TOWARD A SHARED OBJECTIVE FOR OMICS FOUNDATION MODELS: THE PRIMO BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Omics foundation models are trained and promoted as tools for drug development, but they are rarely evaluated in this context. Existing benchmarks mostly operate at the level of the cell, the gene, or the batch, but the decisions that advance a drug program are made at the level of the patient the sample came from: which indication to pursue for a given target, whether a candidate should progress into the clinic, and which patients should be enrolled to observe a treatment effect. In addition, tasks and datasets are often preprocessed and built differently from one study to another, making results difficult to compare. We introduce PRIMO (Patient Representations in Multi-Omics), an open benchmark of sample-level tasks, with a first version focused on bulk and single-cell transcriptomics. We curate 63 tasks from 45 public cohorts on 30 diseases, most of them in four therapeutic areas, covering these decisions through five task categories: treatment outcome, clinical activity, prognosis, endotype and perturbation response, including cross-species and cross-indication transfer. Every task follows the same curation process: a clear drug-development question, a public cohort that can answer it, data cleaning, and tests ensuring the contained signal is linked to the biology rather than to technical artifacts. Under a common evaluation protocol with subject-grouped, label-stratified cross-validation, we evaluate 11 foundation models trained on bulk, single-cell or both modalities in the zero-shot setting: their weights stay frozen, and a shared linear probe reads their embeddings. All entries are rated against a raw-expression baseline, so that every rating reads directly as the value a representation adds over the raw data. Since foundation models are fine-tuned in practice, these zero-shot ratings are a lower bound on their value. Even in this setting, the strongest foundation models reach or exceed the raw-expression baseline, while most remain below it. We release PRIMO as a public leaderboard on curated tasks, with ratings per therapeutic area, giving model developers, drug developers and clinicians a shared reference to track progress and optimize for.

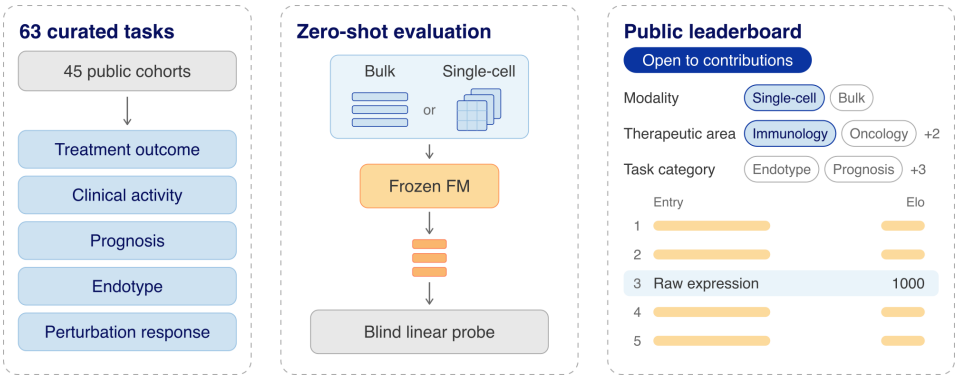

Figure 1: Overview of PRIMO. (i) PRIMO contains 63 tasks curated thoroughly from 45 public cohorts, spanning 5 task categories. (ii) Foundation models producing sample-level representations are evaluated in a zero-shot setting. (iii) PRIMO is a modular public leaderboard open to contributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

A drug program advances through concrete decisions: which indication to pursue for a given target, whether a drug candidate should progress into the clinic, and which patients to enroll to observe a treatment effect. The number of foundation models trained on large omics corpora has exploded in recent years (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Kang et al., 2026), and these models are increasingly offered as tools to inform exactly these decisions. Yet, they are rarely evaluated in this context. In comparison, other fields turned their promise into measurable progress by fixing a shared objective: common tasks, data and metrics that let the whole community measure progress on the same terms (Deng et al., 2009; Wang et al., 2018; Kryshtafovych et al., 2023). Omics foundation models have no such open objective yet for questions in clinical drug development. Existing omics benchmarks mostly focus on technical aspects such as data integration (Luecken et al., 2022), or cell-level predictions and perturbations (Luecken et al., 2025; Roohani et al., 2025). While these are important problems, they do not test the subject-level questions that arise across the clinical pipeline, such as whether a patient will respond to a treatment. This mismatch between what is measured and what is decided is one reason the clinical impact of AI has so far been limited (Bender et al., 2026).

Subject-level evaluation is notoriously difficult. Clinical scores and endpoints use different instruments and metrics across studies; cohorts are small and often contain several study arms as well as potential confounders. In addition, published comparisons often conflate the model itself with data preprocessing (such as normalisation or gene selection) and evaluation setup (such as classifier choice and hyperparameter search), so a reported gain may belong to the pipeline rather than to the model. Consistent with this, recent studies find that complex transcriptomic deep learning models do not reliably generalize beyond the tasks of their original papers, whether the target is a perturbation effect (Ahlmann-Eltze et al., 2025), a cell-level property (Kedzierska et al., 2025), or a patient outcome (Elmarakeby et al., 2025). This calls for a public evaluation protocol shared by the whole field, like the one TabArena provides for tabular foundation models (Erickson et al., 2025).

We introduce **PRIMO** (Patient Representations in Multi-Omics), an open community benchmark for foundation-model representations of sample-level omics, focused on questions that matter in the clinical drug development pipeline, which it turns into scored tasks. PRIMO aims to cover the whole drug discovery and development pipeline, from target identification to the clinic, across omics modalities. Its first release focuses on transcriptomics and on clinical and translational questions. Our contributions are:

- We curate a set of 63 tasks in five task categories that follow the drug development pipeline: treatment outcome, clinical activity, prognosis, endotype and perturbation response, including translation from mouse to human;
- We evaluate 11 bulk and single-cell RNA-seq foundation models in a conservative zero-shot setting, where a linear probe reads frozen embeddings, against a strong raw-expression baseline;
- We release a public leaderboard, with ratings per modality, therapeutic area and task category, and provide code to benchmark new methods with a label-blind evaluation: participants never see the labels, the folds or the probe;
- We give explicit curation and inclusion rules (Section 3.1) that contributors can follow to propose new tasks, and a protocol that can extend to other omics modalities, since entries only submit one embedding per sample.

The first version of PRIMO shows that foundation models already add sample-level value on some tasks, without any fine-tuning. It also shows that the best representation depends on the therapeutic area, which is why PRIMO reports ratings per area. Sample-level tasks remain far from saturated, leaving real room for the next generation of omics foundation models to show their value on the decisions that matter to clinical drug development pipelines. We release PRIMO as an open and evolving benchmark: additional omics modalities, multi-omics tasks, new task categories and therapeutic areas will follow so that PRIMO stands as a shared, growing reference for the community to track progress.

2 RELATED WORK

Transcriptomic foundation models. Single-cell models are pretrained on tens to hundreds of millions of cells (Cui et al., 2024; Theodoris et al., 2023; Chen et al., 2026; Hao et al., 2024; Kalfon et al., 2025b;a; Dong et al., 2026). Their papers mostly evaluate them at the level of the cell or the gene, for example on cell-type annotation, batch integration, perturbation prediction or gene-network inference. Bulk models are pretrained on tumour cohorts or on large bulk compendia (Gélard et al., 2025; Gélard et al., 2026; Kang et al., 2026), sometimes with a clinical fine-tuning stage (Shen et al., 2026), and are mostly evaluated on tasks close to their pretraining data, such as cancer-type classification and survival in TCGA. EVA is trained on both modalities (Bandasack et al., 2026). In every case, the developers choose the tasks, the splits and the preprocessing. Published results are therefore hard to compare, and self-assessment tends to favour the new method (Norel et al., 2011).

Benchmarks. Several benchmarks score cells, genes or perturbations (Luecken et al., 2022; 2025; Roohani et al., 2025), and PertEval-scFM probes frozen single-cell embeddings to predict perturbation effects in cells (Wenteler et al., 2025). PRIMO's design draws on prior work in omics and in other fields. Every entry is rated against a raw-expression baseline; the perturbation metric removes the response shared across subjects, since standard metrics are inflated by systematic differences between perturbed and control cells (Viñas Torné et al., 2026) (Appendix B); embeddings are scored by fixed downstream models, as in MTEB (Muennighoff et al., 2023); labels stay hidden from participants, as in CASP (Kryshtafovych et al., 2023); ratings are anchored to a reference entry and the board keeps accepting new entries, as in TabArena (Erickson et al., 2025); and the curation of every task is documented, in the spirit of datasheets for datasets (Gebru et al., 2021).

Sample-level evaluation. A few studies compare representations on sample-level tasks: Elmarakeby et al. (2025) evaluate single-cell foundation models on cancer outcomes, SPARE compares sample representations built from single-cell data (Shitov et al., 2025), and PaSCient is compared with single-cell foundation models on disease classification (Liu et al., 2026). Liang et al. (2026) find that bulk transcriptomic models of immunotherapy response perform near chance on independent cohorts. Each of these studies uses its own tasks and protocol. To our knowledge, no shared benchmark scores sample-level representations across bulk and single-cell data with a blind protocol. PRIMO provides one, with 63 tasks on 30 diseases and a public leaderboard that stays open to new models.

3 THE PRIMO BENCHMARK

PRIMO is built around three parts: the curation process that turns public data into a task (Section 3.1), the evaluation protocol that scores every model under the same rules (Section 3.2), and the public leaderboard that ranks models and makes results open to the community (Section 3.3).

3.1 CURATED TASKS TO ASSESS SAMPLE-LEVEL REPRESENTATION VALUE IN DRUG DEVELOPMENT

Current evaluations mirror the way drug development is organised: in silos. Preclinical research predicts the molecular effect of a drug, for instance how a compound shifts gene expression in a model system, while clinical development measures disease activity and treatment outcome in patients. Omics foundation models are promoted as the bridge between these two stages, yet their benchmarks focus on the preclinical side, scoring cells, genes or cell-level perturbations, without evaluating whether the same representation captures both the molecular effect of a drug and the clinical state of the patient. PRIMO scores one representation at both ends: perturbation-response tasks, including mouse-to-human transfer, probe the molecular effect of a drug, while treatment outcome, clinical activity, prognosis and endotype tasks probe the patient.

Key tasks in a drug development program. PRIMO admits a task only if it plays an established role in a drug program. In this first version, we instantiate this criterion through five task categories:

- **Treatment outcome** predicts, from a pre-treatment profile, whether a patient reaches the clinical endpoint of a treatment, the question behind responder enrichment in trial design and the choice of indication for a drug;

- **Clinical activity** predicts a validated disease-activity score, of the kind used as trial endpoints, from a molecular readout, so that a representation can serve as a molecular proxy for the clinical outcomes on which drugs are approved;
- **Prognosis** predicts, from a profile taken at diagnosis or at baseline, whether a later clinical event occurs, such as death, relapse or disease progression, a prediction used to enrich trials for patients likely to have events and to size them;
- **Endotype** classifies molecular subtypes of a disease, which define the subpopulations and indications where a mechanism of action is expected to work;
- **Perturbation response** predicts the post-minus-pre profile change induced by drug exposure, the pharmacodynamic readout that links a drug's preclinical mechanism to its effect in patients.

We chose prediction targets for their role in a drug program, across therapeutic areas. Clinical activity uses validated indices, such as the SES-CD endoscopic score in Crohn's disease (Daperno et al., 2004), the NAFLD activity score in steatohepatitis (Kleiner et al., 2005), the ISUP grade in prostate cancer (Epstein et al., 2016) and the SOFA score in sepsis (Vincent et al., 1996). Endotypes range from synovial pathotypes in rheumatoid arthritis (Lewis et al., 2019) to NPM1 and FLT3-ITD mutations in acute myeloid leukaemia.

Data curation. For every dataset, we apply a shared cleaning procedure, including label harmonisation and manual annotation when needed, quality control and filtering. These steps guard against common, silent errors and fix explicit rules for turning a raw dataset into a benchmark task.

- **Genes**: we use NCBI Gene IDs, with gene symbols alongside. Counts of genes that map to the same ID are summed.
- **Counts**: we use raw counts whenever the authors deposited them, and flag datasets that are already preprocessed. We exclude datasets profiled with targeted gene panels, identified by a low number of detected genes per sample and a high share of counts in the 500 most expressed genes.
- **Labels**: we conduct manual annotation when needed, for instance, when a dataset's GEO metadata disagrees with the original publication describing the study.
- **Single-cell quality control**: see Appendix E.

Inclusion criteria. Once a candidate dataset for one of the five task categories has passed the curation procedure above, we check it against the following inclusion criteria before it enters PRIMO:

- **License**: the data must be redistributable; cohorts that cannot be redistributed are excluded and recorded as a data-availability constraint (Gebru et al., 2021). Source studies are available in Appendix D.
- **Number of samples**: minimum 15 for single-cell tasks, minimum 30 for bulk tasks, and 20 pre- and post-treatment pairs for perturbation tasks.
- **Number of patients**: enough patients so that each of the 3 folds (Section 3.2) contains at least 5 samples.
- **Signal**: the raw-expression baseline should score above a label-permutation null. A task that misses this bar can be kept if it answers a central drug-development question; it then stays as an open problem.
- **Confounder checks**: the dataset should have limited confounder effects, such as a biopsy site that only exists at advanced stages, or a treatment only severe patients receive (Appendix F). We also check for technical confounding: sequencing depth, the number of detected genes and, when available, the sequencing batch are each used alone to predict the label.

3.2 Evaluation design protocol

This section describes the evaluation protocol: how representations are probed, how folds are built, and which metric each task reports.

Probing frozen representations. PRIMO evaluates models zero-shot (Kedzierska et al., 2025): no model is fine-tuned on a task, and a linear probe is fitted on its frozen embeddings, as in MTEB (Muennighoff et al., 2023). Scores are therefore a lower bound on a model's value: in practice, foundation models are adapted to each task, which this protocol deliberately excludes to keep the evaluation blind and comparable. The blind protocol requires this setting, since participants never see the labels, and it keeps entries comparable. Entries submit one embedding per sample; they never submit predictions. For single-cell data, pooling cells into a sample vector is part of the method under test. Each submitted embedding is first scaled to unit ℓ_2 norm; the raw-expression baseline keeps its $\log_2(\mathrm{CPM}+1)$ scale. Features are then standardised on the training fold, and the same regularised linear probe is fitted on the same frozen partitions for every entry. Classification uses ℓ_2-penalised logistic regression, with the inverse penalty selected from 13 values log-spaced over $[10^{-6}, 10^{6}]$ by AUROC (weighted one-vs-rest for multiclass targets) under an inner repeated stratified k-fold, $k = \min(5, n_{\text{minority}})$, three repeats. Regression uses ridge regression with the penalty selected from 25 values log-spaced over $[10^{-6}, 10^{6}]$ by efficient leave-one-out error. Perturbation response uses multi-output ridge on the same grid, with standardised test features clipped at ± 20. Two non-linear probes, an RBF-SVM and an MLP, do not beat the linear probe (Appendix I).

Cross-fold evaluation. Each task is scored by subject-grouped 3-fold cross-validation on ten *partitions*: a partition assigns every subject to one of three folds, so that all samples from one subject fall in the same fold (Figure 2). The three transfer tasks have a single partition instead: they train on a source cohort and are scored once on a disjoint target cohort (mouse to human, or one skin disease to another). Classification folds are stratified by class and regression folds by label tertiles. Partitions are frozen before any model is scored: for each task we draw candidate partitions from a fixed list of 16 seeds, discard any in which a test fold lacks a class, and retain the ten with the most even label distribution across folds. Perturbation folds are drawn by grouped k-fold over subjects without stratification. Embeddings are label-free and produced before scoring, while standardisation, penalty selection and response-gene selection all happen inside the training fold; the inner penalty search is stratified but not subject-grouped. A task with constant predictions, a single-class training fold, or a non-finite fold score is recorded as invalid.

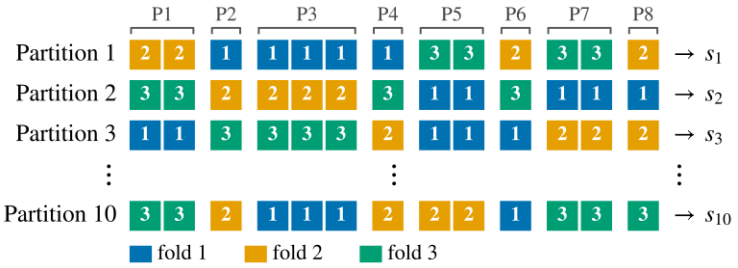

Figure 2: Subject-grouped 3-fold cross-validation on ten partitions, illustrated on a toy task with 13 samples from 8 subjects. Each square is a sample, and brackets group the samples of one subject (P1–P8), which always fall in the same fold. Each partition assigns the subjects to three folds; it is drawn once and shared by all entries. Within a partition, each fold is the test set once while the probe is trained on the other two, and s_p is the mean of the three test-fold scores. The task score is the mean of $s_1, \ldots, s_{10}$. The Elo is computed on a per-partition basis (Section 3.3).

