Pseudo-labeling Semi-supervised Non-negative Matrix Factorization

Semi-Supervised Integrated Learning of Mutational Signatures for Predicting DNA Repair Pathway Deficiencies

Tommaso Tofacchi

Delft University of Technology

Pseudo-labeling Semi-supervised Non-negative Matrix Factorization

Semi-Supervised Integrated Learning of Mutational Signatures for Predicting DNA Repair Pathway Deficiencies

by

Tommaso Tofacchi

Student number:	4933249	
Master’s programme:	Computer Science, Artificial Intelligence Track	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Project duration:	April, 2023 - March, 2024	
Thesis committee:	Dr. Joana Gonçalves	TU Delft, supervisor
	Dr. Gosia Migut	TU Delft
	Dr. Thomas Höllt	TU Delft
	MSc Sander Goossens	TU Delft, daily supervisor

Preface

Since my very first weeks as a TU Delft student, one aspect of the teaching methodology has resonated with me every day: the *learning by exploring* approach. During my academic years, I have been surrounded by a motivating environment that stimulates curiosity and the exploration of ideas, technologies, and problems branching out of the courses' syllabi. This comes with great satisfaction, as the ability to explore and possess tools applicable to almost any human challenge is what led me to study computer science in the first place. This thesis is an 11-month work that arises from the coupling of *learning by exploring* with my favorite TU Delft's motto "impact for a better society", which brought me to investigate and discover the world of bioinformatics by addressing the relevant problem of improving cancer patients' treatments.

I would like to thank my professor supervisor Joana Gonçalves as well as my daily supervisor Sander Goossens for their inspiring guidance throughout the project: our discussions helped me improve as a researcher, and every meeting's feedback provided invaluable contributions toward the crafting of my final thesis work. I would also like to express my gratitude and affection to the entire research lab, which kept alive the motivating environment that made me fall in love with TU Delft. I would like to thank Martin Skrodzki for joining my initial defense committee, despite unforeseen circumstances preventing us from discussing this thesis together. Concurrently, I would like to acknowledge my deepest gratitude to Gosia Migut and Thomas Höllt for stepping in on such short notice and allowing my graduation procedure to follow the agreed schedule.

Finally, the countless - yet undoubtedly in order - acknowledgments to friends and family. Starting with the former, an immense thank you to all my friends: I am sure that you are still asking yourselves what the heck my thesis is about; nonetheless, I am just grateful for the support and laughs that we have shared over the years. A special shout-out to my roommate Thijs, to whom I owe 90% of my Dutch culture exploration. Thanks to my parents, who did not let a day go by without providing me with the best conditions to cultivate my sparking interests; even if away from home. Thanks to my brother, a relentless source of inspiration for his determination put in every accomplished project. Thanks to my aunt, uncle, and cousins for always pushing me to strive and dream big. Last and certainly not least, thanks to my significant other Carlotta: despite almost a year of listening to my ramblings and brainstorming together, she is still hanging out with me every day with unconditional love.

Tommaso Tofacchi
Delft, March 2024

Semi-Supervised Integrated Learning of Mutational Signatures for Predicting DNA Repair Pathway Deficiencies

Tommaso Tofacchi*, Sander Goossens*, and Joana Gonçalves*
* Pattern Recognition and Bioinformatics, EEMCS Faculty, Delft University of Technology, Netherlands

Abstract—**Motivation**. DNA molecules mutate thousands of times every day. Some mutations are harmful to human cells, and may lead to the loss of function in important genes involved in DNA damage repair (DDR) mechanisms. Diseases such as tumors can exploit mutations in important, driver DDR genes to rapidly proliferate. Specific patterns of mutations (or signatures) are insightful indicators for the presence of DDR malfunctioning, which can be exploited to provide targeted treatment (e.g., by leveraging synthetic lethalities). Different methods have been developed to successfully extract relevant mutational signatures from the genomes of tumor patients. Most approaches are unsupervised and thus do not optimize toward distinguishing DDR deficiencies (DDRd). Supervised approaches achieve this, but rely on labeled in vitro data from tumor cell line genomes during training, due to the lack of DDRd ground truth for tumor patient genomes. Semi-supervised learning could bridge the gap and jointly exploit labeled cell line and unlabeled patient mutation profiles to generalize to patient tumors and provide more clinically relevant DDRd mutational signatures.
Results. We propose Pseudo-labeling Semi-Supervised NMF (PSS-NMF), a novel integrated signature extraction and label prediction method, which extends supervised non-negative matrix factorization (NMF) with the ability to incorporate unlabeled samples into the training via pseudo-labeling. Models learned using PSS-NMF were benchmarked on two different tasks, cancer type and DDRd prediction. PSS-NMF consistently improved prediction for patient tumors over the supervised NMF baseline for both tasks, learning signatures that better transferred to the patient tumor domain: the models achieved Macro F1-scores of 0.3842 and 0.1331 respectively for cancer type prediction, and 0.4928 vs 0.4704 for DDRd prediction. We further validated that PSS-NMF identified DDRd signatures were biologically relevant, by comparing them to known DDRd-related mutational signatures curated in COSMIC and investigating their exposures in patient tumor genomes.

