ScanMap: Supervised Confounding Aware Non-negative Matrix Factorization for Polygenic Risk Modeling

Yuan Luo YUAN.LUO@NORTHWESTERN.EDU
Department of Preventive Medicine
Northwestern University
Chicago, IL, USA

Chengsheng Mao CHENGSHENG.MAO@NORTHWESTERN.EDU
Department of Preventive Medicine
Northwestern University
Chicago, IL, USA

Editor: Editor's name

Abstract

Molecular mechanisms are important to inform targeted intervention and are often encoded in gene sets or pathways. Existing machine learning approaches often face challenges in simultaneously reducing the high dimensionality and learning effective features that are discriminative in predicting the disease types with the usual presence of confounding variables. We aim to improve accuracy and interpretability of prediction models by introducing Supervised Confounding Aware Non-negative Matrix Factorization for Polygenic Risk Modeling (ScanMap) for genetic studies. ScanMap selects informative groups of genes that embody multiple interacting molecular functions by using a supervised model that integrates both groups of genes and confounding variables in predicting disease type and status. The learned groups of genes reflect interacting molecular mechanisms, which are suitable features for polygenic risk modeling. These learned features are then used in training a softmax classifier for disease type and status prediction. We evaluated ScanMap against multiple state-of-the-art unsupervised and supervised matrix factorization models using large scale NGS datasets. ScanMap outperformed all comparison models significantly ($p < 0.05$). Feature analysis was performed to illuminate the insights and benefits of gene groups learned by ScanMap in disease risk prediction.

1. Introduction

Recent advances in machine learning have opened avenues towards more effective mining and modeling of large scale genetic and clinical data to facilitate translational research. Traditional machine learning in genetics usually treats genes or variants as features but multiple genes often form pathways and act together to achieve molecular or biological functions. Intuitively, genetic variants are standard to extract and have robust statistical properties, but are less informative and interpretable as they often do not directly speak about molecular mechanisms, at least not alone. In contrast, the genetic pathways are more expressive and informative, but their curation and selection are certainly non-trivial. In addition, the current genetic pathway databases are still actively evolving with the increasingly accelerated growth of genetic discoveries, and are far from being complete. Thus we need automated

tools to identify functionally related genes in order to expand our knowledge on molecular mechanisms. To this end, dimensionality reduction tools are usually explored, among which non-negative matrix factorization (NMF) has achieved much success and is a frequent choice due to the added interpretation advantage from the non-negativity constraints when working on genetic count data (e.g, (Alexandrov et al., 2013)).

Most of the dimensionality reduction methods, including NMF, belong to the unsupervised learning category as no label information is used. In many real world genetic applications, dimensionality reduction is just an intermediate step toward the final goals, such as disease type or status classification and survival or time-to-event regression. Separating the dimensionality reduction and model learning into two steps may not be optimal for the classification or regression goals as we have no guarantee that the learned features in reduced dimensions will be discriminative regarding the tasks at hand. To tackle this problem, supervised dimensionality reduction methods are needed to use the cancer labels to focus the attention on more discriminative genes and mutations.

In this work, we propose a new framework named ScanMap: Supervised Confounding Aware Non-negative Matrix Factorization for Polygenic Risk Modeling, to jointly model the dimensionality reduction problem and confounding aware supervised polygenic risk modeling problem. ScanMap relies on the intuition that the mechanisms by which a clinically meaningful group of disrupted genes act together are usually effective in characterizing disease subtypes (e.g., cancer types (Bailey et al., 2018)). Our source code is available at https://github.com/yuanluo/scanmap. Our contributions are as follows:

- To the best of our knowledge, ScanMap is the first study of simultaneously reducing the dimensions of genetic variants and building a supervised polygenic risk model that is aware of confounding variables.
- Applications on Next Generation Sequencing (NGS) datasets on multiple cancer types show significant performance improvements by ScanMap over multiple state-of-the-art baselines.
- ScanMap has a GPU implementation and runs fast. Feature analysis shows insights from ScanMap on identifying interacting molecular mechanisms of disease genetic risks.

Generalizable Insights about Machine Learning in the Context of Healthcare

Unlike previous supervised NMF methods, ScanMap incorporates confounders via a supervised learning step. This robust mechanism for dimensionality reduction additionally has generalizable utility. While individual genes have been studied, our understanding of genetic pathways in oncogenesis is limited. ScanMap allows for deeper exploration of these pathways. Although only studied on the four most prevalent cancers in the TCGA database in this paper, being able to expand understanding of molecular mechanisms could be useful in other cancers as well as a host of other genetic conditions.

2. Related Work

Nonnegative Matrix Factorization (NMF) refers to the set of problems on approximating a nonnegative matrix as the product of lower rank nonnegative matrices. Since the intro-

duction in (Lee and Seung, 1999), people have been working on NMF from various aspects. (Ding et al., 2005) showed the equivalence between NMF and K-means/spectral clustering. (Inderjit Dhillon, 2005) extended NMF to the case when the matrix approximation loss is measured by Bregman divergence, which is a general loss function that has both Frobenius norm and KL divergence as its special cases. On the solution procedure aspect, (Berry et al., 2007) reviewed the general algorithms, and categorized three classes of algorithms. The first class uses multiplicative updates (Lee and Seung, 1999; Ding et al., 2005), the second class uses gradient based methods such as (Pauca et al., 2006; Lin, 2007; Kim and Park, 2011), the third class uses the alternating least squares (ALS) algorithm (Paatero, 1999; Langville et al., 2014). More recently, (Sun and Fevotte, 2014) adopted the alternating direction method of multipliers (ADMM) to solve the NMF with beta-divergence. As NMF formulation gets more complex, new optimization algorithms often need to be devised from scratch. In this work, we adopt ADAM (Kingma and Ba, 2015), a generic and efficient optimizer, and the autograd utility from PyTorch platform to automate the optimization of ScanMap.