Metrics. Each category is scored with one native metric: Pearson correlation for clinical activity, AUROC for endotype, prognosis and treatment outcome, and a centred Spearman correlation for perturbation response. The perturbation target is the post- minus pre-treatment difference in $\log_2(\mathrm{CPM}+1)$ expression, restricted to the 600 genes with the largest absolute mean response in the training fold. For each test sample, the training-fold mean response is subtracted from both the observed and predicted responses, and the two are compared by Spearman correlation across genes; samples with a constant ranking score zero, and the task score is the mean over test samples. This removes the response shared across subjects, so the metric rewards only the patient-specific part of the effect and reads zero when there is none. For the cross-species transfer task, the target cohort's gene-wise mean is removed instead, separately from observations and predictions (Appendix B).

3.3 PUBLIC LEADERBOARD

PRIMO ships as a public, continuously growing leaderboard, open to external submissions under guidelines. In this section, we present the blind submission system, the baseline model, and the ranking strategy employed in the leaderboard.

Blind submission. PRIMO holds everything downstream of the representation fixed, so two entries can differ only through the vectors they submit. Participants download data with opaque sample identifiers, run their own frozen encoder, and return one vector per sample; they never see the labels, the folds, or the probe. This paper describes each task, but the row-level key that maps an opaque sample to a public record stays in the private registry. The source studies are public and cited in Appendix D, so the protocol cannot rule out that a participant re-identifies samples; like other public benchmarks, it relies on good faith. The blind protocol still limits the risk that a participant tunes on evaluation targets. Repeated submissions could still exploit the per-task scores, so the board shows how many times each entry was submitted.

Baseline. Every board is anchored to one label-free baseline, which we call the raw-expression baseline: for each dataset, counts are normalised to $\log_2(\mathrm{CPM}+1)$ and the 1,200 genes with the highest variance over all samples form the embedding (for single-cell datasets, each cell is first normalised to $\log_2(\mathrm{CP10k}+1)$ and cells are averaged per sample). The baseline goes through the same evaluator as a submission and sets the rating at 1000. Linear probing on top of gene expression is the usual control in evaluations of omics foundation models (Kedzierska et al., 2025; Ahlmann-Eltze et al., 2025); the number of genes is the setting used by scGPT for several of its downstream tasks (Cui et al., 2024). This number of genes is fixed in the first release of PRIMO and not tuned on the current tasks. Appendix H compares it with other label-free baselines: all genes give the same score within noise, and PCA projections score clearly lower.

Ranking. Because tasks use different metrics, their scores cannot simply be averaged. We therefore compare entries pairwise within each task, where the metric is shared, and fit all wins and ties jointly with a regularised Bradley–Terry model (Appendix G). We report one Elo-scale rating per entry, anchored to a reference entry as in TabArena (Erickson et al., 2025), independent of the order in which entries or tasks are compared. Within a task, two entries are compared on each of its frozen partitions, and the task counts as one comparison whose outcome is the share of partitions won. Every task therefore has the same weight, and partition noise does not count as extra evidence. We pin the baseline model at 1000: a rating then reads as the value a representation adds over the raw data. Ratings count wins, not margins: a gap of 100 points means that the higher entry is expected to beat the lower one on about 64% of partitions. The rating stays interpretable as new entries arrive. However, adding an entry or a task moves every number, so ratings from two snapshots are not comparable. The leaderboard fits these ratings on all tasks of a modality, and separately on the tasks of each therapeutic area and of each task category since, as Section 4 shows, the best representation depends on the therapeutic area.

4 RESULTS

4.1 PRIMO V0.1

PRIMO currently contains 63 curated tasks: 52 bulk tasks described in Table 1, and 11 single-cell tasks described in Table 2. The tasks cover four therapeutic areas: immunology and inflammation (25 tasks), oncology (23), infectious diseases (7) and metabolic diseases (4). The four other tasks come from diseases that do not yet form an area: kidney transplantation, chronic obstructive pulmonary disease, fibrotic interstitial lung disease and graft-versus-host disease. They count only in the overall ratings, until enough tasks exist to form their area.

4.2 BENCHMARKING MODELS FROM THE LITERATURE

We evaluated four bulk models (**COMPASS** (Shen et al., 2026), **BulkFormer** (Kang et al., 2026), **BulkRNABert** (Gélard et al., 2025), **MOJO** (Gélard et al., 2026)), five single-cell models (**scGPT** (Cui et al., 2024), **scPRINT** (Kalfon et al., 2025b), **scPRINT-2** (Kalfon et al., 2025a), **Stack-Large**

Table 1: PRIMO bulk task suite. N is the number of samples, or of paired pre- and post-treatment observations (*) for perturbation tasks. The samples of one subject always fall in the same fold. Area: therapeutic area; I&I is immunology and inflammation, and Other marks a disease that does not yet form an area.

Category	Task	Disease	Area	Tissue	N
Treatment outcome (9)	Anti-TNF, endoscopic remission	Inflammatory bowel disease	I&I	Digestive	98
	Adalimumab, endoscopic remission	Inflammatory bowel disease	I&I	Digestive	63
	Infliximab, endoscopic remission	Inflammatory bowel disease	I&I	Digestive	35
	T-DM1 + pertuzumab	HER2-positive breast cancer	Oncology	Breast	242
	Neoadjuvant chemotherapy	Breast cancer	Oncology	Breast	218
	First-line platinum chemotherapy response	High-grade serous ovarian cancer	Oncology	Ovary	88
	Induction chemotherapy response	Acute myeloid leukaemia	Oncology	Haematopoietic	182
	Durable response to BCG	Non-muscle-invasive bladder cancer	Oncology	Bladder	40
	Atezolizumab response	Metastatic urothelial carcinoma	Oncology	Bladder	298
Clinical activity (20)	SES-CD endoscopic severity	Crohn's disease	I&I	Digestive	694
	GHAS-7 histologic activity	Crohn's disease	I&I	Digestive	458
	Harvey–Bradshaw Index	Crohn's disease	I&I	Digestive	694
	SCCAI clinical activity	Ulcerative colitis	I&I	Digestive	249
	Endoscopic Mayo score	Ulcerative colitis	I&I	Digestive	249
	Nancy histologic index	Ulcerative colitis	I&I	Digestive	104
	EASI severity	Atopic dermatitis	I&I	Skin	223
	SCORAD severity	Atopic dermatitis	I&I	Skin	197
	PASI severity	Psoriasis	I&I	Skin	90
	Swollen joint count (SJC28)	Rheumatoid arthritis	I&I	Synovial	162
	Tender joint count (TJC28)	Rheumatoid arthritis	I&I	Synovial	162
	SOFA score at presentation	Sepsis	Infectious	Blood	266
	ISUP grade, diagnostic biopsy	Prostate cancer	Oncology	Prostate	176
	ISUP grade, prostatectomy	Prostate cancer	Oncology	Prostate	135
	Fibrosis stage	Steatohepatitis	Metabolic	Liver	206
	NAFLD activity score	Steatohepatitis	Metabolic	Liver	206
	Fibrosis stage, second cohort	Steatohepatitis	Metabolic	Liver	112
	NAFLD activity score, second cohort	Steatohepatitis	Metabolic	Liver	78
	Days out of hospital	COVID-19	Infectious	Blood	100
	EDSS disability score	Multiple sclerosis	I&I	Blood $CD4^+$ T cells	113
Prognosis (7)	Death after emergency admission	Sepsis	Infectious	Blood	96
	Biochemical relapse, diagnostic biopsy	Prostate cancer	Oncology	Prostate	176
	Biochemical relapse after prostatectomy	Prostate cancer	Oncology	Prostate	104
	Progression in household contacts	Tuberculosis	Infectious	Blood	428
	Progression in infected adolescents	Tuberculosis	Infectious	Blood	248
	Recurrence after surgery	Lung adenocarcinoma	Oncology	Adjacent lung	91
	Relapse after first remission	Acute myeloid leukaemia	Oncology	Haematopoietic	268
Endotype (9)	Blood pathotype	Rheumatoid arthritis	I&I	Blood	49
	Synovial pathotype	Rheumatoid arthritis	I&I	Synovial	251
	NPM1 mutation	Acute myeloid leukaemia	Oncology	Haematopoietic	199
	FLT3-ITD	Acute myeloid leukaemia	Oncology	Haematopoietic	199
	DNMT3A mutation	Acute myeloid leukaemia	Oncology	Haematopoietic	246
	IDH2 mutation	Acute myeloid leukaemia	Oncology	Haematopoietic	246
	CpG island methylator phenotype	T-cell acute lymphoblastic leukaemia	Oncology	Haematopoietic	108
	High tumour mutation burden	Metastatic urothelial carcinoma	Oncology	Bladder	271
	Graft rejection	Kidney transplantation	Other	Kidney	101
Perturbation response (7)	Adalimumab, HS $\rightarrow$ Pso	Inflammatory skin disease	I&I	Skin	32^*
	Adalimumab, Pso $\rightarrow$ HS	Inflammatory skin disease	I&I	Skin	32^*
	Dupilumab, mouse $\rightarrow$ human	Atopic dermatitis	I&I	Skin	55^*
	Inhaled corticosteroids	Chronic obstructive pulmonary disease	Other	Bronchial epithelium	34^*
	Galunisertib	Rectal cancer	Oncology	Rectum	36^*
	Neoadjuvant enzalutamide	Prostate cancer	Oncology	Prostate	39^*
	Neoadjuvant chemotherapy, on-treatment	Breast cancer	Oncology	Breast	23^*

(Dong et al., 2026), **Geneformer-V2** (Theodoris et al., 2023; Chen et al., 2026)), and two sizes of the **EVA** model (Bandasack et al., 2026), which are trained on both bulk and single-cell data.

We generated every entry in Tables 3 and 4. EVA, BulkRNABert, and MOJO outputs were generated in collaboration with the authors. Every model is used zero-shot, with its pretrained weights: only the input preprocessing and the pooling into one vector per sample are chosen for PRIMO. Details for each model are provided in Appendix A.

Reading the ratings. PRIMO scores foundation models in their most demanding setting, frozen and read by a linear probe, whereas in practice they are fine-tuned to each task: the ratings are a lower bound on their value. Every entry, the baseline included, is read by the same probe, so the ratings compare representations, not a foundation model with a linear model: a rating near 1000

Table 2: PRIMO single-cell task suite. N is the number of samples. Areas as in Table 1.

Category	Task	Disease	Area	Tissue	N
Treatment outcome (4)	Adalimumab, remission	Ulcerative colitis	I&I	Digestive	32
	Corticosteroid resistance	Acute graft-versus-host disease	Other	PBMC	37
	Checkpoint-inhibitor response	Melanoma	Oncology	Tumour	18
	Neoadjuvant endocrine therapy response	ER-positive breast cancer	Oncology	Breast	27
Clinical activity (2)	COVID-19 severity	COVID-19	Infectious	PBMC	102
	COVID-19 WHO outcome grade	COVID-19	Infectious	Blood	41
Endotype (5)	Ileal vs ileocolonic endotype	Crohn's disease	I&I	Ileum	20
	Diffuse vs limited cutaneous subtype	Systemic sclerosis	I&I	Skin	56
	Active, potential or treated disease	Coeliac disease	I&I	Duodenum	32
	Fibrotic HP vs IPF	Fibrotic interstitial lung disease	Other	BAL	15
	Sjögren's disease vs sicca	Sjögren's disease	I&I	PBMC	17

means that a representation never trained on the task already carries as much sample-level signal as the raw expression, and a rating above 1000 that it carries more.

Table 3: Bulk RNA-seq results on 52 tasks from 34 cohorts. Ratings are anchored to the raw-expression baseline at 1000. The 95% interval resamples cohorts, taking all tasks of a sampled cohort together (Appendix G). Area columns are ratings fitted on the tasks of one therapeutic area only, as on the public leaderboard; their intervals are in Table 7. The two bulk tasks outside these areas count only in the overall rating. Best per column in bold.

	All tasks		By therapeutic area (number of tasks)			
Entry	Elo	95% CI (by cohort)	Immunology & inflammation (20)	Oncology (21)	Infectious diseases (5)	Metabolic diseases (4)
Raw genes (baseline)	**1000**	–	1000	**1000**	1000	**1000**
EVA 300M	985	[904, 1057]	**1027**	986	**1054**	944
EVA 60M	839	[746, 913]	953	778	951	385
BulkFormer	782	[687, 859]	835	740	820	787
COMPASS	745	[658, 812]	768	726	839	480
MOJO	733	[627, 822]	801	722	710	444
BulkRNABert	638	[501, 744]	710	648	462	404

Table 4: Single-cell RNA-seq results on 11 tasks from 11 cohorts, presented as in Table 3. The two single-cell tasks outside these areas count only in the overall rating. The oncology and infectious-disease areas have two tasks each, so their ratings are only indicative. Best per column in bold.

	All tasks		By therapeutic area (number of tasks)		
Entry	Elo	95% CI (by cohort)	Immunology & inflammation (5)	Oncology (2)	Infectious diseases (2)
EVA 300M	**1048**	[915, 1211]	**1172**	994	944
Stack-Large	1045	[911, 1214]	1046	**1085**	**1522**
Raw genes (baseline)	1000	–	1000	1000	1000
EVA 60M	991	[875, 1110]	1068	854	1064
Geneformer-V2	959	[874, 1043]	1000	1028	902
scPRINT	881	[800, 939]	827	908	858
scGPT	870	[728, 975]	768	1025	998
scPRINT-2	864	[733, 975]	938	829	615

Bulk RNA-seq. EVA 300M is the highest-rated bulk foundation model by a wide margin (Table 3): 203 rating points above the best model outside the EVA family, BulkFormer (782), so it is expected to beat BulkFormer on about three partitions in four. It is also the only bulk model to reach the level of the raw-expression baseline without any adaptation (985, 95% CI 904–1057). Rating increases

with model size, from EVA 60M (839) to EVA 300M, and the other models rate lower, down to BulkRNABert (638). Rankings change with the therapeutic area (Table 7). In immunology and inflammation, a domain for which it was specifically developed, EVA 300M rates above the baseline (1027, CI 871–1158). In oncology, the largest area (21 tasks from 15 cohorts), the baseline leads, and the models pretrained on tumour cohorts (BulkRNABert, MOJO and COMPASS) rate no higher than in immunology and inflammation, so tumour-derived sample-level tasks remain an open problem for current models. The metabolic area, with four tasks from three cohorts, is still too small to rank models. Across task categories (Table 6), EVA 300M and 60M match the baseline on perturbation response (0.339 and 0.338 against 0.340), while the other models remain below it.

Single-cell RNA-seq. Table 4 ranks the eight entries on the 11 single-cell tasks. EVA 300M (1048, CI 915–1211), which embeds one pseudo-bulk profile per sample as in its pretraining, and Stack-Large (1045, CI 911–1214), trained on contextualized single-cell data, rate above the baseline at the same level, with intervals that still include it given 11 cohorts. The two models lead in different areas: EVA 300M in immunology and inflammation (1172, CI 992–1579), and Stack-Large in infectious diseases, on the two COVID-19 tasks (1522). EVA 60M and Geneformer-V2 rate close to the baseline, and scPRINT, scGPT and scPRINT-2 below it.

5 Conclusion

We introduced PRIMO, an open initiative that turns the promise of foundation models for clinical drug development into measurable progress through a shared objective. We defined key questions that arise at several stages of the drug development pipeline and turned them into five task categories: treatment outcome, clinical activity, prognosis, endotype and perturbation response. In this first version, focused on transcriptomics, we curated 63 tasks on 30 diseases, split into 11 single-cell tasks and 52 bulk tasks. We made explicit the preprocessing and evaluation choices that often account for part of the reported performance despite having nothing to do with the model itself.

We benchmarked zero-shot representations of 11 foundation model spanning a range of architectures, trained on bulk, single-cell, or both modalities, against the gene expression space. Foundation models already add value without any adaptation in some settings, and the best model depends on the therapeutic area. Every result is therefore displayed openly on a public leaderboard, broken down by modality, therapeutic area and task category, so that model developers, drug developers, and clinicians can each compare models on the setting that matters to them.

This first version makes deliberate choices, which also set its limits. It evaluates frozen representations with a linear probe, which keeps the evaluation blind and comparable but gives a lower bound on what models reach once fine-tuned, as they are in practice; measuring that gap is the next step. Its inclusion rules keep every task tied to a drug-development decision, at the cost of fewer and smaller public cohorts: intervals are wide, and single-task rankings should be read with care. New cohorts from the community will narrow these intervals, which is why PRIMO is released as an open initiative. We produced the entries of seven of the eleven models ourselves; their authors can now submit their own through the blind protocol. Finally, this version covers only transcriptomics, one facet of a patient's biology, and mainly clinical questions. We are seeking collaborations to extend PRIMO to other omics modalities, to multi-omics tasks, and to new task categories built on patient data, such as target identification. As it grows, PRIMO is meant to give model developers, drug developers and clinicians the same shared reference: one leaderboard to track progress against and optimize for.

AI use statement

We used LLM-based assistants to help with coding, data cleaning, analyses, literature search, feedback on the evaluation design and writing. We did not use them to generate data or labels. We designed the benchmark, made all scientific decisions and checked all AI-assisted work. We take full responsibility for the content of this paper.

