DOES YOUR MODEL UNDERSTAND GENES? A BENCH-MARK OF GENE PROPERTIES FOR BIOLOGICAL AND TEXT MODELS

Anonymous authors

006

008 009 010

011

013

014

015

016

017

018

019

021

024

025

026

027

028

029

031

032

034

039

040

041

042

Paper under double-blind review

ABSTRACT

The application of deep learning for biology, including foundation models, has increased significantly in recent years. Some models are text-based, while others are trained on the underlying biological data, especially omics data of various modalities. Consistently comparing the performance of deep learning models for biology has proven challenging due to the diversity of training data and downstream tasks. Here, we utilize the fact that many models operate on the level of genes and propose a unifying benchmark by defining hundreds of tasks based on ground-truth gene properties collected from professionally curated bioinformatics databases. We collect properties of five types: (1) genomic properties, including predicting which genes can be methylated or which are dose-dependent; (2) regulatory functions, evaluating how the genes participate in cellular regulatory processes; (3) localization, including identification of differential expression in different tissues or sub-cellular localization; (4) biological processes, including predicting gene involvement in pathways or disease prognostics; and (5) protein properties, including prediction of functional domains or post-translational modifications. These properties are used to define binary, multi-label and multi-class classification tasks. To create an architecture-agnostic benchmark we extract gene representation vectors from each model, including single-cell RNA-seq (scRNA) foundation models, large language models, protein language models, DNA foundation models, and classical baselines, and use them to train simple predictive models on the tasks. Depending on the model, we utilize the model's token-level embeddings of gene symbols or transform the gene symbol to an input appropriate for the model, i.e. a description of the gene for text models, the gene sequence for DNA models or amino acid sequences for the protein models. Using these embeddings on the benchmark tasks, we create a detailed assessment of the relative performance of the different models. In general, we find that text-based models and protein language models outperform the expression-based models on tasks related to genomic properties and regulatory functions, while expression-based models tend to outperform the others on localization tasks. We also observe performance for the classical bag-of-words baseline that is similar to the large language models for many tasks. By enabling broad systematic evaluation of diverse deep learning models in biology, this benchmark can help direct future research in artificial intelligence toward improved biological understanding and accelerated therapeutic discoveries. The code and benchmark data can be extended to more models and tasks and is available at GitHub.

1 INTRODUCTION

Recent successes in the application of self-supervised learning in natural language processing have given rise to foundation models, which are trained on a large unlabeled dataset and useful on a broad range of tasks (Bommasani et al., 2021). The potential to realize similar advances in biology has given rise to a new and rapidly growing cohort of biological foundation models, either as specialized language models or new models trained on biological modalities such as DNA sequences (Ji et al., 2021), amino acid sequences (Rao et al., 2021), electronic health records (Yang et al., 2022) or other

^{*}These authors contributed equally to this work



083 foundation models in specific modalities such as single cell models (Liu et al., 2023). However, no benchmarks have been proposed that compare foundation models across modalities or that compare 084 text models against models trained on biological data directly. Part of the difficulty is that the 085 downstream tasks that are used to compare scRNA FMs such as cell-type annotation, batch correction, and perturbation prediction (Ding et al., 2024) are very different from the benchmarking tasks used to 087 evaluate NLP models, such as question-answering or sentence completion. The need for a benchmark 088 that can work with text models and foundation models is particularly important in light of recent works 089 such as Cell2Sentence (Levine et al., 2023), GenePT (Chen & Zou, 2023) and scInterpreter (Li et al., 090 2024), which have shown that text-models can be repurposed to work directly with transcriptomic 091 data.

092 Here we propose to use *gene embeddings* to create a new benchmark that enables comparison of 093 biological foundation models across modalities and against text models. Gene embeddings are an in-094 herent component of expression-based foundation models built on Transformer architectures (Vaswani et al., 2017), parallel to word embeddings in text models. They can also be produced using the gene 096 symbol or gene description with a language model supporting text embedding. Smaller models such as gene2vec (Du et al., 2019) or even bag-of-words models on textual descriptions of the gene can 098 also produce gene embeddings. As with text embedding benchmarks, it is assumed that the models producing better gene embeddings are learning the ground truth more faithfully (Muennighoff et al., 099 2022). 100

To evaluate the gene embeddings, we compile a wide range of ground truth biological knowledge about genes including their genomic properties, regulatory functions, localization, their involvement in biological processes, and their protein properties (Table 1). We connect the gene embeddings to the relevant tasks, and evaluate their performance as illustrated in Figure 1. Though each task captures only a small part of the biology involving the gene, collectively they offer a multi-faceted, panoramic view of the gene. Superior performance on this collection of tasks thus implies that the model's learned embeddings are more inherently meaningful and thus useful for diverse downstream tasks even without seeing labeled data for these tasks. 108 We apply this benchmark to evaluate several families of models. These include text-based models, 109 where the embeddings utilize large language models to encode a textual description of the genes, 110 scRNA foundation models trained on multi-omics data, models that are based on protein or DNA 111 sequence, and classical ML methods to act as a baseline comparison on text and gene expression 112 data. Our analysis shows that text models outperform the other model families for most gene-related tasks, even when the information is not explicitly in the text. This result underscores the need and the 113 potential to continue and improve knowledge integration into gene embeddings, thus improving gene 114 target identification and all downstream tasks. 115

The benchmark platform is available (under an Apache 2.0 license) at GitHub, and includes scripts
 for task data downloads, as well as examples and documentation for using the benchmark on new
 models.

119 120

2 BENCHMARK TASKS

121 122

Following decades of work in bioinformatics and related fields, large amounts of structured data regarding genes have been compiled through projects such as Reactome (Milacic et al., 2023), Human
Protein Atlas (Uhlen et al., 2010; Human Protein Atlas), OpenTargets (Ochoa et al., 2022) and HUGO
Gene Nomenclature Committee at the University of Cambridge (Seal et al., 2022). This enables us to compile a wide variety of validated properties to use to test the quality of gene embeddings. Our benchmarking package allows defining the tasks in general terms, which allows simple addition of new tasks with multiple identifier types (see S6.1. Notably, the benchmarking package is not limited to gene-tasks and can be easily extended to other modalities.

130 131 132

158

2.1 TASKS DESCRIPTION

We compiled 312 gene properties, which we used to define evaluation tasks. Most of the tasks are
 based on single gene properties, while some are based on gene-pairs or links between genes and
 diseases.

- For simplicity, we sort the properties into the following five families:
- Genomic properties This family of tasks evaluates the ability to predict properties inherent to the gene sequence, including predicting which genes can be methylated, and which genes are dose-dependent (their expression depends on the number of copies in the genome) There is a total of 7 tasks in this family. See Table S1 for a full description.
- Regulatory functions This family of tasks evaluates how the genes interact with other genes through
 the cellular regulatory processes and consists of a total of 6 tasks. These include predicting
 which genes are transcription factors, the number of connections in the gene-regulatory
 network, etc. See Table S3 for a full description.
- Localization This family includes tasks for identifying differential expression and activity in different tissues or sub-cellular localization. That includes predicting protein levels found in blood, correctly assigning genes to expression clusters derived from various tissue samples, sub-cellular localization, etc. There are a total of 30 tasks in this family. See Table S4 for a full description.
- Biological processes This family evaluates the biological functionality of the gene by evaluating tasks such as involvement in pathways, being prognostic of survival, and being associated with a disease. This family consists of 29 tasks, representing the most diverse set of questions. See Table S5 for a full description.
- protein properties This family focuses on properties of the protein product of the gene, including
 its functional domains, post-translational modifications, and its ligands.
- These properties cover many of the biological roles that genes play, providing an indication of how well a given pre-trained model has captured various aspects of gene representation, allowing for differentiation between various types of models and training data. Users can use the performance on different task families to select pre-trained models for their use-case.

Table 1: A breakdown of the number of tasks per prediction type and task family. Numbers in
 parenthesis represent the number of binary classification tasks that can be extracted from the multi label tasks.

| Task family | | Numbe | er of tasks (and | l sub-tasks) | |
|----------------------|--------|------------|------------------|--------------|-----------|
| | Binary | Multiclass | Multi-label | Regression | Total |
| Genomic properties | 3 | 1 | 3 (+79) | - | 7 (+79) |
| Regulatory functions | 5 | - | _ | 1 | 6 |
| Localization | - | 21 | 1 (+70) | 7 | 30 (+70) |
| Biological processes | 3 | 21 | 3 (+91) | 2 | 29 (+91) |
| Protein properties | - | - | 3 (+53) | - | 3 (+53) |
| Total | 11 | 43 | 10 (+293) | 10 | 71 (+293) |

2.2 TASK ORIGIN

178 179

185

187

To benchmark the pre-trained models, we aimed to collect properties that are as diverse as possible, capturing the many roles that genes and the proteins they code play in biology. For reliability and reproducibility we have opted for gene properties that are manually validated by hand and freely available, see Section S6.1.6 for data availability details.

- **Reactome** The pathways tasks were curated by taking the full list of genes from the Human Genome Nomenclature Committee (HGNC) downloadable files (including protein-coding gene, noncoding RNA, pseudogene and other tables) (Seal et al., 2022) and labeling each symbol by its inclusion in a top-level pathway S6 from Reactome (Milacic et al., 2023).
- Human Protein Atlas Protein atlas (Human Protein Atlas) tasks were created by compiling the protein atlas file v23 (Protein Atlas Data v23). We selected columns that contained features regarding gene properties, sorting them to binary, multiclass, multi-label, or regression tasks. We removed rows with missing data or genes that had no symbol name.
- Open Targets The gene-disease association (Ochoa et al., 2022) task was compiled by downloading
 the overall association score of genes and diseases from the open targets platform (we used
 the direct file interface using the files published on 2023-09-21 version 23).
- Uniprot Three protein properties tasks were created using data from the Universal Protein Knowl edgebase (UniProt) by assigning binary labels to gene symbols if a given keyword value is
 present for any of the protein products of that gene symbol. The keyword categories used
 were Domain, Ligand and Post-transcriptional modification (Consortium, 2022).
 - Publications Since several papers have used gene properties to evaluate pretrained models, we included those tasks in our benchmark (Chen & Zou, 2023; Fang et al., 2024; Lambert et al., 2018). There are 9 tasks derived directly from publications.

These tasks cover a wide variety of gene roles and provide a well-grounded assessment of gene representation quality. Because research in this domain evolves quickly, we have provided the ability to extend our benchmark with more tasks, at GitHub.

207 208

209

200

201

202

203

2.3 TASK DEFINITION

Independent of the task family, each task evaluates a specific outcome type: a) a binary , or b) a
multi-label assignment, or c) a multiclass, or d) a regression task. We used gene properties to define
tasks only if at least 1% of the covered entities had the label. Binary sub-tasks were derived from
multi-label tasks by selecting specific labels. For all tasks, we used the gene symbol as an identifier;
ensemble stable IDs were converted into symbols using MyGeneInfo (Wu et al., 2012). To simplify
comparisons between models we limited the scope of each task to the gene symbols shared by all encoding models.

| Cross-validation sion, Stratified cross-validation |
|---|
| sion, Stratified cross-validation |
| |
| rest, Stratified cross-validation Accu- |
| , F1, K-fold Iracy |
| n ab- K-fold |
| ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, |

Table 2: Description of the prediction models, evaluation metrics, and cross validation scheme used
 for each of the four task types

2.4 TASK EVALUATION

In contrast to text embedding benchmarks such as MTEB (Muennighoff et al., 2022) where the quality of the model is assessed by its ability to generate similar embeddings for known similar texts, evaluating the biological properties of genes embeddings requires a slightly different approach. Because the genes have many biological properties, we cannot assess the model quality by similarity alone. For this reason, we have primarily adopted classification metrics: the embeddings are provided as inputs to a simple logistic or linear regression model to predict the ground truth properties, and evaluated with 5-fold cross-validation. The benchmark can also be defined using non-linear models, which could detect information in the representation vectors more successfully than a linear model, we discuss this and assess the differences in Section 5.

This setup enables us to evaluate whether the correct information is encoded in the vector without making a-priori assumptions about the underlying information properties of the embedding space.

ENCODING MODELS

We selected several publicly available models for comparison from five major families: Large language models trained on text, deep-learning models trained on gene expression data, deep learning models trained on base pair sequences of genes, deep learning models train on amino acid sequences and classical machine-learning models. We used models that were openly available with weights.
When available, we used top-performing models according to independent leaderboards. Table 3 provides a summary table of model properties, and the following is a brief description of each model. The gene-benchmark allows for simple integration of additional models and tasks (see Supplementary text S6.1)

3.1 TEXT BASED MODELS

For text embedding models, we create an embedding for a gene by extracting the standard sym-bol, full name, and description of the gene from the NCBI Entrez Gene database (Maglott et al., 2010). This information is packed into a textual description that is given to the model as a prompt in the format "Gene symbol <symbol> full name <full name> with the summary <summary description>", and this prompt is embedded using a sentence em-bedding model. As a result, the benchmark that we have defined works seamlessly with any model supported by sentence_transformers (Reimers & Gurevych, 2019). For this as-sessment, we selected the top performing models from the leading embedding benchmark, the MTEB leaderboard (MTEB Leaderboard) and from the sentence transformers leaderboard (Sen-tence Transformers Leaderboard). For simplicity and ease of replication, we limited ourselves to models that did not require to trust remote code as defined by the sentence_transformers

API (trust_remote_code set to false). In addition to compare performance with a simpler non-parametric method, we used a Bag-of-words encoder

- MTEB-L A variant of *Mistral 7B* (Meng et al., 2024) called *SFR-Embedding-Mistral*, which is a transformer based generative LLM with 7.11B parameters. Chosen as the top performing open model on MTEB (MTEB Leaderboard) as of May 2024.
- MTEB-S A compact sentence-embedding model with 335M Parameters (Lee et al., 2024) called
 mxbai-embed-large-v1. Chosen as the top performing small open model (<1B parameters) on MTEB (MTEB Leaderboard) as of May 2024.
- MPNet A transformer-based textual LLM (Song et al., 2020), pre-trained on over 160GB text corpora. Pretraining was done using masked and permuted language modeling learning. It was chosen since it was the top performing model on sentence transformers (Sentence Transformers Leaderboard) as of May 2024.
- Bag-of-words A statistical word-based text model which does not take into account the order of words in the text. The presence of each word is used as an independent feature. We used CountVectorizer from scikit-learn (Pedregosa et al., 2011), with default parameters to select the top informative 1024 words, and used the word counts vector as the embedding for each description. The model was fitted to the text of the gene descriptions.
- 209 290 291

292

306

312

284

276

3.2 GENE EXPRESSION AND TRANSFORMER-BASED MODELS

Inspired by the success of transformer-based LLMs in NLP, these models aim to learn biology as
a "language" over scRNA-seq readings, fitting an embedding to each gene as if it were a 'word'
in an NLP model, and the transformer based architectures integrate the gene expressions into celllevel embeddings. We made use of a recent survey and benchmark to highlight the three best
performing open scRNA foundation models (Liu et al., 2023). The gene embeddings were extracted
from the publicly available model weights. Gene names were taken from the supplementary model
configuration files.

- CellPLM A transformer-based foundation model for single-cell biology with over 80M parameters (Wen et al., 2023). Trained on scRNA-seq and spatially resolved transcriptomic (SRT), adding tissue level information. Trained using MLM variant, on 9 million scRNA-seq cells and 2 million SRT cells. Embedding extracted from the embedder.feat_enc.emb layer in the model downloaded (Wen et al.), with the gene names from the matching configuration file.
- Geneformer A transformer based foundation model for single cell biology with 10.3M parameters. (Theodoris et al., 2023a). This model represents the scRNA expression using a list of genes ranked by their normalized expression levels. This is intended to make the order significant, and allows the use of context-aware attention mechanisms similar to these that work well in NLP. The model is trained on about 30M scRNA-seq readings. Embedding extracted from the embeddings.word_embeddings layer from (Theodoris et al., 2023b)
- **ScGPT** A generative foundation model for single-cell transcriptomics utilizing a self-attention, with 313 53M parameters (Cui et al., 2024). Pretrained using masked language model (MLM) training. 314 Explicitly encoded genes, expression levels and conditions, concatenated to represent each 315 gene in context. Training is performed using a masked language modeling variant, where 316 masking is done with attention masking to accommodate for the non-sequential nature of 317 the data. Embedding extracted from (Cui et al., b) following the instructions in (Cui et al., 318 a), steps 1 and 2. We used two variants, blood (designated ScGPT-B) trained on 10.3 million 319 blood and bone marrow cells and the human model (designated ScGPT-H) trained on 33 320 million normal human cells. 321
- Gene2vec A 200 dimensional concept embedding of the human genes (Du et al., 2019), based on
 the concept of Word2Vec (Mikolov et al., 2013) and learned from co-expression patterns, shared Gene Ontology (GO) annotation, tissue-specific genes, and functional gene sets.

| 326 | | | | | |
|-----|--------------|--------------------|----------------|---------------|-------------|
| 327 | Model | Input type | Model type | Num of params | Output size |
| 328 | MTEB-L | Text | Transformer | 7.1B | 4,096 |
| 323 | MTEB-S | Text | Transformer | 109M | 1024 |
| 330 | MPNET | Text | Transformer | 420 M | 768 |
| 331 | Bag-of-words | Text | Non-parametric | - | 1,024 |
| 332 | CellPLM | ScRNA-seq | Transformer | 85M | 1024 |
| 333 | Geneformer | ScRNA-seq | Transformer | 10.3M | 256 |
| 334 | ScGPT-H | ScRNA-seq | Transformer | 51M | 512 |
| 335 | ScGPT-B | ScRNA-seq | Transformer | 39M | 512 |
| 336 | Gene2Vec | Bulk RNA-seq | Word2Vec | 5M | 200 |
| 337 | DNABERT-2 | Base pair sequence | Transformer | 117M | 768 |
| 338 | ESM-2 | Protein sequence | Transformer | 3B | 2560 |

Table 3: Summary descriptions of the gene-encoding models

341 3.3 BASE-PAIR MODELS

Every gene can be mapped to its DNA sequence. There have been numerous recent advances in foundation models trained on DNA. Though not trained specifically to represent genes, by representing DNA they can generate gene representations as well. DNABERT-2 A BERT based genome foundation model (Zhou et al., 2024) trying to decode a linguistic representation of the genome. In this method they replaced the common k-mer tokenization with a Byte Pair Encoding (BPE) tokenization. This model performed well in a recent DNA model benchmark (Liu et al., 2024).

