A Goemans-Williamson type algorithm for identifying subcohorts in clinical trials

Pratik Worah

Overview & Key Findings:
- ▶ **Algorithm:** Provable SDP method pairing pathology & drug targets
- ▶ **UC Subcohorts:** Two disjoint subgroups (OSM/IL-23 vs. integrin/JAK1)
- ▶ **UC Validation:** Identifies anti-TNF non-responders in adult datasets
- ▶ **Broad Scope:** Extends to sepsis & breast cancer; *hypotheses to verify*

1. Motivation: Why Subcohort Selection vs. Clustering?

Clinical Trial Dilution in Heterogeneous Syndromes: In complex diseases like ulcerative colitis (UC), distinct patient subgroups are driven by contrasting inflammatory pathways. Testing a targeted therapy across an unstratified cohort dilutes treatment benefit when only a 10–20% patient subset carries the active target pathway.

Why Not Standard Unsupervised Clustering (*k*-Means / GMM)?
- ▶ **Forced Partitioning Premise:** Standard clustering assumes *every* patient (100%) belongs to one of k clean clusters. In reality, 75–85% of patients have multifactorial background profiles; forcing them into clusters mixes dissimilar patients and degrades specificity.
- ▶ **Genome-Wide Variance vs. Actionable Panels:** Unsupervised clusters reflect broad transcriptomic variance rather than aligning a focused subgroup with a druggable target pathway on a compact ($\leq$ 20-gene) clinical assay.

A. Global Clustering (Forced Full Partition): Every patient forced into a cluster



B. Subcohort Selection: Extract one homogeneous subgroup; leave background unassigned



Subcohort Selection Principle

Identify a homogeneous 10–20% disease subcohort S defined by a sparse linear threshold $\langle h, x_i \rangle \leq -\theta$ with **near-zero healthy/non-IBD control contamination**, leaving heterogeneous background patients unassigned.

2. Abstract Problem, SDP Algorithm & Guarantees

Abstract Problem Statement: Given gene expression profiles $x_i \in \mathbb{R}^d$ for disease cases U and healthy/non-IBD controls V, find a **sparse linear test** $h \in \mathbb{R}^d$ ($\|h\|_1 \leq s$) with margin $\theta > 0$ defining $S = \{i : \langle h, x_i \rangle \leq -\theta\}$ to simultaneously maximize:

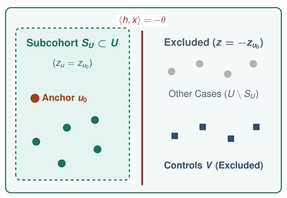
1. **Sensitivity** ($|S_U|/(|U| + |V|)$)**:** Select a large, homogeneous subset of disease cases $S_U = S \cap U$.
2. **Specificity** ($|S_U|/(|S_U| + |S_V|)$)**:** Exclude healthy/non-target controls ($S_V = S \cap V \approx \emptyset$).

Anchored Signed MAX-CUT SDP (Algorithm 1): For an anchor patient $u_0 \in U$ and binary spins $z_i \in \{\pm 1\}$ ($z_i = z_{u_0} \iff i \in S$), we relax $(z_i - z_j)^2 = 2(1 - z_i z_j)$ to unit vectors $v_i \in S^{n-1}$ ($\|v_i\|_2 = 1$) and round via $\gamma \sim \mathcal{N}(0, I_n)$, setting $z_i = \operatorname{sign}\langle v_i, \gamma \rangle$:

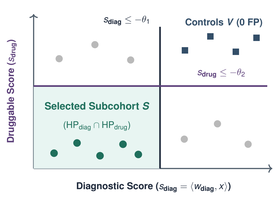
$$\max_{h \in \mathbb{R}^d,\, \|h\|_2 = 1} \sum_{u \in U,\, u' \in V} \frac{2 - 2\langle v_u, v_{u'} \rangle - 2\langle v_{u_0}, v_{u'} \rangle + 2\langle v_u, v_{u_0} \rangle}{3}$$
$$\text{s.t. } \langle h, x_u \rangle \leq -\theta \langle v_u, v_{u_0} \rangle \ (\forall u \in U), \quad \|h\|_1 \leq s,$$
$$\langle h, x_{u'} \rangle \geq \theta(1 - \langle v_{u'}, v_{u_0} \rangle) \ (\forall u' \in V).$$

Dual-Hyperplane Intersection (Algorithm 2): Train independent SDP hyperplanes on **Diagnostic markers** (HP_{diag}) and **Druggable targets** (HP_{drug}), intersect $S_1 = \text{HP}^{(1)}_{\text{diag}} \cap \text{HP}^{(1)}_{\text{drug}}$, and remove S_1 ($U \leftarrow U \setminus S_1$) to identify a disjoint second subcohort S_2.

