5	NOT-A-BINDER-TARGET	Barrett's (early)	LOST tumor suppressor -> restoration/ synthetic-lethal problem, NOT antibody-addressable
GUCY2C	T	13.0	5	5	STRONG	advanced + adjuvant	Surface, luminal-restricted; antibody/ ADC/bispecific/CAR all viable; therapeutic window is the key asset
REG4	I	12.47	4	5	WEAK-BIOMARKER	Barrett's->dysplasia	Secreted lectin; trap-able BUT therapeutic efficacy unproven; mainly progression biomarker
CDH17	T	12.38	4	5	STRONG	advanced	GI-restricted surface cadherin; antibody/ ADC/CAR viable; emerging validation
ERBB2	T	12.29	4	1	CROWDED	advanced (crowded)	Deprioritized per steer (T-DXd approved)
CEACAM5	T	12.0	3	4	STRONG	advanced	Surface; ADC precedent

target	lane	composite	tumor_selectivity	novelty	biologic_druggability	trajectory	druggability_note
							(tusamitamab); moderate window
CD276	T	12.0	3	5	STRONG	advanced	B7-H3 surface; ADC clinical-stage; antibody-tractable but broader normal expression
MET	T	11.55	3	2	GOOD	advanced	Surface RTK; antibody/ADC viable; lower amp frequency
ERBB3	T	11.49	3	3	GOOD	advanced	HER3 surface receptor; antibody/ bispecific viable; less crowded than HER2
DKK1	I	11.37	3	4	GOOD	advanced + progression	Secreted -> ligand-trap/neutralizing antibody viable; DKN-01 precedent
TFF3	I	10.51	3	4	WEAK-BIOMARKER	Barrett's	Secreted trefoil; strong Cytosponge biomarker; not a validated therapeutic target
OLFM4	I	10.46	3	4	WEAK-BIOMARKER	Barrett's->dysplasia	Secreted; interception biomarker; therapeutic role unproven
TACSTD2	T	10.0	3	4	GOOD	advanced	TROP2 surface; ADC-tractable; pan-epithelial (window caution)
MUC1	T	10.0	3	3	MODERATE	advanced	Tumor-glyco-MUC1 epitope needed; complex but antibody-addressable

target	lane	composite	tumor_selectivity	novelty	biologic_druggability	trajectory	druggability_note
EGFR	T	9.88	2	1	CROWDED/POOR-WINDOW	advanced (crowded)	Deprioritized; broad normal expression
CLDN18	T	9.36	4	1	CROWDED	advanced (crowded)	Deprioritized per steer (zolbetuximab approved)

Table S3 — Competitive landscape (EAC target axes)

Source: competitive_landscape.csv. Crowding is the qualitative active-EAC-trial density; approved/lead anchors confirmed against Drugs@FDA. Whitespace verdict drove the novelty steer.

Target/axis	Mechanism	Approved/lead anchor	Crowding (active EAC)	Trajectory position	Whitespace verdict
PD-1 / PD-L1	Immune checkpoint	pembrolizumab, nivolumab (approved, incl. adjuvant CheckMate-577)	VERY HIGH (41 active mentions)	Advanced + adjuvant	CROWDED — backbone, combination-only entry
HER2 / ERBB2	Amplified RTK	trastuzumab, T-DXd (approved GEJ/gastric)	HIGH (12)	Advanced (HER2+ ~15%)	CROWDED — but ADC/bispecific still active
VEGF / VEGFR2	Angiogenesis	ramucirumab (approved gastric/GEJ)	MED (8)	Advanced	MATURE — incremental
Claudin-18.2	Tight-junction surface Ag	zolbetuximab (approved 2024, CLDN18.2+ G/GEJ)	MED (3, fast-growing)	Advanced	RECENTLY VALIDATED — ADC/CAR-T next wave
FGFR2b	Amplified RTK	bemarituzumab (Ph3)	LOW (2)	Advanced (FGFR2b+ subset)	EMERGING — near-term
TIGIT	Immune checkpoint	domvanalimab (Ph2/3)	LOW (1-2)	Advanced	EMERGING IO — crowded pan-tumor
B7-H3 (CD276)	Surface Ag / ADC	ifinatamab deruxtecan (Ph1/2)	VERY LOW (1)	Advanced	WHITESPACE — novel surface ADC target
TROP2	Surface Ag / ADC			Advanced	