- 348 349 2.4
- 350

324

325

339 340

3.4 PROTEIN LANGUAGE MODELS

Representing the protein products of a gene, may be considered a representation of the gene itself.
Following the approach of SATURN (Rosen et al., 2024) and UCE (Rosen et al., 2023), we have represented the gene symbol as the mean of its protein product representation vectors.

Evolutionary Scale Modeling-2 (ESM-2) SOTA general-purpose protein language model. A
 transformer model trained on sequences of natural proteins (Lin et al., 2023) which is able to generate
 novel proteins. The model was trained using the ESM Metagenomic Atlas that contains >617
 million metagenomic protein sequences. We took the model esm2_t36_3B_UR50D with 3 billion
 parameters.

4 RESULTS

360 361 362

359

We report our gene-benchmarks on eleven models, evaluated on all tasks. The benchmark results 363 demonstrate that the various models exhibit different performance patterns on different tasks. When 364 grouping the performance measures by family tasks and averaging across tasks we find that the four text-based models exhibit better performance for the genomic properties and regulatory function 366 families, while the scRNA-based models performed better at the localization and biological process 367 tasks (Figure 2). These trends are consistent when using other evaluation metrics such as F1 (see 368 Figure S1). The calculation time for the benchmark varies depending on the model size. For the largest model, with 7B parameters and embedding size 4096 (Table 3), creating the gene embeddings 369 took approximately 40 minutes on a single NVIDIA A100 80GB. The calculation time for fitting the 370 predictive models is highly dependent on the embedding size, with the smaller embeddings requiring 371 less than an hour to calculate all the benchmarks and the largest requiring 15+ hours on a 48-core 372 Xeon E5. 373

Interestingly, text-based models using transformer architecture only slightly outperform the bag of-words model in most tasks. Furthermore, we do not see an advantage to the model size, where
 the MTEB-L model exhibits comparable performance to the smaller MTEB-S and MPNet models.
 Similarly, in the scRNA-based models the transformers usually slightly outperform the older gene2vec
 model, which is based on word2vec architecture and trained on bulk RNA expression data. ScGPT-H

378 0.81 (0.14) 0.76 (0.13) 0.79 (0.14) 0.67 (0.16) 0.80 (0.17) 0.73 (0.15) Genomic properties 379 0.79 (0.05) 0.73 (0.04) 0.77 (0.05) 0.61 (0.03) 0.60 (0.04) 0.58 (0.03) 0.63 (0.04) 0.59 (0.02) 0.54 (0.02) Protein properties 380 0.74 (0.06) 0.68 (0.06) 0.73 (0.10) 0.66 (0.11) functions 381 0.7 Biological processes - 0.64 (0.13) 0.64 (0.13) 0.65 (0.11) 0.63 (0.13) 0.69 (0.11) 0.69 (0.11) 0.61 (0.10) 0.66 (0.11) 0.67 (0.11) 0.66 (0.11) 0.59 (0.09) 382 0.75 (0.09) 0.84 (0.10) Localization - 0.67 (0.06) 0.71 (0.07) 0.72 (0.08) 0.69 (0.07) 0.85 (0.10) 0.72 (0.09) 383 0.5 MPNet Bag of Words cellPLM Geneformer ScGPT-H Gene2vec MTEB-L MTEB-S ScGPT-B ESM-2-3B DNABert-2 384 Text Description RNA-sequencing 385 Protein sequence Base pair sequence 386

Figure 2: The performance of each model on the task families as measured by average area under the ROC curve. Parentheses show the corresponding standard deviation across all tasks of the same family.

389 390 391

387

388

was the top performer in two different families of tasks. Here, too, we do not see a clear advantage 392 for larger models, with cellPLM exhibiting comparable performance to the smaller ScGPT-H and 393 even smaller Geneformer. ScGPT-H outperformed ScGPT-B significantly, which is likely due to the 394 larger, more diverse, training data. Indeed, it is to be expected that a model trained on expression data 395 from a single tissue would not perform well on tissue localization tasks related to other tissues.

A closer examination of the mean AUC per task (Figure S2, Figures S3, S4, S5, S6, S7, S8, S9) reveals 397 a more complex picture, where within each task family some tasks are dominated by text-based 398 models and others by expression-based models or protein language models. This can also be seen by 399 the high cosine similarity observed between overall task performance amongst models from the same 400 type, as exhibited in Figure S10. 401

The protein models perform best at protein properties, but less so for biological processes and 402 localization. DNABert-2 performs worse than the other models for all but the genomic properties, 403 where it is comparable to the other model families. 404

405 The tasks themselves also show a clustering in performance, as shown in Figure S11, but also show a 406 large range of dissimilarity, suggesting that the benchmark tasks correspond to distinct biological 407 phenomena.

408 Above we used linear models for predicting gene properties from vectors. We consider the possibility 409 that a more expressive model could perform better by training a multilayer perceptron model (MLP) 410 on the binary tasks, comparing the MTEB-L and MTEB-S embeddings. We find that the performance 411 is closely correlated to logistic regression, as seen in Figure S12. For this reason, we prefer the linear 412 models which are not sensitive to hyperparameter selections and thus enable robust comparisons 413 across many thousands of combinations of models and tasks. Nevertheless, the benchmark package 414 code at GitHubsupports the use of any scikit-learn model.

415 One notable result is that text models outperform the scRNA models in most disease involvement tasks 416 except in the Pathology tasks, chromosome, and N1 Network, indicating that there are exceptions to 417 the general rules of model performance we outlined. Similarly, expression-based models outperform 418 in cell-type localization tasks, but under-perform in sub-cellular localization tasks. This is in line 419 with our expectations, given the close relation between cell-type, tissue-type and single cell RNA expression levels. 420

421 422

5 SUMMARY AND DISCUSSION

423 424

427

We present a gene-centric benchmark that includes hundreds of tasks, sorted into functional families. 425 We designed this benchmark to evaluate the gene embeddings provided by pretrained models applied 426 to biology, thus suggesting a common ground for evaluating the potential of the models to provide useful, and potentially novel insights on gene involvement in biological and medical questions. 428

429 We applied the benchmark to a representative set of models trained on text, gene-expression data, protein sequences and DNA sequences and compared their performance. We observed that each family 430 of models exhibits superiority for a different set of tasks, hinting that combining the knowledge 431 that comes from multiple modalities may provide additional benefits. It should be pointed out, however, that our analysis is not intended to be a comprehensive comparison of all text models and all gene-expression models and that our results may not be generalizable to all such models.

The embeddings that we chose to use for the text models were based on a detailed description of 435 the genes, rather than using the gene symbol, which could have evaluated whether the LLM has 436 sufficient useful information on genes from its general training on text. However, while virtually 437 all coding genes have literature in PubMed, the coverage is highly skewed, with the most popular 438 10% of genes accounting for over 60% of the publication (Lee et al., 2019), suggesting that language 439 model pretraining may be skewed towards the more widely studied genes. Instead, in our evaluation 440 of text models, we included a bag-of-words model that can serve as a baseline of sorts, allowing us to 441 examine the ability of LLMs to generalize and contextualize. We find that in most tasks the LLMs 442 provide a modest improvement in performance, indicating that additional work may be needed to create LLMs that are useful for basic biology. 443

444 We point out that most of the models we evaluated here were not designed to provide useful gene 445 embeddings. Some were trained to perform textual tasks, while others were trained to predict 446 properties at the whole-cell level. However, utilizing pretrained gene embeddings from a foundation 447 model to improve performance of a more specialized model has become an active research area 448 with numerous promising results produced in the last year. For example, scFoundation (Hao et al., 449 2023) showed improvement on gene perturbation prediction by injecting their gene embeddings into GEARS (Roohani et al., 2023), SATURN (Rosen et al., 2024) has used ESM (Lin et al., 2023) 450 protein embeddings to enable cross-species cell label propagation and GNN based models such as 451 Otter-Knowledge (Lam et al., 2023) and BioBridge (Wang et al., 2023) have proposed comprehensive 452 biomedical models built on embeddings across multiple domains. Given the interest and promise of 453 this technology, our benchmark can help guide researchers toward more successful application of 454 deep learning to biology. 455

456 457

5.1 LIMITATIONS AND FUTURE WORK

458 We gathered the benchmark tasks from actively maintained professionally curated sources and we 459 have relied on their quality control processes. For many reasons, the entire genome is not studied 460 evenly (Lee et al., 2019) and when genes are studied, the full diversity of human ancestry is not 461 evenly reflected (Fairley et al., 2019). As biological research improves in performance and fairness, 462 we look forward to updating our benchmark tasks accordingly. We used only open source models 463 with released weights excluding models that did not (Zrimec et al., 2022). Though we have explored 464 the benchmark tasks using gene embeddings, the tasks could be utilized in other ways, such as by 465 defining fine-tuning objectives for deep learning models or even as the basis for question answering in text models. Such a strategy, while not applicable for all models, may uncover predictive power 466 that is specific to each model. 467

468 469

470

References

471 Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, 472 Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, 473 Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano 474 Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren 475 Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter 476 Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil 477 Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar 478 Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal 479 Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu 480 Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, 481 Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, 482 Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu 483 Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, 484 Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, 485 Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai

Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi
Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the
opportunities and risks of foundation models. *arXiv*, 2021. doi: 10.48550/ARXIV.2108.07258.
URL https://arxiv.org/abs/2108.07258.

- 490
 491
 491
 491
 491
 491
 491
 492
 492
 493
 493
 493
 494
 494
 495
 495
 495
 496
 496
 497
 497
 498
 498
 499
 499
 499
 490
 491
 491
 491
 492
 493
 493
 494
 494
 495
 495
 496
 496
 497
 498
 498
 499
 499
 499
 499
 499
 499
 490
 490
 490
 491
 491
 492
 493
 494
 494
 495
 495
 496
 496
 497
 497
 498
 498
 498
 499
 498
 498
 499
 499
 499
 499
 490
 490
 491
 491
 492
 493
 494
 494
 494
 495
 496
 496
 497
 498
 498
 498
 498
 498
 499
 499
 499
 499
 499
 499
 499
 490
 490
 491
 491
 491
 492
 493
 494
 494
 494
 495
 496
 496
 497
 498
 498
 498
 498
 498
 498
 498
 498
 498
 498
 498
 498
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
 499
- The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids
 Research, 51(D1):D523–D531, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL
 https://doi.org/10.1093/nar/gkac1052.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang.
 Grn inference on pre-trained model. https://github.com/bowang-lab/scGPT/blob/
 main/tutorials/Tutorial_GRN.ipynb, a. Accessed: 2024-05-27.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang.
 scgpt model trained on a all available human tissue samples. https://drive.google.com/ drive/folders/10Wh_-ZRdhtoGQ2Fw24HP41FgLoomVo-y, b. Accessed: 2024-06-04.
- 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*, pp. 1–11, 2024.
- Jiayuan Ding, Renming Liu, Hongzhi Wen, Wenzhuo Tang, Zhaoheng Li, Julian Venegas, Runze
 Su, Dylan Molho, Wei Jin, Yixin Wang, Qiaolin Lu, Lingxiao Li, Wangyang Zuo, Yi Chang,
 Yuying Xie, and Jiliang Tang. Dance: a deep learning library and benchmark platform for
 single-cell analysis. *Genome Biology*, 25(1), March 2024. ISSN 1474-760X. doi: 10.1186/
 s13059-024-03211-z. URL http://dx.doi.org/10.1186/s13059-024-03211-z.
- Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(S1), February 2019.
 ISSN 1471-2164. doi: 10.1186/s12864-018-5370-x. URL http://dx.doi.org/10.1186/s12864-018-5370-x.
- Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. The international genome sample resource (igsr) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947, October 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz836. URL http://dx.doi.org/10.1093/nar/gkz836.
- Chen Fang, Yidong Wang, Yunze Song, Qingqing Long, Wang Lu, Linghui Chen, Pengfei Wang,
 Guihai Feng, Yuanchun Zhou, and Xin Li. How do large language models understand genes and
 cells. *bioRxiv*, 2024. doi: 10.1101/2024.03.23.586383. URL https://www.biorxiv.org/
 content/early/2024/03/27/2024.03.23.586383.
- Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng
 Wang, Jianzhu Ma, Le Song, and Xuegong Zhang. Large scale foundation model on single cell transcriptomics. *bioRxiv*, May 2023. doi: 10.1101/2023.05.29.542705. URL http:
 //dx.doi.org/10.1101/2023.05.29.542705.
- Human Protein Atlas. Human protein atlas. https://www.proteinatlas.org/. Accessed:
 2024-05-26.
- Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, February 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btab083. URL http://dx.doi.org/10.1093/bioinformatics/btab083.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

573

577

579 580

581

582

583

540 Hoang Thanh Lam, Marco Luca Sbodio, Marcos Martinez Gallindo, Mykhaylo Zayats, Raul 541 Fernandez-Diaz, Victor Valls, Gabriele Picco, Cesar Berrospi Ramis, and Vanessa Lopez. Otter-542 knowledge: benchmarks of multimodal knowledge graph representation learning from different 543 sources for drug discovery. arXiv preprint arXiv:2306.12802, 2023.

- Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai 545 Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. The hu-546 man transcription factors. Cell, 172(4):650-665, 2018. ISSN 0092-8674. doi: https://doi. 547 org/10.1016/j.cell.2018.01.029. URL https://www.sciencedirect.com/science/ 548 article/pii/S0092867418301065. 549
- 550 Kyubum Lee, Mindy Clyne, Wei Yu, Zhiyong Lu, and Muin J. Khoury. Tracking human genes along the translational continuum. npj Genomic Medicine, 4(1), October 2019. ISSN 551 2056-7944. doi: 10.1038/s41525-019-0100-0. URL http://dx.doi.org/10.1038/ 552 s41525-019-0100-0. 553
- 554 Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread -555 new fluffy embeddings model, 2024. URL https://www.mixedbread.ai/blog/ 556 mxbai-embed-large-v1.
- Daniel Levine, Sacha Lévy, Syed Asad Rizvi, Nazreen Pallikkavaliyaveetil, Xingyu Chen, David 558 Zhang, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, 559 Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul M. Dhodapkar, and David van Dijk. Cell2sentence: Teaching large language models the language of 561 biology. bioRxiv, September 2023. doi: 10.1101/2023.09.11.557287. URL http://dx.doi. org/10.1101/2023.09.11.557287. 563
- 564 Cong Li, Meng Xiao, Pengfei Wang, Guihai Feng, Xin Li, and Yuanchun Zhou. scinterpreter: Training large language models to interpret scrna-seq data for cell type annotation. arXiv preprint 565 arXiv:2402.12405, 2024. 566
- 567 Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, 568 Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom 569 Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level 570 protein structure with a language model. Science, 379(6637):1123-1130, March 2023. ISSN 1095-571 9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science. 572 ade2574.
- Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of 574 foundation models in single-cell data analysis. *bioRxiv*, September 2023. doi: 10.1101/2023.09.08. 575 555192. URL http://dx.doi.org/10.1101/2023.09.08.555192. 576
- Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z Li. 578 Genbench: A benchmarking suite for systematic evaluation of genomic foundation models. arXiv preprint arXiv:2406.01627, 2024.
 - D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. Nucleic Acids Research, 39(Database):D52–D57, November 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq1237. URL http://dx.doi.org/10.1093/nar/gkq1237.
- 584 Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-585 embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://blog.salesforceairesearch.com/sfr-embedded-mistral/. 586
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word represen-588 tations in vector space. arXiv, 2013. URL https://doi.org/10.48550/arXiv.1301. 589 3781. 590

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, 591 Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen 592 Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and

595

596

Peter D'Eustachio. The reactome pathway knowledgebase 2024. Nucleic Acids Research, 52(D1):D672–D678, November 2023. ISSN 1362-4962. doi: 10.1093/nar/gkad1025. URL http://dx.doi.org/10.1093/nar/gkad1025.