A. One-Sided Anchored SDP Cut (Algorithm 1)



B. Dual-Hyperplane Intersection (Algorithm 2)



Theoretical Guarantees (Worah, AISTATS 2026)

Goemans-Williamson (1995) rounding guarantees expected sensitivity $\geq$ **0.820 OPT** (Theorem 1), specificity $\Omega(\sqrt{\text{OPT}})$ (Theorem 2), and expected margin preservation within factor $\frac{2}{\pi} \approx 0.64$ (Theorem 3).

3. Ulcerative Colitis (PROTECT): Problem & 20-Gene Panel

Limited Efficacy of First-Line Anti-TNF Therapy (30–40% Remission): In unselected ulcerative colitis (UC) patients, first-line anti-TNF biologics achieve 1-year steroid-free mucosal healing in only 30–40% of patients. **West et al. (*Nature Medicine*, 2017)** screened 64 cytokines across clinical cohorts, stratified patients into **OSM-high vs. OSM-low** (10–15% vs. 69–85% anti-TNF healing), and proposed blocking OSM directly (though clinical anti-OSM trials faced hematologic toxicity). Moreover, stratifying solely on OSM misses a distinct **low-OSM / low-TNF subcohort driven by lymphocyte trafficking and JAK1 signaling (SC2)** and does not score which approved target pathways (IL-23p19, TL1A, JAK1, $\alpha_4\beta_7$) fit each subgroup.

Discovery Dataset & Dual-Hyperplane Setup (GSE109142; Haberman et al., 2019): Publicly available pretreatment rectal biopsy RNA-seq from **206 treatment-naïve pediatric UC patients** and **20 non-IBD controls** (226 total samples, standardized z-scores). We apply Algorithm 2 ($s = 2, \theta = 0.30$) to a curated **20-gene clinical panel**:

Panel Axis	Curated Genes & Biological Role in Colonic Mucosa
10 Diagnostic	**OSM** (stromal cytokine resistance endotype), **S100A8, S100A9** (fecal calprotectin alarmin subunits), **MUC2** (goblet-cell mucin barrier), **IL1B, IFNG, IL10** (mucosal cytokines), **CXCL8** (IL-8 neutrophil recruitment), **CD68** (macrophages), **REG1A** (epithelial crypt regeneration)
10 Druggable	**TNF** (soluble/transmembrane TNF-α cytokine), **IL23A** (IL-23-specific p19 subunit), **IL12B** (shared IL-12/IL-23 p40 subunit), **TNFSF15/TL1A** (mucosal cytokine & fibrosis amplifier), **S1PR1** (sphingosine-1-phosphate lymphocyte egress receptor), **JAK1** (intracellular Janus kinase 1 signaling), **JAK3** (lymphoid Janus kinase 3 γ_c signaling), **ITGA4** ($\alpha_4\beta_7$ gut-homing integrin subunit), **OSMR** (oncostatin M receptor β stromal subunit), **PPP3CA** (calcineurin catalytic α subunit)

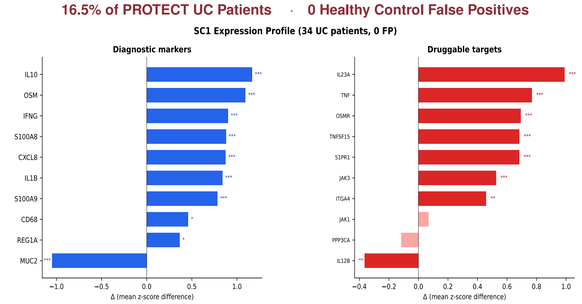
Why Split Diagnostic & Druggable Hyperplanes?