Target/axis	Mechanism	Approved/lead anchor	Crowding (active EAC)	Trajectory position	Whitespace verdict
		sacituzumab tirumotecan (Ph1/2)	VERY LOW (1)		WHITESPACE — ADC expanding into GI
GUCY2C	GI-restricted surface Ag	Ad5-GUCY2C-PADRE vaccine (Ph1/2)	VERY LOW (1)	Advanced/ adjuvant	WHITESPACE — GI-selective, vaccine/CAR/ ADC
CD73 / adenosine	Immunosuppressive axis	quemliclustat (Ph1)	VERY LOW (1)	Advanced	WHITESPACE — TME immuno-metabolic
DKK1	Wnt modulator (secreted)	DKN-01 (Ph2)	VERY LOW (1)	Advanced	WHITESPACE — secreted, low-DKK1 vs high
GATA6 / MYC / CCNE1 amplicons	Amplified TFs / cell-cycle	none (undrugged TFs)	NONE	All	HARD — transcription-factor/ undruggable
TP53 / CDKN2A loss	Tumor-suppressor loss	none directly	NONE	Barrett's->EAC (early)	HARD — synthetic-lethal / interception only
Barrett's/ dysplasia interception	Progression biology	none approved (chemoprevention: PPI+aspirin AspECT)	LOW (203 trials, mostly ablation/ screening)	Precursor	WHITESPACE — molecular interception unexploited

S4 — Detailed computational methods

Target structure preparation

- **GUCY2C:** AlphaFold DB UniProt P25092 (model v6). Extracellular domain isolated (residues 24-430; pLDDT>70 across ~50-425). Intracellular kinase/cyclase domain (~490-1073) excluded.
- **DKK1:** AlphaFold DB UniProt O94907 (model v6). CRD2 domain isolated (residues 178-256), the LRP6-binding functional region; disordered N-terminus and linkers excluded.
- **Hotspot definition (design input):** surface-exposed, high-confidence residue patches by neighbor-count + pLDDT heuristic - GUCY2C ECD [239,240,267,422,424,425]; DKK1 CRD2 [181,182,183,194,195] (isolated-domain numbering).

Design and validation

- **RFdiffusion:** Complex_base checkpoint, noise 0, 8 backbones/target, binder length 70-100 aa (GUCY2C) / 60-90 aa (DKK1).
- **ProteinMPNN:** v_48_020, binder chain only, 8 sequences/backbone, sampling T=0.1; best per backbone carried forward.
- **Boltz-2 fold-back:** target chain with --use_msa_server, binder as single sequence, 5 diffusion samples. Metrics: interface ipTM (pass > 0.5), complex pLDDT (fold > 0.7).
- **Interface residues (Fig 4):** computed post-hoc as target Ca within 8 A of any binder Ca in the predicted complex - GUCY2C 13 residues [130,132,146,192,325,326,327,330,347,348,349,378,379]; DKK1 8 residues [10,11,12,13,28,29,30,31] (complex-file numbering).
- **Hardware:** cloud A100 GPU.

Data sources

- Genomics: cBioPortal esca_tcga_pan_can_atlas_2018 (n=182, adenocarcinoma).
- Target association/tractability: Open Targets (EAC MONDO_0005028; Barrett's MONDO_0013662).
- Landscape: ClinicalTrials.gov (351 EAC interventional / 106 active; 203 Barrett's/dysplasia); Open Targets known-drug sets (106 agents); Drugs@FDA for approved-agent confirmation.