Besides basic NMF methods, there are many variants of constrained NMF that form three groups. The first group enforced constraints into basic NMF to obtain certain desirable charateristics, such as sparsity (Morup et al., 2008) and orthogonality (Ding et al., 2006; Yoo and Choi, 2010). The second group named structured NMF modified the standard formulation of NMF, including weighted NMF (Kim and Choi, 2009), convolutive NMF (O'grady and Pearlmutter, 2006) and nonnegative matrix trifactorization (Yoo and Choi, 2010). The third group is the generalized NMF, including semi-NMF (Ding et al., 2010), matrix-set factorization (Li and Zhang, 2007) and kernel NMF (Zhang et al., 2006). For details, refer to the survey paper (Wang and Zhang, 2013). In this work, we adopt multiple elements in the first group of constrained NMF.

NMF has been an effective unsupervised dimensionality reduction method using single feature modality for data structure exploration (see review (Wang and Zhang, 2013)). NMF has been applied extensively in the biomedical domain: to cluster similar patients (e.g., (Hofree et al., 2013)) and sample cell lines (e.g., (Müller et al., 2008)). Recently, NMF has been applied to study differential cancer risks of genetic mutations with promising successes (e.g., (Alexandrov et al., 2013; Zeng et al., 2019)).

Unsupervised NMF methods cannot guarantee that the prediction ability is retained because the label information is not used to guide the factorization. To tackle this problem, several supervised NMF methods were proposed. They can be classified into two categories. The first category including (Hyekyoung Lee, 2010; Liping Jing and Ng, 2012) uses a Frobenius loss for supervision and is suitable for the regression problem. The second category including (MacMillan and Wilson, 2017; Bisot et al., 2017) uses the cross-entropy loss for supervision and is suitable for the classification problem. (MacMillan and Wilson, 2017) introduced the weakly topic supervised non-negative matrix factorization method (wsNMF) to enable the use of labeled example documents to promote the discovery of meaningful semantic structures of a corpus. (Bisot et al., 2017) introduced the non-negative formulation for task-driven dictionary learning to combine NMF and classification into a joint optimization problem, and called their method TNMF. None of these methods simultaneously consider the confounding variables in the supervised dimensionality reduction problem. Instead they assume that the NMF is performed on all raw features. However, in genetics,

confounding variables such as age, gender, and race are often of distinct nature compared with raw features (genetic mutations in this case), and often have much lower dimensions than raw features. Thus we need to separate confounding variables from NMF input but consider them in the supervised modeling, which is a need unmet by existing supervised NMF methods.

Our proposed framework ScanMap belongs to the second category of supervised NMF models, in which we use the cross entropy loss as classification constraint to explicitly guarantee the predictive utility of the learned gene groups. A distinct feature of ScanMap, compared to previous supervised NMF methods, is that ScanMap is confounding aware in that we merge confounding variables with NMF-derived features in the supervised learning component. Another feature is the orthogonality constraints applied on NMF-derived features. This is motivated by the fact in the field of genetics that we often need to assess the effect sizes of the features (predictors), and decoupling correlations among features is often desirable in related statistic analysis.

3. Methods

We develop a supervised feature learning framework in order to build machine learning models that are both more accurate and more interpretable for genetics. The framework uses supervised learning for polygenic risk modeling to guide the Non-negative Matrix Factorization and is capable of considering a variety of confounding variables ranging from subjects' demographics to comorbidities. We name our model Supervised Confounding Aware Non-negativeMatrix Factorization for Polygenic Risk Modeling, which is abbreviated as ScanMap.

3.1. ScanMap workflow

We first outline the workflow of ScanMap in Fig. 1, referring to Table 1 for symbols used throughout this paper. This study considers both genetic pathway features and genetic variant features. For genetic variants, we first annotate the variants and then keep the deleterious variants. The variants are of high dimensionality, and we choose to aggregate their counts according to the affected genes to avoid impractically large matrices. Thus we aggregate genetic variant count at gene level and abuse terminology to use "gene" to really mean "variants in the gene" in the following text. We filter redundant pathways and those that are too small. We then devise an occurrence counting scheme to construct the *subject* $\times$ *pathway* and the *subject* $\times$ *gene* matrices. We further perform supervised constrained NMF that is capable of considering a variety of confounding variables. Finally we combine the learned subject factor matrix together with confounding variables for disease type and status classification. We next explain each step in detail.

3.2. Annotation-based variant filtering and deleterious variant selection

For annotation-based variant filtering, we use the ANNOVAR toolkit (Wang et al., 2010) to comprehensively annotate called variants. ANNOVAR integrates a wide array of information regarding genetic variants, including their hosting genes annotated with several gene models including RefSeq, UCSC Known Gene, Ensembl Gene; the variant function and its

Table 1: Common notations used throughout the paper. When describing the model formulation that is applicable to both pathway count matrix and gene count matrix, we drop the g and p in the subscripts and superscripts to avoid clutter.