Requiring every selected patient to satisfy **both** $\langle w_{\text{diag}}, x_{\text{diag}} \rangle \leq -\theta_1$ and $\langle w_{\text{drug}}, x_{\text{drug}} \rangle \leq -\theta_2$ guarantees that each subcohort combines a coherent mucosal pathology state with a uniform, actionable target profile (0/20 non-IBD controls).

**Caveat: Values and tables in this box are retrospective predictions from our algorithm and need clinical verification.*

4. UC Subcohort 1 (SC1): Parallel Inflammation Bypasses Anti-TNF

16.5% of PROTECT UC Patients · 0 Healthy Control False Positives



Summary: Why Blocking TNF Alone Is Predicted to Be Insufficient in SC1. Although these patients in the PROTECT dataset exhibit elevated TNF expression, blocking TNF alone is predicted to be insufficient for mucosal healing because a parallel inflammatory pathway (**Oncostatin M / OSM**) is simultaneously upregulated:
- ▶ **Mucosal Pathology Profile (Blue Bars, Left):** Marked Oncostatin M–driven mucosal inflammation accompanied by depletion of goblet-cell mucin (MUC2), consistent with primary anti-TNF non-response.
- ▶ **Druggable Target Profile (Red Bars, Right):** Alongside TNF, three alternative therapeutic targets (**IL-23p19**, **TL1A**, and the **S1PR1** lymphocyte-trafficking receptor) are upregulated, whereas the shared IL-12/23 subunit (p40) is not upregulated.

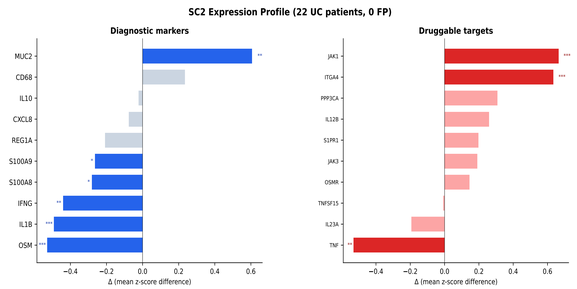
SC1 Summary (Hypothesis to Verify)

Neutralizing TNF leaves the parallel Oncostatin M / IL-23p19 / TL1A cascade active; these patients are candidates for **selective IL-23p19 blockade, anti-TL1A blockade, or S1PR1 modulation**.

**Caveat: Values and figures in this box are retrospective predictions from our algorithm and need clinical verification.*

5. UC Subcohort 2 (SC2): Lymphocyte-Trafficking & JAK1 Colitis (TNF-Low)

10.7% of PROTECT UC Patients · Zero Overlap · 0 Control False Positives



Summary: Lymphocyte-Trafficking & JAK1 Colitis (Low TNF/OSM). This second, disjoint subcohort in the PROTECT dataset exhibits moderate-to-severe active colitis driven by a **distinct non-cytokine-storm mechanism**: baseline TNF and OSM expression are low, while lymphocyte-trafficking and intracellular signaling genes are elevated:
- ▶ **Mucosal Pathology Profile (Blue Bars, Left):** Low Oncostatin M and low acute cytokine expression, with preserved MUC2 goblet-cell barrier expression.
- ▶ **Druggable Target Profile (Red Bars, Right):** Despite low baseline TNF and IL-23 expression, mucosal inflammation is maintained by **gut-homing integrin (ITGA4** / $\alpha_4\beta_7$**)** and intracellular **JAK1/JAK3 signaling**.

SC2 Summary & Combined UC Coverage

Rather than an acute TNF/OSM cytokine cascade, SC2 is characterized by **selective JAK1/3 signaling and gut-homing integrin (ITGA4** / $\alpha_4\beta_7$**)** expression. Together, SC1 and SC2 account for **27.2% of UC patients in PROTECT** (and **17–27% across adult datasets**) with zero control false positives.