Ethics statement

PRIMO only uses public, de-identified data released by the original studies under their own ethics approvals. We redistribute a cohort only when its licence allows it (Section 3), and the released files

carry opaque sample identifiers; labels, subject identifiers and clinical covariates stay in a private registry used only for scoring. PRIMO measures what a representation makes accessible about held-out patients. A high score is not evidence that a model is fit for clinical decisions, which would require prospective validation.

Reproducibility statement

Section 3 and Appendices B and G give the full evaluation protocol: probe, penalty grids, fold construction, metrics and rating fit. Appendix A gives the checkpoint, preprocessing and pooling used for every model, Appendix E the processing of every dataset, and Appendix C every per-task score. The released leaderboard results hold the score of every partition, from which all ratings can be recomputed. The data, the evaluation code and the leaderboard are public, and the labels stay private so that the evaluation remains blind. All numbers in this paper come from one fixed revision of the benchmark repositories. Anonymous code, frozen leaderboard results and the scripts that rebuild Tables 1–4 and Appendix C: `https://anonymous.4open.science/r/primo-benchmark`. Public links will be added after the review.

References

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025.

Richard Ahn, Rashmi Gupta, Kevin Lai, Nitin Chopra, Sarah T Arron, and Wilson Liao. Network analysis of psoriasis reveals biological pathways and roles for coding and long non-coding RNAs. *BMC Genomics*, 17(1):841, 2016. doi: 10.1186/s12864-016-3188-y.

Carmen Argmann, Ruixue Hou, Ryan C Ungaro, Haritz Irizar, Zainab Al-Taie, Ruiqi Huang, Roman Kosoy, Swati Venkat, Won-Min Song, Antonio F Di'Narzo, Bojan Losic, Ke Hao, Lauren Peters, Phillip H Comella, Gabrielle Wei, Ashish Atreja, Milind Mahajan, Alina Iuga, Prerak T Desai, Patrick Branigan, Aleksandar Stojmirovic, Jacqueline Perrigoue, Carrie Brodmerkel, Mark Curran, Joshua R Friedman, Amy Hart, Esi Lamousé-Smith, Jan Wehkamp, Saurabh Mehandru, Eric E Schadt, Bruce E Sands, Marla C Dubinsky, Jean-Frederic Colombel, Andrew Kasarskis, and Mayte Suárez-Fariñas. Biopsy and blood-based molecular biomarker of inflammation in IBD. *Gut*, 72(7): 1271–1287, 2023. doi: 10.1136/gutjnl-2021-326451.

Arjun Baghela, Olga M Pena, Amy H Lee, Beverlie Baquir, Reza Falsafi, Andy An, Susan W Farmer, Andrew Hurlburt, Alvaro Mondragon-Cardona, Juan Diego Rivera, Andrew Baker, Uriel Trahtemberg, Maryam Shojaei, Carlos Eduardo Jimenez-Canizales, Claudia C dos Santos, Benjamin Tang, Hjalmar R Bouma, Gabriela V Cohen Freue, and Robert E W Hancock. Predicting sepsis severity at first clinical presentation: The role of endotypes and mechanistic signatures. *eBioMedicine*, 75: 103776, 2022. doi: 10.1016/j.ebiom.2021.103776.

Arjun V Balar, Matthew D Galsky, Jonathan E Rosenberg, Thomas Powles, Daniel P Petrylak, Joaquim Bellmunt, Yohann Loriot, Andrea Necchi, Jean Hoffman-Censits, Jose Luis Perez-Gracia, Nancy A Dawson, Michiel S van der Heijden, Robert Dreicer, Sandy Srinivas, Margitta M Retz, Richard W Joseph, Alexandra Drakaki, Ulka N Vaishampayan, Srikala S Sridhar, David I Quinn, Ignacio Durán, David R Shaffer, Bernhard J Eigl, Petros D Grivas, Evan Y Yu, Shi Li, Edward E Kadel III, Zachary Boyd, Richard Bourgon, Priti S Hegde, Sanjeev Mariathasan, AnnChristine Thåström, Oyewale O Abidoye, Gregg D Fine, Dean F Bajorin, and IMvigor210 Study Group. Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *The Lancet*, 389(10064): 67–76, 2017. doi: 10.1016/S0140-6736(16)32455-2. Correction: Lancet 2017;390(10097):848 (PMID 28807535).

Stefanos A Bamopoulos, Aarif M N Batcha, Vindi Jurinovic, Maja Rothenberg-Thurley, Hanna Janke, Bianka Ksienzyk, Julia Philippou-Massier, Alexander Graf, Stefan Krebs, Helmut Blum, Stephanie Schneider, Nikola Konstandin, Maria Cristina Sauerland, Dennis Görlich, Wolfgang E Berdel, Bernhard J Woermann, Stefan K Bohlander, Stefan Canzar, Ulrich Mansmann, Wolfgang Hiddemann, Jan Braess, Karsten Spiekermann, Klaus H Metzeler, and Tobias Herold. Clinical

presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with acute myeloid leukemia. *Leukemia*, 34(10):2621–2634, 2020. doi: 10.1038/s41375-020-0839-4.

Ethan Bandasack, Vincent Bouget, Apolline Bruley, Yannis Cattan, Charlotte Claye, Matthew Corney, Julien Duquesne, Karim El Kanbi, Aziz Fouché, Pierre Marschall, and Francesco Strozzi. EVA: Towards a universal model of the immune system. *arXiv preprint arXiv:2602.10168*, 2026.

Andreas Bender, Morgan C Thomas, Jack W Scannell, David A Shaywitz, Gian Marco Ghiandoni, Joe G Greener, Lavinia-Lorena Pruteanu, Rachel DeVay Jacobson, Koichi Handa, Mariko Hirano, et al. Artificial intelligence in drug discovery—what it is, where we stand and the path forward. *Nature Reviews Drug Discovery*, 2026.

Richard J Bownes, Arran K Turnbull, Carlos Martinez-Perez, David A Cameron, Andrew H Sims, and Olga Oikonomidou. On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. *Breast Cancer Research*, 21(1):73, 2019. doi: 10.1186/s13058-019-1159-3.

Nicholas Cardillo, Douglas Russo, Andreea Newtson, Henry Reyes, Yasmin Lyons, Eric Devor, David Bender, Michael J Goodheart, and Jesus Gonzalez-Bosquet. Identification of Novel lncRNAs in Ovarian Cancer and Their Impact on Overall Survival. *International Journal of Molecular Sciences*, 22(3):1079, 2021. doi: 10.3390/ijms22031079.

Han Chen, Madhavan S. Venkatesh, Javier Gómez Ortega, Siddharth V. Mahesh, Tarak N. Nandi, Ravi K. Madduri, Karin Pelka, and Christina V. Theodoris. Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology. *Nature Computational Science*, 6(5):450–463, 2026. doi: 10.1038/s43588-026-00972-4.

Joan W. Chen, Ryan P. Russell, Trushna Desai, Mary Fiel-Gan, Varun Bhat, Maria de Fátima Dias Gaui, Luis Claudio Amendola, Zilton Vasconcelos, Adam M. Brufsky, Marcia V. Fournier, and Susan H. Tannenbaum. RNA Expression Classifiers from a Model of Breast Epithelial Cell Organization to Predict Pathological Complete Response in Triple Negative Breast Cancer. *Archives of Clinical and Biomedical Research*, 5(3), 2021. doi: 10.26502/acbr.50170170. Preprint: medRxiv, doi:10.1101/2021.02.10.21251517.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938.e58, 2022. doi: 10.1016/j.cell.2022.01.012.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Marco Daperno, Geert D'Haens, Gert Van Assche, Filip Baert, Philippe Bulois, Vincent Maunoury, Raffaello Sostegni, Rodolfo Rocca, Angelo Pera, Annemie Gevers, et al. Development and validation of a new, simplified endoscopic activity score for crohn's disease: the ses-cd. *Gastrointestinal endoscopy*, 60(4):505–512, 2004.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Igor Dolgalev, Hua Zhou, Nina Murrell, Hortense Le, Theodore Sakellaropoulos, Nicolas Coudray, Kelsey Zhu, Varshini Vasudevaraja, Anna Yeaton, Chandra Goparaju, Yonghua Li, Imran Sulaiman, Jun-Chieh J Tsay, Peter Meyn, Hussein Mohamed, Iris Sydney, Tomoe Shiomi, Sitharam Ramaswami, Navneet Narula, Ruth Kulicke, Fred P Davis, Nicolas Stransky, Gromoslaw A Smolen, Wei-Yi Cheng, James Cai, Salman Punekar, Vamsidhar Velcheti, Daniel H Sterman, J T Poirier, Ben Neel, Kwok-Kin Wong, Luis Chiriboga, Adriana Heguy, Thales Papagiannakopoulos, Bettina Nadorp, Matija Snuderl, Leopoldo N Segal, Andre L Moreira, Harvey I Pass, and Aristotelis Tsirigos. Inflammation in the tumor-adjacent lung as a predictor of clinical outcome in lung adenocarcinoma. *Nature Communications*, 14(1):6764, 2023. doi: 10.1038/s41467-023-42327-x.

Mingze Dong, Abhinav Adduri, Dhruv Gautam, Christopher Carpenter, Rohan Shah, Chiara Ricci-Tam, Yuval Kluger, Dave P Burke, and Yusuf Husein Roohani. Stack: In-context learning of single-cell biology. biorxiv. 2026.

Haitham Elmarakeby, Ahmed Roman, Shreya Johri, and Eliezer M Van Allen. Empirical evaluation of single-cell foundation models for predicting cancer outcomes. *bioRxiv*, 2025.

Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *American Journal of Surgical Pathology*, 40(2):244–252, 2016. doi: 10.1097/PAS.0000000000000530.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. arXiv:2506.16791.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723.

Maxence Gélard, Guillaume Richard, Thomas Pierrot, and Paul-Henry Cournède. BulkRNABert: Cancer prognosis from bulk RNA-seq based language models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 384–400. PMLR, 15–16 Dec 2025.

Maxence Gélard, Hakim Benkirane, Thomas Pierrot, Guillaume Richard, and Paul-Henry Cournède. Bimodal masked language modeling for bulk RNA-seq and DNA methylation representation learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jesus Gonzalez Bosquet, Eric J Devor, Andreea M Newtson, Brian J Smith, David P Bender, Michael J Goodheart, Megan E McDonald, Terry A Braun, Kristina W Thiel, and Kimberly K Leslie. Creation and validation of models to predict response to primary treatment in serous ovarian cancer. *Scientific Reports*, 11(1):5957, 2021. doi: 10.1038/s41598-021-85256-9.

Olivier Govaere, Simon Cockell, Dina Tiniakos, Rachel Queen, Ramy Younes, Michele Vacca, Leigh Alexander, Federico Ravaioli, Jeremy Palmer, Salvatore Petta, Jerome Boursier, Chiara Rosso, Katherine Johnson, Kristy Wonders, Christopher P Day, Mattias Ekstedt, Matej Orešič, Rebecca Darlay, Heather J Cordell, Fabio Marra, Antonio Vidal-Puig, Pierre Bedossa, Jörn M Schattenberg, Karine Clément, Michael Allison, Elisabetta Bugianesi, Vlad Ratziu, Ann K Daly, and Quentin M Anstee. Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. *Science Translational Medicine*, 12(572):eaba4448, 2020. doi: 10.1126/scitranslmed.aba4448.

Jason I Griffiths, Jinfeng Chen, Patrick A Cosgrove, Anne O'Dea, Priyanka Sharma, Cynthia Ma, Meghna Trivedi, Kevin Kalinsky, Kari B Wisinski, Ruth O'Regan, Issam Makhoul, Laura M Spring, Aditya Bardia, Frederick R Adler, Adam L Cohen, Jeffrey T Chang, Qamar J Khan, and Andrea H Bild. Serial single-cell genomics reveals convergent subclonal evolution of resistance as patients with early-stage breast cancer progress on endocrine plus CDK4/6 therapy. *Nature Cancer*, 2(6):658–671, 2021. doi: 10.1038/s43018-021-00215-7.

Jason I Griffiths, Feng Chi, Elena Farmaki, Eric F Medina, Patrick A Cosgrove, Kimya L Karimi, Jinfeng Chen, Vince K Grolmusz, Frederick R Adler, Qamar J Khan, Aritro Nath, Jeffrey T Chang, and Andrea H Bild. Blocking cancer-fibroblast mutualism inhibits proliferation of endocrine therapy resistant breast cancer. *Molecular Systems Biology*, 21(7):825–855, 2025a. doi: 10.1038/s44320-025-00104-6.

Jason I Griffiths, Patrick A Cosgrove, Eric F Medina, Aritro Nath, Jinfeng Chen, Frederick R Adler, Jeffrey T Chang, Qamar J Khan, and Andrea H Bild. Cellular interactions within the immune microenvironment underpins resistance to cell cycle inhibition in breast cancers. *Nature Communications*, 16(1):2132, 2025b. doi: 10.1038/s41467-025-56279-x.

Rashmi Gupta, Richard Ahn, Kevin Lai, Elizabeth Mullins, Maya Debbaneh, Michelle Dimon, Sarah Arron, and Wilson Liao. Landscape of Long Noncoding RNAs in Psoriatic and Healthy Skin. *Journal of Investigative Dermatology*, 136(3):603–609, 2016. doi: 10.1016/j.jid.2015.12.009.

Chamutal Gur, Shuang-Yin Wang, Fadi Sheban, Mor Zada, Baoguo Li, Fadi Kharouf, Hagit Peleg, Suhail Aamar, Adam Yalin, Daniel Kirschenbaum, Yolanda Braun-Moscovici, Diego Adhemar Jaitin, Tomer Meir-Salame, Efrat Hagai, Bjørt K Kragesteen, Batia Avni, Sigal Grisariu, Chamutal Bornstein, Shir Shlomi-Loubaton, Eyal David, Rony Shreberk-Hassidim, Vered Molho-Pessach, Dalit Amar, Tomer Tzur, Rottem Kuint, Moshe Gross, Oren Barboy, Adi Moshe, Liat Fellus-Alyagor, Dana Hirsch, Yoseph Addadi, Shlomit Erenfeld, Moshe Biton, Tehila Tzemach, Anat Elazary, Yaakov Naparstek, Reut Tzemach, Assaf Weiner, Amir Giladi, Alexandra Balbir-Gurman, and Ido Amit. LGR5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell*, 185(8):1373–1388.e20, 2022. doi: 10.1016/j.cell.2022.03.011.

Roisin Hambly, Solene Gatault, Conor M Smith, Luis F Iglesias-Martinez, Sean Kearns, Helen Rea, Vivien Marasigan, Kate Lynam-Loane, Shivashini Kirthi, Rosalind Hughes, Jean M Fletcher, Walter Kolch, and Brian Kirby. B-cell and complement signature in severe hidradenitis suppurativa that does not respond to adalimumab. *British Journal of Dermatology*, 188(1):52–63, 2023. doi: 10.1093/bjd/ljac007.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024.

Tobias Herold, Vindi Jurinovic, Aarif M N Batcha, Stefanos A Bamopoulos, Maja Rothenberg-Thurley, Bianka Ksienzyk, Luise Hartmann, Philipp A Greif, Julia Phillippou-Massier, Stefan Krebs, Helmut Blum, Susanne Amler, Stephanie Schneider, Nikola Konstandin, Maria Cristina Sauerland, Dennis Görlich, Wolfgang E Berdel, Bernhard J Wörmann, Johanna Tischer, Marion Subklewe, Stefan K Bohlander, Jan Braess, Wolfgang Hiddemann, Klaus H Metzeler, Ulrich Mansmann, and Karsten Spiekermann. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica*, 103(3):456–465, 2018. doi: 10.3324/haematol.2017.178442.

Stephen A Hoang, Abdul Oseini, Ryan E Feaver, Banumathi K Cole, Amon Asgharpour, Robert Vincent, Mohammad Siddiqui, Mark J Lawson, Nathan C Day, Justin M Taylor, Brian R Wamhoff, Faridoddin Mirshahi, Melissa J Contos, Michael Idowu, and Arun J Sanyal. Gene Expression Predicts Histological Severity and Reveals Distinct Molecular Profiles of Nonalcoholic Fatty Liver Disease. *Scientific Reports*, 9(1):12541, 2019. doi: 10.1038/s41598-019-48746-5.

Tu Hu, Tanja Todberg, Daniel Andersen, Niels Banhos Danneskiold-Samsøe, Sofie Boesgaard Neestrup Hansen, Karsten Kristiansen, David Adrian Ewald, Susanne Brix, Joel Correa da Rosa, Ilka Hoof, Lone Skov, and Thomas Litman. Profiling the Atopic Dermatitis Epidermal Transcriptome by Tape Stripping and BRB-seq. *International Journal of Molecular Sciences*, 23 (11):6140, 2022. doi: 10.3390/ijms23116140.