Notation	Definition
$\mathbf{X}^{(g)}$	*Subject* $\times$ *gene* matrix
$\mathbf{X}^{(p)}$	*Subject* $\times$ *pathway* matrix
$\mathbf{F}$	Subject factor matrix
$\mathbf{G}^{(g)}$	Gene factor matrix
$\mathbf{G}^{(p)}$	Pathway factor matrix
$\mathbf{F}_r$	r^{th} column of $\mathbf{F}$
n	Number of subjects
m_g	Number of genes
m_p	Number of pathways
k	Number of latent groups
$\circ$	Outer product

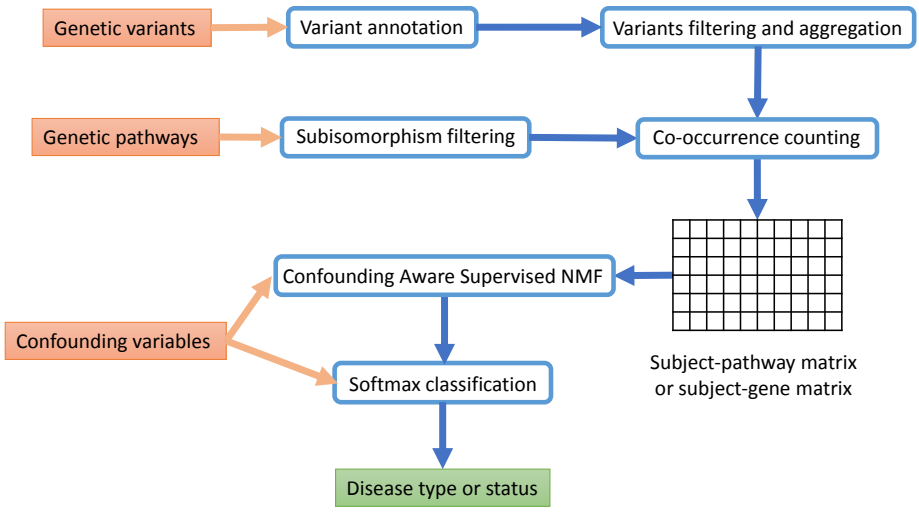


Figure 1: Workflow for the framework of Supervised Confounding Aware Non-negative Matrix Factorization for Polygenic Risk Modeling (ScanMap). Square boxes are data, round corner boxes are steps. The framework takes Variant Call Format (VCF) files as input for genetic variants. A row in a VCF file specify a particular variant (e.g., Single Nucleotide Polymorphism or insertion/deletion), its chromosomal location, and whether the variant occurs in 0, 1 or both strands of the chromosome, among other characteristics.

putative likelihood of pathogenicity based on scores including PolyPhen2, SIFT, CADD, and other meta predictors; variant's minor allele frequency and its phenotype associations evidenced by ClinVar and HGMD.

To address issues of reference mis-annotation, we resort to the recently released Exome Aggregation Consortium (ExAC) exome dataset (Lek et al., 2016), which aims to aggregate exome sequencing data and derive population level variant frequencies from a wide range of large-scale sequencing projects. For variants whose allele frequencies are observed to be over 90% among the 60,706 individuals aggregated by ExAC, we filter them out as they are less likely to be culprits in disease onset and development. There have been extensive debates in the genetics field on whether a disease is only due to rare variants or rare + common variants (see (Gibson, 2012)). To be inclusive, we included rare and modestly common variants (using ExAC reported statistics). We further focus on deleterious variants, which include frame-shift insertion, frame-shift deletions, nonsense variants, nonsynonymous variants and splice site alterations. Focusing on deleterious variants allows one to select likely truly harmful variants instead of including too many variants such as those that are synonymous (i.e. do not alter the resultant amino acids), and is an important step in genetic analysis. Following the common practice in machine learning to exclude extremely rare features, we exclude the genes that have very rare variants ($< 1\%$ subjects in the training data of respective experiments).

3.3. Pathway Collecting and Pruning

In this work, we also evaluate comparison models using the pathway directly as features as an approximation of the disruptions to molecular functions. We use the REACTOME database (Croft et al., 2010) to obtain a comprehensive collection of known and curated genetic pathways in a best effort attempt, while acknowleding that our current compiled knowledge about pathways is not complete. REACTOME is a database of biological pathways curated by expert biologists with evidence from literature. For this work, we primarily focus on human pathways. Sifting through the pathways, we found that some smaller pathways are part of larger pathways. As (Holmans, 2010) pointed out, small pathways often exhibit large single-gene or single-SNP effects, and lead to false positive associations with disease phenotypes. Thus we choose to keep only the larger pathways when encountering such pairs. The part-of relation between pathways is usually formulated as the problem of graph subisomorphism. Formally, let $G_1 = (N_1, E_1, l_1)$ and $G_2 = (N_2, E_2, l_2)$ be two graphs, where N_1 and N_2 are the sets of nodes, E_1 and E_2 are the sets of edges and l_1 and l_2 are the labeling functions for nodes and edges. G_1 is subisomorphic to G_2 if the following conditions are satisfied:

- Node agreement - There exists an injective mapping $\mathcal{M}$ from each node n_1 in G_1 to a counterpart $\mathcal{M}(n_1)$ in G_2 that shares the same label

$$l_1(n_1) = l_2(\mathcal{M}(n_1)) \tag{1}$$

- Edge agreement - Under the condition of node agreement, for a mapping $\mathcal{M}$, each edge (n_1, n_2) in G_1 should also have a corresponding edge in $(\mathcal{M}(n_1), \mathcal{M}(n_2))$ in G_2

Algorithm 1 Detecting subisomorphism in a set of pathways. Adj(·) denotes adjacency matrix of a pathway (graph). $N_{P_s}(P_b)$ denotes node subsetting of a graph, only retaining the nodes (and associated edges) in P_b if they are also in P_s.