S5 — Epitope patches and cross-reactivity

- Design hotspots were chosen by structural heuristic (exposure + confidence), **not** by experimental functional-site mapping. It remains to be confirmed that the DKK1 binder occludes the LRP6 interface and that the GUCY2C binder engages a tumor-exposed epitope (open question O-5).
- Cross-reactivity screen (epitope_cross_reactivity.csv): **no high-similarity binder pairs detected** across the design set, indicating diverse solutions rather than a single repeated motif.

Table S4 / Figure S1 — Normal-tissue selectivity (GTEx v8)

Median TPM profiling of the two leads and four comparators across GTEx v8 tissues. GI fraction = share of summed median expression in intestine/colon/stomach/esophagus tissues. **This is normal-tissue data and establishes tissue selectivity only — it does NOT establish EAC tumor-cell expression** (a flagged gap for GUCY2C). GUCY2C and CDH17 are sharply GI-restricted; TROP2 (TACSTD2) is pan-epithelial (high in skin), illustrating the poor-window designation. See Figure S1.

Footnote: DKK1's top GTEx row, Cells_Cultured_fibroblasts (852.8 TPM), is a cell-culture artifact and not a physiological normal-tissue value; DKK1's solid-tissue levels are low, consistent with a stromal/paracrine source. In Barrett's tissue the biologically relevant DKK1 is likely stromal/paracrine — a pattern compatible with (and arguably favorable to) a secreted-ligand neutralizing-trap strategy.

gene	top_tissue	top_median_TPM	GI_fraction	top_nonGI_tissue	top_n
GUCY2C	Small_Intestine_Terminal_Ileum	28.2	0.85	Skin_Sun_Exposed_Lower_leg	1.7
DKK1	Cells_Cultured_fibroblasts	852.8	0.0	Cells_Cultured_fibroblasts	852.8
CDH17	Colon_Transverse	141.1	0.98	Artery_Tibial	0.4
CEACAM5	Esophagus_Mucosa	293.4	0.84	Vagina	78.9
TACSTD2	Esophagus_Mucosa	1419.1	0.29	Skin_Not_Sun_Exposed_Suprapubic	708.7
EGFR	Skin_Sun_Exposed_Lower_leg	78.3	0.15	Skin_Sun_Exposed_Lower_leg	78.3
ERBB2	Nerve_Tibial	130.1	0.17	Nerve_Tibial	130.1

Table S5 — Binder sequence-composition liability screen

Computed directly from the ProteinMPNN designed sequences. pctA = % alanine; high alanine content and long homopolymer runs are the signature of a generic low-complexity scaffold that can fold with spuriously high ipTM. The top GUCY2C design by ipTM (gucy2c_bb2) carries a high-alanine liability (44% Ala, 8-residue run); the DKK1 lead (dkk1_bb1) is clean on both interface confidence and composition. Orthogonal fold-back and monomer/decoy specificity controls (O-1) remain the decisive next check and have not been run.

Binder diversity (length-penalized pairwise sequence identity, all 16 designs): maximum pairwise identity 21.4% within the GUCY2C set (top pair gucy2c_bb5/bb3) and 32.9% within the DKK1 set (top pair dkk1_bb2/bb3, both alanine-rich); set means ≈15%. No pair approaches a high-similarity threshold; the designs are independent solutions. *Note:* %alanine is reported against no external benchmark distribution here — designs are flagged relative to an internal 15%-alanine caution line, and benchmarked calibration against published RFdiffusion/ProteinMPNN composition distributions (e.g. Cao et al. 2022; Bennett et al. 2023) is recommended before external submission.