- 597 598 MTEB Leaderboard. Mteb leaderboard - a hugging face space by mteb. https://huggingface. co/spaces/mteb/leaderboard. Accessed: 2024-05-27.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding
 benchmark. arXiv preprint arXiv:2210.07316, 2022. URL https://doi.org/10.48550/
 arXiv.2210.07316.
- 603 David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, 604 Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, Manuel Bernal-Llinares, Kirill 605 Tsukanov, Helena Cornu, Konstantinos Tsirigos, Olesya Razuvayevskaya, Annalisa Buniello, Jeremy Schwartzentruber, Mohd Karim, Bruno Ariano, Ricardo Esteban Martinez Osorio, Javier 607 Ferrer, Xiangyu Ge, Sandra Machlitt-Northen, Asier Gonzalez-Uriarte, Shyamasree Saha, Santosh Tirunagari, Chintan Mehta, Juan María Roldán-Romero, Stuart Horswell, Sarah Young, Maya 608 Ghoussaini, David G Hulcoop, Ian Dunham, and Ellen M McDonagh. The next-generation open 609 targets platform: reimagined, redesigned, rebuilt. Nucleic Acids Research, 51(D1):D1353–D1359, 610 November 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac1046. URL http://dx.doi.org/ 611 10.1093/nar/gkac1046. 612
- Open Medical-LLM Leaderboard. Open medical-llm leaderboard a hugging face space
 by openlifescienceai. https://huggingface.co/spaces/openlifescienceai/
 open_medical_llm_leaderboard. Accessed: 2024-05-26.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Protein Atlas Data v23. Data from the human protein atlas in tab-separated format. https: //v23.proteinatlas.org/download/proteinatlas.tsv.zip. Accessed: 2024-05-28.
- Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang (eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 8844–8856. PMLR, 18–24 Jul 2021. URL https://proceedings. mlr.press/v139/rao21a.html.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
 In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
 Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.
 10084.
- Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multi gene perturbations with gears. *Nature Biotechnology*, August 2023. ISSN 1546-1696. doi: 10.1038/
 s41587-023-01905-6. URL http://dx.doi.org/10.1038/s41587-023-01905-6.
- Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918. URL https://www.biorxiv.org/content/early/ 2023/11/29/2023.11.28.568918.
- Yanay Rosen, Maria Brbić, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Toward universal cell embeddings: integrating single-cell rna-seq datasets across species with saturn. *Nature Methods*, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02191-z. URL http://dx.doi.org/10.1038/s41592-024-02191-z.
- Ruth L Seal, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Susan Tweedie, Liora Haim Vilmovsky, and Elspeth A Bruford. Genenames.org: the HGNC resources in 2023. Nucleic Acids
 Research, 51(D1):D1003–D1009, 10 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac888. URL
 https://doi.org/10.1093/nar/gkac888.

688

689

- Sentence Transformers Leaderboard. Pretrained models sentence-transformers documentation. https://sbert.net/docs/pretrained_models.html. Accessed: 2024-05-27.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.
- Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C
 Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, Shirley Liu X, and Patrick T Ellinor.
 Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023a.
 ISSN 0028-0836. doi: 10.1038/s41586-023-06139-9.
- Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C
 Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, Shirley Liu X, and Patrick T Ellinor.
 Geneformer model. https://huggingface.co/ctheodoris/Geneformer/tree/
 pr146_branch, 2023b. Accessed: 2024-06-05.
- Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. Foundation models in healthcare: Opportunities, risks &; strategies forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI
 ACM, April 2023. doi: 10.1145/3544549.3583177. URL http://dx.doi.org/10. 1145/3544549.3583177.
- Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based human protein atlas. *Nature Biotechnology*, 28(12):1248–1250, December 2010. ISSN 1546-1696. doi: 10.1038/nbt1210-1248. URL http://dx.doi.org/10.1038/nbt1210-1248.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/ file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N. Ioannidis, Huzefa Rangwala, and Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graphs, 2023. URL https://arxiv.org/abs/2310.03320.
- Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang
 Tang. Cellplm model 20230926_85m.best.ckpt. https://www.dropbox.com/scl/fo/
 i5rmxgtqzg7iykt2e9uqm/h?rlkey=08hi0xads9ol07o48jdityzv1&e=2&dl=0.
 Accessed: 2024-05-20.
- Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang.
 Cellplm: Pre-training of cell language model beyond single cells. *bioRxiv*, October 2023. doi: 10.
 1101/2023.10.03.560734. URL http://dx.doi.org/10.1101/2023.10.03.560734.
 - Chunlei Wu, Ian MacLeod, and Andrew I. Su. BioGPS and MyGene.info: organizing online, genecentric information. *Nucleic Acids Research*, 41(D1):D561–D565, 11 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1114. URL https://doi.org/10.1093/nar/gks1114.
- Ki Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, et al. A large language model for electronic health records. *npj Digital Medicine*, 5(1), December 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00742-2. URL http://dx.doi.org/10.1038/s41746-022-00742-2.
- Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert 2: Efficient foundation model and benchmark for multi-species genome, 2024. URL https: //arxiv.org/abs/2306.15006.
- Jan Zrimec, Xiaozhi Fu, Azam Sheikh Muhammad, Christos Skrekas, Vykintas Jauniskis, Nora K
 Speicher, Christoph S Börlin, Vilhelm Verendel, Morteza Haghir Chehreghani, Devdatt Dubhashi,
 Verena Siewers, Florian David, Jens Nielsen, and Aleksej Zelezniak. Controlling gene expression
 with deep generative design of regulatory DNA. *Nat. Commun.*, 13(1):5099, August 2022.

702 6 APPENDIX

704 S6.1 THE GENE-BENCHMARK PACKAGE

This freely available package was developed to facilitate easy access to the tasks and efficient use of them for benchmarking. It includes three main modules as well as notebooks and scripts that demonstrate the package's usability. The main flow of task evaluation using the package is described in Figure 1. Below we review the main modules used in the package, which is available at GitHub.

711 S6.1.1 TASKS

712 This module contains two main parts. The first is the means to load the task definition according to 713 task name in a generic format into a designated, easy-to-use object. The class allows easy access 714 to the entity identifiers (usually gene symbols) and their outcomes. They will usually be a single 715 columned data frame, but if the task includes multiple genes per instance (for example, gene-to-gene 716 interaction), it will include multiple entities in a column structure. In the multi-label case, the output 717 is also a multi-columned data frame. The second part is a pipeline class that manages the process 718 from a task name to description (in the case of text-based models) to encoding, training a simple 719 prediction model in a cross-validation fashion, and creating a report. Adding additional tasks is designed to be simple, all it requires is saving the task descriptions in a specific format. 720

721

710

722 S6.1.2 DESCRIPTOR

The module manages the transition from an entity identifier into a text description. For gene symbols, we retrieve the description fields from NCBI using the *MyGene.Info* services and construct a description sentence. We allow predefined descriptions by creating a descriptor that loads the descriptions from a CSV file. This feature enabled us to download the disease description from open targets without needing to integrate with their service and facilitate easy introduction of new descriptions. We are also able to construct multi identifier types descriptors, thus enabling the creation of a descriptor that can describe tasks with different identifiers, such as in gene-disease association.

- 730 731
- S6.1.3 ENCODER

This module manages the encoding of either the entity identifier or its textual summary. We enable
encoding using any HuggingFace sentence transformer supporting module. In addition, we enable
encoding using a pre-computed encoder by loading the encodings from a precomputed CSV file.
This enables us to pre-compute the encoding from scRNA-based models. In addition, we enable
the creation of a multi-entity type encoder that enables encoding each type of entity differently. For
example, in the case of Gene-Disease association, we can encode the genes using pre-computed
encoding and the disease using a sentence transform encoder.

740 S6.1.4 BASE MODELS 741

The package supports any scikit-learn model. For the manuscript, we explored linear and logistic regression with the default scikit-learn parameters, and an MLP with three hidden layers of size 100 and 500 max iterations.

745

746 S6.1.5 SCRIPTS AND NOTEBOOKS

To efficiently create benchmarks the package includes a command line interface. Enabling benchmarking multiple models (described in YAML format) on multiple tasks (supplied in the command line or in YAML) and output a single report in CSV format. An additional script is supplied that can extract the embedding of the given identifiers list. The package also includes a notebook demonstrating how the package can be used and how to create figures, as displayed in this manuscript.

- 752
- 753 S6.1.6 DATA AVAILABILITY AND LICENSING 754
- All of the data is from publicly available sources and the steps required to download and prepare the tasks for benchmarking are implemented in our GitHub repository. We did not produce the task data

756 0.57 (0.28) 0.57 (0.29) 0.45 (0.33) 0.59 (0.30) 0.47 (0.31) 0.8 0.72 (0.07) 0.62 (0.05) 0.69 (0.06) 0.43 (0.06) 0.42 (0.07) 0.40 (0.06) 0.47 (0.06) 0.39 (0.05) 0.27 (0.09) 758 0.72 (0.15) 0.62 (0.32) 0.68 (0.20) 0.65 (0.22) 0.63 (0.18) 0.59 (0.16) 0.67 (0.15) 0.63 (0.16) 0.38 (0.29) 759 760 0.2 761 Localization -0.45 (0.24) 0.48 (0.26) 0.46 (0.26) 0.47 (0.26) 0.58 (0.24) 0.60 (0.25) 0.51 (0.26) 0.58 (0.25) 0.55 (0.27) 0.49 (0.27) 0.44 (0.28) 0.0 cellPLM 762 MTEB-L MTEB-S MPNet Bag of Words Geneformer ScGPT-B ScGPT-H Gene2vec ESM-2-3B DNABert-2 RNA-sequencing Text Description Protein sequence Base pair 764

Figure S1: The performance of each model on the task families as measured by average f1 score. Parentheses show the corresponding standard deviation across all tasks of the same family.

and do not redistribute the data used for the benchmark tasks. To reproduce the results shown here,
we provide code to populate the benchmark task directly from the public sources. We do not own the
task data and refer the users to the licenses of the data owners.

- **Reactome** The current task retrieval code downloads that pathway directly from reactome's current server, Reactome data is robustly backed at third party servers. Reactome content is readily accessible for download from its website, GitHub, and various aggregators like NCBI and EMBL-EBI with data availability path in case of of funding loss. see Reactome digital preservation for further details.
- The human protein atlas Tasks are created using files saved at the human protein atlas, the data is accessible using programmatic access as well. Previous versions of the data files are accessible as well.
- **Open Targets** open targets is committed to open source and supporting open access research. with multiple data download and retrieval options see data access for further details provides data from community contributions, with the original data owners retaining ownership and rights. There are no additional restrictions on the use or redistribution of this data, Open Targets allow does not guarantee the accuracy or suitability of the data or services provided
- Uniprot tasks UniProt conforms with EMBL-European Bioinformatics Institute's data preservation policies. Uniprot has applied a CC-BY-4.0 license to the copyrightable parts of their database. For more info see Uniprot license.
- **Publication tasks** The data used for the creation of these tasks comes from the cited publications. For the HLA task we the data was derived from the HGNC web site. For the Tf vs non-tf task the data was derived from the The Human Transcription Factors web-site. That made the data publicly available via files but did not make clear data availability commitment. Scripts detailing exactly how to obtain these datasets and to construct the tasks as utilized here are provided at GitHub.
- 787 788 789

791

792

793

794

765

766

767 768

771 772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

- 799 800
- 801

- 804
- 805
- 806
- 807
- 808 809

| 810 | Table S1: Detailed descrip | tion | of the genomic ta | sks used for benchmarking |
|-----|----------------------------|------|---|---------------------------|
| 811 | 1 | | C | e |
| 812 | | | s s | |
| 813 | | | atla | |
| 814 | | | B. B. | |
| 815 | | | ner ote ner ote | |
| 816 | | | | |
| 817 | | gin | nefo nefo nefo nan | |
| 818 | | Ori | Ter | |
| 819 | | 0 | | |
| 820 | | | ne | |
| 821 | | | ac | |
| 822 | | | is ted | |
| 823 | | | t ha lec' | |
| 824 | | | lan ss i sel | |
| 825 | | | ival ppie | |
| 826 | | | f cc ing | |
| 827 | | | is i r of ord | |
| 828 | | | or be | |
| 829 | | | tion ion | |
| 830 | | | lati nyls du | |
| 831 | | | thy / th netl | |
| 832 | | | ne ne ne | |
| 833 | | | gh ys. | |
| 834 | | | ou ffec the | |
| 835 | | | thr n aj of 1 | |
| 836 | | | go three ss) | |
| 837 | | _ | ne ne go ss(e | |
| 838 | | ion | son son ne ge clas | |
| 839 | | ipt | the be a ge | |
| 840 | | SCI | ote ote | |
| 841 | | Ď | E: Å Ď º Č Ď | |
| 842 | | 0 | ~ + ~ - + | |
| 843 | | Size | 133 1787 1737 1737 | |
| 844 | | •1 | 19 | |
| 845 | | | | |
| 846 | | | ica | |
| 847 | | | i la | |
| 848 | | ype | ina. ina. ina. | |
| 849 | | H | a saa a | |
| 850 | | | q | |
| 851 | | | late | |
| 852 | | | hyl | |
| 853 | | | net | |
| 854 | | | iive n-1 | |
| 855 | | | ted nsit | |
| 856 | | | yla 1sei 1 vs | |
| 857 | | | eth s ir utec | |
| 858 | | | -m e v iyla | |
| 859 | | | e eth | |
| 860 | | Je | /S T Omo ens: m-/ ass | |
| 861 | | lan | nt v nosc e sc nly i cl | |
| 862 | | X | ale om age 4-0 teir | |
| 863 | | [as] | Dos Jos | |
| | | | | |

Table S2: Detailed description of the protein structural tasks used for benchmarking UniProt UniProt UniProt Origin sis (e.g., phosphorylation, glycosylation). Predicts specific molecules or ions the protein can bind to, impact-Predicts which chemical modifications occur after protein syntheing its function (e.g., ATP, zinc). Predicts structural or functional regions within the protein, like specific repeats or conserved motifs (e.g., transmembrane do-main). Description Size multilabel multilabel multilabel Type UniProt Keyword Domain UniProt Keyword Ligand UniProt Keyword PTM

Task Name

| 918 919 | Table S3: Detailed descr | iptic | on of the regulatory tasks used for benchmarking |
|------------|--------------------------|---------------|--|
| 920 | | | 1 1 |
| 921 | | | s s s |
| 922 | | | to de |
| 923 | | | |
| 924 | | | Jelle Jelle |
| 925 | | | and C diagonal diag |
| 926 | | | ar and and ar |
| 927 | | | atl atl anse anse anse anse anse anse atl |
| 928 | | | - De Gegerande |
| 929 | | | |
| 930 | | ц | Hur do sta do do |
| 931 | | . <u>19</u> . | ene |
| 932 | | ō | |
| 933 | | | |
| 934 | | | iph iph |
| 935 | | | l rai |
| 936 | | | u d nd n |
| 937 | | | (lc (lc) |
| 938 | | | get con |
| 939 | | | Du 1 lin |
| 940 | | | e N internet |
| 941 | | | th the volume |
| 942 | | | ans in the second secon |
| 943 | | | a tr sets |
| 944 | | | a as as |
| 945 | | | n n a multiple and a |
| 946 | | | n go ow ow ow of the search of |
| 947 | | | fin strong |
| 948 | | | h a l 1 g f f a ow ow |
| 949 | | | wit wit wit wit wit wit |
| 950 | | _ | khn ar ar ar |
| 951 | | ion | ine hat fille |
| 952 | | ript | of of tiva per tion of the second sec |
| 953 | | esc | the energy of the start of the |
| 955 | | Ω | |
| 956 | | e | $\left 0 \otimes 4 \tilde{\omega} - \tilde{v} \right $ |
| 957 | | Siz | 2003 134 110 228 28 28 276 |
| 958 | | | |
| 959 | | | |
| 960 | | | |
| 961 | | မ | |
| 962 | | Ŋ | |
| 963 | | | |
| 964 | | | tt l |
| 965 | | | nge |
| 966 | | | t ra |
| 967 | | e | tt khon |
| 968 | | am | gen sets si |
| 969 | | Z | le2 le2 v s n rrac v s n v s n |
| 970 | | [as] | IF VI NI nte |
| 971 | | | |
| | | | |