Input: $\mathcal{P}$ - set of pathways
Output: S - list of discovered subisomorphisms in $\mathcal{P}$

```
Let S = []
stable sort P in ascending order of #nodes
for i = 1 to length(P) −1 do
  for j = i + 1 to length(P) do
    P_s = P[i]; P_b = P[j]
    if nodes(P_s) ⊂ nodes(P_b) then
      if Adj(P_s) == Adj(N_{P_s}(P_b)) then
        Append (P_s, P_b) to S
      end if
    end if
  end for
end for
return S
```

such that

$$l_1(n_1, n_2) = l_2(\mathcal{M}(n_1), \mathcal{M}(n_2)) \tag{2}$$

In our case, the definition that one graph (pathway) is subisomorphic to another graph simply means that the latter contains the former. That is, $\mathcal{M}$ is the identity mapping and l_1 and l_2 agree on the common nodes and edges between the two graphs. This greatly simplifies subisomorphism comparison between two pathways. Moreover, we have used heuristics to prune unnecessary subisomorphism comparisons, for example, pathway P_s can only be subisomorphic to pathway P_b if #nodes(P_s) < #nodes(P_b) (size heuristic) and nodes$(P_s) \subset$ nodes(P_b) (using faster pre-check of set containment), as shown in Algorithm 1.

3.4. Constructing the Matrices

We build the *subject* × *pathway* matrix $\mathbf{X}^{(p)}$ and the *subject* × *gene* matrix $\mathbf{X}^{(g)}$ that are the input in Fig. 2. The matrix entry records the occurrence count of variants in a gene or in a pathway. In Fig. 2, the factor matrix $\mathbf{F}$ is the *subject* × *subject group* matrix, $\mathbf{G}$ the *gene group* × *gene* matrix.

For the *subject* × *gene* matrix $\mathbf{X}^{(g)}$, the entry $\mathbf{X}^{(g)}_{i,k}$ denotes the count of variants hitting gene k in subject i and is defined as

$$\mathbf{X}^{(g)}_{i,k} = \sum_{v \in \mathcal{V}_i \text{ and } v \in Span_k} c(v) \tag{3}$$

where $\mathcal{V}_i$ is the set of variants hitting subject i, $Span_k$ is the basepair position spans of gene k, and $c(v)$ is the allele count of variant v.

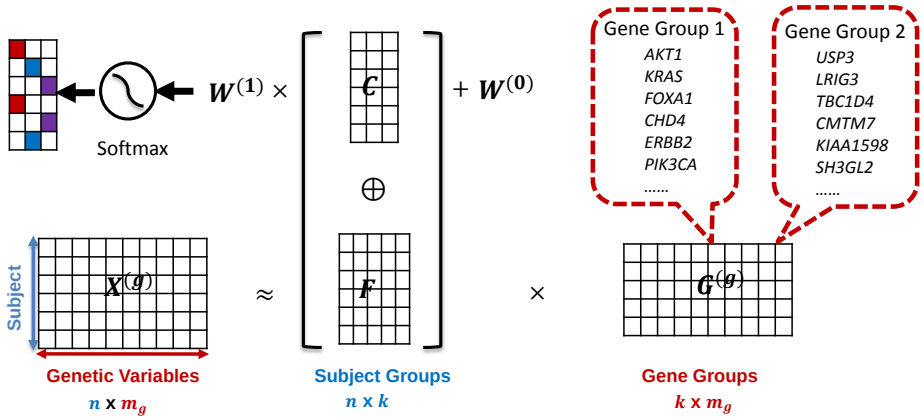


Figure 2: ScanMap's matrix factorization scheme for using confounding aware supervised learning to guide the gene group discovery. Analogous mechanism works for pathway count matrix as well. $\bigoplus$ denotes concatenation of two matrices horizontally. $\mathbf{C}$ is the matrix of confounding variables. $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(0)}$ are classification weights and bias. Other symbols are defined in Table 1.

For the *subject* × *pathway* matrix $\mathbf{X}^{(p)}$, the entry $\mathbf{X}^{(p)}_{i,j}$ denotes the count of variants hitting pathway j in subject i and is defined as

$$\mathbf{X}^{(p)}_{i,j} = \sum_{k \in \mathcal{P}_j} \mathbf{X}^{(g)}_{i,k} \tag{4}$$

where $\mathcal{P}_j$ is the set of genes in pathway j and $\mathbf{X}^{(g)}_{i,k}$ is defined in equation 3. There are certainly alternative ways to define the gene and pathway occurrence counts, but we observe that the definitions in equations 3 and 4 work well in our experiments.

3.5. Factorization in ScanMap

In this section, the formulation applies to both pathway count matrix and gene count matrix, hence we drop the g and p in the subscripts and superscripts to aovid clutter. Let k denote the number of latent groups as in Fig. 2. We combine each mode's vectors into corresponding factor matrices as in

$$\begin{aligned} \mathbf{F} &= [\mathbf{F}_1 \mid ... \mid \mathbf{F}_r \mid ... \mid \mathbf{F}_k] \in \mathbb{R}^{n \times k} \\ \mathbf{G} &= [\mathbf{G}_1 \mid ... \mid \mathbf{G}_r \mid ... \mid \mathbf{G}_k] \in \mathbb{R}^{m \times k} \end{aligned} \tag{5}$$

We define the outer product of the $r^{th}(1 \leq r \leq k)$ vectors from matrices $\mathbf{F}, \mathbf{G}$ as the following rank-one matrix

$$\mathbf{M} = \mathbf{F}_r \circ \mathbf{G}_r \tag{6}$$

where the entries are $\mathbf{M}_{i,j} = \mathbf{F}_{i,r}\mathbf{G}_{j,r}$.