design	target	len	iptm	pctA	pct_hydrophobic	max_homopolymer_run	frac_charged
dkk1_bb1	DKK1	63	0.858	17.5	25.4	2	55.6
dkk1_bb4	DKK1	63	0.844	27.0	39.7	3	46.0
dkk1_bb5	DKK1	60	0.821	50.0	63.3	10	21.7
dkk1_bb2	DKK1	62	0.745	61.3	77.4	9	6.5
dkk1_bb6	DKK1	89	0.743	38.2	62.9	5	28.1
dkk1_bb7	DKK1	85	0.709	27.1	44.7	5	48.2

design	target	len	iptm	pctA	pct_hydrophobic	max_homopolymer_run	frac_charged
dkk1_bb3	DKK1	76	0.571	63.2	78.9	7	6.6
dkk1_bb0	DKK1	89	0.524	9.0	23.6	4	65.2
gucy2c_bb2	GUCY2C	71	0.915	43.7	56.3	8	21.1
gucy2c_bb5	GUCY2C	80	0.861	48.8	68.8	3	20.0
gucy2c_bb4	GUCY2C	90	0.811	27.8	44.4	2	48.9
gucy2c_bb1	GUCY2C	73	0.715	13.7	28.8	3	56.2
gucy2c_bb3	GUCY2C	84	0.619	17.9	42.9	2	45.2
gucy2c_bb0	GUCY2C	87	0.536	14.9	32.2	3	46.0
gucy2c_bb7	GUCY2C	81	0.521	28.4	50.6	3	37.0
gucy2c_bb6	GUCY2C	72	0.401	6.9	22.2	3	62.5

Table S6 — Interface-residue mapping to canonical UniProt positions

Interface residues (target Cα within 8 Å of any binder Cα) in complex-file numbering and their canonical UniProt positions. Offsets: GUCY2C complex residue + 23 = UniProt P25092 position; DKK1 complex residue + 177 = UniProt O94907 position.

Target	Complex-file residues	UniProt positions	Domain (range)
GUCY2C	130,132,146,192,325,326,327,330,347,348,349,378,379	153,155,169,215,348,349,350,353,370,371,372,401,402	ECD (P25092 24–430)
DKK1	10,11,12,13,28,29,30,31	187,188,189,190,205,206,207,208	CRD2 (O94907 178–256)

S6 — Development gates and open questions

Reproduced from the scientific development plan and program dossier.

Go/no-go decision gates

Gate	Criterion
G1 developability	Leads pass ESM / biophysics liability screen -> else redesign
G2 binding	KD <= 50 nM + confirmed epitope -> else affinity maturation
G3 function + selectivity	Potent in disease-relevant assay + >=10x normal-tissue window -> else kill/reformat
G4 in-vivo PoC	Tumor control (treatment) / dysplasia regression (interception) at tolerated dose
G5 tox	Acceptable GI (GUCY2C) / skeletal+bone-GI systemic (DKK1) safety margin -> IND

Open questions

ID	Question	Kind
O-1	ESM developability/liability screen + orthogonal fold-back (Chai-1/AF2) on leads	computational (near-term)
O-2	GUCY2C normal-gut therapeutic window (human IHC + TDCC selectivity; mouse cross-reactivity)	preclinical - make-or-break
O-3	GUCY2C T-cell-engager GI toxicity manageable at efficacious dose	clinical
O-4	DKK1 interception surrogate endpoint (dysplasia regression) acceptance + payer	clinical/regulatory - make-or-break
O-5	Epitope hotspots validated as functional (LRP6-block for DKK1; tumor-exposed for GUCY2C)	computational/experimental

S7 — Strategy briefs (accompanying documents)

The following program documents accompany this manuscript as separate files and contain the clinical/regulatory/commercial reasoning, whose forward-looking statements are flagged there as assumptions requiring expert confirmation: - scientific_development_plan.md - dual-arm development plan, assay cascade, gates - regulatory_brief.md - EAC precedent, treatment/interception pathways, assumptions A-1..A-4 - commercial_financing_brief.md - epidemiology, market thesis, staged financing, assumptions C-1..C-4