983

984

985

986

987 988

990

1000

1001

1002 1003 1004

1005 1006 1007

1008

1009

1010 1011 1012

1013

1014

1015

1016

1017

1018

1019 1020 1021

1022

1023 1024 1025

Table S4: Detailed description of the localization tasks used for benchmarking protein atlas human protein atlas human protein atlas protein atlas human protein atlas protein atlas human protein atlas numan protein atlas protein atlas human protein atlas human protein atlas human protein atlas protein atlas human protein atlas protein atlas human protein atlas human protein atlas human protein atlas numan protein atlas human protein atlas human 1 human 1 human human human human human Origin numan numan human numan human Fold change between highest expression the second highest ex-Assignment to a cluster according to single-cell expression profile Assignment to a cluster according to bulk RNA expression profile Fold change between highest expression the second highest ex-Fold change between highest expression the second highest ex-Subcellular Localization according to immunocytochemistry/IF Classification of spatial distribution of gene by histochemistry Level of differential expression compared to other tissues The tissues in which the genes is significantly expressed Cluster assignment in blood-derived expression data Cluster assignment in brain-derived expression data Cluster assignment in cell-line expression data Concentration of protein in blood stream across tissues across tissues Description pression pression oression oression pression pression 19016 Size 7590 19784 2238 19784 19784 19784 595 19784 6655 6655 16595 16595 4798 19784 19784 13039 18355 438 2697 19167 19784 19784 19784 19761 19761 13957 4653 3351 2171 categorical categorical categorical categorical categorical categorical categorical categorical multi label categorical categorical categorical categorical categorical categorical categorical categorical categorical multi label categorical categorical regression categorical categorical regression egression regression regression regression regression Type RNA mouse brain regional distribution RNA mouse brain regional specificity RNA single cell type specificity score Blood concentration - conc. blood im RNA brain regional specificity score RNA blood lineage specificity score RNA pig brain regional distribution RNA pig brain regional specificity Cell line expression cluster (HPA) RNA single cell type distribution RNA blood cell specificity score RNA tissue cell type enrichment RNA single cell type specificity RNA brain regional distribution Blood expression cluster (HPA) Brain expression cluster (HPA) RNA blood lineage distribution RNA brain regional specificity RNA cell line specificity score RNA blood lineage specificity Single cell expression cluster RNA tissue specificity score RNA blood cell distribution RNA blood cell specificity RNA cell line distribution **Fissue expression cluster** line specificity RNA tissue distribution **RNA** tissue specificity Subcellular location Task Name **RNA** cell

| 1027 | | | | | | | | | | | | | | | | |
|------|---------|--------|-----------------|---------|------------------------|------|----------|------------------|-------------|--------------|-------------|----------------|--------------|-----------|------------|----|
| 1028 | Table S | S5: | Deta | ailed d | lescrip | tio | n of th | e biologi | cal tas | ks | use | d fo | r be | nch | markir | ıg |
| 1029 | _ | | | | | | | | | | | | | | | |
| 1030 | | | SI | SI | SI | SI | | IS | SI | | SI | SI | S | SI | SI | |
| 1031 | | | atlé | atl | atla | atlâ | | atla | atl | | atlâ | atl | alle | atlâ | atla | |
| 1032 | | | н. | II. | ц. | in. | | E. Q | II. | | ц. | . <u>I</u> I. | E | ш. | .g | |
| 1033 | | | ote | ote | ote | ote | ets | ofe G | ote | | ote | ote | ole | ote | ote | |
| 1034 | | | pr | pr | pr | pr | arg | T/H | pr | ne | pr | pr | pr | pr | pr | |
| 1035 | . | EI | nan | nan | nan | nan | n ti | iPT Jan | nan | .toi | nan | nan | nan | nan | lan | |
| 1036 | | Ē | un | unu | unu | un | be | Sco | unu | eac | un | un | uni | un | un | |
| 1037 | ľ | \sim | بك ر | 1 | 4 | 4 | 0 | | 1 | Ч | <u>ب</u> کر | - ب | 9 | 1 | - | |
| 1038 | | | al | I | Je | μ | sts | A) ar | es | (s | | | × | | | |
| 1030 | | | Gi | ŭ (| n tl | цh | rge | HL./ | typ | vay | , F | | ste | | | |
| 1039 | | | olo | õ | u ii | ed | ı ta | arti | er | thv | nist | se - | ne | ц | | |
| 1040 | | | r bi | 1-2 | Sio | ľov | per | r p: | anc | pa | Jen | ns: | a E | tio | | |
| 1041 | | | ılaı | ПС | es | Idd | e o | nei | 7 Ci | 29 | och | - E | g | cre | | |
| 1042 | | | ticı | 8 | ıdx | ⊀ a | the | , cc o th | , T | Je (| nist | heı | co | .se | | |
| 1043 | | | par | Ĩ. | e V | Ď | mc | lity le t | ıl ir | gei | ž | ot | Se | ılar | | |
| 1044 | | | 1a] | le I | N | dΕ | fr | idi du | 'iva | he | le b | 1 tc | the | ellu | | |
| 1045 | | | nt ir | 1 th | Ř | an | ved | ins | urv | le t | gen | arec | | с С | | |
| 1046 | | | ner | s ii | þ | er, | eriv | om ote | of s | Juc | of | ğ. | SSIC | nt ii | | |
| 1047 | | | ver | ein | les | anc | sde | pr pr | ic c | inc | nc | cor | ore | ner | (<u>f</u> | |
| 1048 | | | vol | rot | gei | ů, | e e | His t to | ost | nat | uti | on (| exi | ven | fai | |
| 1049 | | | Ë. | d (, | $\widehat{\mathbf{O}}$ | ase. | COL | l el P ro | gn | 's tl | rib | ssic | SSI | <u>ío</u> | | |
| 1050 | | | ing | Ð | Q | Sea | 1 SC | f tl niF | pro | vay | list | Jres | gne | iŋ | | |
| 1051 | | | cat | Ŭ | ပ္စ | Ģ | ioi | , U | 3 D D | thv | al c | , idx | Ē | of | Cret | |
| 1052 | | | ndi | int | ent L lir | foi | ciat | ass I by | Jeij | pa | ati | ll e: | čen | ion | sec | |
| 1053 | | | ls i | Jde | cell | sp. | SOC | r cl | orl | vel | Sp | ntia | Me | tati | lar | |
| 1054 | | | Orc | pei | Spe | VOI | as | 1 o ligito | il fe | le | lo | rer | Del | nno | n li | |
| 1055 | | = | уw | de | C de | sy | ene | ass ast | eve u | top | ior | iffe | 60 | lar | ဦ | |
| 1056 | . | tio. | ke | cle | u-, | t k | š | n rds | s) la | ne | cat | fd | u an | nal | g | |
| 1057 | . | Ξ. | rol | c À | SQ CC C | ro | ase | orr voi | aske cr | tor | sifi | - i 10 | Si Cl | itio | site | |
| 1058 | | esc | lin | ell el | ell O | lin | ise | len ey | vid t | eac | las | eve | old rest | nn | le | |
| 1059 | , | | D 5 | E O E | UE | D | , D ta | g d X 8 | 山 山 に | ² | C | Ц, | д | Ē | F | |
| 1060 | | e | 9 | 6 | - | ~ | 6 | 4 - | 0 | 6 | × | ~ ~ | 1 | 9 | | |
| 1061 | | Siz | 6L) | 42 | 63 | 83 | 56 | $^{+}60$ | - | 96 | 58 | 58 | Ŕ | 73 | 92 | |
| 1062 | | | 10 | 1 | 1 | Ś | 411 | 10 | | 10 | 19 | 61 0 | 7 | 2 | | |
| 1063 | | | | | | | 4 | | | | | | | | | |
| 1064 | | | el | | | el | ų | el | cal | e | cal | al | Ę | al | [a] | |
| 1065 | | | lab | ~ | ~ | lab | sic | / lab | DTic | lab | DTiC | . jric | SIC | ij | - I II | |
| 1066 | | be | lti | ıary | lary | lti | res | lary Ilti | ege | ılti | eg G | eg | res | go | eg | |
| 1067 | | 진 | nu | bin | bin | mu | reg | bin mu | cat | mu | cat | cat | leg | cat | cat | |
| 1007 | | | | | | | | | | | | | | | - | |
| 1000 | | | | | | | | | | | | | ore | | | |
| 1009 | | | | | | | ų | | | | ц | | sč | | | |
| 1070 | | | | | | | atic | | S | | tio | ity. | 11 | | | |
| 10/1 | | | s | | | ent |) Cié | s ii n | stic | | ibu | ific | IIIC | nc | g | |
| 10/2 | | | ces. | | | em | ISSC | las tio | gno | | stri | ec. | Sec | ctic | atic | |
| 1073 | | | roc | | ipt | ЫV | ie a | /s c | 301 | | r di | r sp | r sp | 'n | ö | |
| 1074 | | le | al p | ein | scri | лv | eas | ur f | Уp | ~ | lcei | Icel | ice | le f | le l | |
| 1075 | , | ar. | gici | rot | ans | šē i | dis | ass ula | 08. | ays | can | can | car | om | OIC | |
| 1076 | : | K | golo | d p | d tr | eat | ne- | a cl lec | hol | hw | Ā | <u>N</u> | Ā | ret | cret | |
| 1077 | | Las | Bio | ЗС | Cc | Dis | Jer | Hla Mo | Pat | Dati | Z | ZZ Z | Z | Sec | Sec | |
| 1078 | | - | | 0 | 0 | Ι | \cup | | I | Н | щ | | | | | |

| 1081 | | |
|------|--|----------------------------|
| 1001 | | |
| 1002 | | |
| 1083 | | |
| 1084 | | |
| 1085 | | |
| 086 | | |
| 1087 | | |
| 088 | | |
| 089 | | |
| 090 | | |
| 091 | | |
| 092 | Table S6: The list of ton level reactome nathway | names and unique identifie |
| 093 | Table 50. The list of top level reactome pathway | names and unique identifie |
| 094 | D. d. | Dette i tentifican |
| 005 | Pathway name | Pathway identifier |
| 006 | Autophagy | R-HSA-9612973 |
| 090 | Cell Cycle | R-HSA-1640170 |
| 097 | Cell-Cell communication | R-HSA-1500931 |
| 098 | Cellular responses to stimuli | R-HSA-8953897 |
| 099 | Chromatin organization | R-HSA-4839726 |
| 100 | Circadian Clock | R-HSA-400253 |
| 101 | Developmental Biology | R-HSA-1266738 |
| 102 | Digestion and absorption | R-HSA-8963743 |
| 103 | Disease | R-HSA-1643685 |
| 104 | DNA Repair | R-HSA-73894 |
| 105 | DNA Replication | R-HSA-69306 |
| 106 | Drug ADME | R-HSA-9748784 |
| 107 | Extracellular matrix organization | R-HSA-1474244 |
| 102 | Gene expression (Transcription) | R-HSA-74160 |
| 100 | Hemostasis | R-HSA-109582 |
| 109 | Immune System | R-HSA-168256 |
| 110 | Metabolism | R-HSA-1430728 |
| 111 | Metabolism of proteins | R-HSA-392499 |
| 112 | Metabolism of RNA | R-HSA-8953854 |
| 113 | Muscle contraction | R-HSA-397014 |
| 114 | Neuronal System | R-HSA-112316 |
| 115 | Organelle biogenesis and maintenance | R-HSA-1852241 |
| 116 | Programmed Cell Death | R-HSA-5357801 |
| 117 | Protein localization | R-HSA-9609507 |
| 118 | Reproduction | R-HSA-1474165 |
| 119 | Sensory Perception | R-HSA-9709957 |
| 120 | Signal Transduction | R-HSA-162582 |
| 121 | Transport of small molecules | R-HSA-382551 |
| 121 | Vesicle-mediated transport | R-HSA-5653656 |
| 122 | | |
| 123 | | |
| 124 | | |
| 125 | | |
| 126 | | |
| 127 | | |
| 128 | | |
| 129 | | |
| 130 | | |
| 131 | | |
| 132 | | |
| 133 | | |
| | | |