1 Introduction

DNA is located in the nuclei of cells and is organized into two connected strands of complementary bases, which encode the instructions required to correctly synthesize and regulate proteins. During the life cycle of a cell, DNA can be subjected to damage caused by exogenous or endogenous factors at a rate of ~10.000 occurrences per day [1]. DNA base sequences that deviate from their regular structures are referred to as *mutations*. Mutations can be either driver or passenger: the former are non-silent, loss of function, or deleterious mutations that specifically affect driver genes for cancers; the latter comprise all mutations that do not affect driver genes. To detect and correct mutations, cells employ multiple DNA damage repair (DDR) mechanisms, each targeting distinct types of mutations [2]. To function, DDR mechanisms rely on the expression of specific genes. If genomic mutations happen in regions of the DNA relevant to the codification of such genes, DDR mechanisms can be affected and prevented from functioning optimally, leading to the further spreading of undetected mutations.

Detecting malfunctioning DDR mechanisms is crucial for limiting damages associated with the proliferation of genomic instabilities, including those involved in cancer. The direct correction of such malfunctioning behaviors is not possible with currently available medical therapies. However, knowledge of DDRd status in patient tumors can be used for treatment indication using tailored DDR-targeting therapies, including those exploiting known synthetic lethalities [3]. A common approach entails identifying the presence of driver mutations in genes that regulate DNA repair mechanisms [4]. However, these methods are limited by the knowledge of the genes involved in the various DDR pathways, which is incomplete, and gene loss of function can happen indirectly via other processes that are more difficult to uncover. DDR deficiencies (DDRd) also leave specific patterns of passenger mutations in the DNA according to which repair mechanism is affected. DDRd in human genomes can thus be predicted by checking for the presence of such mutation patterns (or *mutational signatures*), without the need to know the exact genes involved.

1.1 Mutational signature extraction methods

Multiple methods have been developed to extract mutational signatures from genomes. Such techniques factorize the mutational profiles (tabular representation of the frequencies of occurrence of different mutation types occurring in genomes) into two matrices: a signature matrix capturing a series of mutation patterns (or mutational signatures), and an exposure matrix weighting the contribution of each signature towards the profile of each genome (their exposures to the extracted mutational signatures).

Unsupervised matrix decomposition. The most successful approaches rely on non-negative matrix factorization (NMF), a matrix decomposition technique that imposes a positivity constraint on the entries of the product matrices. The non-negativity requirement facilitates the biological interpretation of results, since signatures can be interpreted as linear (additive) combinations of probability of occurrences of different mutation types, whereas genome mutation profiles are the result of a linear (additive) mixture of signatures weighed by the corresponding exposures.

Unsupervised NMF signature extraction methods have been effective in finding patterns that can be used to predict DDR deficiencies in genomes [5], [6]. However, unsupervised signature learning focuses on latent patterns while ignoring existing knowledge of the DDRd status of genomes. As a result, the extracted signatures are not optimized to specifically discern signals from DDR deficiencies [7].

Supervised matrix decomposition. By treating the exposure matrix as the input to a prediction model and jointly optimizing mutation profile decomposition and classification, a label-informed (*integrated*) supervised version of NMF can learn mutational signatures that more effectively capture and discriminate between DDRd patterns. One limitation of supervised methods is the scarce availability of labeled genomes with known DDR status. Supervised NMF (S-NMF) [8] tackles the issue by resorting to *in vitro* cell lines with induced gene knockouts. Inhibiting the expression of genes that relate to specific DDR pathways results in mutation profiles obtained from specific known DDRd, which can serve as labeled data for the training of S-NMF.