Tu Hu, Tanja Todberg, David Adrian Ewald, Ilka Hoof, Joel Correa da Rosa, Lone Skov, and Thomas Litman. Assessment of Spatial and Temporal Variation in the Skin Transcriptome of Atopic Dermatitis by Use of 1.5 mm Minipunch Biopsies. *Journal of Investigative Dermatology*, 143(4): 612–620.e6, 2023. doi: 10.1016/j.jid.2022.10.004.

Frances Humby, Myles Lewis, Nandhini Ramamoorthi, Jason A Hackney, Michael R Barnes, Michele Bombardieri, A Francesca Setiadi, Stephen Kelly, Fabiola Bene, Maria DiCicco, Sudeh Riahi, Vidalba Rocher, Nora Ng, Ilias Lazarou, Rebecca Hands, Désirée van der Heijde, Robert B M Landewé, Annette van der Helm-van Mil, Alberto Cauli, Iain McInnes, Christopher Dominic Buckley, Ernest H Choy, Peter C Taylor, Michael J Townsend, and Costantino Pitzalis. Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. *Annals of the Rheumatic Diseases*, 78(6):761–772, 2019. doi: 10.1136/annrheumdis-2018-214539.

Frances Humby, Patrick Durez, Maya H Buch, Myles J Lewis, Hasan Rizvi, Felice Rivellese, Alessandra Nerviani, Giovanni Giorli, Arti Mahto, Carlomaurizio Montecucco, Bernard Lauwerys, Nora Ng, Pauline Ho, Michele Bombardieri, Vasco C Romão, Patrick Verschueren, Stephen Kelly, Pier Paolo Sainaghi, Nagui Gendi, Bhaskar Dasgupta, Alberto Cauli, Piero Reynolds, Juan D Cañete, Robert Moots, Peter C Taylor, Christopher J Edwards, John Isaacs, Peter Sasieni, Ernest Choy, Costantino Pitzalis, and R4RA collaborative group. Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16-week outcomes of a stratified, biopsy-driven, multicentre, open-label, phase 4 randomised controlled trial. *The Lancet*, 397(10271):305–317, 2021. doi: 10.1016/S0140-6736(20)32341-2.

Jérémie Kalfon, Gabriel Peyré, and Laura Cantini. scPRINT-2: Towards the next-generation of cell foundation models and benchmarks. *bioRxiv*, 2025a.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16(1):3607, 2025b.

Boming Kang, Rui Fan, Meizheng Yi, Chunmei Cui, and Qinghua Cui. BulkFormer: A large-scale foundation model for bulk transcriptomes. *Cell Systems*, 17(7):101657, 2026. doi: 10.1016/j.cels.2026.101657.

Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Qamar J. Khan, Anne O'Dea, Aditya Bardia, Kevin Kalinsky, Kari Braun Wisinski, Ruth O'Regan, Yuan Yuan, Cynthia X. Ma, Mohammad Jahanzeb, Meghna S. Trivedi, Laura Spring, Issam Makhoul, Jamie L. Wagner, Onalisa Winblad, Amanda Leigh Amin, Sibel Blau, Gregory James Crane, Manana Elia, Mia Hard, and Priyanka Sharma. Letrozole + ribociclib versus letrozole + placebo as neoadjuvant therapy for ER+ breast cancer (FELINE trial). *Journal of Clinical Oncology*, 38(15_suppl):505, 2020. doi: 10.1200/JCO.2020.38.15_suppl.505. ASCO Annual Meeting abstract 505 (FELINE trial, NCT02712723).

Kicheol Kim, Anne-Katrin Pröbstel, Ryan Baumann, Julia Dyckow, James Landefeld, Elva Kogl, Lohith Madireddy, Rita Loudermilk, Erica L Eggers, Sneha Singh, Stacy J Caillier, Stephen L Hauser, Bruce A C Cree, UCSF MS-EPIC Team, Lucas Schirmer, Michael R Wilson, and Sergio E Baranzini. Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis. *Brain*, 144(2):450–461, 2021. doi: 10.1093/brain/awaa421.

Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi, Jung-Il Lee, Yeon-Lim Suh, Bo Mi Ku, Hye Hyeon Eum, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications*, 11(1):2285, 2020.

David E. Kleiner, Elizabeth M. Brunt, Mark Van Natta, Cynthia Behling, Melissa J. Contos, Oscar W. Cummings, Linda D. Ferrell, Yao-Chang Liu, Michael S. Torbenson, Aynur Unalp-Arida, Matthew Yeh, Arthur J. McCullough, and Arun J. Sanyal. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*, 41(6):1313–1321, 2005. doi: 10.1002/hep.20701.

Anastasiya A Kobelyatskaya, Elena A Pudova, Irina V Katunina, Anastasiya V Snezhkina, Maria S Fedorova, Vladislav S Pavlov, Anastasiya O Kotelnikova, Kirill M Nyushko, Boris Y Alekseev, George S Krasnov, and Anna V Kudryavtseva. Transcriptome Profiling of Prostate Cancer, Considering Risk Groups and the TMPRSS2-ERG Molecular Subtype. *International Journal of Molecular Sciences*, 24(11):9282, 2023. doi: 10.3390/ijms24119282.

Adam Kornberg, Theo Botella, Christine S Moon, Samhita Rao, Jared Gelbs, Liang Cheng, Jonathan Miller, Alyssa M Bacarella, Javier A García-Vilas, Justin Vargas, Xuechen Yu, Izabela Krupska, Erin Bush, Reuben Garcia-Carrasquillo, Benjamin Lebwohl, Suneeta Krishnareddy, Suzanne Lewis, Peter H R Green, Govind Bhagat, Kelley S Yan, and Arnold Han. Gluten induces rapid reprogramming of natural memory $\alpha\beta$ and $\gamma\delta$ intraepithelial T cells to induce cytotoxicity in celiac disease. *Science Immunology*, 8(85):eadf4312, 2023. doi: 10.1126/sciimmunol.adf4312.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023. doi: 10.1002/prot.26617.

Raymond J Langley, Ephraim L Tsalik, Jennifer C van Velkinburgh, Seth W Glickman, Brandon J Rice, Chunping Wang, Bo Chen, Lawrence Carin, Arturo Suarez, Robert P Mohney, Debra H Freeman, Mu Wang, Jinsam You, Jacob Wulff, J Will Thompson, M Arthur Moseley, Stephanie Reisinger, Brian T Edmonds, Brian Grinnell, David R Nelson, Darrell L Dinwiddie, Neil A Miller, Carol J Saunders, Sarah S Soden, Angela J Rogers, Lee Gazourian, Laura E Fredenburgh, Anthony F Massaro, Rebecca M Baron, Augustine M K Choi, G Ralph Corey, Geoffrey S Ginsburg, Charles B Cairns, Ronny M Otero, Vance G Fowler, Emanuel P Rivers, Christopher W Woods, and Stephen F Kingsmore. An integrated clinico-metabolomic model improves prediction of death in sepsis. *Science Translational Medicine*, 5(195):195ra95, 2013. doi: 10.1126/scitranslmed.3005893.

Audrey Le Floc'h, Jeanne Allinne, Kirsten Nagashima, George Scott, Dylan Birchard, Seblewongel Asrat, Yu Bai, Wei Keat Lim, Joel Martin, Tammy Huang, Terra B Potocky, Jee H Kim, Ashique Rafique, Nicholas J Papadopoulos, Neil Stahl, George D Yancopoulos, Andrew J Murphy, Matthew A Sleeman, and Jamie M Orengo. Dual blockade of IL-4 and IL-13 with dupilumab, an IL-4Rα antibody, is required to broadly inhibit type 2 inflammation. *Allergy*, 75(5):1188–1204, 2020. doi: 10.1111/all.14151.

Sophie Le Grand, Yannick Marie, Delphine Bouteiller, Margo Fernandez, Gwendolyn Marguerit, Marion Lambert, Émeline Mundwiller, Émilie Robert, Régis Peffault de Latour, Gérard Socié, Nicolas Vallet, and David Michonneau. Corticosteroid resistance is predetermined by early immune response dynamics at acute graft-versus-host disease onset. *Science Translational Medicine*, 18 (844):eaeb7834, 2026. doi: 10.1126/scitranslmed.aeb7834.

Fernando Sergio Leitao Filho, Hiroto Takiguchi, Kentaro Akata, Seung Won Ra, Ji-Yong Moon, Hyun Kuk Kim, Yuji Cho, Kei Yamasaki, Stephen Milne, Julia Yang, Cheng Wei Tony Yang, Xuan Li, Corey Nislow, Stephan F van Eeden, Tawimas Shaipanich, Stephen Lam, Janice M Leung, and Don D Sin. Effects of Inhaled Corticosteroid/Long-Acting β_2-Agonist Combination on the Airway Microbiome of Patients with Chronic Obstructive Pulmonary Disease: A Randomized Controlled Clinical Trial (DISARM). *American Journal of Respiratory and Critical Care Medicine*, 204(10): 1143–1152, 2021. doi: 10.1164/rccm.202102-0289OC.

Myles J Lewis, Michael R Barnes, Kevin Blighe, Katriona Goldmann, Sharmila Rana, Jason A Hackney, Nandhini Ramamoorthi, Christopher R John, David S Watson, Sarah K Kummerfeld, et al. Molecular portraits of early rheumatoid arthritis identify clinical and treatment response phenotypes. *Cell Reports*, 28(9):2455–2470, 2019. doi: 10.1016/j.celrep.2019.07.091.

Bingshan Li, Lam C Tsoi, William R Swindell, Johann E Gudjonsson, Trilokraj Tejasvi, Andrew Johnston, Jun Ding, Philip E Stuart, Xianying Xing, James J Kochkodan, John J Voorhees, Hyun M Kang, Rajan P Nair, Goncalo R Abecasis, and James T Elder. Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms. *Journal of Investigative Dermatology*, 134(7):1828–1838, 2014. doi: 10.1038/jid.2014.28.

Zheqi Li, Otto Metzger Filho, Giuseppe Viale, Patrizia dell'Orto, Leila Russo, Marie-Anne Goyette, Avni Kamat, Denise A Yardley, Vandana Gupta Abramson, Carlos L Arteaga, Laura M Spring, Kami Chiotti, Carol Halsey, Adrienne G Waks, Tari A King, Susan C Lester, Jennifer R Bellon, Eric P Winer, Paul T Spellman, Ian E Krop, and Kornelia Polyak. HER2 heterogeneity and treatment response-associated profiles in HER2-positive breast cancer in the NCT02326974 clinical trial. *Journal of Clinical Investigation*, 134(7):e176454, 2024. doi: 10.1172/JCI176454.

Yuheng Liang, Lucy Chhuo, Ahmadreza Argha, Nona Farbehi, Lu Chen, Roohallah Alizadehsani, Mehdi Hosseinzadeh, Min Yang, Thantrira Porntaveetusm, Youqiong Ye, and Hamid Alinejad-Rokny. Transcriptomic models for immunotherapy response prediction show limited cross-cohort generalisability. *arXiv preprint arXiv:2604.05478*, 2026.

Simon Linder, Marlous Hoogstraat, Suzan Stelloo, Nils Eickhoff, Karianne Schuurman, Hilda de Barros, Maartje Alkemade, Elise M Bekers, Tesa M Severson, Joyce Sanders, Chia-Chi Flora Huang, Tunc Morova, Umut Berkay Altintas, Liesbeth Hoekman, Yongsoo Kim, Sylvan C Baca, Martin Sjöström, Anniek Zaalberg, Dorine C Hintzen, Jeroen de Jong, Roelof J C Kluin, Iris de Rink, Claudia Giambartolomei, Ji-Heui Seo, Bogdan Pasaniuc, Maarten Altelaar, René H Medema, Felix Y Feng, Amina Zoubeidi, Matthew L Freedman, Lodewyk F A Wessels, Lisa M Butler, Nathan A Lack, Henk van der Poel, Andries M Bergman, and Wilbert Zwart. Drug-Induced Epigenomic Plasticity Reprograms Circadian Rhythm Regulation to Drive Prostate Cancer toward Androgen Independence. *Cancer Discovery*, 12(9):2074–2097, 2022. doi: 10.1158/2159-8290.CD-21-0576.

Tianyu Liu, Edward De Brouwer, Archit Verma, Alsu Missarova, Tony Kuo, Nathaniel Diamant, Hanchen Wang, Minsheng Hao, Tommaso Biancalani, Héctor Corrada Bravo, Gabriele Scalia, Aviv Regev, and Graham Heimberg. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. *Cell Systems*, 17(5):101570, 2026. doi: 10.1016/j.cels.2026.101570.

Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Mueller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Malte D. Luecken, Scott Gigante, Daniel B. Burkhardt, Robrecht Cannoodt, Daniel C. Strobl, Nikolay S. Markov, Luke Zappia, Giovanni Palla, Wesley Lewis, Daniel Dimitrov, et al. Defining and benchmarking open problems in single-cell analysis. *Nature Biotechnology*, 43(7):1035–1040, 2025. doi: 10.1038/s41587-025-02694-w.

Sushma Chowdary Maddipatla, Vasantha L Kolachala, Suresh Venkateswaran, Anne F Dodd, Ranjit Singh Pelia, Duke Geem, Hong Yin, Yutong Sun, Congmin Xu, Angela Mo, Astrid Kosters, Junkai Yang, Jason D Matthews, Eliver Ghosn, Subra Kugathasan, and Peng Qiu. Assessing Cellular and Transcriptional Diversity of Ileal Mucosa Among Treatment-Naïve and Treated Crohn's Disease. *Inflammatory Bowel Diseases*, 29(2):274–285, 2023. doi: 10.1093/ibd/izac201.

Sanjeev Mariathasan, Shannon J Turley, Dorothee Nickles, Alessandra Castiglioni, Kobe Yuen, Yulei Wang, Edward E Kadel III, Hartmut Koeppen, Jillian L Astarita, Rafael Cubas, Suchit Jhunjhunwala, Romain Banchereau, Yagai Yang, Yinghui Guan, Cecile Chalouni, James Ziai, Yasin Şenbabaoğlu, Stephen Santoro, Daniel Sheinson, Jeffrey Hung, Jennifer M Giltnane, Andrew A Pierce, Kathryn Mesh, Steve Lianoglou, Johannes Riegler, Richard A D Carano, Pontus Eriksson, Mattias Höglund, Loan Somarriba, Daniel L Halligan, Michiel S van der Heijden, Yohann Loriot, Jonathan E Rosenberg, Lawrence Fong, Ira Mellman, Daniel S Chen, Marjorie Green, Christina Derleth, Gregg D Fine, Priti S Hegde, Richard Bourgon, and Thomas Powles. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature*, 554(7693): 544–548, 2018. doi: 10.1038/nature25501.

Maxwell McDermott, Wenyi Li, Yin-Hu Wang, Allen Y Chen, Rodrigo Lacruz, Bettina Nadorp, and Stefan Feske. Machine learning approach to single cell transcriptomic analysis of Sjogren's disease reveals altered activation states of B and T lymphocytes. *Journal of Autoimmunity*, 154: 103419, 2025. doi: 10.1016/j.jaut.2025.103419.

Otto Metzger Filho, Giuseppe Viale, Shayna Stein, Lorenzo Trippa, Denise A Yardley, Ingrid A Mayer, Vandana G Abramson, Carlos L Arteaga, Laura M Spring, Adrienne G Waks, Eileen Wrabel, Michelle K DeMeo, Aditya Bardia, Patrizia Dell'Orto, Leila Russo, Tari A King, Kornelia Polyak, Franziska Michor, Eric P Winer, and Ian E Krop. Impact of HER2 Heterogeneity on Treatment Response of Early-Stage HER2-Positive Breast Cancer: Phase II Neoadjuvant Clinical Trial of T-DM1 Combined with Pertuzumab. *Cancer Discovery*, 11(10):2474–2487, 2021. doi: 10.1158/2159-8290.CD-20-1557.

Stephen Milne, Xuan Li, Chen Xi Yang, Fernando Sergio Leitao Filho, Ana I Hernández Cordero, Cheng Wei Tony Yang, Tawimas Shaipanich, Stephan F van Eeden, Janice M Leung, Stephen Lam, and Don D Sin. Inhaled corticosteroids downregulate SARS-CoV-2-related genes in COPD: results from a randomised controlled trial. *European Respiratory Journal*, 58(1):2100130, 2021. doi: 10.1183/13993003.00130-2021.

Lena Möbus, Elke Rodriguez, Inken Harder, Dora Stölzl, Nicole Boraczynski, Sascha Gerdes, Andreas Kleinheinz, Susanne Abraham, Annice Heratizadeh, Christiane Handrick, Eva Haufe, Thomas Werfel, Jochen Schmitt, Stephan Weidinger, and TREATgermany study group. Atopic dermatitis displays stable and dynamic skin transcriptome signatures. *Journal of Allergy and Clinical Immunology*, 147(1):213–223, 2021. doi: 10.1016/j.jaci.2020.06.012.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023. doi: 10.18653/v1/2023.eacl-main.148.

Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. The self-assessment trap: can we all be better than average? *Molecular Systems Biology*, 7:537, 2011. doi: 10.1038/msb.2011.70.