| 135 | | | | | Genon | nic pro | nerties | | | | | | |
|-----|--|--------|--|------|---------|--------------|--|------------------|------------|-------|-------|-------------------|-----|
| 136 | bivalent vs non-methylated - | 0.88 | 0.90 | 0.92 | 0.90 | 0.89 | 0.84 | 0.57 | 0.90 | 0.69 | 0.94 | 0.94 | |
| 137 | Chromosome - | 0.59 | 0.59 | 0.57 | 0.58 | 0.67 | 0.73 | 0.53 | 0.62 | 0.64 | 0.65 | 0.69 | |
| 138 | - dosage sensitive vs insensitive TF - lys4-only-methylated vs non-methylated | 0.90 | 0.85 | 0.87 | 0.89 | 0.83 | 0.92 | 0.86 | 0.96 | 0.89 | 0.98 | 0.90 | |
| 130 | - Protein class | 0.76 | 0.76 | 0.71 | 0.92 | 0.56 | 0.56 | 0.54 | 0.54 | 0.53 | 0.94 | 0.52 | |
| 139 | | | | | Prote | n prop | erties | | | | | | |
| 140 | UniProt keyword Domain - | 0.84 | 0.83 | 0.77 | 0.81 | 0.62 | 0.60 | 0.58 | 0.64 | 0.59 | 0.91 | 0.53 | |
| 141 | UniProt keyword Ligand - UniProt keyword PTM - | 0.85 | 0.83 | 0.74 | 0.79 | 0.56 | 0.55 | 0.55 | 0.57 | 0.56 | 0.84 | 0.52 | |
| 142 | of the Reyword 1 th | 0.71 | 0.72 | 0.07 | Regula | torv fu | nctions | 0.01 | 0.07 | 0.01 | 0.77 | 0.57 | |
| 143 | Gene2Gene - | 0.95 | 0.86 | 0.82 | 0.86 | 0.79 | 0.71 | 0.73 | 0.75 | 0.71 | 0.89 | 0.69 | |
| 144 | long vs short range TF - | 0.60 | 0.75 | 0.57 | 0.51 | 0.65 | 0.71 | 0.70 | 0.77 | 0.67 | 0.70 | 0.50 | |
| 1/5 | NI hetwork - N1 targets - | 0.63 | 0.83 | 0.75 | 0.67 | 0.80 | 0.82 | 0.71 | 0.79 | 0.80 | 0.75 | 0.74 | - 1 |
| 145 | TF vs non-TF - | 0.96 | 0.93 | 0.93 | 0.94 | 0.74 | 0.76 | 0.68 | 0.74 | 0.75 | 0.96 | 0.77 | |
| 146 | | | | | Biologi | cal pro | cesses | | | | | | |
| 147 | Biological process - | 0.82 | 0.81 | 0.73 | 0.77 | 0.55 | 0.56 | 0.54 | 0.58 | 0.55 | 0.71 | 0.52 | |
| 148 | CCD Protein - CCD Transcript - | 0.57 | 0.59 | 0.61 | 0.60 | 0.56 | 0.56 | 0.54 | 0.53 | 0.56 | 0.63 | 0.54 | |
| 149 | Disease involvement - | 0.69 | 0.67 | 0.62 | 0.64 | 0.55 | 0.55 | 0.54 | 0.56 | 0.54 | 0.59 | 0.50 | |
| 150 | HLA class I vs class II - | 0.82 | 0.83 | 0.64 | 0.85 | 0.80 | 0.81 | 0.87 | 0.93 | 0.74 | 0.90 | 0.65 | |
| 154 | Molecular function - | 0.89 | 0.87 | 0.75 | 0.82 | 0.53 | 0.54 | 0.53 | 0.56 | 0.53 | 0.80 | 0.52 | |
| ICI | Pathology prognostics - Breast cancer - | 0.55 | 0.55 | 0.59 | 0.53 | 0.66 | 0.64 | 0.55 | 0.64 | 0.60 | 0.56 | 0.55 | |
| 52 | Pathology prognostics - Cervical cancer - Pathology prognostics - Colorectal cancer - | - 0.52 | 0.50 | 0.57 | 0.52 | 0.64 | 0.65 | 0.59 | 0.59 | 0.60 | 0.59 | 0.59 | |
| 53 | Pathology prognostics - Endometrial cancer - | 0.52 | 0.54 | 0.53 | 0.52 | 0.65 | 0.64 | 0.56 | 0.62 | 0.63 | 0.54 | 0.58 | |
| 54 | Pathology prognostics - Glioma - | 0.53 | 0.51 | 0.56 | 0.56 | 0.65 | 0.65 | 0.57 | 0.63 | 0.57 | 0.57 | 0.57 | |
| 55 | Pathology prognostics - Head and neck cancer - | 0.58 | 0.54 | 0.57 | 0.52 | 0.69 | 0.65 | 0.59 | 0.64 | 0.66 | 0.56 | 0.56 | |
| 56 | Pathology prognostics - Liver cancer - Pathology prognostics - Lung cancer - | 0.59 | 0.66 | 0.68 | 0.65 | 0.80 | 0.81 | 0.74 | 0.78 | 0.78 | 0.69 | 0.70 | |
| 00 | Pathology prognostics - Melanoma - | 0.59 | 0.60 | 0.66 | 0.57 | 0.69 | 0.68 | 0.57 | 0.62 | 0.68 | 0.67 | 0.61 | |
| 57 | Pathology prognostics - Ovarian cancer - | 0.52 | 0.53 | 0.55 | 0.55 | 0.67 | 0.64 | 0.56 | 0.60 | 0.64 | 0.59 | 0.56 | |
| 58 | Pathology prognostics - Pancreatic cancer - | 0.55 | 0.55 | 0.56 | 0.53 | 0.76 | | 0.60 | 0.72 | 0.70 | 0.58 | 0.59 | |
| 59 | Pathology prognostics - Prostate cancer - | 0.60 | 0.59 | 0.68 | 0.60 | 0.67 | 0.69 | 0.61 | 0.60 | 0.73 | 0.65 | 0.61 | |
| 60 | Pathology prognostics - Renar cancer - | 0.55 | 0.59 | 0.59 | 0.59 | 0.74 | 0.75 | 0.59 | 0.74 | 0.70 | 0.60 | 0.56 | |
| 61 | Pathology prognostics - Testis cancer - | 0.61 | 0.58 | 0.70 | 0.58 | 0.76 | 0.81 | 0.53 | 0.56 | 0.68 | 0.70 | 0.71 | |
| 01 | Pathology prognostics - Thyroid cancer - | 0.55 | 0.53 | 0.57 | 0.49 | 0.62 | 0.55 | 0.52 | 0.56 | 0.56 | 0.55 | 0.53 | |
| 62 | Pathology prognostics - Urothelial cancer - | 0.54 | 0.55 | 0.57 | 0.54 | 0.68 | 0.67 | 0.60 | 0.66 | 0.66 | 0.59 | 0.54 | |
| 63 | Pathways - RNA cancer distribution - | 0.79 | 0.75 | 0.66 | 0.72 | 0.55 | 0.55 | 0.55 | 0.58 | 0.54 | 0.67 | 0.50 | - (|
| 64 | RNA cancer specificity - | - 0.72 | | | | 0.90 | 0.91 | 0.81 | 0.90 | 0.89 | 0.80 | 0.79 | |
| 65 | Secretome function - | 0.93 | 0.92 | 0.92 | 0.88 | 0.68 | 0.67 | 0.57 | 0.63 | 0.66 | 0.90 | 0.62 | |
| 66 | Secretome location - | 0.88 | 0.87 | 0.87 | 0.83 | 0.78 | 0.77 | 0.62 | 0.71 | 0.73 | 0.86 | 0.66 | |
| 67 | Rlood expression cluster - | 0.55 | 0.56 | 0.60 | Lo | calizat | ion 0.76 | 0.72 | 0.71 | 0.73 | 0.61 | 0.60 | |
| C0 | Brood expression cluster - Brain expression cluster - | 0.59 | 0.60 | 0.64 | 0.59 | 0. <u>80</u> | 0.79 | 0.62 | 0.80 | 0.70 | 0.67 | 0.66 | |
| ÖÖ | Cell line expression cluster - | 0.63 | 0.63 | 0.68 | 0.62 | 0.81 | 0.81 | 0.66 | 0.74 | 0.78 | 0.69 | 0.68 | |
| 69 | RNA blood cell distribution - | 0.66 | 0.71 | 0.72 | 0.70 | 0.88 | 0.91 | 0.85 | 0.89 | 0.88 | 0.75 | 0.72 | |
| 70 | RNA blood cell specificity - | 0.68 | 0.75 | 0.76 | 0.74 | 0.89 | 0.91 | 0.86 | 0.89 | 0.89 | 0.79 | 0.77 | |
| 71 | RNA blood lineage distribution - RNA blood lineage specificity - | 0.69 | 0.74 | 0.74 | 0.75 | 0.90 | 0.92 | 0.86 | 0.91 | 0.89 | 0.78 | 0.78 | |
| 72 | RNA brain regional distribution - | 0.75 | 0.79 | 0.79 | | 0.91 | 0.93 | 0.76 | 0.90 | 0.89 | 0.83 | 0.80 | |
| 72 | RNA brain regional specificity - | 0.73 | | 0.79 | | 0.89 | 0.91 | 0.71 | 0.86 | 0.86 | 0.82 | 0.79 | |
| 13 | RNA cell line distribution - | 0.72 | | | | 0.90 | 0.93 | 0.85 | 0.91 | 0.92 | 0.81 | 0.82 | - (|
| 74 | - RNA cell line specificity - RNA mouse brain regional distribution | 0.71 | 0.77 | 0.78 | 0.76 | 0.89 | 0.90 | 0.82 | 0.90 | 0.90 | 0.81 | 0.81 | |
| 75 | RNA mouse brain regional specificity - | 0.73 | | | | 0.90 | 0.92 | 0.77 | 0.90 | 0.88 | 0.83 | 0.80 | |
| 76 | RNA pig brain regional distribution - | 0.71 | | | 0.73 | 0.88 | 0.90 | 0.77 | 0.89 | 0.86 | 0.81 | 0.78 | |
| 77 | RNA pig brain regional specificity - | 0.72 | 0.74 | 0.77 | 0.72 | 0.87 | 0.88 | 0.74 | 0.85 | 0.85 | 0.81 | 0.78 | |
| 78 | RNA single cell type distribution - | 0.68 | 0.73 | 0.73 | 0.72 | 0.89 | 0.94 | 0.84 | 0.92 | 0.89 | 0.78 | 0.76 | |
| 70 | - RNA single cell type specificity RNA tissue cell type enrichment | - 0.55 | 0.65 | 0.67 | 0.53 | 0.59 | 0.82 | 0.70 | 0.64 | 0.56 | 0.68 | 0.50 | |
| 19 | - RNA tissue distribution | 0.71 | 0.76 | 0.76 | 0.74 | 0.90 | 0.93 | 0.85 | 0.92 | 0.91 | 0.80 | 0.79 | |
| 80 | RNA tissue specificity - | 0.66 | 0.72 | 0.74 | 0.71 | 0.84 | 0.86 | | 0.85 | 0.83 | 0.75 | 0.74 | |
| 81 | Single cell expression cluster - | 0.67 | 0.67 | 0.71 | 0.65 | 0.90 | 0.92 | 0.75 | 0.89 | 0.79 | 0.72 | 0.66 | |
| 82 | Subcellular location - | 0.58 | 0.60 | 0.57 | 0.60 | 0.54 | 0.54 | 0.53 | 0.55 | 0.53 | 0.59 | 0.52 | |
| 192 | lissue expression cluster - | 80.0 | 80.0 | 0.73 | 0.66 | 0.86 | -0.88 | 0.08 | 0.83 | 0.80 | 0.74 | 0.07 | - (|
| 00 | | 18th | and and a second | Net | | 15 . JN | a contraction of the second se | » م ^ر | s'x | -Je | 3 | ر _{يم} د | |
| | | ~~ | ~~ | NX · | .10 | 111 | · · · · | (Å | <u>(</u> X | | | 202 | |
| 84 | | WIE | WIL | Put. | orno | celli | eneton | SCOR | scor , | Senel | ESMIL | NABEI | |



| 1189 | | | | | | | | | | | | | | | |
|------|------|--|--------|------|---------|-------|--------|--------|-----------|-------|-------|------|------------|-------|----------|
| 1190 | | | | | | | | | | | | | | | |
| 1191 | | | | | | | | | | | | | | | |
| 1192 | | | | | | | | | | | | | | | |
| 1193 | | | | | | | | | | | | | | | |
| 1194 | | | | | | | | | | | | | | | |
| 1195 | | | | | | | | | | | | | | | |
| 1196 | | | | | | | | | | | | | | | |
| 1107 | | | | | | | | | | | | | | | |
| 1108 | | | | | | | | | | | | | | | |
| 1100 | | | | | | | | | | | | | | | |
| 1200 | | | | | | | | | | | | | | | |
| 1200 | | | | | | | | | | | | | | | |
| 1201 | | | | | | | | | | | | | | | |
| 1202 | | | | | | | | | | | | | | | |
| 1203 | | | | | | | | | | | | | | | |
| 1204 | | Protein class-FDA approved drug targets - | 0.93 | 0.92 | 0.90 | 0.88 | 0.71 | 0.71 | 0.60 | 0.70 | 0.70 | 0.87 | 0.65 | | 1.0 |
| 1205 | | Protein class-Predicted membrane proteins - | 0.88 | 0.90 | 0.88 | 0.90 | | 0.73 | 0.66 | 0.77 | 0.68 | 0.96 | 0.67 | | |
| 1206 | | Protein class-Human disease related genes - | 0.83 | 0.87 | 0.86 | 0.86 | 0.64 | 0.62 | 0.56 | 0.63 | 0.61 | 0.69 | 0.58 | | |
| 1207 | | Protein class-Transcription factors - | 0.99 | 0.99 | 0.98 | 0.98 | 0.81 | 0.82 | 0.70 | 0.83 | 0.80 | 0.99 | 0.81 | - | 0.9 |
| 1208 | | Protein class-Predicted intracellular proteins - | 0.84 | 0.89 | 0.87 | 0.87 | | 0.78 | 0.70 | 0.80 | 0.73 | 0.96 | 0.72 | | |
| 1209 | | Protein class-Cancer-related genes - | 0.78 | 0.81 | 0.81 | | 0.73 | 0.72 | 0.69 | 0.72 | 0.71 | 0.76 | 0.60 | | |
| 1210 | | Protein class-CD markers - | 0.98 | 0.97 | 0.97 | 0.96 | 0.91 | 0.90 | 0.83 | 0.90 | 0.87 | 0.98 | 0.69 | | 0.0 |
| 1211 | ylic | Protein class-Plasma proteins - | 0.66 | 0.73 | 0.73 | 0.71 | 0.71 | 0.72 | 0.68 | 0.73 | 0.69 | 0.77 | 0.60 | | 0.0 Y |
| 1212 | Fan | Protein class-Ribosomal proteins - | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 0.94 | 0.97 | 0.99 | 1.00 | 0.82 | | IN AI |
| 1213 | lask | Protein class-Predicted secreted proteins - | 0.96 | 0.95 | 0.94 | 0.94 | 0.86 | 0.86 | 0.80 | 0.87 | 0.82 | 0.97 | 0.77 | | Mea |
| 1214 | | Protein class-Disease related genes - | 0.80 | 0.85 | 0.84 | 0.85 | 0.63 | 0.62 | 0.56 | 0.63 | 0.60 | 0.69 | 0.59 | - | 0.7 |
| 1215 | | Protein class-Metabolic proteins - | 0.95 | 0.95 | 0.94 | 0.93 | 0.71 | 0.71 | 0.63 | 0.72 | 0.69 | 0.93 | 0.61 | | |
| 1216 | | Protein class-Potential drug targets - | 0.82 | 0.86 | 0.85 | 0.83 | 0.67 | 0.66 | 0.60 | 0.65 | 0.64 | 0.80 | 0.60 | | |
| 1217 | | Protein class-RAS pathway related proteins - | 0.98 | 0.97 | 0.95 | 0.92 | 0.67 | 0.62 | 0.60 | 0.63 | 0.73 | 0.94 | 0.64 | - | 0.6 |
| 1218 | | Protein class-G-protein coupled receptors - | 1.00 | 1.00 | 1.00 | 1.00 | 0.79 | 0.87 | 0.64 | 0.84 | 0.87 | 1.00 | 0.86 | | |
| 1219 | | Protein class-Transporters - | 0.88 | 0.89 | 0.89 | 0.87 | 0.68 | 0.66 | 0.59 | 0.68 | 0.66 | 0.90 | 0.60 | | |
| 1220 | | Protein class-Enzymes - | 0.96 | 0.96 | 0.94 | 0.93 | 0.66 | 0.67 | 0.60 | 0.66 | 0.65 | 0.96 | 0.63 | | 0.5 |
| 1221 | | | ~~ | Ś | . let | | 85 JN | ' | × ^^ ا | × | . '.e | , . | ໍ່.າ | , - | 0.5 |
| 1222 | | | MILL | MILL | ANPIT . | of WO | cellPt | etorn | yce? | 4CR | enely | SN?? | NABert | | |
| 1223 | | | | | \$ | ৵ | Ċ | et. | | | 6 | < | <u>)</u> , | | |
| 1224 | | | | | | | | | | | | | | | |
| 1225 | Fig | ure S3: Model performance meas | ured | by m | nean . | AUC | for b | oinary | / task | s dei | rived | from | the 1 | nulti | ilabel |
| 1226 | tasl | k 'protein class' | | • | | | | | | | | | | | |
| 1227 | | | | | | | | | | | | | | | |
| 1227 | | | | | | | | | | | | | | | |
| 1220 | | | | | | | | | | | | | | | |
| 1229 | | | | | | | | | | | | | | | |
| 1230 | | | | | | | | | | | | | | | |
| 1231 | | | | | | | | | | | | | | | |
| 1232 | | | | | | | | | | | | | | | |
| 1233 | | | | | | | | | | | | | | | |
| 1234 | | | | | | | | | | | | | | | |
| 1235 | | | | | | | | | | | | | | | |
| 1236 | | | | | | | | | | | | | | | |
| 1237 | | | | | | | | | | | | | | | |
| 1238 | | | | | | | | | | | | | | | |
| 1239 | | | | | | | | | | | | | | | |
| 1240 | | | | | | | | | | | | | | | |
| 1241 | | | | | | | | | | | | | | | |