Lab-grown cells can only partially simulate the complexity of human (*in vivo*) cells, as their DNA was exposed just to a specific set of alterations. Conversely, human cells accumulate mutations from endogenous and exogenous processes over multiple years, resulting in more convoluted mutational profiles [9]. The mutational signatures solely learned from in vitro samples may fail to correctly capture the mutational burden left by DDRd on patient genomes, limiting their clinical applications.

The main limitations of existing DDRd prediction models are twofold: non-integrated and unsupervised approaches lack discrimination for DDRd specific signatures; integrated supervised approaches suffer from limited generalizability to tumor patient data due to exclusive training on in vitro cell line samples. We aim to develop a semi-supervised model that can learn jointly from labeled cell line data and unlabeled patients' genomes, to maximize the discriminability of the extracted mutational signatures as well as their generalizability as predictors of DDRd in patient genomes.

1.2 Pseudo-labeling Semi-Supervised NMF

To address the gap in literature, we first introduce Semi-Supervised NMF (SS-NMF), a learning procedure where both labeled and unlabeled samples are used for matrix decomposition, but only labeled samples contribute to the classification optimization [10], [11].

To further bridge the generalizability between cell line and tumor data, we propose Pseudo-labeling Semi-Supervised NMF (PSS-NMF) as an extension of SS-NMF. Pseudo-labeling is a semi-supervised learning technique that aims to exploit the latent information contained in unlabeled samples to improve the training of a prediction model. It achieves so by incorporating tumor samples in the supervised training procedure using their predicted labels (or *pseudo-labels*) [12]. By iteratively applying pseudo-labeling during training, tumor samples are incrementally included in the computation of the classification loss while contributing to the integrated factorization of input mutation profiles into exposures and signatures.

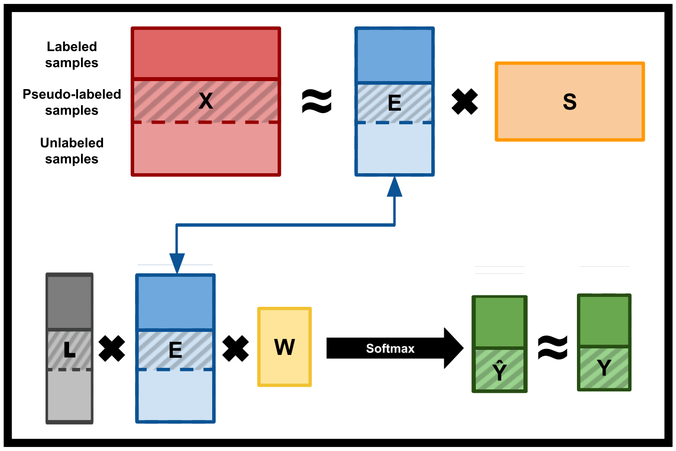


Fig. 1: PSS-NMF components organized in the matrix factorization and label prediction pipeline. $\boldsymbol{X}$ in red is the input matrix; $\boldsymbol{E}$ in blue is the exposure matrix; $\boldsymbol{S}$ in orange is the signature matrix; $\boldsymbol{l}$ in gray is the auxiliary label vector; $\boldsymbol{W}$ in yellow are the multinomial regressor's weights; $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ in green are respectively the predicted and ground truth labels. Darker colors in $\boldsymbol{X}$, $\boldsymbol{E}$, $\boldsymbol{l}$, $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ indicate entries related to labeled samples; diagonal stripes indicate entries related to pseudo-labeled samples; lighter colors indicate entries related to unlabeled samples.

The absence of DDRd ground truth for patient tumors limits the evaluation of (P)SS-NMF's generalization performance. The models are therefore also benchmarked on a cancer type prediction problem with a dataset composed of both cell line and patient tumor samples. Cancer type labels are known for both cell line and tumor samples, enabling an informative and reliable assessment of the methods.

2 Methods

PSS-NMF aims to incorporate (labeled) cell line data with (unlabeled) patient tumor samples in the supervised learning phase, to improve the generalization ability of the model when predicting on patient samples.