Katherine A Overmyer, Evgenia Shishkova, Ian J Miller, Joseph Balnis, Matthew N Bernstein, Trenton M Peters-Clarke, Jesse G Meyer, Qiuwen Quan, Laura K Muehlbauer, Edna A Trujillo, Yuchen He, Amit Chopra, Hau C Chieng, Anupama Tiwari, Marc A Judson, Brett Paulson, Dain R Brademan, Yunyun Zhu, Lia R Serrano, Vanessa Linke, Lisa A Drake, Alejandro P Adam, Bradford S Schwartz, Harold A Singer, Scott Swanson, Deane F Mosher, Ron Stewart, Joshua J Coon, and Ariel Jaitovich. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Systems*, 12(1):23–40.e7, 2021. doi: 10.1016/j.cels.2020.10.003.

Lorena Pantano, George Agyapong, Yang Shen, Zhu Zhuo, Francesc Fernandez-Albert, Werner Rust, Dagmar Knebel, Jon Hill, Carine M Boustany-Kari, Julia F Doerner, Jörg F Rippmann, Raymond T Chung, Shannan J Ho Sui, Eric Simon, and Kathleen E Corey. Molecular characterization and cell type composition deconvolution of fibrosis in NAFLD. *Scientific Reports*, 11(1):18045, 2021. doi: 10.1038/s41598-021-96966-5.

Michael Rade, Markus Kreuz, Angelika Borkowetz, Ulrich Sommer, Conny Blumert, Susanne Füssel, Catharina Bertram, Dennis Löffler, Dominik J Otto, Livia A Wöller, Carolin Schimmelpfennig, Ulrike Köhl, Ann-Cathrin Gottschling, Pia Hönscheid, Gustavo B Baretton, Manfred Wirth, Christian Thomas, Friedemann Horn, and Kristin Reiche. A reliable transcriptomic risk-score applicable to formalin-fixed paraffin-embedded biopsies improves outcome prediction in localized prostate cancer. *Molecular Medicine*, 30(1):19, 2024. doi: 10.1186/s10020-024-00789-9.

Venkatesh Rajamanickam, Noah D Simons, Wesley Rosales, Anton Kravchenko, Tomoko Yamazaki, Brady Bernard, Brian Piening, Enric Domingo, Timothy Maughan, Charlems Alvarez-Jimenez, Thomas Desilvio, Satish Viswanath, Mark Whiteford, Amanda Hayman, David O'Brien, Maria X Kiely, Rehan Ahmad, Michael J Gough, Marka R Crittenden, and Kristina H Young. CMS subtypes correlate with complete response in trial of neoadjuvant Galunisertib plus chemoradiation in rectal cancer. *Translational Oncology*, 66:102690, 2026. doi: 10.1016/j.tranon.2026.102690.

Felice Rivellese, Anna E A Surace, Katriona Goldmann, Elisabetta Sciacca, Cankut Çubuk, Giovanni Giorli, Christopher R John, Alessandra Nerviani, Liliane Fossati-Jimack, Georgina Thorborn, Manzoor Ahmed, Edoardo Prediletto, Sarah E Church, Briana M Hudson, Sarah E Warren, Paul M McKeigue, Frances Humby, Michele Bombardieri, Michael R Barnes, Myles J Lewis, Costantino Pitzalis, and R4RA collaborative group. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nature Medicine*, 28(6):1256–1268, 2022. doi: 10.1038/s41591-022-01789-0.

Harry Robertson, Hani Jieun Kim, Jennifer Li, Nicholas Robertson, Paul Robertson, Elvira Jimenez-Vera, Farhan Ameen, Andy Tran, Katie Trinh, Philip J O'Connell, Jean Y H Yang, Natasha M Rogers, and Ellis Patrick. Decoding the hallmarks of allograft dysfunction with a comprehensive pan-organ transcriptomic atlas. *Nature Medicine*, 30(12):3748–3757, 2024. doi: 10.1038/s41591-024-03030-6.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, et al. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

Jonathan E Rosenberg, Jean Hoffman-Censits, Tom Powles, Michiel S van der Heijden, Arjun V Balar, Andrea Necchi, Nancy Dawson, Peter H O'Donnell, Ani Balmanoukian, Yohann Loriot, Sandy Srinivas, Margitta M Retz, Petros Grivas, Richard W Joseph, Matthew D Galsky, Mark T Fleming, Daniel P Petrylak, Jose Luis Perez-Gracia, Howard A Burris, Daniel Castellano, Christina Canil, Joaquim Bellmunt, Dean Bajorin, Dorothee Nickles, Richard Bourgon, Garrett M Frampton, Na Cui, Sanjeev Mariathasan, Oyewale Abidoye, Gregg D Fine, and Robert Dreicer. Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *The Lancet*, 387(10031):1909–1920, 2016. doi: 10.1016/S0140-6736(16)00561-4.

Moshe Sade-Feldman, Keren Yizhak, Stacey L Bjorgaard, John P Ray, Carl G de Boer, Russell W Jenkins, David J Lieb, Jonathan H Chen, Dennie T Frederick, Michal Barzily-Rokni, Samuel S Freeman, Alexandre Reuben, Paul J Hoover, Alexandra-Chloé Villani, Elena Ivanova, Andrew Portell, Patrick H Lizotte, Amir R Aref, Jean-Pierre Eliane, Marc R Hammond, Hans Vitzthum, Shauna M Blackmon, Bo Li, Vancheswaran Gopalakrishnan, Sangeetha M Reddy, Zachary A Cooper, Cloud P Paweletz, David A Barbie, Anat Stemmer-Rachamimov, Keith T Flaherty, Jennifer A Wargo, Genevieve M Boland, Ryan J Sullivan, Gad Getz, and Nir Hacohen. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell*, 175(4):998–1013.e20, 2018. doi: 10.1016/j.cell.2018.10.038. Erratum: Cell 2019;176(1–2):404 (PMID 30633907).

Stefan Salcher, Gregor Sturm, Lena Horvath, Gerold Untergasser, Christiane Kuempers, et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell*, 40(12):1503–1520.e8, 2022. doi: 10.1016/j.ccell.2022.10.008.

J Alexa Sanders, Connor Frasier, Justin T Matulay, Nury M Steuerwald, Jason Zhu, Claud M Grigg, James T Kearns, Stephen B Riggs, Kris E Gaston, Cory R Brouwer, R Tucker Burks, Aaron L Hartman, David M Foureau, Earle F Burgess, and Peter E Clark. Genomic analysis of response to bacillus Calmette-Guérin (BCG) treatment in high-grade stage 1 bladder cancer patients. *Translational Andrology and Urology*, 10(7):2998–3009, 2021. doi: 10.21037/tau-21-158.

Fernanda Schäfer Hackenhaar, Nina Refhagen, Melanie Hagleitner, Frank van Leeuwen, Hanne Vibeke Marquart, Hans Ole Madsen, Mattias Landfors, Pia Osterman, Kjeld Schmiegelow, Trond Flaegstad, Ólafur Jónsson, Jukka Kanerva, Jonas Abrahamsson, Mats Heyman, Ulrika Norén Nyström, Magnus Hultdin, and Sofie Degerman. CpG island methylator phenotype classification improves risk assessment in pediatric T-cell acute lymphoblastic leukemia. *Blood*, 145(19):2161–2178, 2025. doi: 10.1182/blood.2024026027.

Carolin Schimmelpfennig, Michael Rade, Susanne Füssel, Dennis Löffler, Conny Blumert, Catharina Bertram, Angelika Borkowetz, Dominik J Otto, Sven-Holger Puppel, Pia Hönscheid, Ulrich Sommer, Gustavo B Baretton, Ulrike Köhl, Manfred Wirth, Christian Thomas, Friedemann Horn, Markus Kreuz, and Kristin Reiche. Characterization and evaluation of gene fusions as a measure of genetic instability and disease prognosis in prostate cancer. *BMC Cancer*, 23(1):575, 2023. doi: 10.1186/s12885-023-11019-6.

Wanxiang Shen, Intae Moon, Thinh H. Nguyen, Michelle M. Li, Yepeng Huang, Nitya Nair, Daniel Marbach, and Marinka Zitnik. Generalizable AI predicts immunotherapy outcomes across cancers and treatments. *Nature Medicine*, 32(8):3010–3022, 2026. doi: 10.1038/s41591-026-04502-7.

Vladimir Shitov, Mohammad Moghareh Dehkordi, and Malte D. Luecken. Benchmarking sample representations from single-cell data: Metrics for biologically meaningful embeddings. In *ICLR 2025 Workshop on Learning Meaningful Representations of Life*, 2025. URL `https://openreview.net/forum?id=IoRv5afWtb`.

Julie Sølberg, Stine B Jacobsen, Jeppe D Andersen, Thomas Litman, Nina H Ulrich, Malin G Ahlström, Marie-Louise Kampmann, Niels Morling, Jacob P Thyssen, and Jeanne D Johansen.

The stratum corneum transcriptome in atopic dermatitis can be assessed by tape stripping. *Journal of Dermatological Science*, 101(1):14–21, 2021. doi: 10.1016/j.jdermsci.2020.10.011.

Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, Masahiro Yoshida, Natsuhiko Kumasaka, Katarzyna Kania, Justin Engelbert, Bayanne Olabi, Jarmila Stremenova Spegarova, Nicola K Wilson, Nicole Mende, Laura Jardine, Louis C S Gardner, Issac Goh, Dave Horsfall, Jim McGrath, Simone Webb, Michael W Mather, Rik G H Lindeboom, Emma Dann, Ni Huang, Krzysztof Polanski, Elena Prigmore, Florian Gothe, Jonathan Scott, Rebecca P Payne, Kenneth F Baker, Aidan T Hanrath, Ina C D Schim van der Loeff, Andrew S Barr, Amada Sanchez-Gonzalez, Laura Bergamaschi, Federica Mescia, Josephine L Barnes, Eliz Kilich, Angus de Wilton, Anita Saigal, Aarash Saleh, Sam M Janes, Claire M Smith, Nusayhah Gopee, Caroline Wilson, Paul Coupland, Jonathan M Coxhead, Vladimir Yu Kiselev, Stijn van Dongen, Jaume Bacardit, Hamish W King, Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration, Anthony J Rostron, A John Simpson, Sophie Hambleton, Elisa Laurenti, Paul A Lyons, Kerstin B Meyer, Marko Z Nikolić, Christopher J A Duncan, Kenneth G C Smith, Sarah A Teichmann, Menna R Clatworthy, John C Marioni, Berthold Göttgens, and Muzlifah Haniffa. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27(5):904–916, 2021. doi: 10.1038/s41591-021-01329-2.

Sara Suliman, Ethan G Thompson, Jayne Sutherland, January Weiner, Martin O C Ota, Smitha Shankar, Adam Penn-Nicholson, Bonnie Thiel, Mzwandile Erasmus, Jeroen Maertzdorf, Fergal J Duffy, Philip C Hill, E Jane Hughes, Kim Stanley, Katrina Downing, Michelle L Fisher, Joe Valvo, Shreemanta K Parida, Gian van der Spuy, Gerard Tromp, Ifedayo M O Adetifa, Simon Donkor, Rawleigh Howe, Harriet Mayanja-Kizza, W Henry Boom, Hazel M Dockrell, Tom H M Ottenhoff, Mark Hatherill, Alan Aderem, Willem A Hanekom, Thomas J Scriba, Stefan H E Kaufmann, Daniel E Zak, Gerhard Walzl, and GC6-74 cohort study team, The ACS cohort study team. Four-Gene Pan-African Blood Signature Predicts Progression to Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 197(9):1198–1208, 2018. doi: 10.1164/rccm.201711-2340OC.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Tom Thomas, Matthias Friedrich, Charlotte Rich-Griffin, Mathilde Pohin, Devika Agarwal, Julia Pakpoor, Carl Lee, Ruchi Tandon, Aniko Rendek, Dominik Aschenbrenner, Ashwin Jainarayanan, Alexandru Voda, Jacqueline H Y Siu, Raphael Sanches-Peres, Eloise Nee, Dharshan Sathananthan, Dylan Kotliar, Peter Todd, Maria Kiourlappou, Lisa Gartner, Nicholas Ilott, Fadi Issa, Joanna Hester, Jason Turner, Saba Nayar, Jonas Mackerodt, Fan Zhang, Anna Jonsson, Michael Brenner, Soumya Raychaudhuri, Ruth Kulicke, Danielle Ramsdell, Nicolas Stransky, Ray Pagliarini, Piotr Bielecki, Noah Spies, Brian Marsden, Stephen Taylor, Allon Wagner, Paul Klenerman, Alissa Walsh, Mark Coles, Luke Jostins-Dean, Fiona M Powrie, Andrew Filer, Simon Travis, Holm H Uhlig, Calliope A Dendrou, and Christopher D Buckley. A longitudinal single-cell atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. *Nature Immunology*, 25(11):2152–2165, 2024. doi: 10.1038/s41590-024-01994-8.

Clément Triaille, Gaëlle Tilman, Tatiana Sokolova, Axelle Loriot, Joelle Marchandise, Stéphanie De Montjoye, Adrien Nzeusseu-Toukap, Laurent Méric de Bellefon, Caroline Bouzin, Christine Galant, Patrick Durez, Bernard R Lauwerys, and Nisha Limaye. Disease activity drives transcriptomic heterogeneity in early untreated rheumatoid synovitis. *Annals of the Rheumatic Diseases*, 82(12):1538–1546, 2023. doi: 10.1136/ard-2023-224068.

Ephraim L Tsalik, Raymond J Langley, Darrell L Dinwiddie, Neil A Miller, Byunggil Yoo, Jennifer C van Velkinburgh, Laurie D Smith, Isabella Thiffault, Anja K Jaehne, Ashlee M Valente, Ricardo Henao, Xin Yuan, Seth W Glickman, Brandon J Rice, Micah T McClain, Lawrence Carin, G Ralph Corey, Geoffrey S Ginsburg, Charles B Cairns, Ronny M Otero, Vance G Fowler, Emanuel P Rivers, Christopher W Woods, and Stephen F Kingsmore. An integrated transcriptome and expressed variant analysis of sepsis survival and death. *Genome Medicine*, 6(11):111, 2014. doi: 10.1186/s13073-014-0111-5.

Bram Verstockt, Sare Verstockt, Jonas Dehairs, Vera Ballet, Helene Blevi, Willem-Jan Wollants, Christine Breynaert, Gert Van Assche, Séverine Vermeire, and Marc Ferrante. Low TREM1 expression in whole blood predicts anti-TNF response in inflammatory bowel disease. *eBioMedicine*, 40:733–742, 2019. doi: 10.1016/j.ebiom.2019.01.027.

Bram Verstockt, Sare Verstockt, Marisol Veny, Jonas Dehairs, Kaline Arnauts, Gert Van Assche, Gert De Hertogh, Séverine Vermeire, Azucena Salas, and Marc Ferrante. Expression Levels of 4 Genes in Colon Tissue Might Be Used to Predict Which Patients Will Enter Endoscopic Remission After Vedolizumab Therapy for Inflammatory Bowel Diseases. *Clinical Gastroenterology and Hepatology*, 18(5):1142–1151.e10, 2020. doi: 10.1016/j.cgh.2019.08.030.

Sara Villar, Beñat Ariceta, Xabier Agirre, Aura Daniela Urribarri, Rosa Ayala, David Martínez-Cuadrón, Juan Miguel Bergua, Susana Vives, Lorenzo Algarra, Mar Tormo, Pilar Martínez, Josefina Serrano, Catia Simoes, Pilar Herrera, Maria José Calasanz, Ana Alfonso-Piérola, Bruno Paiva, Joaquín Martínez-López, Jesús F San Miguel, Felipe Prósper, and Pau Montesinos. The transcriptomic landscape of elderly acute myeloid leukemia identifies *B7H3* and *BANP* as a favorable signature in high-risk patients. *Frontiers in Oncology*, 12:1054458, 2022. doi: 10.3389/fonc.2022.1054458.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Jean-Louis Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C. K. Reinhart, Peter M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. doi: 10.1007/BF01709751.

Christopher J Walker, Krzysztof Mrózek, Hatice Gulcin Ozer, Deedra Nicolet, Jessica Kohlschmidt, Dimitrios Papaioannou, Luke K Genutis, Marius Bill, Bayard L Powell, Geoffrey L Uy, Jonathan E Kolitz, Andrew J Carroll, Richard M Stone, Ramiro Garzon, John C Byrd, Ann-Kathrin Eisfeld, Albert de la Chapelle, and Clara D Bloomfield. Gene expression signature predicts relapse in adult patients with cytogenetically normal acute myeloid leukemia. *Blood Advances*, 5(5):1474–1482, 2021. doi: 10.1182/bloodadvances.2020003727.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446.

Yin-Hu Wang, Wenyi Li, Maxwell McDermott, Ga-Yeon Son, George Maiti, Fang Zhou, Anthony Y Tao, Dimitrius Raphael, Andre L Moreira, Boheng Shen, Martin Vaeth, Bettina Nadorp, Shukti Chakravarti, Rodrigo S Lacruz, and Stefan Feske. IFN-γ-producing T_H1 cells and dysfunctional regulatory T cells contribute to the pathogenesis of Sjögren's disease. *Science Translational Medicine*, 16(778):eado4856, 2024. doi: 10.1126/scitranslmed.ado4856.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 66633–66677. PMLR, 2025.