The matrix factorization approximates the original matrix $\mathbf{X}$ as the sum of a series of rank-one matrices and is expressed as

$$\mathbf{X} \approx \sum_{r=1}^{k} \mathbf{F}_r \circ \mathbf{G}_r = \mathbf{FG} \tag{7}$$

Non-negativity constraints are typically enforced for interpretation advantages. To serve the modeling considerations in this paper, we further constrain the factorization as

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G}} \|\mathbf{X} - \mathbf{FG}\|_F^2 + \lambda_1 \phi(\mathbf{F}) \\ \text{s.t. } \mathbf{F} \geq 0, \mathbf{G} \geq 0 \\ \mathbf{G} \in \{0\} \cup [\gamma_{\mathbf{G}}, +\infty)^{m \times k} \end{aligned} \tag{8}$$

where,

$$\phi(\mathbf{F}) = \left\| \mathbf{I} - k \cdot \mathbf{F}^T\mathbf{F} / \sum \mathbf{F}^T\mathbf{F} \right\|_F^2 \tag{9}$$

Intuitively speaking, $\phi(\mathbf{F})$ is the orthogonality constraints on the subject factor matrix $\mathbf{F}$, which will be used as the feature matrix in the downstream classification step. This is motivated by the need in the field of genetics that we often need to assess the effect sizes of the features (predictors), and decoupling correlations among features is often desirable. Note that equation 9 is intended to be scale-free, in that we only ask for vectors to be orthogonal but not for them to have a norm of 1. On the other hand, we add sparsity constraints on the pathway or gene factor matrices ($\mathbf{G}$) to aid the interpretation of the learned composed high-order features, which aids clinical interpretation.

3.6. Classification for ScanMap

In clinical and genetic applications, there are usually confounding variables whose effects need to be explicitly assessed together with the features of interest to avoid biased model interpretation. These confounding variables typically include at least gender, race, and age, and are explicitly accounted for in the ScanMap classification step. Let $\mathbf{C}$ be the confounding variable matrix, and $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(0)}$ be classification weights and bias. We then concatenate the confounding matrix $\mathbf{C}$ and the learned feature matrix $\mathbf{F}$ and feed them into a softmax classifier

$$\mathbf{Z} = \text{softmax}(\mathbf{W}^{(1)} \, [\mathbf{C} \,|\, \mathbf{F}] \; + \; \mathbf{W}^{(0)}) \tag{10}$$

The loss function is defined as the cross-entropy error over all subjects in all classes as in

$$\mathcal{L} = - \sum_{d \in \mathcal{Y}_D} \sum_{f=1}^{F} \mathbf{Y}_{df} \ln \mathbf{Z}_{df} \tag{11}$$

where $\mathcal{Y}_D$ is the training set of subjects that have labels and F is the dimension of the output labels, which is equal to the number of classes. $\mathbf{Y}$ is the label indicator matrix.

Table 2: Statistics of TCGA experiment data. The table includes the distribution of the four most prevalent cancer types: breast cancer, colorectal cancer, lung cancer and prostate cancer. The dataset is split into a training set, a validation set and a test set according to a 6:2:2 ratio.

Cancer	Total	Train	Validation	Test
Breast	959	575	192	192
Colorectal	728	437	146	145
Lung	440	264	88	88
Prostate	418	251	83	84

To serve the modeling considerations in this paper, we further constrain the factorization as

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G},\mathbf{W}^{(0)},\mathbf{W}^{(1)}} & \|\mathbf{X}-\mathbf{F}\mathbf{G}\|_F^2+\lambda_1\phi(\mathbf{F})+\lambda_2\mathcal{L} \\ \text{s.t. } & \mathbf{F}\geq 0,\mathbf{G}\geq 0 \\ & \mathbf{G}\in\{0\}\cup[\gamma_{\mathbf{G}},+\infty)^{m\times k} \end{aligned} \tag{12}$$

where $\mathcal{L}$ is defined in equation 11 and $\phi(\mathbf{F})$ is defined as equation 9.

We base the ScanMap implementation on PyTorch, and design it to entirely run on GPU. We train ScanMap for a maximum of 4000 iterations using Adam (Kingma and Ba, 2015) and stop training if the validation loss does not decrease for 10 consecutive epochs. Parameters λ_1, λ_2 are tuned on validation dataset. ScanMap computes the sparse factor matrices $\mathbf{G}$ using a threshold $\gamma_{\mathbf{G}}$. This threshold provides a way to adjust the sparsity of the candidate pathway groups and gene groups. In this work, we also numerically tuned the sparsity threshold $\gamma_{\mathbf{G}}$ using the validation dataset.

4. Experiment on Cancer Type Prediction with NGS Data

Personalized medicine is becoming increasingly popular in cancer, which utilizes genetic profiles of tumors to guide early screening, preventive measures and clinical decisions on intervention options. The high throughput DNA sequencing technology has made genetic variants data in cancer increasingly accessible. Understanding the association between genetics and disease is important for understanding the underlying pathophysiologic onset and progression. We focus on using germline variants (variants that are inherited from a parent) to differentiate among different cancer types, which can inform early screening strategy and even targeted therapy for specific cancer types (Bertelsen et al., 2019). We use the proposed ScanMap framework to effectively explore the landscape of germline mutations and their genetic pathways to predict cancer types.