2.1 PSS-NMF model

PSS-NMF is an extension of supervised NMF (S-NMF) that further includes unlabeled samples during training as in traditional semi-supervised NMF methods (see Figure 1). We add an element of self-supervised pseudo-labeling to iteratively assign unlabeled data to predicted classes and incorporate them in the labeled dataset during training. An original regularization technique that accounts for unlabeled data in the computation of the classification loss is proposed as an add-on component to the model.

2.1.1 PSS-NMF model components and optimization

NMF models encode input data as a matrix $\boldsymbol{X} \in [0,1]^{N \times M}$, where N is the number of profiled samples and M is the number of mutation types used as features. In the experiments hereby conducted, 96 single-base-substitution (SBS) mutations are considered as features, therefore $M = 96$ [7]. When considering both labeled and unlabeled data, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_L} \\ \boldsymbol{X_U} \end{bmatrix}$, where $\boldsymbol{X_L} \in [0,1]^{N_L \times M}$ and $\boldsymbol{X_U} \in [0,1]^{N_U \times M}$ with N_L and N_U respectively as the number of labeled and unlabeled samples. Each $x_{n,m}$ entry in $\boldsymbol{X}$ contains the probability of mutation m in the profile of sample n.

Annotations for labeled data are encoded in one-hot vectors and concatenated in $\boldsymbol{Y_L} \in \{0,1\}^{N_L \times O}$, where O is the number of output classes. As (P)SS-NMF incorporates unlabeled data in training, $\boldsymbol{Y_U} \in \{0\}^{N_U \times O}$ is constructed so that $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y_L} \\ \boldsymbol{Y_U} \end{bmatrix}$; otherwise, $\boldsymbol{Y} = \boldsymbol{Y_L}$.

To ensure that only labeled entries are utilized for classification optimization, we provide the auxiliary vector $\boldsymbol{l} \in \{0,1\}^N$. Entries l_n in $\boldsymbol{l}$ take value 1 if the associated mutation profile $\boldsymbol{x}_{n,*} \in \boldsymbol{X_L}$, or 0 if $\boldsymbol{x}_{n,*} \in \boldsymbol{X_U}$, where $\boldsymbol{x}_{n,*}$ indicates the mutation profile for a sample n.

Matrix factorization. Mutation profiles in $\boldsymbol{X}$ are approximated by the product of the *exposure* matrix $\boldsymbol{E} \in [0,1]^{N \times K}$ and the *signature* matrix $\boldsymbol{S} \in [0,1]^{K \times M}$, where K is the chosen fixed number of mutational signatures to be found.

$$\boldsymbol{X} \approx \boldsymbol{E}\boldsymbol{S}$$

Non-negative matrix factorization (NMF) [7], [13] is used to extract $\boldsymbol{E}$ and $\boldsymbol{S}$. Update rules for the matrices entries are derived by optimizing the matrix reconstruction error measured using the Frobenius loss $\mathcal{L}_r$ (eq. 1).

$$\mathcal{L}_r = ||\boldsymbol{X} - \boldsymbol{E}\boldsymbol{S}||_F^2 \tag{1}$$

Label prediction. The prediction of sample classes is performed with a multinomial regression component. A function is learned so that the exposure vectors of samples are mapped to one of the O available output classes. The multinomial regression is performed as a linear combination of entries in $\boldsymbol{E}$ with the weight matrix $\boldsymbol{W} \in \mathbb{R}^{K \times O}$; a total of O prediction values are obtained for every sample, and their softmax is taken to obtain class prediction probabilities $\hat{\boldsymbol{Y}} \in \{0,1\}^{N \times O}$ (eq. 2). Finally, samples are classified to the class predicted with the highest probability.

$$\hat{\boldsymbol{Y}} = softmax(\boldsymbol{E}\boldsymbol{W}) \tag{2}$$

The class label prediction is optimized via minimization of the classification loss $\mathcal{L}_c$, a categorical cross-entropy loss that quantifies the errors in class prediction (eq. 3).

$$\mathcal{L}_c = -\sum_{n=1}^{N}\sum_{o=1}^{O} l_n y_{n,o} \log \hat{y}_{n,o} + \lambda_{L2} \sum_{w \in \boldsymbol{W}} w^2 \tag{3}$$

In the categorical cross-entropy loss, l_n denotes entries in the $\boldsymbol{l}$ auxiliary vector. The indicator $\boldsymbol{l}$ ensures that only labeled data from $\boldsymbol{X_L}$ are considered for classification optimization, since semi-supervised NMF models also include unlabeled samples in the input matrix $\boldsymbol{X}$. An L_2 regularization term is added to mitigate overfitting on training data, with its effect controlled by hyperparameter λ_{L2} [14].