Tomoko Yamazaki, Andrew J Gunderson, Miranda Gilchrist, Mark Whiteford, Maria X Kiely, Amanda Hayman, David O'Brien, Rehan Ahmad, Jeffrey V Manchio, Nathaniel Fox, Kayla McCarty, Michaela Phillips, Evelyn Brosnan, Gina Vaccaro, Rui Li, Miklos Simon, Eric Bernstein, Mary McCormick, Lena Yamasaki, Yaping Wu, Ashley Drokin, Trevor Carnahan, Yy To, William L Redmond, Brian Lee, Jeannie Louie, Eric Hansen, Matthew C Solhjem, Julie Cramer, Walter J Urba, Michael J Gough, Marka R Crittenden, and Kristina H Young. Galunisertib plus neoadjuvant chemoradiotherapy in patients with locally advanced rectal cancer: a single-arm, phase 2 trial. *The Lancet Oncology*, 23(9):1189–1200, 2022. doi: 10.1016/S1470-2045(22)00446-6.

Daniel E Zak, Adam Penn-Nicholson, Thomas J Scriba, Ethan Thompson, Sara Suliman, Lynn M Amon, Hassan Mahomed, Mzwandile Erasmus, Wendy Whatney, Gregory D Hussey, Deborah Abrahams, Fazlin Kafaar, Tony Hawkridge, Suzanne Verver, E Jane Hughes, Martin Ota, Jayne Sutherland, Rawleigh Howe, Hazel M Dockrell, W Henry Boom, Bonnie Thiel, Tom H M Ottenhoff, Harriet Mayanja-Kizza, Amelia C Crampin, Katrina Downing, Mark Hatherill, Joe Valvo, Smitha Shankar, Shreemanta K Parida, Stefan H E Kaufmann, Gerhard Walzl, Alan Aderem, Willem A Hanekom, and ACS and GC6-74 cohort study groups. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*, 387(10035):2312–2322, 2016. doi: 10.1016/S0140-6736(15)01316-1.

Amy Y Zhao, Avraham Unterman, Nebal S Abu Hussein, Prapti Sharma, Fadi Nikola, Jasper Flint, Xiting Yan, Taylor S Adams, Aurelien Justet, Tomokazu S Sumida, Jiayi Zhao, Jonas C Schupp, Micha Sam B Raredon, Farida Ahangari, Giuseppe Deluliis, Yingze Zhang, Ivette Buendia-Roldan, Ayodeji Adegunsoye, Anne I Sperling, Antje Prasse, Changwan Ryu, Erica Herzog, Moises Selman, Annie Pardo, and Naftali Kaminski. Single-Cell Analysis Reveals Novel Immune Perturbations in Fibrotic Hypersensitivity Pneumonitis. *American Journal of Respiratory and Critical Care Medicine*, 210(10):1252–1266, 2024. doi: 10.1164/rccm.202401-0078OC.

A MODELS

The models benchmarked in PRIMO are described in Table 5.

Table 5: Models evaluated on PRIMO, grouped by the board(s) where they are scored. *With author*: the embeddings were produced by, or with, the model's authors. *Params*: as reported by the authors; † counted from the released weights. *Input genes*: genes seen in one forward pass. "Fixed" means a fixed gene panel, with missing genes filled; HVG means highly variable genes. *Sample vector*: how one vector per sample is built. On single-cell data, the sample vector is the mean of the cell vectors, except for EVA, which embeds one pseudo-bulk profile per sample. ‡ These cohorts include IMvigor210, the cohort of the atezolizumab-response task (Table 1).

Model	Emb. dim.	With author	Params	Pretraining data	Input genes	Architecture	Sample vector
Bulk							
BulkRNABert	256	✓	6M	11k TCGA samples (bulk)	17,116 (fixed)	Transformer (4 layers)	mean of gene tokens
MOJO	512	✓	52M	11.8k TCGA samples (bulk RNA + methylation)	17,116 (fixed)	Bimodal transformer (8 layers)	mean of output tokens
COMPASS	44		1.0M	10k TCGA tumours (bulk); fine-tuned on 1,133 ICI patients‡	15,672 (fixed)	Performer (1 layer) + concept bottleneck	44 concept scores
BulkFormer	643		150M	582k human bulk profiles	20,010 (fixed)	GCN (1 layer) + Performer (12 layers)	max over genes
Bulk and single-cell							
EVA 60M	256	✓	60M	545k samples, human + mouse (bulk, microarray, pseudo-bulk)	5,000 HVG	Transformer (24 layers)	CLS token
EVA 300M	768	✓	300M	545k samples, human + mouse (bulk, microarray, pseudo-bulk)	5,000 HVG	Transformer (32 layers)	CLS token
Single-cell							
scGPT	512		51M†	33M cells (CELLxGENE)	1,200 (HVG per dataset)	Transformer (12 layers)	mean of all tokens
Geneformer-V2 (316M)	1,152		316M	104M cells (Genecorpus-104M)	$\leq$4,094 (rank per cell)	BERT (18 layers)	CLS token
Stack-Large	1,600		217M	149M cells (scBaseCount)	15,012 (fixed)	Tabular attention (9 layers, 256 cells)	100 tokens $\times$ 16, flattened
scPRINT (medium v1.5)	256		20M	CELLxGENE (2024 census)	2,000 (top expressed per cell)	Transformer (8 layers)	mean of class tokens
scPRINT-2 (small)	64		20M	350M cells, 16 organisms	4,000 (sampled per cell) + 6 neighbour cells	Transformer (8 + 6 layers)	cell-type token

A.1 COMPASS

Model and code. We used the partially fine-tuned COMPASS checkpoint `finetuner_pft_all.pt` (pretrained on TCGA and fine-tuned on all 1,133 immune checkpoint inhibitor-treated patients; https://www.immuno-compass.com/download/) with the COMPASS code (https://github.com/mims-harvard/COMPASS) to run inference. The 44 concept scores are computed by the latent projector, which this checkpoint fine-tunes on the treatment response of these 1,133 patients. These patients include the IMvigor210 cohort of the atezolizumab-response task (Table 1), so COMPASS's score on this task is not a held-out result.

COMPASS reads a fixed panel of 15,672 protein-coding genes, identified by HGNC symbols that we mapped to NCBI Entrez gene IDs with COMPASS's own `gene_id_map.tsv`. Genes that are absent are zero-filled. Counts are first converted to TPM with NCBI gene lengths, as the authors recommend; genes without a length are dropped. Its internal `Datascaler` then applies $\log_2(x+1)$ followed by the min–max scaler fitted during pretraining. The cancer-type token, which COMPASS requires as the first input column, is set to code 0 (ACC) for every sample. COMPASS has no usable code for non-tumour tissue: `NORMAL`$= -1$ indexes outside the 33-row cancer embedding.

Method to obtain sample-level embeddings. Each sample is encoded by COMPASS's `extract` method. We use the 44 concept scores as the sample representation, which gives a sample embedding of size 44. The concepts are 44 biologically grounded immune concepts representing immune cell states, tumor–microenvironment interaction and signaling pathways.

A.2 BULKFORMER

Model and code. We used the original BulkFormer release (Zenodo 10.5281/zenodo.15559368). We followed the preprocessing and encoding steps described in the notebook of the official repository (https://github.com/KangBoming/BulkFormer).

Raw counts are converted to $\log(\mathrm{TPM}+1)$, which is BulkFormer's reference input. The counts are then mapped to BulkFormer's fixed vocabulary of 20,010 protein-coding genes, matched on NCBI gene IDs. Missing genes are filled with -10, BulkFormer's "unmeasured" placeholder.

Method to obtain sample-level embeddings. Each sample is embedded in one forward pass. The per-gene representations are read from the last GBFormer block. Sample-level embeddings are obtained by max pooling over all 20,010 gene positions, including the -10-filled ones, as in the BulkFormer preprint; the example notebook of the repository uses the mean instead. Each position has 643 values: the 640-dimensional gene representation and three sample-level statistics that the model appends (the share of missing genes, the mean expression and the share of non-zero genes). This yields a sample embedding of size $d = 643$.

A.3 MOJO

This model was benchmarked in collaboration with the authors.

Model. MOJO (Gélard et al., 2026) is a bimodal encoder for bulk RNA-seq and DNA methylation, pretrained using masked language modelling on TCGA pan-cancer samples. It consists of a convolutional stem that downsamples the input sequence, followed by a stack of eight transformer layers with embedding dimension 512 and 16 attention heads.

Gene panel. Because MOJO was trained on *matched* RNA-seq and methylation samples, its expression channel is restricted to genes available for both modalities in TCGA. This results in a panel of **17,116 genes**. This panel is used as the reference for both models evaluated here: it is inherited from the bimodal setting, and the RNA-seq-only model described below was pretrained on exactly the same gene set. The two encoders therefore receive identical inputs.

Preprocessing and tokenisation. PRIMO expression profiles were restricted to the 17,116-gene panel by matching uppercased gene symbols; genes absent from a profile were zero-filled. Expression values were transformed as $\log_{10}(1+\mathrm{TPM})$, divided by the normalisation factor used during

pretraining (5.5279), and discretised into the checkpoint's original expression bins. This reproduces the pretraining input pipeline exactly. Each sample is represented as a sequence of 17,116 gene tokens, padded to 17,152 tokens.

Handling the missing modality. As this version of PRIMO only contains RNA-seq data, DNA methylation is unavailable at inference time. We therefore used the **MOJO-Extended-Pretraining** variant, which was explicitly pretrained to accommodate missing modalities. Following its training-time convention, the methylation sequence was filled with <MASK> tokens over its entire length, while the RNA-seq branch received the observed expression tokens.

Sample-level embedding. We obtained a sample-level representation by average pooling the output of the transformer tower over the downsampled sequence axis, yielding a fixed-size embedding.

A.4 BULKRNABERT

This model was benchmarked in collaboration with the authors.

Model. BulkRNABert (Gélard et al., 2025) is the RNA-seq-only counterpart to MOJO, pretrained using masked language modelling on TCGA pan-cancer bulk RNA-seq data. To avoid confounding the comparison with MOJO through differences in gene coverage, the authors retrained an equivalent model on the same **17,116-gene** panel. The model comprises four transformer layers, with embedding dimension 256, eight attention heads, and 64 expression bins. The panel is not stored with the checkpoint; instead, gene order was recovered from the column order of the training matrix used for the run, which defines the model's input order.

Preprocessing and tokenisation. Preprocessing was identical to that used for MOJO. Because the two checkpoints share the same gene panel, normalisation factor, and 64 expression bins, each sample was represented by the same tokenised expression sequence.

Sample-level embedding. Each tokenised sample was processed by the encoder using full self-attention over the complete gene panel. The final-layer token embeddings were average-pooled over the gene axis, yielding a sample embedding of dimension $d = 256$.

A.5 SCGPT

Model and code. We used the scGPT whole-human checkpoint (Cui et al., 2024; 33M cells; 12 layers, 8 heads, $d = 512$, 51 expression bins). Genes are matched to the vocabulary on NCBI gene IDs. For each dataset, we compute the mean and variance of $\log(1 + x)$ across all cells and keep the 1,200 in-vocabulary genes with the highest variance-to-mean ratio, among genes with non-zero mean. Each cell's raw counts are quantile-binned over all of its non-zero genes. This gives 50 bins, with zero kept as bin 0. The bins are computed before the gene subset is taken, so the subset does not move the bin boundaries.

Method to obtain sample-level embeddings. Each cell is encoded as a <cls> token followed by its 1,200 selected genes, zero-expressed ones included. A cell embedding of size $d = 512$ is the average of the final-layer outputs over all non-padding positions, including <cls>. Sample-level embeddings are the average over all cells of the sample. This yields a sample embedding of size $d = 512$.

A.6 SCPRINT

Both scPRINT models use Ensembl IDs instead of NCBI Gene IDs. We first mapped gene IDs to Ensembl IDs. We could map from 63.7% to 89.4% of the genes, depending on the dataset.

scPRINT. We used the *medium-v1.5* checkpoint from the official Hugging Face repository. Following the methods described by the authors, each cell is represented by its 2,000 most-expressed genes, given as raw counts. Cells with fewer than 2,000 detected genes were padded to that length with

zero-expression genes drawn at random from the model's vocabulary. We took the cell embedding as the model's default *cell_emb*, which averages all nine class tokens. This cell embedding has size $d = 256$.

scPRINT-2. We used the *small-v2* checkpoint from the official Hugging Face repository and used the authors' own Preprocessor and Embedder, with `how="random expr"` and `max_len=4000`. We built the kNN graph separately within each sample, so that no cell's embedding contains expression from another sample.

Method to obtain sample-level embeddings. A sample's embedding is the arithmetic mean of its cells' embeddings.

A.7 STACK

Model and code. We used the Stack-Large checkpoint from the Hugging Face repository (https://huggingface.co/arcinstitute/Stack-Large, commit 6e1f352) and the Stack code on GitHub (https://github.com/ArcInstitute/stack, release v0.1.3) to run inference.

Raw transcript counts are zero-filled onto Stack's fixed gene vocabulary (15,012 genes; between 74.5% and 95.7% are found, depending on the dataset). Matching is on uppercased gene symbols, a duplicated symbol resolves to its last column.

Method to obtain sample-level embeddings. Let N be the number of cells in one sample. Let K be Stack's context size, the number of cells in one input sequence ($K = 256$ for Stack-Large). If $N < K$, cells are repeated (cycled) to fill the context. If $N \geq K$, a sample is divided into $\lfloor \frac{N+K-1}{K} \rfloor$ sequences of K cells, where the last sequence takes the last K cells, possibly repeating cells from the previous sequence.

Each sequence is embedded through the same forward pass as Stack's get_latent_representation (log1p, _reduce_and_tokenize, _run_attention_layers), reimplemented so that context composition stays under our control and verified to match the reference implementation numerically.

Each cell is represented by a set of n_{hidden} tokens of size $token_dim$. The contextualized token embeddings are flattened to obtain one embedding per cell of size $d = n_{hidden} \times token_dim$ ($d = 1600$ for Stack-Large, with $n_{hidden} = 100$ and $token_dim = 16$).

Sample-level embeddings are obtained by an average pooling over all cells' embeddings, computed without double-counting cells that fall in overlapping or padded context windows. This yields a sample embedding of size d.

A.8 GENEFORMER-V2

Model and code. We used the Geneformer-V2-316M checkpoint from the Hugging Face repository (https://huggingface.co/ctheodoris/Geneformer, commit fcd26c4). We used its `gc104M` token and gene-median dictionaries. Tokenisation is reimplemented following Geneformer's `TranscriptomeTokenizer`.

Dataset genes (NCBI gene IDs) are mapped to Ensembl IDs and then to Geneformer vocabulary (20,271 genes). Out-of-vocabulary genes are kept for the library size. Each cell's raw counts are divided by the cell's total count over all genes and multiplied by 10^4. They are then divided by each gene's non-zero median across Genecorpus-104M. Non-zero genes are ranked by decreasing normalised expression, truncated to 4,094 genes, and wrapped with `<cls>` and `<eos>`.

Method to obtain sample-level embeddings. Hidden states are read from the penultimate layer, as in Geneformer's `EmbExtractor` default `emb_layer`$= -1$. A cell embedding is the output at the `<cls>` position, as in `EmbExtractor`'s default `emb_mode="cls"` for V2 models. Sample-level embeddings are the average over all cells of the sample. This yields a sample embedding of size $d = 1152$.

A.9 EVA

This model was benchmarked in collaboration with the authors.

Model and code. The 60M model corresponds to the checkpoint from the Hugging Face repository (https://huggingface.co/ScientaLab/eva-rna, commit 2407734) and the 300M model corresponds to a private checkpoint. Genes are matched to the vocabulary on NCBI Gene IDs, and out-of-vocabulary genes are dropped.

Bulk inference. Counts are converted to FPKM, and values below 1 are set to 0, as in pretraining. Each sample is then rescaled to a total of 10^6 (TPM) and transformed as $\log_2(1 + x)$, over all genes. The 5,000 genes with the highest variance across the samples of the dataset are then kept as input, selected without labels. The sample embedding is the final-layer output at the `<cls>` position, of size $d = 256$ for EVA 60M and $d = 768$ for EVA 300M.

Single-cell inference. EVA was pretrained on pseudo-bulk profiles, not on single cells: in its pretraining corpus, each profile sums the raw counts of all cells of one donor and tissue. We do the same for each sample: the raw counts of all its cells are summed into one pseudo-bulk profile, normalised to $\log_2(\mathrm{CPM} + 1)$. As for bulk, the 5,000 genes with the highest variance across the pseudo-bulk profiles of the dataset are then kept as input, selected without labels. The sample embedding is the `<cls>` output.