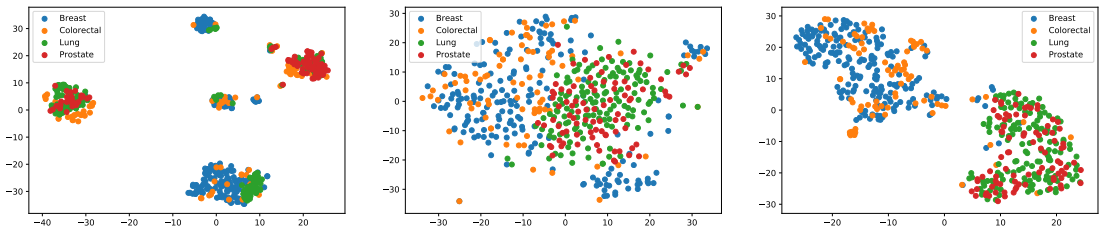
In this experiment, we have used the dataset from The Cancer Genome Atlas (TCGA) and focus on the top four prevalent cancers, including breast cancer, lung cancer, colorectal cancer and prostate cancer (Siegel et al., 2019). We recalibrate aligned sequencing data from

Table 3: Test accuracy of cancer type classification task on TCGA dataset. ScanMap significantly outperforms comparison models based on permutation test ($p < 0.05$). For the models directly using counts from genes and/or pathways, as well as models using unsupervised NMF produced features, we used logistic regression with l_2 regularization as the classifier. This is to be consistent with our motivation of developing supervised NMF to better suit regression analysis in genetics, also consistent with the choices of classifiers by other supervised NMF models.

Model	k	Test Accuracy
Gene count	-	0.7525
Pathway count	-	0.7387
Gene+pathway count	-	0.7544
Gene count (confounding)	-	0.8016
Pathway count (confounding)	-	0.7701
Gene+pathway count (confounding)	-	0.7682
NMF_{gene}	250	0.7485
$\text{NMF}_{\text{pathway}}$	100	0.7112
NMF_{gene} (confounding)	400	0.8173
$\text{NMF}_{\text{pathway}}$ (confounding)	400	0.7819
$\text{wsNMF}_{\text{gene}}$	200	0.7308
$\text{wsNMF}_{\text{pathway}}$	400	0.6051
$\text{wsNMF}_{\text{gene}}$ (confounding)	350	0.7800
$\text{wsNMF}_{\text{pathway}}$ (confounding)	100	0.7446
$\text{TNMF}_{\text{gene}}$	150	0.7505
$\text{TNMF}_{\text{pathway}}$	400	0.5992
$\text{TNMF}_{\text{gene}}$ (confounding)	250	0.7957
$\text{TNMF}_{\text{pathway}}$ (confounding)	500	0.7603
$\text{ScanMap}_{\text{gene}}$	400	**0.8468**
$\text{ScanMap}_{\text{pathway}}$	150	0.7957

blood or adjacent normal tissues, and call variants using HaplotypeCaller in GATK package with assembly hg19, which produces germline variants from Whole Exome Sequencing data. We partition the included 2545 total subjects with a 6:2:2 ratio, stratified by mortality, into a 1527-subjects training set, a 509-subjects validation set and a 509-subjects held-out test set, as shown in Table 2. Our dataset has 626 pathways and 684 genes from the filtering steps described in the Methods section.

The number of groups k in NMF models, including our model in equation 12, needs to be empirically tuned. We tune this parameter using the validation set and consider a range of group numbers from 50 to 500, at an increment of 50. For ScanMap, λ_1, λ_2 are tuned using validation set and are set to 1 and 0.1 respectively. The sparsity threshold $\gamma_{\mathbf{G}}$ is tuned



(*a*) ScanMap$_{\text{gene}}$ on test set (*b*) wsNMF$_{\text{gene}}$ (confounding) on test set (*c*) TNMF$_{\text{gene}}$ (confounding) on test set

Figure 3: The t-SNE visualization of the learned test set subject features in TCGA.

on validation set and set to 0.001. NMF based models are often randomly initialized, thus we also run initializations 10 times and select the best one using the validation set.

In order to assess whether performance changes are due to simply adding confounding variables or integrative consideration of confounding variables in the model itself, we have evaluated different comparison models under both without coufounding variables and with confounding variables modes. The accuracy scores of ScanMap and comparison models on the held-out test data are shown in Table 3. Comparing all the models and baselines, we can see that the raw gene and/or pathway count has an accuracy at best 0.8016 when combined with confounding variables. Note that without confounding variables, raw count features see clear performance drop and the best AUC is only 0.7544. This is consistent with the intuition that confounding variables indeed carry useful information in cancer type classification. Also note that simply concatenating the gene count and pathway count matrix does not improve accuracy. When adding confounding variables as additional features to NMF produced features from gene count matrix, accuracy improves to 0.8173. Without confounding variables as additional features, NMF also does not produce improved accuracy. For supervised NMF based models, wsNMF and TNMF in fact mostly suffer from accuracy decrease from the unsupervised NMF models, with or without confounding as additional features. This is possibly due to the fact that they do not take confounding variables into their supervised training and may have learned gene and/or pathway groups that are idiosyncratic to the training set. The model based on our ScanMap-derived subject groups on gene count matrix has the best performance, with an accuracy of 0.8468, significantly better ($p < 0.05$ by random permutation test (Noreen, 1989)) than all state-of-the-art models with a notable margin. Also note that in general, we have some performance drop across models when working on pathway count matrix instead of gene count matrix. This likely reflects that our current knowledge on pathway is still growing, and echoes our intuition that ScanMap on gene level data may still provide useful and discriminative features beyond known molecular mechanisms.

Document Visualization. We give an illustrative visualization of the subjects' features learned by ScanMap and comparison models, using t-SNE (Maaten and Hinton, 2008). Fig. 3 shows the visualization of the learned features corresponding to representative models in Table 3. We observe that ScanMap can learn more discriminative subject group features by jointly modeling genetic variants and confounding variables in a supervised manner, compared with state-of-the-art supervised matrix factorization models.

Table 4: Top genetic pathways associated with different cancer risks.