| 1242 | | | | | | | | | | | | | | | |
|------|-----------|--|------|--------------------|------|---------|----------|---------|------|--------|--------|--------------------|--------|---|-------|
| 1243 | | Biological process-Innate immunity - | 0.87 | 0.84 | 0.87 | 0.82 | 0.76 | 0.76 | 0.68 | 0.72 | 0.73 | 0.85 | 0.64 | | - 1.0 |
| 1244 | | Biological process-Antiviral defense - | 0.90 | 0.92 | 0.93 | 0.90 | 0.83 | 0.83 | 0.68 | 0.69 | 0.78 | 0.89 | 0.62 | | |
| 1245 | | Biological process-DNA repair - | 0.94 | 0.94 | 0.95 | 0.92 | 0.79 | 0.79 | 0.72 | 0.78 | 0.77 | 0.91 | 0.68 | | |
| 1246 | F | Biological process-Wht signaling pathway - | 0.92 | 0.88 | 0.90 | 0.87 | 0.69 | 0.71 | 0.60 | 0.64 | 0.66 | 0.86 | 0.66 | | |
| 1247 | в | iological process-Transcription regulation - | 0.92 | 0.00 | 0.90 | 0.07 | 0.78 | 0.71 | 0.68 | 0.78 | 0.00 | 0.00 | 0.75 | | |
| 1248 | D | Biological process-manscription regulation - | 0.90 | 0.90 | 0.90 | 0.95 | 0.70 | 0.77 | 0.00 | 0.70 | 0.74 | 0.90 | 0.75 | | |
| 1249 | | Biological process-biological mythins - | 0.87 | 0.84 | 0.87 | 0.70 | 0.68 | 0.73 | 0.60 | 0.69 | 0.05 | 0.84 | 0.63 | | |
| 1250 | | Biological process-Host-virus interaction - | 0.76 | 0.76 | 0.79 | 0.72 | 0.74 | 0.73 | 0.67 | 0.70 | 0.74 | 0.76 | 0.60 | | |
| 1251 | | Biological process-mRNA transport - | 0.90 | 0.93 | 0.93 | 0.92 | 0.78 | 0.79 | 0.68 | 0.70 | 0.81 | 0.92 | 0.73 | | |
| 1252 | | Biological process-Lipid metabolism - | 0.97 | 0.97 | 0.97 | 0.95 | 0.77 | | 0.64 | 0.76 | 0.73 | 0.96 | 0.64 | - | - 0.9 |
| 1253 | Biologica | al process-Cilium biogenesis/degradation - | 0.92 | 0.91 | 0.90 | 0.91 | 0.83 | 0.84 | 0.68 | 0.74 | 0.76 | 0.89 | 0.59 | | |
| 1254 | | Biological process-DNA damage - | 0.91 | 0.92 | 0.93 | 0.90 | 0.77 | 0.78 | 0.68 | 0.74 | | 0.91 | 0.67 | | |
| 1255 | | Biological process-Cell cycle - | 0.88 | 0.89 | 0.89 | 0.85 | 0.74 | 0.75 | 0.66 | 0.72 | 0.73 | 0.86 | 0.64 | | |
| 1256 | | Biological process-Olfaction - | 0.97 | 0.98 | 0.98 | 0.98 | 0.83 | 0.92 | 0.57 | 0.86 | 0.97 | 0.98 | 0.94 | | |
| 1257 | | Biological process-Autophagy - | 0.90 | 0.87 | 0.87 | 0.80 | 0.71 | 0.67 | 0.61 | 0.68 | 0.67 | 0.82 | 0.57 | | |
| 1258 | | Biological process-Translation regulation - | 0.93 | 0.91 | 0.91 | 0.85 | 0.68 | 0.69 | 0.60 | 0.66 | 0.73 | 0.89 | 0.70 | | |
| 1259 | | Biological process-Lipid biosynthesis - | 0.97 | 0.97 | 0.96 | 0.93 | 0.72 | 0.74 | 0.62 | | 0.72 | 0.94 | 0.57 | | |
| 1260 | | Biological process-Symport - | 1.00 | 0.99 | 0.99 | 0.98 | 0.76 | 0.80 | 0.55 | 0.79 | 0.91 | 1.00 | 0.66 | | |
| 1201 | | Biological process-Neurogenesis - | 0.83 | 0.84 | 0.87 | 0.80 | 0.75 | | 0.59 | 0.71 | 0.69 | 0.80 | 0.67 | | - 0.8 |
| 1202 | | Biological process-Lipid transport - | 0.96 | 0.96 | 0.96 | 0.95 | 0.72 | 0.71 | 0.54 | 0.63 | 0.72 | 0.93 | 0.63 | | |
| 1203 | В | iological process-Inflammatory response - | 0.91 | 0.90 | 0.91 | 0.89 | 0.80 | | 0.68 | 0.77 | | 0.89 | 0.66 | | |
| 1204 | | Biological process-Sensory transduction - | 0.98 | 0.97 | 0.97 | 0.95 | 0.83 | 0.86 | 0.62 | | 0.84 | 0.88 | 0.81 | | |
| 1265 | Mily | Biological process-Transcription - | 0.95 | 0.96 | 0.96 | 0.95 | 0.78 | | 0.69 | | 0.73 | 0.95 | 0.74 | | AUC |
| 1267 | ж Б | Biological process-Fatty acid metabolism - | 0.97 | 0.98 | 0.98 | 0.95 | 0.76 | 0.73 | 0.56 | 0.72 | 0.80 | 0.95 | 0.60 | | ean |
| 1268 | Ta | Biological process-Ion transport - | 0.99 | 0.98 | 0.98 | 0.98 | 0.77 | | 0.67 | | | 0.98 | 0.71 | | Σ |
| 1269 | | Biological process-Spermatogenesis - | 0.80 | 0.80 | 0.86 | 0.85 | 0.73 | 0.82 | 0.59 | 0.68 | 0.68 | 0.78 | 0.57 | | |
| 1270 | | Biological process-Potassium transport - | 0.99 | 1.00 | 0.99 | 0.98 | 0.76 | 0.76 | 0.65 | 0.76 | 0.89 | 0.99 | 0.74 | | |
| 1271 | | Biological process-Apoptosis - | 0.85 | 0.85 | 0.86 | 0.82 | 0.64 | 0.62 | 0.58 | 0.59 | 0.64 | 0.77 | 0.56 | | - 0.7 |
| 1272 | Bio | logical process-Ubl conjugation pathway - | 0.97 | 0.97 | 0.96 | 0.95 | 0.68 | 0.72 | 0.64 | 0.69 | 0.68 | 0.97 | 0.64 | | |
| 1273 | | Biological process-Sodium transport - | 1.00 | 0.99 | 0.99 | 0.99 | 0.79 | 0.76 | 0.65 | 0.73 | 0.85 | 0.99 | 0.67 | | |
| 1274 | | Biological process-Mitosis - | 0.90 | 0.91 | 0.92 | 0.86 | 0.78 | 0.78 | 0.70 | | 0.78 | 0.89 | 0.68 | | |
| 1275 | | Biological process-Protein transport - | 0.93 | 0.93 | 0.92 | 0.92 | 0.74 | 0.72 | 0.63 | 0.74 | 0.72 | 0.93 | 0.66 | | |
| 1276 | | Biological process-Protein biosynthesis | 0.55 | 0.95 | 0.92 | 0.92 | 0.81 | 0.72 | 0.05 | 0.74 | 0.72 | 0.95 | 0.71 | | |
| 1277 | | Biological process Coll division | 0.50 | 0.00 | 0.00 | 0.96 | 0.01 | 0.04 | 0.75 | 0.75 | 0.05 | 0.50 | 0.66 | | |
| 1278 | | Biological process-Cell division - | 0.00 | 0.91 | 0.90 | 0.00 | 0.75 | 0.77 | 0.00 | 0.72 | 0.70 | 0.00 | 0.00 | | |
| 1279 | | Biological process-Steroid metabolism - | 0.95 | 0.95 | 0.95 | 0.92 | 0.74 | 0.77 | 0.00 | 0.71 | 0.72 | 0.90 | 0.55 | | |
| 1280 | | Biological process-Endocytosis - | 0.93 | 0.92 | 0.91 | 0.88 | 0.67 | 0.72 | 0.62 | 0.73 | 0.63 | 0.92 | 0.63 | | - 0.6 |
| 1281 | | Biological process-mRNA splicing - | 0.95 | 0.95 | 0.95 | 0.94 | 0.85 | 0.85 | 0.74 | 0.79 | 0.83 | 0.96 | 0.77 | | |
| 1282 | | Biological process-Vision - | 0.95 | 0.96 | 0.98 | 0.90 | 0.83 | 0.86 | 0.59 | 0.78 | | | 0.67 | | |
| 1283 | | Biological process-Angiogenesis - | 0.87 | 0.85 | 0.90 | 0.83 | 0.71 | 0.72 | 0.62 | 0.73 | 0.77 | 0.81 | 0.66 | | |
| 1284 | | Biological process-Differentiation - | 0.78 | 0.79 | 0.81 | 0.77 | 0.68 | 0.69 | 0.59 | 0.62 | 0.63 | 0.75 | 0.62 | | |
| 1285 | | Biological process-mRNA processing - | 0.96 | 0.96 | 0.95 | 0.94 | 0.84 | 0.82 | | 0.80 | 0.81 | 0.96 | 0.76 | | |
| 1286 | | Biological process-Immunity - | 0.88 | 0.87 | 0.89 | 0.85 | 0.80 | 0.80 | 0.72 | | | 0.86 | 0.67 | | |
| 1287 | | Biological process-Cell adhesion - | 0.95 | 0.95 | 0.95 | 0.93 | 0.82 | 0.83 | 0.74 | 0.82 | 0.82 | 0.95 | 0.74 | | |
| 1288 | | Biological process-Adaptive immunity - | 0.91 | 0.92 | 0.93 | 0.89 | 0.86 | 0.86 | 0.77 | 0.84 | 0.85 | 0.89 | 0.72 | | |
| 1289 | | Biological process-Transport - | 0.95 | 0.95 | 0.94 | 0.94 | 0.70 | 0.70 | 0.60 | 0.69 | 0.68 | 0.94 | 0.60 | | - 0 5 |
| 1290 | | | | . (8 ⁵⁾ | oNet | کہ | 05 101 M | , ne | 2 S | , AN | , ver | , ' ₃ s | s kir | | 0.0 |
| 1291 | | | MIL | MIL | ly, | .00t NU | cell. | enetori | SCO | scor , | sener, | ESM. K | ONABE. | | |
| 1292 | | | | | Ś | ð~' | Ģ | > | | | | | - | | |
| 1293 | | | | | | | | | | | | | | | |

Figure S4: Model performance measured by mean AUC for binary tasks derived from the multi label task 'biological process'

| 1297 | | | | | | | | | | | | | | |
|-------|-------|--|------|------|------------|--------------|---------|---------|------|-------------------|---------|--------|--------|------------|
| 1298 | | | | | | | | | | | | | | |
| 1299 | | | | | | | | | | | | | | |
| 1300 | | | | | | | | | | | | | | |
| 1301 | | | | | | | | | | | | | | |
| 1302 | | | | | | | | | | | | | | |
| 1303 | | | | | | | | | | | | | | |
| 1204 | | | | | | | | | | | | | | |
| 1205 | | | | | | | | | | | | | | |
| 1305 | | | | | | | | | | | | | | |
| 1300 | | | | | | | | | | | | | | |
| 1000 | | | | | | | | | | | | | | |
| 1308 | | | | | | | | | | | | | | |
| 1309 | | | | | | | | | | | | | | |
| 1310 | | | | | | | | | | | | | | |
| 1311 | | | | | | | | | | | | | | |
| 1312 | | | | | | | | | | | | | | - 1.0 |
| 1313 | | Disease involvement-FDA approved drug targets Disease involvement-Neuropathy | 0.90 | 0.88 | 0.87 | 0.85 | 0.68 | 0.68 | 0.59 | 0.66 | 0.68 | 0.86 | 0.65 | |
| 1314 | | Disease involvement-Hereditary spastic paraplegia | 0.82 | 0.89 | 0.85 | 0.72 | 0.72 | 0.68 | 0.57 | 0.63 | 0.63 | 0.78 | 0.59 | |
| 1315 | | Disease involvement-Disease variant | 0.70 | 0.74 | | 0.71 | 0.62 | 0.60 | 0.55 | 0.60 | 0.59 | 0.69 | 0.56 | - 0.0 |
| 1316 | | Disease involvement-Deamess | 0.76 | 0.73 | 0.76 | 0.85 | 0.56 | 0.56 | 0.50 | 0.52 | 0.54 | 0.82 | 0.52 | - 0.9 |
| 1317 | | Disease involvement-Non-syndromic deafness | 0.86 | 0.84 | 0.84 | 0.70 | 0.56 | 0.57 | 0.43 | 0.56 | 0.61 | 0.70 | 0.59 | |
| 1318 | | Disease involvement-Obesity Disease involvement-Primary mitochondrial disease | 0.92 | 0.86 | 0.82 | 0.78 | 0.74 | 0.58 | 0.54 | 0.61 | 0.58 | 0.68 | 0.62 | |
| 1319 | ₹ | Disease involvement-Epilepsy | 0.74 | 0.76 | 0.80 | 0.69 | 0.74 | 0.75 | 0.58 | 0.63 | 0.62 | 0.73 | 0.57 | - 0.8 y |
| 1320 | Fan | Disease involvement-Charcot-Marie-Tooth disease | 0.95 | 0.93 | 0.94 | 0.92 | 0.90 | 0.88 | 0.69 | 0.87 | 0.83 | 0.90 | 0.63 | an Al |
| 1321 | Task | Disease involvement-Neurodegeneration | 0.82 | 0.80 | 0.81 | 0.68 | 0.68 | 0.67 | 0.60 | 0.62 | 0.63 | 0.68 | 0.58 | Mea |
| 1322 | | Disease involvement-Intellectual disability | 0.75 | 0.74 | 0.77 | 0.73 | 0.76 | 0.76 | 0.64 | 0.72 | 0.69 | 0.77 | 0.65 | - 0.7 |
| 1323 | | Disease involvement-Cancer-related genes Disease involvement-Cardiomyopathy | 0.79 | 0.80 | 0.81 | 0.77 | 0.72 | 0.70 | 0.66 | 0.68 | 0.68 | 0.75 | 0.59 | |
| 1324 | | Disease involvement-Tumor suppressor | 0.82 | 0.83 | 0.84 | 0.80 | 0.62 | 0.61 | 0.55 | 0.58 | 0.61 | 0.72 | 0.54 | |
| 1325 | | Disease involvement-Autism spectrum disorder | 0.75 | 0.66 | 0.78 | 0.60 | 0.77 | 0.69 | 0.53 | 0.71 | 0.71 | 0.76 | 0.66 | - 0.6 |
| 1326 | | Disease involvement-Cataract | 0.80 | 0.75 | 0.81 | 0.67 | 0.64 | 0.68 | 0.63 | 0.66 | 0.59 | 0.69 | 0.50 | |
| 1327 | | Disease involvement-Diabetes mellitus | 0.88 | 0.88 | 0.86 | 0.77 | 0.68 | 0.66 | 0.51 | 0.54 | 0.52 | 0.61 | 0.56 | |
| 1328 | | Disease involvement-Dwarfism | 0.79 | 0.79 | 0.79 | 0.76 | 0.66 | 0.65 | 0.54 | 0.56 | 0.58 | 0.68 | 0.53 | - 0.5 |
| 1329 | | | 1HB) | 1483 | MPNet | NOT | 92 IBIN | am | * A* | CAL'H | 2ver | | , seti | |
| 1330 | | | tr. | 4. | <i>A</i> . | and a second | се С | enero | SC | 5 ⁰⁰ , | serre . | ESN' V | JNAL | |
| 1331 | | | | | ¢ | ,0 | | | | | | | | |
| 1332 | Eig | ura \$5. Madal parformanaa maasura | dhu | maa | n 11 | IC fo | r hir | 0.000 1 | ocko | dari | und f | rom | tha m | ulti labal |
| 1333 | - Fig | disease involvement' | u by | mea | II AU | | | iary i | asks | uerr | veu i | 10111 | the m | uni naber |
| 1334 | tas | C disease involvement | | | | | | | | | | | | |
| 1335 | | | | | | | | | | | | | | |
| 1336 | | | | | | | | | | | | | | |
| 1337 | | | | | | | | | | | | | | |
| 1338 | | | | | | | | | | | | | | |
| 1220 | | | | | | | | | | | | | | |
| 12/0 | | | | | | | | | | | | | | |
| 12/14 | | | | | | | | | | | | | | |
| 1041 | | | | | | | | | | | | | | |
| 1342 | | | | | | | | | | | | | | |
| 1343 | | | | | | | | | | | | | | |
| 1344 | | | | | | | | | | | | | | |
| 1345 | | | | | | | | | | | | | | |
| 1346 | | | | | | | | | | | | | | |
| 1347 | | | | | | | | | | | | | | |
| 1348 | | | | | | | | | | | | | | |