B Evaluation details

Centred Spearman for perturbation tasks. Most of the post-minus-pre change is shared by every treated subject, because it is the average drug effect. A plain correlation would reward an entry that only reproduces this common response, which needs no patient-specific information. We therefore remove the shared part before correlating. Let $\delta_i \in \mathbb{R}^G$ be the true response of test pair i over the $G = 600$ response genes selected on training rows, and $\hat{\delta}_i$ the probe's prediction. The task score is

$$S \;=\; \frac{1}{n}\sum_{i=1}^{n} \rho\Big(\delta_i - \bar{\delta},\;\; \hat{\delta}_i - \bar{\delta}\Big)\,,$$

where ρ is the Spearman correlation across the G genes and $\bar{\delta}$ is a reference response subtracted from both sides. The score stays in $[-1, 1]$. Zero means the entry adds nothing to the shared response; a positive value means it ranks a subject's deviation from that response better than chance. Removing the common effect is what keeps these values small, so they should not be compared with uncentred correlations.

Two reference responses are used. Within-cohort and same-species transfer tasks subtract the mean response over the training rows, so an entry that predicts nothing but the average training response scores exactly zero. The mouse-to-human task subtracts the per-gene mean over the test rows instead, because a mouse training mean is not on the human scale; there any constant prediction scores exactly zero. Both variants keep the same scale and the same zero point and change only how hard the task is. A pair left constant on either side by the centring scores zero.

C Per-task results

Table 6 gives the mean score of each task category, and Table 7 the interval of each area rating. Table 8 reports every bulk task individually, grouped by category and labelled with that category's metric. Metrics are not comparable across task categories, so the ratings compare entries within each task only. The category means can be recomputed from Tables 8 and 9 alone. The ratings also need the score of every partition, released with the leaderboard results, and the area of each task, given in Tables 1 and 2. The cohort-bootstrap intervals also need the map from task to cohort, released with the code.

Table 9 reports the 11 single-cell tasks in the same way, for the entries of Table 4.

Table 6: Mean task score of each task category, in its native metric (C. Spearman is centred Spearman). Best per column in bold, within each modality. The single-cell tasks include no prognosis or perturbation task.

Entry	Treatment outcome (AUROC)	Clinical activity (Pearson)	Prognosis (AUROC)	Endotype (AUROC)	Perturbation response (C. Spearman)
Bulk RNA-seq					
Tasks	9	20	7	9	7
Raw genes (baseline)	0.719	**0.483**	0.714	**0.792**	**0.340**
EVA 300M	**0.727**	0.479	**0.725**	0.772	0.339
EVA 60M	0.660	0.452	0.702	0.730	0.338
BulkFormer	0.696	0.438	0.702	0.704	0.288
COMPASS	0.672	0.415	0.689	0.716	0.291
MOJO	0.659	0.420	0.684	0.692	0.318
BulkRNABert	0.657	0.405	0.656	0.687	0.172
Single-cell RNA-seq					
Tasks	4	2	0	5	0
EVA 300M	0.638	0.337	–	**0.903**	–
Stack-Large	0.661	**0.497**	–	0.819	–
Raw genes (baseline)	0.682	0.346	–	0.814	–
EVA 60M	0.598	0.353	–	0.859	–
Geneformer-V2	0.651	0.299	–	0.808	–
scPRINT	0.638	0.305	–	0.749	–
scGPT	**0.686**	0.345	–	0.683	–
scPRINT-2	0.666	0.150	–	0.737	–

Table 7: Ratings by therapeutic area (Tables 3 and 4) with their 95% cohort-bootstrap intervals (5,000 draws). Bold: the interval excludes 1000, the baseline. Areas with fewer than five cohorts get no interval, since a bootstrap over so few cohorts is not informative.

Entry	Immunology & inflammation	Oncology	Infectious diseases	Metabolic diseases
Bulk RNA-seq				
Tasks, cohorts	20 tasks, 9 cohorts	21 tasks, 15 cohorts	5 tasks, 5 cohorts	4 tasks, 3 cohorts
EVA 300M	1027 [871, 1158]	986 [874, 1075]	1054 [907, 1215]	944
EVA 60M	953 [856, 1033]	**778 [627, 901]**	951 [859, 1035]	385
BulkFormer	**835 [679, 947]**	**740 [554, 885]**	**820 [752, 862]**	787
COMPASS	**768 [616, 861]**	**726 [591, 821]**	839 [509, 1132]	480
MOJO	**801 [597, 946]**	**722 [559, 849]**	**710 [482, 866]**	444
BulkRNABert	**710 [425, 907]**	**648 [462, 789]**	**462 [143, 624]**	404
Single-cell RNA-seq				
Tasks, cohorts	5 tasks, 5 cohorts	2 tasks, 2 cohorts	2 tasks, 2 cohorts	–
EVA 300M	1172 [992, 1579]	994	944	–
Stack-Large	1046 [891, 1217]	1085	1522	–
EVA 60M	1068 [949, 1270]	854	1064	–
Geneformer-V2	1000 [887, 1139]	1028	902	–
scPRINT	**827 [627, 940]**	908	858	–
scGPT	**768 [481, 942]**	1025	998	–
scPRINT-2	938 [697, 1126]	829	615	–

Table 8: Per-task scores on the 52 bulk RNA-seq tasks. Bold marks the best score; underlining marks the next distinct score. COMPASS was fine-tuned on the cohort of the atezolizumab-response task (Appendix A), so its score on that task is not held out.

Task	Raw genes	BulkFormer	BulkRNABert	COMPASS	EVA 60M	EVA 300M	MOJO
Treatment outcome (AUROC)							
Anti-TNF, endoscopic remission	0.808	0.806	0.766	0.688	0.744	**0.843**	0.683
Adalimumab, endoscopic remission	0.757	0.815	0.795	0.721	0.744	**0.843**	0.706
Infliximab, endoscopic remission	0.775	0.781	0.691	0.697	0.718	**0.819**	0.666
T-DM1 + pertuzumab	0.597	0.570	0.571	**0.613**	0.538	0.498	0.555
Neoadjuvant chemotherapy	**0.678**	0.648	0.639	0.674	0.654	0.672	0.661
First-line platinum chemotherapy response	0.678	0.677	0.500	0.643	0.582	**0.684**	0.642
Induction chemotherapy response	**0.712**	0.577	0.633	0.677	0.571	0.671	0.615
Durable response to BCG	0.777	0.758	0.671	0.588	0.722	**0.810**	0.739
Atezolizumab response	0.693	0.635	0.647	**0.751**	0.666	0.704	0.666
Clinical activity (Pearson)							
SES-CD endoscopic severity	0.469	**0.475**	0.472	0.461	0.452	0.464	0.455
GHAS-7 histologic activity	**0.638**	0.601	0.587	0.627	0.637	0.626	0.626
Harvey–Bradshaw Index	**0.321**	0.160	0.145	0.162	0.261	0.277	0.143
SCCAI clinical activity	**0.550**	0.464	0.417	0.476	0.526	0.527	0.486
Endoscopic Mayo score	0.715	0.679	0.630	0.700	**0.716**	0.704	0.683
Nancy histologic index	**0.754**	0.730	0.695	0.697	0.740	0.737	0.698
EASI severity	0.292	0.280	0.308	0.302	**0.316**	0.270	0.304
SCORAD severity	0.129	0.108	**0.276**	0.072	0.180	0.220	0.191
PASI severity	**0.435**	0.295	0.193	0.327	0.356	0.391	0.270
Swollen joint count (SJC28)	0.259	0.295	0.202	0.264	0.254	0.293	**0.298**
Tender joint count (TJC28)	0.249	0.228	0.222	0.238	**0.265**	**0.265**	0.256
SOFA score at presentation	0.328	0.296	0.173	0.207	**0.340**	0.326	0.281
ISUP grade, diagnostic biopsy	**0.471**	0.401	0.363	0.370	0.463	0.465	0.396
ISUP grade, prostatectomy	0.313	0.292	**0.447**	0.181	0.275	0.335	0.275
Fibrosis stage	**0.713**	0.655	0.586	0.626	0.586	0.684	0.613
NAFLD activity score	0.578	0.562	0.503	0.451	0.507	**0.593**	0.476
Fibrosis stage, second cohort	**0.713**	0.672	0.651	0.647	0.640	0.696	0.657
NAFLD activity score, second cohort	**0.735**	0.665	0.562	0.625	0.558	0.659	0.566
Days out of hospital	0.733	0.710	0.635	0.669	0.711	**0.744**	0.657
EDSS disability score	0.274	0.191	0.034	0.189	0.252	**0.308**	0.076
Prognosis (AUROC)							
Death after emergency admission	0.759	0.734	0.700	0.735	0.743	**0.794**	0.666
Biochemical relapse, diagnostic biopsy	0.686	0.681	0.673	0.645	0.627	**0.704**	0.653
Biochemical relapse after prostatectomy	0.735	0.718	0.716	0.736	0.772	**0.778**	0.765
Progression in household contacts	0.653	0.636	0.614	**0.752**	0.649	0.665	0.649
Progression in infected adolescents	**0.737**	0.699	0.590	0.697	0.720	0.708	0.689
Recurrence after surgery	0.680	**0.760**	0.689	0.615	0.707	0.710	0.706
Relapse after first remission	**0.750**	0.688	0.612	0.645	0.693	0.719	0.656
Endotype (AUROC)							
Blood pathotype	0.647	0.596	0.632	0.541	0.635	**0.653**	0.564
Synovial pathotype	0.844	0.809	0.789	0.825	0.844	**0.850**	0.838
NPM1 mutation	**0.870**	0.756	0.662	0.787	0.812	0.865	0.715
FLT3-ITD	**0.811**	0.772	0.713	0.715	0.762	0.801	0.710
DNMT3A mutation	**0.780**	0.559	0.577	0.660	0.643	0.723	0.591
IDH2 mutation	**0.701**	0.515	0.610	0.581	0.530	0.640	0.553
CpG island methylator phenotype	**0.925**	0.888	0.764	0.804	0.832	0.907	0.764
High tumour mutation burden	0.676	0.639	0.644	0.663	0.658	**0.680**	0.661
Graft rejection	**0.870**	0.800	0.792	0.867	0.857	0.831	0.830
Perturbation response (C. Spearman)							
Adalimumab, HS → Pso	0.354	0.354	0.290	0.311	0.353	**0.365**	0.287
Adalimumab, Pso → HS	0.621	0.433	-0.117	0.661	0.679	0.661	**0.685**
Dupilumab, mouse → human	**0.091**	0.082	0.025	-0.015	0.080	0.069	0.073
Inhaled corticosteroids	**0.387**	0.364	0.319	0.385	0.365	0.349	0.363
Galunisertib	0.353	0.343	0.346	0.321	**0.385**	0.347	0.349
Neoadjuvant enzalutamide	0.207	0.138	0.143	0.130	0.212	**0.220**	0.147
Neoadjuvant chemotherapy, on-treatment	**0.366**	0.302	0.195	0.242	0.295	0.362	0.321

Table 9: Per-task scores on the 11 single-cell RNA-seq tasks. Bold marks the best score; underlining marks the next distinct score. The tasks and their sizes are in Table 2.

Task	Raw genes	EVA 60M	EVA 300M	Geneformer-V2	scGPT	scPRINT	scPRINT-2	Stack-Large
Treatment outcome (AUROC)								
Adalimumab, remission	0.557	0.492	0.526	0.469	0.563	0.528	**0.679**	0.570
Corticosteroid resistance	0.554	0.497	0.516	0.529	0.534	0.540	**0.567**	0.478
Checkpoint-inhibitor response	0.883	0.895	0.956	0.886	0.844	0.860	0.868	**0.968**
Neoadjuvant endocrine therapy response	0.735	0.506	0.553	0.721	**0.804**	0.623	0.550	0.628
Clinical activity (Pearson)								
COVID-19 severity	0.453	0.536	0.495	0.488	0.505	0.393	0.340	**0.633**
COVID-19 WHO outcome grade	0.238	0.170	0.179	0.109	0.184	0.217	-0.041	**0.361**
Endotype (AUROC)								
Ileal vs ileocolonic endotype	0.678	0.769	**0.837**	0.644	0.452	0.573	0.589	0.570
Diffuse vs limited cutaneous subtype	**0.838**	0.824	0.825	0.831	0.707	0.716	0.835	0.826
Active, potential or treated disease	0.739	0.821	**0.903**	0.803	0.680	0.705	0.652	0.847
Fibrotic HP vs IPF	0.978	0.978	**1.000**	0.897	0.856	0.967	0.869	0.933
Sjögren's disease vs sicca	0.835	0.901	**0.948**	0.866	0.719	0.784	0.739	0.921

D SOURCE STUDIES

The benchmark tasks are built from the 52 public studies listed in Table 10. A dataset can pool several studies, and datasets that share subjects form one cohort: these 52 studies give the 48 datasets and 45 cohorts of PRIMO. To keep the evaluation blind, rows are deliberately written in random order. For each study the table gives the archive records the data were taken from, the publication(s) that generated them, and related trial or cohort publications.

Table 10: Source studies of the PRIMO datasets, in random order.

#	Data records	Publication(s)	Related publications
1	GSE54456	Li et al. (2014)	–
2	E-MTAB-11611	Rivellese et al. (2022)	Humby et al. (2021)
3	GSE225731	Triaille et al. (2023)	–
4	GSE199046	Hu et al. (2022)	–
5	GSE79362	Zak et al. (2016)	–
6	CZ CELLxGENE collection 8f126edf-5405-4731-8374-b5ce11f53e82, Zenodo record 6120249	COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium (2022)	CZI Cell Science Program et al. (2025)
7	GSE195452	Gur et al. (2022)	–
8	GSE185263	Baghela et al. (2022)	–
9	GSE213761	Hambly et al. (2023)	–
10	GSE202052	Maddipatla et al. (2023)	–
11	GSE318268, NCT02688712	Rajamanickam et al. (2026)	Yamazaki et al. (2022)
12	GSE137143	Kim et al. (2021)	–
13	GSE74697	Gupta et al. (2016)	Ahn et al. (2016)
14	GSE229904	Kobelyatskaya et al. (2023)	–
15	GSE261892	Robertson et al. (2024)	–
16	GSE282122, Zenodo record 14007626	Thomas et al. (2024)	–
17	IMvigor210CoreBiologies R package, NCT02108652, NCT02951767	Mariathasan et al. (2018)	Rosenberg et al. (2016); Balar et al. (2017)
18	GSE94438	Suliman et al. (2018)	–
19	GSE208218	Villar et al. (2022)	–
20	GSE140900	Le Floc'h et al. (2020)	–
21	GSE162120, NCT02833480	Milne et al. (2021)	Leitao Filho et al. (2021)
22	GSE229705	Dolgalev et al. (2023)	–
23	GSE63042, NCT00258869	Tsalik et al. (2014)	Langley et al. (2013)
24	GSE243375, NCT02326974	Li et al. (2024)	Metzger Filho et al. (2021)
25	E-MTAB-6141	Lewis et al. (2019)	Humby et al. (2019)
26	GSE122630	Bownes et al. (2019)	–
27	GSE197780, NCT03297385	Linder et al. (2022)	–
28	E-MTAB-7914	none (deposit by B. Verstockt, KU Leuven)	Verstockt et al. (2019; 2020)

#	Data records	Publication(s)	Related publications
29	GSE253568	Wang et al. (2024); McDermott et al. (2025)	–
30	GSE165430	Walker et al. (2021)	–
31	GSE160501	Sølberg et al. (2021)	–
32	GSE156699	Gonzalez Bosquet et al. (2021)	Cardillo et al. (2021)
33	GSE193677	Argmann et al. (2023)	–
34	GSE176178	Sanders et al. (2021)	–
35	GSE252762	Kornberg et al. (2023)	–
36	GSE220095	Rade et al. (2024); Schimmelpfennig et al. (2023)	–
37	GSE229733, GSE306441, GitLab SophieLG/gvh_transcriptomic	Le Grand et al. (2026)	–
38	GSE271789	Zhao et al. (2024)	–
39	GSE162694	Pantano et al. (2021)	–
40	GSE193309	Hu et al. (2023)	–
41	CZ CELLxGENE collection ddfad306-714d-4cc0-9985-d9072820c530, E-MTAB-10026	Stephenson et al. (2021)	CZI Cell Science Program et al. (2025)
42	GSE130970	Hoang et al. (2019)	–
43	GSE120575	Sade-Feldman et al. (2018)	–
44	GSE163882	Chen et al. (2021)	–
45	GSE146173	Bamopoulos et al. (2020)	–
46	E-MTAB-7604	Verstockt et al. (2019)	–
47	GSE272023	Schäfer Hackenhaar et al. (2025)	–
48	GSE157103	Overmyer et al. (2021)	–
49	GSE211434, NCT02712723	Griffiths et al. (2025b;a)	Griffiths et al. (2021); Khan et al. (2020)
50	GSE106291	Herold et al. (2018)	–
51	GSE135251	Govaere et al. (2020)	–
52	GSE157194	Möbus et al. (2021)	–

E Data processing details

E.1 Bulk RNA-seq cohorts

PRIMO has 37 bulk datasets. Each is rebuilt from the public deposit of its source study (GEO, ArrayExpress or the authors' own release) with a common builder.