Breast Cancer	Colorectal Cancer
Regulation of IFN-γ signaling	Tryptophan metabolism
Pyrimidine metabolism	TNF receptor signaling pathway
Synthesis of PIPs at the Golgi membrane	Metabolism of lipids
Activation of HOX genes during differentiation	Cholesterol biosynthesis
Regulation of FZD by ubiquitination	TNFR1 Signaling Pathway
Lung Cancer	**Prostate Cancer**
Regulation of TP53 Activity through Acetylation	Signaling by GPCR
Activation of HOX genes during differentiation	Olfactory transduction
PI5P Regulates TP53 Acetylation	Elevation of cytosolic Ca2+ levels
E2F transcription factor network	Regulation of FZD by ubiquitination
p73 transcription factor network	G alpha (s) signaling events

5. Discussion and Future Work

Discovering pathway groups. We identify the top gene groups that are associated with different cancers as follows. For each class f, we rank the classifier weights $\mathbf{W}_f^{(1)}$ and pick the index of top weight, say r. We then take the gene group vector $\mathbf{G}_r$ and pick the indices of genes with nonzero weights (recall we enforce sparsity constraints in gene factor matrix $\mathbf{G}$) associated with class f. We then perform gene set enrichment analysis (Subramanian et al., 2005) for the top gene group associated with each cancer, and the top gene sets associated with different cancer risks are in Table 4. From the table we see that many of the listed gene sets reflect innate key events in the development of individual or multiple types of cancers, consistent with knowledge from wet lab (e.g., Activation of HOX genes during differentiation (Alanee et al., 2012), Regulation of FZD by ubiquitination (Ueno et al., 2013)). Of note, the gene sets listed in Table 4 for each cancer type are all linked to the same top gene group, thus ScanMap additionally connects the gene sets that likely function together in tumorigenesis. Interestingly, interferon signaling and pyrimidine metabolism have been linked in vitro (Lucas-Hourani et al., 2013) and are both connected to breast cancer development and progression (Brown et al., 2017; Mojic et al., 2018); tryptophan and lipid metabolism act in conjunction with TNF receptor signaling in colorectal cancer onset and progression (Zhang et al., 2019; Pavlova and Thompson, 2016); both p53 (encoded by TP53) and p73 are tumor suppressors and p73 network disruption is linked to chronic infections and inflammation of the lungs (Marshall et al., 2016), which together likely predispose an individual to lung cancer; activation of olfactory receptors is documented to be linked to activation of GPCR signaling pathway and intracellular Ca2+ increase, and is in turn associated with proliferation of prostate cancer cells (Neuhaus et al., 2009). These analyses suggest that besides providing useful and discriminative features, ScanMap on gene level data can still provide insights into functional and molecular mechanisms by linking together

multiple pathways that may function together and contribute to cancer development and progression.

Besides significantly improved accuracy and added interpretability, ScanMap also runs fast. For example, on TCGA data, with $k = 500$, ScanMap runs in 16.8 sec. However, our study has limitations. The confounding matrix C is typically of very low dimension (e.g., age, gender, race) for most existing genetic datasets, thus we did not consider the need for factorizing C in this study. When future genetic datasets are accompanied with rich phenotype data, collective matrix factorization methods (Singh and Gordon, 2008; Gunasekar et al., 2016) can be considered. An increasing number of studies now come with both deep genotyping data and deep phenotyping data, and the need for dimensionality reduction is present for both modalities. In these cases, joint matrix factorization is needed where both factorizations share the same subject matrix. This is currently beyond ScanMap's capability but will be future work. In addition, we did not attempt to build a Bayesian version of ScanMap to account for uncertainty and give probablistic interpretaions. In future work, we plan to incorporate priors on matrix structural parameters. Those priors can be used to specify our knowledge or confidence level on putative driver mutations for cancers and other diseases (Bailey et al., 2018). We also plan to perform ablation studies on the changes in interpretation where the classifier used to regularize is from a different hypothesis family. Other NIH dbGaP datasets such as cardiovascular disease related datasets may be used to validate the algorithm, but TCGA dataset is currently one of the largest public genetic dataset and hence the choice in this study. In the future, as even larger genetic datasets will be collected through NIH programs such as All of Us and TopMed, we plan to build a full generative model of the data and evaluate whether the constrained NMF setup is recovered under such generative assumptions on the data.

6. Conclusions

We proposed a novel framework of Supervised Confounding Aware Non-negative Matrix Factorization for Polygenic Risk Modeling (ScanMap) for genetic studies. ScanMap is designed for using supervised learning for polygenic risk modeling to guide the Non-negative Matrix Factorization while capable of considering a variety of confounding variables ranging from subjects' demographics to comorbidities. We showed that ScanMap improves the accuracy of the learned model and provides insights on disease type and status prediction. Confounding aware supervision effectively guided the learning of the groups of genes that are discriminative features and that embody multiple interacting molecular mechanisms (gene sets or pathways). This led to better accuracy with added interpretability. We compared ScanMap with multiple state-of-the-art unsupervised and supervised matrix factorization models, as well as different configurations of genes and pathways as features, with and without confounding variables. ScanMap outperformed all the comparison models significantly ($p < 0.05$). Feature analysis of the learned gene groups that are generated by ScanMap offered more clinical insights about multiple molecular mechanisms that interact with each other and are associated with disease types and status, which were automatically identified from the data.

References

Shaheen Alanee, Fergus Couch, and Kenneth Offit. Association of a hoxb13 variant with breast cancer. *The New England journal of medicine*, 367(5):480, 2012.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500 (7463):415–421, 2013.

Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173 (2):371–385, 2018.

Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

Birgitte Bertelsen, Ida Viller Tuxen, Christina Westmose Yde, et al. High frequency of pathogenic germline variants within homologous recombination repair in patients with advanced cancer. *NPJ genomic medicine*, 4(1):1–11, 2019.

Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. Feature learning with matrix factorization applied to acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1216–1229, 2017.

Kristin K Brown, Jessica B Spinelli, John M Asara, and Alex Toker. Adaptive reprogramming of de novo pyrimidine synthesis is a metabolic vulnerability in triple-negative breast cancer. *Cancer discovery*, 7(4):391–399, 2017.

David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39 (suppl_1):D691–D697, 2010.

Chris Ding, Xiaofeng He, and Horst D, Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, 2005.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2010.

Greg Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13 (2):135–145, 2012.

Suriya Gunasekar, Joyce C Ho, Joydeep Ghosh, Stephanie Kreml, Abel N Kho, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Phenotyping using structured collective matrix factorization of multi–source ehr data. *arXiv preprint arXiv:1609.04466*, 2016.

Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, 2013.

Peter Holmans. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In *Advances in genetics*, volume 72, pages 141–179. Elsevier, 2010.

Seungjin Choi Hyekyoung Lee, Jiho Yoo. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2010.

Suvrit Sra Inderjit Dhillon. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in neural information processing systems (nips)*, pages 283–290, DEC 2005.

Jingu Kim and Haesun Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

Yong-Deok Kim and Seungjin Choi. Weighted nonnegative matrix factorization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1541–1544. IEEE, 2009.

DP Kingma and JL Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Amy N Langville, Carl D Meyer, Russell Albright, James Cox, and David Duling. Algorithms, initializations, and convergence for the nonnegative matrix factorization. *arXiv preprint arXiv:1407.7299*, 2014.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616): 285–291, 2016.

Le Li and Yu-Jin Zhang. Non-negative matrix-set factorization. In *Image and Graphics, 2007. ICIG 2007. Fourth International Conference on*, pages 564–569. IEEE, 2007.

Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

Chao Zhang Liping Jing and Michael K. Ng. Snmfca: Supervised nmf-based image classification and annotation. *IEEE Transactions on Image Processing*, 21(11):4508–4521, 2012.

Marianne Lucas-Hourani, Daniel Dauzonne, Pierre Jorda, Gaëlle Cousin, Alexandru Lupan, Olivier Helynck, Grégory Caignard, Geneviève Janvier, Gwénaëlle André-Leroux, Samira Khiar, et al. Inhibition of pyrimidine biosynthesis pathway suppresses viral growth through innate immunity. *PLoS pathogens*, 9(10), 2013.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov): 2579–2605, 2008.

Kelsey MacMillan and James D Wilson. Topic supervised non-negative matrix factorization. *arXiv preprint arXiv:1706.05084*, 2017.

Clayton B Marshall, Deborah J Mays, J Scott Beeler, Jennifer M Rosenbluth, Kelli L Boyd, Gabriela L Santos Guasch, Timothy M Shaver, Lucy J Tang, Qi Liu, Yu Shyr, et al. p73 is required for multiciliogenesis and regulates the foxj1-associated gene network. *Cell reports*, 14(10):2289–2300, 2016.

Marija Mojic, Kazuyoshi Takeda, and Yoshihiro Hayakawa. The dark side of ifn-γ: its role in promoting cancer immunoevasion. *International journal of molecular sciences*, 19(1): 89, 2018.

Morten Morup, Kristoffer Hougaard Madsen, and Lars Kai Hansen. Approximate l 0 constrained non-negative matrix and tensor factorization. In *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pages 1328–1331. IEEE, 2008.

Franz-Josef Müller, Louise C Laurent, Dennis Kostka, Igor Ulitsky, Roy Williams, Christina Lu, In-Hyun Park, Mahendra S Rao, Ron Shamir, Philip H Schwartz, et al. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, 455(7211):401–405, 2008.

Eva M Neuhaus, Weiyi Zhang, Lian Gelis, Ying Deng, Joachim Noldus, and Hanns Hatt. Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *Journal of Biological Chemistry*, 284(24):16218–16225, 2009.

Eric W Noreen. *Computer-intensive methods for testing hypotheses*. Wiley New York, 1989.

Paul D O'grady and Barak A Pearlmutter. Convolutive non-negative matrix factorisation with a sparseness constraint. In *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pages 427–432. IEEE, 2006.

Pentti Paatero. The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, 8(4):854–888, 1999.

V Paul Pauca, Jon Piper, and Robert J Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.

Natalya N Pavlova and Craig B Thompson. The emerging hallmarks of cancer metabolism. *Cell metabolism*, 23(1):27–47, 2016.

Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658, 2008.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43):15545–15550, 2005.

Dennis L. Sun and Cedric Fevotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6201–6205, 2014.

Koji Ueno, Hiroshi Hirata, Yuji Hinoda, and Rajvir Dahiya. Frizzled homolog proteins, micrornas and wnt signaling in cancer. *International journal of cancer*, 132(8):1731–1740, 2013.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

Jiho Yoo and Seungjin Choi. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5):559–570, 2010.

Zexian Zeng, Andy H Vo, Chengsheng Mao, Susan E Clare, Seema A Khan, and Yuan Luo. Cancer classification and pathway discovery using non-negative matrix factorization. *Journal of biomedical informatics*, 96:103247, 2019.

Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Non-negative matrix factorization on kernels. In *Pacific Rim International Conference on Artificial Intelligence*, pages 404–412. Springer, 2006.

Hong-lian Zhang, Ai-hua Zhang, Jian-hua Miao, Hui Sun, Guang-li Yan, Fang-fang Wu, and Xi-jun Wang. Targeting regulation of tryptophan metabolism for colorectal cancer therapy: a systematic review. *RSC advances*, 9(6):3072–3080, 2019.