| 1350 | | | | | | | | | | | | | 1.0 |
|-------|--|------------------|------|-------|--|-----------|--------|------|----------------|---------|------|------|-------|
| 1351 | Molecular function-Ribonucleoprotein - | 0.98 | 0.98 | 0.98 | 0.97 | 0.93 | 0.92 | 0.84 | 0.88 | 0.91 | 0.97 | 0.81 | - 1.0 |
| 1352 | Molecular function-Protease - | 0.98 | 0.98 | 0.96 | 0.95 | 0.64 | 0.68 | 0.59 | 0.67 | 0.61 | 0.98 | 0.60 | |
| 1353 | Molecular function-Isomerase - | 0.96 | 0.96 | 0.92 | 0.89 | 0.70 | 0.68 | 0.58 | 0.68 | 0.70 | 0.95 | 0.60 | |
| 1354 | Molecular function-Nuclease | 0.97 | 0.98 | 0.96 | 0.89 | 0.63 | 0.65 | 0 59 | 0.55 | 0.66 | 0.96 | 0.62 | |
| 1355 | Molecular function C protein coupled recenter | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.03 | 0.55 | 0.00 | 0.00 | 1.00 | 0.96 | |
| 1356 | Molecular function-G-protein coupled receptor - | 1.00 | 1.00 | 1.00 | 1.00 | 0.80 | 0.67 | 0.65 | 0.65 | 0.67 | 1.00 | 0.00 | |
| 1357 | Molecular function-Calmodulin-binding - | 0.94 | 0.92 | 0.92 | 0.86 | 0.71 | 0.65 | 0.52 | 0.68 | 0.71 | 0.93 | 0.66 | |
| 1358 | Molecular function-Chromatin regulator - | 0.93 | 0.92 | 0.93 | 0.90 | 0.79 | 0.79 | 0.71 | | 0.77 | 0.94 | 0.70 | |
| 1359 | Molecular function-Serine/threonine-protein kinase - | 1.00 | 0.99 | 0.99 | 0.98 | 0.73 | 0.68 | 0.60 | 0.68 | 0.70 | 0.99 | 0.63 | |
| 1361 | Molecular function-RNA-binding | 0.94 | 0.94 | 0.95 | 0.92 | 0.80 | | 0.72 | | | 0.95 | 0.75 | - 0.9 |
| 1362 | Molecular function-Metalloprotease - | 0.97 | 0.98 | 0.96 | 0.92 | 0.63 | 0.71 | 0.56 | 0.65 | 0.71 | 0.99 | 0.65 | |
| 1363 | Molecular function-Serine protease - | 0.99 | 1.00 | 0.98 | 0.97 | 0.76 | | 0.65 | | | 0.98 | 0.73 | |
| 1364 | Molecular function-Motor protein - | 1.00 | 1.00 | 1.00 | 0.98 | 0.81 | | 0.72 | 0.79 | 0.88 | 0.99 | 0.59 | |
| 1365 | Molecular function-DNA-binding | 0.97 | 0.97 | 0.97 | 0.96 | 0.79 | 0.78 | 0.69 | 0.79 | 0.76 | 0.97 | 0.75 | |
| 1366 | Molecular function-Transferase - | 0.97 | 0.97 | 0.94 | 0.94 | 0.63 | 0.65 | 0.60 | 0.65 | 0.62 | 0.96 | 0.62 | |
| 1367 | Molecular function-Recentor | n 99 | 0 99 | 0 98 | 0.08 | 0.79 | 0.80 | 0.68 | 0.80 | 0.79 | 0 98 | 0.76 | |
| 1368 | Molecular function Chaptered | 0.55 | 0.07 | 0.50 | 0.50 | 0.75 | | 0.00 | 0.00 | | 0.50 | 0.70 | |
| 1369 | Molecular function-chaperone - | 0.96 | 0.97 | 0.95 | 0.96 | 0.76 | 0.78 | 0.72 | 0.75 | | 0.95 | 0.72 | |
| 1370 | Molecular function-Guanine-nucleotide releasing factor - | 0.98 | 0.98 | 0.98 | 0.97 | 0.77 | 0.72 | 0.62 | 0.68 | | 0.98 | 0.68 | - 0.8 |
| 1371 | Molecular function-Methyltransferase - | 0.97 | 0.98 | 0.97 | 0.98 | 0.69 | 0.68 | 0.60 | 0.66 | 0.72 | 0.98 | 0.56 | |
| 1372 | Molecular function-Transducer - | 1.00 | 1.00 | 0.99 | 0.99 | 0.78 | 0.83 | 0.64 | 0.79 | 0.85 | 1.00 | 0.83 | |
| 1373 | > Molecular function-Repressor | 0.89 | 0.89 | 0.90 | 0.88 | 0.72 | 0.72 | 0.62 | 0.72 | 0.69 | 0.90 | 0.68 | Ч |
| 1374 | 면 Molecular function-Protein phosphatase - | 0.99 | 0.99 | 0.98 | 0.97 | 0.62 | 0.66 | 0.47 | 0.63 | 0.79 | 0.98 | 0.56 | an A |
| 1375 | Molecular function-Thiol protease - | 0.98 | 0.99 | 0.97 | 0.95 | 0.63 | 0.74 | 0.57 | 0.65 | 0.73 | 0.99 | 0.61 | Me |
| 1376 | Molecular function-Helicase - | 0.99 | 0.99 | 0.99 | 0.98 | 0.84 | 0.82 | 0.78 | 0.81 | 0.84 | 0.99 | 0.72 | |
| 1377 | Molecular function-Lyase - | 0.95 | 0.94 | 0.92 | 0.89 | 0.67 | 0.68 | 0.56 | 0.59 | 0.67 | 0.95 | 0.60 | |
| 1370 | Molecular function-Glycosyltransferase - | 0.99 | 0.99 | 0.98 | 0.96 | 0.67 | 0.65 | 0.62 | 0.65 | 0.65 | 0.98 | 0.61 | - 0 7 |
| 1380 | Molecular function-Oxidoreductase | 0 99 | 0.98 | 0.97 | 0 94 | 0.76 | 0.72 | 0.63 | 0.70 | 0.75 | 0 98 | 0.63 | 0.7 |
| 1381 | Molecular function Protoase inhibitor | 0.07 | 0.00 | 0.07 | 0.06 | 0.79 | 0.76 | 0.64 | 0.90 | 0.90 | 0.00 | 0.76 | |
| 1382 | Molecular function-protease infibition - | 0.97 | 0.90 | 0.97 | 0.90 | 0.76 | 0.70 | 0.04 | 0.60 | 0.60 | 0.90 | 0.70 | |
| 1383 | Molecular function-Kinase - | 0.99 | 0.99 | 0.98 | 0.98 | 0.68 | 0.64 | 0.58 | 0.64 | 0.64 | 0.99 | 0.62 | |
| 1384 | Molecular function-Acyltransferase - | 0.98 | 0.96 | 0.95 | 0.91 | 0.65 | 0.60 | 0.59 | 0.60 | 0.66 | 0.97 | 0.60 | |
| 1385 | Molecular function-Voltage-gated channel - | 0.99 | 0.99 | 0.99 | 0.99 | 0.73 | 0.80 | 0.60 | 0.74 | 0.87 | 0.99 | 0.74 | |
| 1386 | Molecular function-Growth factor - | 0.99 | 0.99 | 0.98 | 0.97 | 0.79 | | 0.64 | | 0.79 | 0.97 | 0.73 | |
| 1387 | Molecular function-Activator | 0.90 | 0.90 | 0.91 | 0.90 | 0.72 | 0.71 | 0.63 | 0.72 | 0.64 | 0.92 | 0.68 | |
| 1388 | Molecular function-lon channel - | 0.99 | 1.00 | 0.99 | 0.99 | 0.77 | 0.81 | 0.68 | | 0.81 | 1.00 | 0.74 | - 0.6 |
| 1389 | Molecular function-Hydrolase | 0.97 | 0.96 | 0.93 | 0.92 | 0.62 | 0.62 | 0.56 | 0.60 | 0.62 | 0.97 | 0.60 | |
| 1390 | Molecular function-Cytokine - | 0.99 | 0.99 | 0.98 | 0.98 | 0.86 | 0.82 | 0.69 | 0.77 | 0.85 | 0.97 | 0.76 | |
| 1391 | Molecular function-Actin-binding | 0.97 | 0.96 | 0.97 | 0.95 | 0.78 | 0.75 | 0.66 | | 0.73 | 0.97 | 0.62 | |
| 1392 | - Molecular function-GTPase activation - | 0.99 | 0.97 | 0.98 | 0.97 | 0.77 | 0.70 | 0.64 | 0.71 | 0.78 | 0.99 | 0.65 | |
| 1393 | Molecular function Developmental protein | 0.95 | 0.97 | 0.00 | 0.07 | 0.72 | 0.75 | 0.67 | 0.60 | 0.60 | 0.95 | 0.70 | |
| 1394 | Molecular function-bevelopmentar protein | 0.00 | 0.00 | 0.00 | 0.05 | 0.75 | | 0.05 | 0.05 | 0.05 | 0.05 | 0.70 | |
| 1395 | Molecular function-Ligase - | 0.98 | 0.98 | 0.96 | 0.95 | 0.75 | 0.77 | 0.62 | 0.72 | 0.76 | 0.97 | 0.63 | |
| 1390 | Molecular function-Ribosomal protein - | 1.00 | 1.00 | 0.99 | 0.99 | 0.98 | 0.97 | 0.95 | 0.97 | 0.98 | 0.99 | 0.82 | |
| 1398 | Molecular function-Tyrosine-protein kinase - | 0.99 | 0.99 | 0.98 | 0.97 | 0.72 | 0.68 | 0.55 | 0.66 | 0.68 | 0.98 | 0.64 | - 0.5 |
| 1399 | | 148 ¹ | 1485 | APNet | NO | ds up IM | rme | AT R | (gr),H | 2ver | | cet? | - |
| 1400 | | 4. | 4, | 4, | ad a large state of the second state of the se | ري. رو | enero. | SCO | s ^c | sene. < | SWIC | MAD | |
| 1401 | | | | \$ | - | | | | | | | | |
| 1 400 | | | | | | | | | | | | | |

Figure S6: Model performance measured by mean AUC for binary tasks derived from the multi label
 task 'molecular location'

| Pathways-Chromatin organi | zation - 0.94 | 0.93 | 0.93 | 0.92 | 0.78 | 0.78 | 0.72 | 0.73 | 0.77 | 0.94 | 0.68 | 1.0 |
|---|---|---|---|---|--|--|--|---|--|---|--|--|
| Pathways-Vesicle-mediated tra | nsport - 0.92 | 0.90 | 0.87 | 0.87 | 0.67 | 0.68 | 0.63 | 0.68 | 0.67 | 0.90 | 0.61 | |
| Pathways-Muscle contr Pathways-Immune S | ystem - 0.85 | 0.94 | 0.92 | 0.90 | 0.79 | 0.79 | 0.68 | 0.74 | 0.75 | 0.92 | 0.58 | |
| Pathways-DNA | Repair - 0.93 | 0.92 | 0.92 | 0.88 | 0.79 | 0.78 | 0.71 | 0.76 | 0.77 | 0.91 | 0.67 | - 0.9 |
| Pathways-Cellular responses to s | stimuli - 0.86 | 0.91 | 0.91 | 0.88 | 0.73 | 0.79 | 0.67 | 0.72 | 0.70 | 0.90 | 0.67 | |
| Pathways-Metabolism of pr Pathways-DNA Benli | oteins - 0.85 | 0.84 | 0.80 | 0.81 | 0.65 | 0.66 | 0.62 | 0.66 | 0.64 | 0.83 | 0.62 | |
| Pathways-Hemo | ostasis - 0.87 | 0.86 | 0.83 | 0.81 | 0.72 | 0.70 | 0.67 | 0.70 | 0.68 | 0.86 | 0.59 | - 0.8 |
| Pathways-Metabolism o | of RNA - 0.95 zation - 0.96 | 0.93 | 0.93 0.96 | 0.93 0.92 | 0.88 0.85 | 0.88 0.85 | 0.83 0.74 | 0.87 0.82 | 0.85 0.83 | 0.93 0.96 | 0.79 | U, |
| Pathways-Metal | polism - 0.93 | 0.92 | 0.91 | 0.90 | 0.73 | 0.72 | 0.64 | 0.73 | 0.70 | 0.90 | 0.61 | an A |
| Pathways-Programmed Cell | uction - 0.85 | 0.85 | 0.85 | 0.85 | 0.72 | 0.70 | 0.65 | 0.69 | 0.71 | 0.88 | 0.61 | Å. |
| Pathways-Transport of small mol | ecules - 0.94 | 0.93 | 0.91 | 0.91 | 0.69 | 0.68 | 0.61 | 0.71 | 0.72 | 0.91 | 0.58 | - 0.7 |
| Pathways-D | isease - 0.83 | 0.94 | 0.95 | 0.78 | 0.62 | 0.62 | 0.59 | 0.63 | 0.63 | 0.93 | 0.58 | |
| Pathways-Sensory Perc Pathways-Developmental B | eption - 0.95 | 0.94 | 0.92 | 0.89 | 0.77 | 0.78 | 0.61 0.61 | 0.72 | 0.76 | 0.89 | 0.75 | |
| Pathways-Auto | phagy - 0.93 | 0.89 | 0.88 | 0.84 | 0.74 | 0.71 | 0.62 | 0.70 | 0.71 | 0.89 | 0.66 | - 0.6 |
| Pathways-Cell Pathways-Organelle biogenesis and mainte | Cycle - 0.92 nance - 0.91 | 0.92 | 0.91 0.87 | 0.89 0.86 | 0.83 | 0.82 | 0.76 | 0.81 0.76 | 0.82 | 0.91 | 0.70 | |
| Pathways-Gene expression (Transcri | iption) - 0.88 | 0.89 | 0.88 | 0.86 | 0.73 | 0.71 | 0.65 | 0.72 | 0.74 | 0.87 | 0.70 | |
| Pathways-Cell-Cell Communi | | 0.91 | 0.89 | 0.87 | 0.71 | 0.65 | 0.63 (| 0.69 , X | 0.69 | 0.92 | 0.60 | - 0.5 |
| | MEBU | MIEB | MPNet | , wor | ellbrig. | torne | GAN | GAN | elve | W.2.31 | Bert | |
| | x | × | 6 | 20 ⁰ | Č c | ene | 5 | ⁵⁰ (| -ser . | \$~ < | 2Nr. | |
| | | | Ŷ | | | | | | | | | |
| Figure S7: Model performance m | easured by | v me | an Al | UC f | or bii | narv | tasks | deri | ved f | from | the m | ulti label |
| task 'pathways' | ••••••••••••••••••••••••••••••••••••••• | , | | | | J | | | | | | |
| 1 · | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | Pathways-Chromatin organi Pathways-Vesicle-mediated tra Pathways-Muscle contr Pathways-Immune S Pathways-ONA Pathways-Protein locaii Pathways-Cellular responses to o Pathways-Cellular matrix organi Pathways-Metabolism of Pathways-Metabolism of Pathways-Metabolism of Pathways-Sensory Perc Pathways-Signal Transd Pathways-Signal Transd Pathways-Sensory Perc Pathways-Sensory Perc Pathways-Sensory Perc Pathways-Cellogenesis and mainte Pathways-Cell Pathways-Cell Pathways-Cell communi Pathways-Cell communi | Pathways-Chromatin organization 0.94 Pathways-Vesicle-mediated transport 0.92 Pathways-Muscle contraction 0.95 Pathways-Protein localization 0.92 Pathways-Cellular responses to stimul Pathways-Cellular responses to stimul Pathways-Metabolism of proteins 0.95 Pathways-Metabolism of Proteins 0.95 Pathways-Metabolism 07 RNA 0.95 Pathways-Metabolism 07 RNA 0.95 Pathways-Metabolism 07 RNA 0.95 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.96 Pathways-Signal Transduction 0.95 Pathways-Signal Transduction 0.95 Pathways-Cell-Cell communication 0.96 Pathways-Cell-Cell communication 0.96 Pathways-Cell-Cell communication 0.98 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.98 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.99 Pathways-Cell-Cell communication 0.99 Pathways-Cell- | Pathways-Chromatin organization 0.94 0.93 Pathways-Vesicle-mediated transport 0.92 0.90 Pathways-Mice contraction 0.95 0.94 Pathways-Mice contraction 0.95 0.94 Pathways-Mice contraction 0.92 0.91 Pathways-Mice contraction 0.92 0.91 Pathways-Potein localization 0.92 0.91 Pathways-Mice contraction 0.98 0.98 Pathways-Metabolism of proteins 0.95 0.94 Pathways-Mice contraction 0.96 0.95 Pathways-Mice contraction 0.96 0.95 Pathways-Mice contraction 0.96 0.95 Pathways-Forgamed Cell Death 0.90 0.85 Pathways-Forgamed Cell Death 0.90 0.85 Pathways-Forgamed Cell Death 0.90 0.85 Pathways-Signal Transduction 0.95 0.94 Pathways-Signal Transduction 0.95 0.94 Pathways-Signal Transduction 0.95 0.94 Pathways-Signal Transduction 0.95 0.94 Pathways-Cell Scell mathways-Cell 0.92 0.92 Pathways-Compare Cell Death 0.90 0.95 Pathways-Cell Scell Cycle 0.92 0.92 Pathways-Gene expression (Transcription) 0.88 0.89 Pathways-Cell-Cell communication 0.90 0.91 Turkeys-Gene expression (Transcription) 0.90 0.91 Figure S7: Model performance measured by metatask 'pathways' | Pathways-Vesicle-mediated transport Pathways-Vesicle-mediated transport Pathways-Muscle contraction Pathways-Immune System O35 0.85 0.84 Pathways-DNA Repir Contraction Pathways-Petholism of proteins Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Hemostasis Pathways-Metabolism of RNA O35 0.85 0.86 Pathways-Hemostasis Pathways-Hemostasis Pathways-Metabolism of NNA Pathways-Selfare Cell Death O35 0.96 0.93 Pathways-Metabolism of 0.96 0.95 0.96 Pathways-Metabolism of 0.96 0.95 0.96 Pathways-Selfare Cell Death O35 0.96 0.93 0.93 Pathways-Metabolism 0.95 0.96 0.93 0.93 Pathways-Metabolism 0.95 0.96 0.93 0.93 Pathways-Metapolism 0.95 0.96 0.93 0.93 Pathways-Metapolism 0.95 0.94 0.92 Pathways-Senger Pathways-Surence Pathways-Surence Pathways-Surence Pathways-Surence Pathways-Surence Pathways-Cell-Cell communication Pathways-Cell-Cell communication Pathways-Cell-Cell communication Pathways-Senger Struct Pathways-Senger Struct Pathways-Cell-Cell communication Pathways-Senger Struct Pathways-Senger Struct Pathways-Cell-Cell communication Pathways-Senger Struct Pathways-Senger Struct Pathways-Senger Struct Pathways-Senger Struct Pathways-Cell-Cell communication Pathways-Senger Struct Pathways-Senger Stru | Pethways-Chromatin organization 0.94 0.93 0.93 0.92 Pathways-Wesicle-mediated transport 0.92 0.90 0.87 0.87 Pathways-Muscle contraction 0.95 0.94 0.92 0.90 0.87 0.87 Pathways-Muscle contraction 0.95 0.94 0.92 0.90 0.92 0.91 0.91 0.92 0.90 0.92 0.93 0.92 0.93 0.92 0.93 0.92 0.93 0.92 0.93 0.93 0.93 0.92 0.94 0.98 0.98 0.98 0.98 0.98 0.98 0.93 0.92 0.98 0.98 0.98 0.98 0.93 0.92 0.98 0.89 0.99 0.99 0.99 0.99 | Pathways-Chromatin organization 0.94 0.93 0.93 0.92 0.76 Pathways-Muscle contraction 0.92 0.90 0.67 0.68 0.61 0.68 0.61 0.68 0.61 0.68 0.61 0.68 0.61 0.61 0.62 0.60 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 0.61 | Pathways-Chromatin organization 0.94 0.93 0.93 0.92 0.67 0.67 0.68 Pathways-Wesicle-mediated transport 0.92 0.90 0.87 0.78 0.67 0.68 Pathways-Muscle contraction 0.95 0.94 0.92 0.90 0.97 0.78 0.72 0.70 Pathways-Poteiologianted 0 0.80 0.81 0.80 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 0.81 | Pathways-Chromatin organization 0.04 0.93 0.93 0.92 0.78 | Pathways-Chromatin organization 0.94 0.93 0.92 0.96 0.72 0.76 0.66 0.63 0.68 Pathways-Vesicle-mediated transport 0.92 0.94 0.97 0.87 0.67 0.66 0.63 0.68 0.64 0.66 0.63 0.68 0.74 0.76 0.66 0.63 0.68 0.74 0.76 0.68 0.74 0.76 0.68 0.72 0.78 0.68 0.71 0.76 0.78 0.68 0.72 0.78 0.66 0.72 0.78 0.66 0.71 0.76 0.78 0.66 0.72 0.76 0.71 0.76 0.76 0.71 0.76 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.71 0.76 0.76 0.71 0.76 0.71 0.76 0.76 0.71 0.76 0.70 0.76 0.70 0.76 0.76 0.71 0.76 0.71 0.76 0.76 0.77 0.77 <t< th=""><th>Pathways-Chromatin organization 094 0.93 0.92 0.78 0.70 0.72 0.73 0.77 Pathways-Vesicle mediated transport 0.93 0.94 0.94 0.97 0.67 0.68 0.63 0.64 0.64 0.64 0.67 0.67 0.68 0.63 0.64 0.71 0.71 0.68 0.67 0.72 0.71 0.68 0.70 0.71 0.74 0.64 0.70 0.71 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.78 0.78 0.78 0.78 0.78 0.78 0.78 <td< th=""><th>Pathways-Chromatin organization Pathways-Vesicle-mediated transport Pathways-Nesicle contraction Pathways-Mucle contraction Pathways-Coll Coll Contrunucle Contrunucle Pathways-Coll Coll Contraction Pathw</th><th>Pathways-Chomath organization 0.5 0.51 0.57 <</th></td<></th></t<> | Pathways-Chromatin organization 094 0.93 0.92 0.78 0.70 0.72 0.73 0.77 Pathways-Vesicle mediated transport 0.93 0.94 0.94 0.97 0.67 0.68 0.63 0.64 0.64 0.64 0.67 0.67 0.68 0.63 0.64 0.71 0.71 0.68 0.67 0.72 0.71 0.68 0.70 0.71 0.74 0.64 0.70 0.71 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.76 0.77 0.78 0.78 0.78 0.78 0.78 0.78 0.78 0.78 0.78 <td< th=""><th>Pathways-Chromatin organization Pathways-Vesicle-mediated transport Pathways-Nesicle contraction Pathways-Mucle contraction Pathways-Coll Coll Contrunucle Contrunucle Pathways-Coll Coll Contraction Pathw</th><th>Pathways-Chomath organization 0.5 0.51 0.57 <</th></td<> | Pathways-Chromatin organization Pathways-Vesicle-mediated transport Pathways-Nesicle contraction Pathways-Mucle contraction Pathways-Coll Coll Contrunucle Contrunucle Pathways-Coll Coll Contraction Pathw | Pathways-Chomath organization 0.5 0.51 0.57 < |