**Caveat: Values and figures in this box are retrospective predictions from our algorithm and need clinical verification.*

6. Therapeutic Hypotheses in Ulcerative Colitis

Summary: Contrasting Molecular Subcohorts in Active Ulcerative Colitis. Moderate-to-severe UC is not a single uniform molecular state. Retrospective analysis of the PROTECT dataset identifies **two disjoint UC subcohorts** driven by contrasting pathways that suggest different targeted therapies:
- ▶ **Subcohort 1: Parallel Cytokine Cascade Bypasses Anti-TNF (SC1, 16.5%).** Baseline TNF is elevated, but a parallel inflammatory cascade (Oncostatin M, IL-23p19, and TL1A) is simultaneously upregulated and sustains colitis despite TNF blockade.
- ▶ **Subcohort 2: Lymphocyte-Trafficking & JAK1 Endotype (SC2, 10.7%).** Acute mucosal cytokines (OSM, TNF, IL-1β) are low; active colitis is instead associated with lymphocyte trafficking into the colon (ITGA4 / $\alpha_4\beta_7$ integrin) and intracellular JAK1/3 signaling.

Subunit Specificity Within a Pathway (IL-23p19 vs. Shared IL-12/23p40):
- ▶ The cytokine IL-23 comprises a **specific p19 subunit** and a **shared p40 subunit** (also present in IL-12). In SC1, only the **IL-23-specific p19 subunit** is upregulated while the shared p40 subunit is not upregulated, favoring **selective IL-23p19 blockade** over shared p40 blockade.

Subcohort	Primary Molecular Driver	Candidate Pathways to Test
SC1 (16.5%)	Parallel OSM / IL-23p19 / TL1A cascade sustains colitis despite TNF blockade	**Selective IL-23p19 blockade, Anti-TL1A blockade, or S1PR1 modulation**
SC2 (10.7%)	Low TNF/OSM; active colitis driven by gut-homing lymphocytes & JAK signaling	**Selective JAK1 inhibition or gut-homing integrin (ITGA4) blockade**

Clinical Trial Enrichment Hypothesis

Stratifying with a 20-gene biopsy signature identifies primary anti-TNF non-responders with parallel OSM/IL-23p19/TL1A inflammation (SC1) as candidates for selective IL-23p19 or TL1A trials, while separating a low-cytokine, integrin/JAK1-driven subcohort (SC2) as candidates for JAK1 or $\alpha_4\beta_7$ integrin trials.

**Caveat: Values and tables in this box are retrospective predictions from our algorithm and need clinical verification.*

7. Retrospective UC Validation: Disease Severity & Adult Datasets

Retrospective Disease Severity Check: Both Subcohorts Have Moderate-to-Severe Colitis. Does **SC2** (preserved MUC2, low acute cytokines) merely capture mild disease? Recorded clinical metadata in **PROTECT (GSE109142)** and the adult datasets show comparable baseline severity:
- ▶ **Histologic & Inflammatory Burden:** In the PROTECT dataset, **SC1 and SC2 had comparable baseline luminal inflammation** (fecal calprotectin $\sim 50\times$ normal) and **identical median histologic severity**, and most patients in both groups required **systemic corticosteroids at diagnosis** (94% in SC1, 77% in SC2).
- ▶ **Clinical & Endoscopic Severity:** In PROTECT metadata, **SC2 matches the overall cohort in baseline Total Mayo severity** (median 8.0; 86% moderate-to-severe). Similarly, in the adult datasets with recorded Mayo scores (**GSE92415, GSE73661**), **SC1 and SC2 had statistically indistinguishable baseline Mayo scores**.

Retrospective Evaluation on Four Adult UC Datasets: We evaluated the **20-gene scoring rule trained on PROTECT** retrospectively on pretreatment colon biopsy microarrays from **four adult UC studies (Arijs et al., *Gut* 2009 [GSE12251]; Arijs et al., *PLoS One* 2009 [GSE16879]; Sandborn et al., *Gastroenterology* 2014 [GSE92415]; Arijs et al., *Gut* 2018 [GSE73661])** without modifying weights:
- ▶ **Subcohort Proportions & Specificity:** Across all four adult datasets, the rule trained on PROTECT recovers **SC1 (9–17%)** and **SC2 (8–10%)** (**GSE12251:** 17.4% SC1, 8.7% SC2; **GSE16879:** 16.7% SC1, 8.3% SC2; **GSE92415:** 9.2% SC1, 8.0% SC2; **GSE73661:** 16.4% SC1, 10.4% SC2) with **zero false positives among 39 healthy controls**, reproducing their distinct profiles (**OSM/TNF/IL23A-high** in SC1 vs. **OSM/IL1B-low, ITGA4/JAK1-high** in SC2).
- ▶ **Recorded Anti-TNF Outcomes:** Using the treatment outcomes recorded in those four studies, patients classified into **SC1 had markedly lower anti-TNF response rates**: **0% mucosal healing in GSE16879 (0/4) and GSE73661 (0/6), and 25% response in GSE12251 (1/4) and GSE92415 (1/4) (under 17% overall** across GSE12251, GSE16879, GSE92415, and GSE73661, compared to **over 50%** among the remaining non-SC1 UC patients).