Input matrix. We start from raw counts: those deposited by the authors, or those NCBI recomputes from the deposited reads. When a cohort pools several series, their matrices are concatenated and only genes measured in every series are kept. Three datasets are only public on another scale and are used as released: d033 (non-integer counts), d035 (author-normalised counts) and d036 (author log2-normalised values, converted back to a linear scale as $2^x - 1$).

Gene identifiers. As for single-cell data, genes are mapped to NCBI Gene IDs, and counts of genes that map to the same ID are summed.

Samples and labels. The prediction unit is the sample, one biopsy or blood draw, linked to its subject; all folds are grouped by subject. For each task, samples without a label are removed, as are values that are not a class; in classification, a class with a single sample is removed. A dataset can carry several tasks, and its public file holds the union of their samples. For perturbation tasks, the public file holds pre-treatment expression only; the post-minus-pre responses stay private.

Public release. Each dataset is published as a samples $\times$ genes AnnData file, with opaque sample identifiers, NCBI Gene IDs as columns and gene symbols alongside. Labels, subject identifiers and provenance are kept in the private registry.

E.2 SINGLE-CELL RNA-SEQ COHORTS

PRIMO has 11 single-cell datasets. Each is rebuilt from the public deposit of its source study (GEO, CELLxGENE or Zenodo) with a dedicated, versioned script, following a common protocol. The released files contain unnormalised expression when available; any normalisation is left to the model being evaluated.

Input matrix. We start from the count matrix released by the authors. We never use their normalised or log-transformed values; where a deposit ships both, we read the raw-count layer. The only exception is the Smart-seq2 cohort (d046), whose only public expression is $\log_2(\mathrm{TPM}+1)$. We convert it back to linear TPM ($2^x - 1$, clipped at 0) and flag the scale in the dataset manifest.

Cell quality control. Cells are kept if they have at least 200 detected genes and fewer than 15% of their counts from mitochondrial genes. Mitochondrial genes are identified by the `MT-` prefix, and the fraction is measured on the source gene symbols, before gene identifiers are remapped. There are three exceptions: d046 (Smart-seq2, no mitochondrial threshold), d041 (MARS-seq, we keep the authors' single-cell wells, meaning wells annotated with one cell and a cell type, instead of our thresholds), d048 (we also remove the cells the authors annotated as doublets).

Gene identifiers. Gene symbols or Ensembl identifiers are mapped to NCBI Gene IDs. Genes without an NCBI ID are dropped, which removes the mitochondrially encoded genes, and counts of genes that map to the same NCBI ID are summed. Genes detected in fewer than three cells are removed. The resulting gene spaces range from 15,498 to 30,272 genes.

Samples. The prediction unit is the collection sample, one biopsy or blood draw. Each sample is linked to its donor, and all cross-validation folds are grouped by donor. Samples with fewer than 300 cells after quality control are discarded; the threshold is 200 cells for the low-throughput Smart-seq2 cohort. Each sample is then capped at 5,000 cells, drawn uniformly at random without replacement with a fixed seed (42).

Public release. Each dataset is published as a sparse cells $\times$ genes AnnData file. Columns are NCBI Gene IDs, with gene symbols provided alongside. Cell and sample identifiers are replaced by opaque ones, and `obs` holds a single column, `sample_id`, linking each cell to its collection sample. Labels, donor identifiers, clinical covariates and provenance are kept in a private repository. Table 11 lists the eleven datasets; there are 1,186,883 cells from 397 samples in total.

ID	Source (technology)	Samples	Cells	Genes	Cells/sample (median, range)
d038	10x	102	362,647	18,851	4,184 (324–5,000)
d039	10x	32	130,386	30,272	4,678 (1,595–5,000)
d040	10x	20	70,433	30,272	3,783 (513–5,000)
d041	MARS-seq	56	37,310	23,567	577 (302–3,086)
d042	10x	32	31,508	15,498	763 (360–3,328)
d043	10x	15	47,949	21,956	4,526 (434–5,000)
d044	10x	37	137,751	21,399	4,335 (343–5,000)
d045	10x	17	79,470	30,272	5,000 (2,404–5,000)
d046	Smart-seq2, TPM	18	5,765	24,225	331 (229–452)
d047	snRNA-seq	27	93,351	20,516	3,527 (836–5,000)
d048	10x	41	190,313	19,232	5,000 (2,349–5,000)

Table 11: Single-cell datasets in PRIMO, after preprocessing.

F CONFOUNDER CHECKS

A variable c is a confounder for a target y when two conditions hold together: c is recoverable from the expression profile, and c on its own predicts y at least as well as the profile does. Either condition alone is harmless. If the profile encodes c but c does not predict the label, c only adds noise, not bias. If c predicts the label but the profile does not encode it, a model cannot use c, since it only sees

the profile. When both conditions hold, a model can reach the reported score by recovering c and mapping it to y, without learning anything about the biology the task claims to probe.

For each candidate task we test the confounders specific to that task, as far as the available metadata allows. Typical confounders can be study, site, platform, tissue, biopsy origin, batch, treatment arm and sex. A task whose source deposited little metadata is therefore screened more weakly than one with rich metadata. Every variable is tested with the probe, metric and partitions of Section 3.2, and compared against the expression arm: a linear baseline fitted on the 1200 most variable genes. We reject a candidate only when both conditions hold for the same variable: the profile recovers c far above chance, and c alone matches or beats the expression arm on y.

A rejected task. We illustrate with a task we built and then discarded: estimating UICC stage from a single-cell lung adenocarcinoma biopsy. The cohort is one study of a lung cancer atlas (Kim et al., 2020; Salcher et al., 2022), 44 donors and 150,172 cells, restricted to a single study so that study and platform are constant by construction. The task is reformulated as a regression task, evaluated with Pearson correlation: the target scores stage I as 1.0, stage II and III as 2.5 (because too few samples are present for each label individually) and stage IV as 4.0. The candidate confounder is the biopsy origin: one value per donor among "normal", "normal adjacent", "primary tumour" and "metastasis". It is unevenly distributed across stage (see Table 12): no donor whose biopsy was normal or normal adjacent tissue reaches stage III or IV, and 12 of the 14 are stage I.

The analysis shows that both conditions hold. The profile recovers the origin well above chance (see Table 13), and the origin alone predicts stage better than the profile does (Table 14). Collapsed to a single indicator (whether the donor gave normal tissue or tumour), the confounder still matches the expression profile: 0.693 against 0.674. That margin is inside the noise of a 44-donor design, so we read it as a tie rather than a win. We tried dropping the 14 donors whose biopsy was normal or normal adjacent tissue, leaving 30 donors. But this does not help as the origin still beats expression: 0.694 against 0.577. We therefore removed the task.

Table 12: Donor counts by biopsy origin and UICC stage. The target merges stages II and III, so the three label levels hold 17, 7 and 20 donors.

Biopsy origin	I	II	III	IV	Total
Normal	8	2	0	0	10
Normal adjacent	4	0	0	0	4
Primary tumour	5	1	2	3	11
Metastasis	0	0	2	17	19
Total	17	3	4	20	44

Table 13: The origin is recoverable from expression. AUROC, mean over ten partitions with the standard deviation in parentheses.

Target	Donors	Expression
Normal versus tumour	44	0.846 (0.015)
Origin, four classes	44	0.924 (0.024)
Primary versus metastasis	30	0.883 (0.074)

Table 14: The origin predicts stage better than expression does, on the full cohort and after restricting it to tumour biopsies. Pearson correlation, mean over ten partitions with the standard deviation in parentheses.

Arm	44 donors	30 tumour donors
Expression	0.674 (0.075)	0.577 (0.073)
Origin	0.824 (0.013)	0.694 (0.014)
Origin, binarised	0.693 (0.006)	-

G LEADERBOARD DETAILS

For each task and pair of entries, we compare the two entries on each frozen partition of the task: ten, or one for the three transfer tasks. The outcome y_{ij}^t is the share of the partitions of task t won by entry i, a tie counting one half. All outcomes are fitted jointly by maximum likelihood with an ℓ_2 penalty. We minimise

$$\sum_t \sum_{i<j} \left[\log\left(1 + e^{\theta_i - \theta_j}\right) - y_{ij}^t \left(\theta_i - \theta_j\right)\right] + \frac{\lambda}{2}\|\theta\|^2,$$

with $\lambda = 0.1$ and the baseline's θ fixed at 0, and report the rating $R = 1000 + (400/\ln 10)\,\theta$. Under this model, entry i beats entry j on a partition with probability $1/(1 + 10^{-(R_i - R_j)/400})$. Each task has weight one per pair of entries, whatever its number of partitions.

Fitting all comparisons at once makes the result invariant to ordering. Partition scores are compared as stored, at four decimals. Refitting a rating therefore needs the per-partition scores, which are released with the leaderboard results; the per-task means of Appendix C are not enough. The baseline is fixed during, rather than after, optimisation, so that the penalty is centred on the baseline rating of 1000. Without the penalty, the modality ratings would change by less than 3 points, but ratings on boards with few tasks can change much more.

The 95% intervals resample cohorts with replacement (5,000 draws), so that all tasks of a cohort are drawn together. In each draw, each task is weighted by the number of times its cohort is drawn, and all ratings are refitted with the baseline fixed. The interval spans the 2.5th to the 97.5th percentile of the refitted ratings. The baseline therefore has no interval, and an interval that excludes 1000 marks a gap to the baseline beyond noise.

H CHOICE OF THE BASELINE

PRIMO anchors every board to the raw-expression baseline, HVG-1,200 (Section 3.3). The anchor is the reference of the rating scale (rating 1000) and the bar a model has to clear, so it must be fixed in advance and strong enough. The number of genes, 1,200, is the setting used by scGPT for several of its downstream tasks (Cui et al., 2024); it was fixed in the first release of PRIMO and not tuned on the current tasks. This appendix compares the anchor with other label-free baselines. The anchor was not changed after these results.

Protocol. Every baseline is label-free, computed once per dataset on all samples, and scored by the board's own evaluator with the same probe and partitions as a submission. We compare the anchor with all genes of the dataset, with a PCA projection of the same normalised expression (32 components on all genes), and with a random 64-dimensional Gaussian embedding, which gives the noise floor. For each baseline we report the board score (the mean over task categories of the mean task score, a simple summary that mixes metrics and is not used to rank entries) and test its gap to the anchor with a 95% bootstrap interval over cohorts (52 bulk tasks on 34 cohorts, 11 single-cell tasks on 11 cohorts, 5,000 draws).

Results. Table 15 gives the results. All genes score the same as the anchor within noise: gap -0.009 $[-0.023, +0.005]$ on bulk and $+0.008$ $[-0.014, +0.027]$ on single-cell. With all genes as the anchor, the models keep the same order on both boards. The PCA projection is far below the anchor, by 0.057 on bulk and 0.119 on single-cell; PCA with more components (64 to 256) scores lower still. Scoring the anchor with L2-normalised rows, as submissions are, changes nothing (0.000 on bulk, -0.009 $[-0.029, +0.006]$ on single-cell). We keep the pre-declared anchor: choosing the best baseline on the same tasks would select on the test labels, for gains that are within noise.

Baseline	Bulk (52 tasks)	Single-cell (11 tasks)
HVG-1,200 (anchor)	0.610	0.614
All genes	0.601	0.622
All genes → PCA-32	0.553*	0.495*
Random, 64 dimensions	0.307*	0.379*

Table 15: Board score of each label-free baseline. *: gap to the anchor beyond noise (its 95% cohort-bootstrap interval excludes zero). PCA and random embeddings are scored like submissions (rows L2-normalised).

I CHOICE OF THE LINEAR PROBE

To evaluate the impact of choosing a linear probe, we trained two non-linear probes: an SVM with an RBF kernel and an MLP with one hidden layer.

Both alternative probes replace only the supervised head of the evaluation protocol; the frozen representations, the task definitions, the outer partitions (10 partitions × 3 folds, fixed and shared across all probes and all models) and the reported metrics are identical to those of the linear probe. In every outer fold, features are standardized to zero mean and unit variance using training-fold statistics only, and hyperparameters are selected on an inner cross-validation of the training fold alone, using the same selection criterion as the linear probe: mean AUROC (weighted one-vs-rest for multi-class targets) over a repeated stratified 5-fold split (3 repeats, the number of folds reduced when the rarest class is too small) for classification, and mean squared error over a repeated 5-fold split (5 folds, 3 repeats) for regression. The selected configuration is then refit on the complete training fold and applied once to the held-out fold.

I.1 SVM

Method. The SVM probe is a support-vector machine with an RBF kernel, tuned over a grid of 7 logarithmically spaced values of the penalty C, from 1e-3 to 1e3, crossed with 4 values of the kernel width gamma, from 0.01/d to 10/d for embedding dimension d (28 configurations). For regression, an epsilon-insensitive SVR with epsilon = 0.1 is fit on z-scored targets and its predictions are mapped back to the original scale.

Results. We compared results on all tasks, except perturbation response, and all models. Average runtime per model submission on the 45 bulk tasks is 17 minutes (versus 4 minutes for the linear probe) and 2 minutes versus 1 minute for the 11 single-cell tasks. Overall, SVM is worse than linear in both modalities (Table 16).

Table 16: RBF-SVM versus linear probe on frozen representations. Scores are means over task–model pairs, each averaged over 10 partitions. Δ = SVM − linear; wins = pairs where the SVM scores higher. p: two-sided Wilcoxon signed-rank test on paired scores; bold indicates $p < 0.05$. Pairs sharing a task or a model are not independent, so p-values should be read as descriptive. The *All* rows pool categories with different metrics, so only paired differences are reported.

	Category	Metric	Tasks	Pairs	Linear	SVM	Δ	SVM wins	p
Bulk	Clinical activity	r	20	120	0.434	0.401	−0.034	25 (21%)	$\mathbf{1.3 \times 10^{-13}}$
	Endotype	AUROC	9	54	0.716	0.704	−0.011	15 (28%)	$\mathbf{6.4 \times 10^{-4}}$
	Prognosis	AUROC	7	42	0.692	0.684	−0.009	14 (33%)	**0.027**
	Treatment outcome	AUROC	9	54	0.679	0.682	+0.003	19 (35%)	0.22
	All		45	270			−0.018	73 (27%)	$\mathbf{5.8 \times 10^{-16}}$
Single-cell	Clinical activity	r	2	14	0.326	0.305	−0.021	6 (43%)	0.63
	Endotype	AUROC	5	35	0.794	0.789	−0.005	13 (37%)	0.25
	Treatment outcome	AUROC	4	28	0.648	0.620	−0.028	7 (25%)	$\mathbf{5.3 \times 10^{-4}}$
	All		11	77			−0.017	26 (34%)	**0.0026**

I.2 MLP

Method. The MLP probe is a single-hidden-layer network with ReLU activations and dropout on the hidden units, trained full-batch with Adam at a learning rate of 1e-3 and an L2 penalty applied to the weight matrices only. Its hyperparameters are drawn by random search over 30 trials: hidden width uniform on the integers 64 to 128, dropout uniform on [0.1, 0.5], and L2 coefficient log-uniform on [1e-5, 1e-1]. The combinations are sampled once under a fixed seed and reused for every task, fold and representation, so that all models are compared over an identical search space. The number of epochs is treated as a fourth hyperparameter selected from {50, 100, 200, 500} by evaluating each trial at the corresponding checkpoints of a single 500-epoch run, yielding 120 candidate configurations at the cost of 30 training runs.

Results. We compared results on all tasks, except perturbation response, and all models. Average runtime per model submission on an H100 GPU is 51 minutes for the 45 bulk tasks and 15 minutes for the 11 single-cell tasks. Overall, the MLP does not beat the linear probe: it scores slightly lower on bulk and about the same on single-cell (Table 17).

Table 17: MLP versus linear probe on frozen representations. Scores are means over task–model pairs, each averaged over 10 partitions. Δ = MLP $-$ linear; wins = pairs where the MLP scores higher. p: two-sided Wilcoxon signed-rank test on paired scores; bold indicates $p < 0.05$. Pairs sharing a task or a model are not independent, so p-values should be read as descriptive. The *All* rows pool categories with different metrics, so only paired differences are reported.

	Category	Metric	Tasks	Pairs	Linear	MLP	Δ	MLP wins	p
Bulk	Clinical activity	r	20	120	0.434	0.411	-0.024	28 (23%)	$\mathbf{3.9 \times 10^{-10}}$
	Endotype	AUROC	9	54	0.716	0.709	-0.006	26 (48%)	0.35
	Prognosis	AUROC	7	42	0.692	0.688	-0.004	18 (43%)	0.50
	Treatment outcome	AUROC	9	54	0.679	0.691	$+0.012$	31 (57%)	0.11
	All		45	270			-0.010	103 (38%)	$\mathbf{7.0 \times 10^{-6}}$
Single-cell	Clinical activity	r	2	14	0.326	0.331	$+0.005$	9 (64%)	0.71
	Endotype	AUROC	5	35	0.794	0.795	$+0.001$	14 (40%)	0.98
	Treatment outcome	AUROC	4	28	0.648	0.653	$+0.005$	14 (50%)	0.65
	All		11	77			$+0.003$	37 (48%)	0.53