| 1459 | | | | | | | | | | | | | | | |
|-------|-------|--|------------------|--------------|----------------------|--------------|----------------------|----------------------|--------------|----------------------|----------------------|--------------|--------------|-------|------|
| 1460 | | | | | | | | | | | | | | | |
| 1461 | | | | | | | | | | | | | | | |
| 1462 | | | | | | | | | | | | | | | |
| 1463 | | DNA ticsue cell type enrichment Prostate - Fibroblasts | 0.67 | 0.63 | 0.68 | 0.66 | 0.88 | 0.88 | 0.74 | 0.84 | 0.83 | 0.73 | 0.66 | _ | 1.0 |
| 1/6/ | | RNA tissue cell type enrichment-Stomach - Gastric mucous cells | 0.62 | 0.62 | 0.68 | 0.60 | 0.94 | 0.88 | 0.73 | 0.89 | 0.85 | 0.69 | 0.63 | | |
| 1404 | | RNA tissue cell type enrichment-Heart muscle - Fibroblasts RNA tissue cell type enrichment-Prostate - Smooth muscle cells | - 0.65 - 0.64 | 0.65 0.65 | 0.71 0.71 | 0.65 0.63 | 0.90 0.90 | 0.89 0.81 | 0.79 0.67 | 0.84 0.88 | 0.84 0.80 | 0.72 0.70 | 0.70 0.63 | | |
| 1400 | | RNA tissue cell type enrichment-Liver - Hepatocytes - RNA tissue cell type anrichment Skelatal muscle - Masraphages | 0.65 | 0.71 | 0.72 | 0.68 | 0.85 | 0.89 | 0.70 | 0.86 | 0.85 | 0.74 | 0.59 | | |
| 1466 | | RNA tissue cell type enrichment-Skin - Mitotic cells (Skin) | 0.82 | 0.75 | 0.80 | 0.81 | 0.99 | 0.93 | 0.84 | 0.95 | 0.95 | 0.90 | 0.71 | | |
| 1467 | | RNA tissue cell type enrichment-Testis - Late spermatids RNA tissue cell type enrichment-Breast - Breast glandular cells | - 0.56 - 0.56 | 0.58 | 0.61 0.54 | 0.57 0.55 | 0.69 0.84 | 0.91 | 0.62 0.60 | 0.69 | 0.69 0.68 | 0.64 0.58 | 0.64 | | |
| 1468 | | RNA tissue cell type enrichment-Stomach - Fibroblasts | 0.70 | 0.70 | 0.74 | 0.64 | 0.90 | 0.87 | 0.74 | 0.83 | 0.84 | 0.79 | 0.73 | | |
| 1469 | | RNA tissue cen type enrichment-Adipose viscerar - Adipose progenitor cens - RNA tissue cell type enrichment-Stomach - Macrophages - | 0.38 | 0.39 | 0.83 | 0.69 | 0.87 | 0.85 | 0.71 | 0.81 | 0.92 | 0.80 | 0.83 | | |
| 1470 | | RNA tissue cell type enrichment-Skin - Eccrine sweat gland cells RNA tissue cell type enrichment-Adipose visceral - Macrophages | - 0.60 - 0.69 | 0.62 | 0.67 0.76 | 0.61 0.67 | 0.83 0.95 | 0.81 0.95 | 0.61 | 0.76 | 0.80 0.90 | 0.70 0.75 | 0.66 0.66 | | |
| 1471 | | RNA tissue cell type enrichment-Prostate - T-cells | 0.83 | 0.84 | 0.85 | 0.75 | 0.98 | 0.98 | 0.92 | 0.96 | 0.97 | 0.87 | 0.72 | - | 0.9 |
| 1472 | | RNA tissue cell type enrichment-Thyroid - Thyroid glandular cells | 0.57 | 0.62 | 0.64 | 0.60 | 0.98 | 0.97 | 0.92 | 0.95 | 0.97 | 0.82 | 0.70 | | |
| 1473 | | RNA tissue cell type enrichment-Heart muscle - Macrophages RNA tissue cell type enrichment-Breast - T-cells | 0.77 | 0.74 | 0.82 | 0.72 | 0.94 | 0.94 | 0.89 | 0.93 | 0.90 | 0.76 | 0.70 | | |
| 1474 | | RNA tissue cell type enrichment-Adipose visceral - Endothelial cells | 0.70 | 0.66 | 0.72 | 0.67 | 0.95 | 0.94 | 0.77 | 0.92 | 0.77 | 0.69 | 0.63 | | |
| 1475 | | RNA tissue cell type enrichment-Colon - Smooth muscle cells RNA tissue cell type enrichment-Skin - Fibroblast_2 | 0.72 | 0.70 | 0.77 | 0.64 0.64 | 0.93 0.88 | 0.87 0.85 | 0.70 | 0.90 0.84 | 0.81 0.80 | 0.75 | 0.59 | | |
| 1/76 | | RNA tissue cell type enrichment-Skin - Keratinocyte (granular) | 0.57 | 0.57 | 0.64 | 0.59 | 0.78 | 0.74 | 0.58 | 0.71 | 0.75 | 0.67 | 0.69 | | |
| 1470 | | RNA tissue cell type enrichment-Thyroid - T-cells | 0.79 | 0.30 | 0.82 | 0.71 | 0.96 | 0.96 | 0.85 | 0.93 | 0.02 | 0.81 | 0.70 | | |
| 14// | | RNA tissue cell type enrichment-Skin - Sebaceous gland cells - RNA tissue cell type enrichment-Kidney - Fibroblasts - | 0.69 | 0.68 | 0.76 | 0.69 0.61 | 0.75 0.92 | 0.69 0.89 | 0.61 | 0.69 | 0.82 | 0.75 | 0.59 | | |
| 1478 | | RNA tissue cell type enrichment-Lung - Macrophages | 0.74 | 0.72 | 0.78 | 0.65 | 0.98 | 0.96 | 0.90 | 0.96 | 0.90 | 0.79 | 0.67 | - | 0.8 |
| 1479 | | RNA tissue cell type enrichment-Adipose subcutaneous - Adipocytes (Subcutaneous) - RNA tissue cell type enrichment-Colon - Enteric glia cells - | - 0.62 | 0.63 | 0.68 | 0.63 | 0.85 | 0.76 | 0.62 | 0.82 | 0.76 | 0.71 | 0.61 | | |
| 1480 | | RNA tissue cell type enrichment-Adipose subcutaneous - Macrophages RNA tissue cell type enrichment-Colon - Endothelial cells | 0.70 | 0.73 | 0.77 | 0.68 | 0.96 | 0.96 | 0.89 | 0.93 | 0.91 | 0.76 | 0.68 | | |
| 1481 | | RNA tissue cell type enrichment-Adipose visceral - Mesothelial cells | 0.54 | 0.59 | 0.65 | 0.59 | 0.88 | 0.81 | 0.71 | 0.82 | 0.80 | 0.69 | 0.64 | | |
| 1482 | dime | RNA tissue cell type enrichment-Breast - Adipocytes (Breast) RNA tissue cell type enrichment-Skeletal muscle - Fibroblasts | 0.59 | 0.63 | 0.68 | 0.58 | 0.85 0.84 | 0.76 | 0.60 | 0.83 | | 0.66 | 0.59 | | AUC |
| 1483 | ask F | RNA tissue cell type enrichment-Liver - Vascular Endothelial cells | 0.61 | 0.57 | 0.64 | 0.58 | 0.88 | 0.87 | 0.73 | 0.82 | 0.76 | 0.67 | 0.66 | | dean |
| 1484 | P | RNA tissue cell type enrichment-Adipose subcutaneous - Adipose progenitor cells | 0.62 | 0.57 | 0.67 | 0.59 | 0.90 | 0.87 | 0.72 | 0.81 | 0.80 | 0.73 | 0.67 | | - |
| 1485 | | RNA tissue cell type enrichment-Adipose visceral - Adipocytes (Visceral) - RNA tissue cell type enrichment-Lung - Alveolar cells type 2 | - 0.64 - 0.64 | 0.69 | 0.72 0.63 | 0.64 0.62 | 0.86 0.95 | 0.78 0.89 | 0.63 | 0.82 0.89 | 0.77 0.78 | 0.73 0.63 | 0.60 | | |
| 1486 | | RNA tissue cell type enrichment-Lung - Fibroblast_2 | 0.73 | 0.71 | 0.76 | 0.69 | 0.92 | 0.91 | 0.79 | 0.86 | 0.89 | 0.79 | 0.72 | | |
| 1/187 | | RNA tissue cell type enrichment-Heart muscle - Endothelial cells | 0.92 | 0.90 | 0.92 | 0.67 | 0.95 | 0.94 | 0.98 | 0.99 | 0.78 | 0.93 | 0.59 | - | 0.7 |
| 1400 | | RNA tissue cell type enrichment-Adipose subcutaneous - T-cells RNA tissue cell type enrichment-Thyroid - Fibroblasts | 0.85 | 0.83 | 0.88 | 0.74 0.60 | 0.99 0.87 | 0.98 0.84 | 0.96 | 0.97 0.83 | 0.96 | 0.86 | 0.70 | | |
| 1400 | | RNA tissue cell type enrichment-Skeletal muscle - Skeletal myocytes | 0.66 | 0.71 | 0.71 | 0.66 | 0.81 | 0.82 | 0.57 | 0.74 | 0.78 | 0.69 | 0.61 | | |
| 1489 | | RNA tissue cell type enrichment-Liver - nepatic stellate cells - RNA tissue cell type enrichment-Testis - Spermatogonia - | 0.59 | 0.63 | 0.75 | 0.64 | 0.92 | 0.93 | 0.81 | 0.88 | 0.88 | 0.69 | 0.72 | | |
| 1490 | | RNA tissue cell type enrichment-Prostate - Urothelial cells RNA tissue cell type enrichment-Stomach - Mitotic cells (Stomach) | 0.63 | 0.58 | 0.69 | 0.57 | 0.88 0.98 | 0.87 0.98 | 0.66 | 0.86 | 0.86 0.96 | 0.72 | 0.66 | | |
| 1491 | | RNA tissue cell type enrichment-Colon - Colon enteroendocrine cells | 0.77 | 0.79 | 0.83 | 0.75 | 0.94 | 0.94 | 0.72 | 0.91 | 0.87 | 0.78 | 0.72 | | |
| 1492 | | RNA tissue cell type enrichment-Prostate - Prostate glandular cells - RNA tissue cell type enrichment-Skin - Keratinocyte (other) - | - 0.57 | 0.59 | 0.63 | 0.59 | 0.80 | 0.69 | 0.57 | 0.74 | 0.68 | 0.60 | 0.59 | | |
| 1493 | | RNA tissue cell type enrichment-Pancreas - Beta cells RNA tissue cell type enrichment-Testis - Early spermatids | 0.70 | 0.70 | 0.78 | 0.67 | 0.94 | 0.95 | 0.67 | 0.87 | 0.89 | 0.73 | 0.72 | | |
| 1494 | | RNA tissue cell type enrichment-Skin - Endothelial cells | 0.61 | 0.61 | 0.67 | 0.59 | 0.90 | 0.87 | 0.77 | 0.83 | 0.81 | 0.67 | 0.66 | _ | 0.6 |
| 1495 | | RNA tissue cell type enrichment-Colon - Colon enterocytes RNA tissue cell type enrichment-Skeletal muscle - Endothelial cells | 0.71 | 0.73 | 0.72 | 0.68 | 0.97 | 0.98 | 0.80 | 0.95 | 0.94 | 0.75 | 0.70 | | |
| 1496 | | RNA tissue cell type enrichment-Kidney - Endothelial cells - RNA tissue cell type enrichment-Pancreas - Alpha cells - | 0.61 | 0.63 | 0.68 | 0.58 | 0.88 | 0.89 | 0.70 | 0.84 | 0.73 | 0.66 | 0.58 | | |
| 1497 | | RNA tissue cell type enrichment-Breast - Macrophages | 0.84 | 0.84 | 0.86 | 0.75 | 0.98 | 0.98 | 0.95 | 0.97 | 0.97 | 0.85 | 0.76 | | |
| 1498 | | RNA tissue cell type enrichment-Breast - Fibroblasts RNA tissue cell type enrichment-Adipose subcutaneous - Smooth muscle cells | - 0.64 - 0.63 | 0.66 | 0.68 | 0.64 0.60 | 0.87 0.89 | 0.83 0.83 | 0.72 | 0.83 0.83 | 0.75 | 0.73 0.68 | 0.60 | | |
| 1/100 | | RNA tissue cell type enrichment-Heart muscle - Cardiomyocytes - | 0.68 | 0.72 | 0.74 | 0.72 | 0.90 | 0.91 | 0.66 | 0.89 | 0.84 | 0.75 | 0.61 | | |
| 1500 | | RNA tissue cell type enrichment-Lung - Respiratory Cillated Cells - RNA tissue cell type enrichment-Kidney - Proximal tubular cells - | 0.68 | 0.76 | 0.76 | 0.00 | 0.93 | 0.92 | 0.73 | 0.91 | 0.92 | 0.72 | 0.65 | | |
| 1500 | | RNA tissue cell type enrichment-Colon - Macrophages - RNA tissue cell type enrichment-Prostate - Endothelial cells - | 0.77 | 0.78 0.68 | 0.82 0. <u>76</u> | 0.69 0.68 | 0.94 0. <u>95</u> | 0.92 0. <u>94</u> | 0.87 | 0.93 0. <u>92</u> | 0.93 0. <u>82</u> | 0.79 0.74 | 0.67 0.67 | | |
| 1001 | | RNA tissue cell type enrichment-Breast - Endothelial cells | 0.64 | 0.67 | 0.73 | 0.66 | 0.93 | 0.92 | 0.78 | 0.89 | 0.81 | 0.69 | 0.64 | | |
| 1502 | | איזא ussue cen type enrichment-Adipose subcutaneous - Endothelial cells - | 0.65 | 0.6/ | 0.71 | 0.64 | 0.92 5 a | 0.92 | 0.77 \$ | -0.90 | 0.77 | 0.70 | 0.03 | _ | 0.5 |
| 1503 | | | MILB | MIEB | NPPHe . | of Work | cellPL | etorne | Scopt." | Corr. | enerve | .SNA2'SU | NABert | | |
| 1504 | | | | | Ś | °, | Ġ | er. | | | ، سی | ~ < | »- | | |
| 1505 | | | | | | | | | | | | | | | |
| 1506 | Fig | gure S8: Model performance measured by mean | AU | C fo | r biı | nary | tas | ks d | eriv | ed f | rom | the | mu | lti 1 | abel |
| 1507 | tas | k 'RNA tissue cell type enrichment' | | | | | | | | | | | | | |





Figure S10: Similarity of performance across models. We construct vectors of the average AUC ROC for every model and task and then use 1 - cosine distance vectors to calculate their proximities, which are then re-scaled to the interval (0,1).



Figure S11: Similarity of performance across tasks. We construct vectors of the average AUC ROC for every model and task and then use 1 - cosine distance vectors to calculate their proximities, which are then re-scaled to the interval (0,1).



Figure S12: The prediction performance as measured by the 5-fold mean AUC-ROC score (with standard deviation in red) of a multiplayer perceptron (MLP) model versus a logistic regression model using the embeddings of MTEB-S and MTEB-L. We can see that in both cases the correlation is high (Pearson's coefficient of 0.92 and 0.97) and significant (p-value<0.001).