Retrospective UC Validation Summary

Across pediatric (**PROTECT, GSE109142**) and adult (**GSE12251, GSE16879, GSE92415, GSE73661**) datasets, **SC1 and SC2 both represent moderate-to-severe active colitis** with **zero control false positives**, and **SC1 patients in the adult datasets had significantly lower anti-TNF response rates**.

**Caveat: All analyses in this box are retrospective evaluations of public datasets; prospective clinical validation is required.*

8. Sepsis, Breast Cancer, Caveats & Future Work

A. Retrospective Analyses in Additional Domains: ICU Sepsis & Breast Cancer.
- ▶ **ICU Sepsis (Discovery & European MARS Datasets):** Applying our algorithm retrospectively to ICU sepsis blood transcriptomes (**GSE185263**, Baghela et al., 2022; **GSE65682**, Scicluna et al., 2017) identifies **two distinct immunosuppressive subgroups** with **zero healthy control false positives**: a **checkpoint-inhibited subgroup (∼10%)** with elevated PD-1/LAG-3 expression on T-cells (suggesting checkpoint blockade), and a **lymphopenic innate-inflammatory subgroup (∼7%)** driven by TLR4/MyD88 signaling (suggesting IL-7 restoration + TLR4 inhibition). Evaluating the rule trained on the discovery dataset (**GSE185263**) on the European MARS dataset (**GSE65682**; Scicluna et al., 2017) reproduces the **same ∼10% checkpoint-inhibited subgroup** with **zero control false positives**.
- ▶ **Breast Cancer (METABRIC Dataset, *N* = 1,904):** In a retrospective split of the METABRIC dataset (Curtis et al., 2012), pairing an epigenetic **promoter-methylation hyperplane** (BRCA2/MSH6 DNA-repair silencing) with a **druggable receptor expression hyperplane** (NR1H2) increases holdout specificity from **57%** (single test) to **70%** (dual test), identifying subcohort-specific epigenetic–receptor coupling not apparent in whole-cohort averages.

Important Caveat: Retrospective Computational Hypotheses

All results on this poster (across UC, sepsis, and breast cancer) are retrospective computational analyses of publicly available gene expression datasets; no clinical trials were conducted in this work, and all therapeutic predictions require prospective clinical verification.

B. Future Work & Clinical Verification Steps:
- ▶ **1. Larger Outcome Datasets for SC2 & Next-Line Pathways:** Expand retrospective biopsy scoring to larger datasets to evaluate SC2 drug outcomes and test differential response to selective IL-23p19, TL1A, JAK1, and $\alpha_4\beta_7$ integrin blockade.
- ▶ **2. Bedside 20-Gene Assay Calibration:** Partner with gastroenterologists to calibrate weights and thresholds on rapid 20-gene clinical biopsy panels.

References: Arijs et al. (*Gut*, 2009; *PLoS One*, 2009; *Gut*, 2018); Sandborn et al. (*Gastroenterology*, 2014); West et al. (*Nat. Med.*, 2017); Haberman et al. (*Nat. Commun.*, 2019); Baghela et al. (*EBioMedicine*, 2022); Scicluna et al. (*Lancet Respir. Med.*, 2017); Curtis et al. (*Nature*, 2012); Goemans & Williamson (*J. ACM*, 1995); Worah (*AISTATS*, 2026).

AI Use Statement: AI was not used to design and implement the abstract SDP algorithm, but it was used to modify and run the pipeline for UC datasets.