Model optimization. The total loss for integrated NMF models is computed as the sum of the reconstruction and classification losses, as in eq. 4

$$\mathcal{L}_{Tot} = \mathcal{L}_r + \lambda_c \mathcal{L}_c \tag{4}$$

Hyperparameter λ_c (*classification strength*) controls the weight of the classification loss in the total loss.

Components of NMF models are optimized using the following iterative updates to accomplish gradient descent on $\mathcal{L}_{Tot}$ (eqs. 5, 6, and 7; derivations of the update formulas at time $t+1$ in Appendix A):

$$\boldsymbol{S}_{t+1} = \boldsymbol{S}_t \odot \frac{\boldsymbol{E}_t^T \boldsymbol{X}}{\boldsymbol{E}_t^T \boldsymbol{E}_t \boldsymbol{S}_t} \tag{5}$$

$$\boldsymbol{E}_{t+1} = \boldsymbol{E}_t \odot \frac{\boldsymbol{X}\boldsymbol{S}_t^T - \frac{\lambda_c}{2} \boldsymbol{l}_t (\hat{\boldsymbol{Y}}_{\boldsymbol{t}} - \boldsymbol{Y}) \boldsymbol{W}_t^T}{\boldsymbol{E}_t \boldsymbol{S}_t \boldsymbol{S}_t^T} \tag{6}$$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu_W [\boldsymbol{E}_t^T \boldsymbol{l}_t (\hat{\boldsymbol{Y}}_t - \boldsymbol{Y}) + 2\lambda_{L2} \boldsymbol{W}_t] \tag{7}$$

In the update formula for $\boldsymbol{W}$ (eq. 7), μ_W is the constant learning rate for the prediction component. Division and dot-multiplication in the formulas are applied element-wise.

2.1.2 Procedures

Pseudo-labeling. Pseudo-labeling is a self-learning technique in which a model iteratively predicts potential labels (*pseudo-labels*) for unlabeled samples and incrementally adds them to the labeled set $\boldsymbol{X_L}$. During training, PSS-NMF predictions on unlabeled data are analyzed and samples are assigned to the most probable pseudo-labeled class [15] according to the two criteria of *confidence* and *balance*, following the pseudo-labeling procedure outlined in Algorithm 1.

Algorithm 1: Pseudo-labeling procedure

```
   Input: $\hat{\boldsymbol{Y}}, \boldsymbol{l}, O, c_t$
   Output: $\boldsymbol{Y}, \boldsymbol{l}$
1  $\boldsymbol{P}_{pool} = \{\hat{y}_{n,*} \in \hat{\boldsymbol{Y}} \mid \boldsymbol{l}_n = 0 \wedge max(\hat{y}_{n,*}) \geq c_t\}$
2  if $\exists\, o \in O$ s.t. $|\boldsymbol{P}_{pool_{*,o}}| = 0$ then
3      return $\boldsymbol{Y}, \boldsymbol{l}$
4  $N_{P_{min}} = min(|\boldsymbol{P}_{pool_{*,o}}| \; \forall\, o \in O)$
5  $N_{P_{min}} = \frac{N_{P_{min}}}{2}$
6  for $o \in O$ do
7      for $\hat{y}_{s,o} \in rand\_sample(\boldsymbol{P}_{pool*,o}, N_{P_{min}})$ do
8          $\boldsymbol{Y}_{s,o} = 1$
9          $\boldsymbol{l}_s = 1$
10 return $\boldsymbol{Y}, \boldsymbol{l}$
```

The confidence criterion is applied in line 1 to construct an initial pool of pseudo-labeling candidate samples $\boldsymbol{P_{pool}}$ out of the unlabeled samples that are inferred with a predicted class probability larger than a given *confidence threshold* c_t [16]. Enforcing a confidence threshold on pseudo-labeling candidates can mitigate the model's supervised learning from being affected by pseudo-labels assigned to samples predicted with low class probabilities, which are characterized by greater uncertainty. However, excessively large c_t values restrict the pseudo-labeling pool to a subset of few very confidently predicted samples. If incorrectly predicted samples from $\boldsymbol{P}_{pool}$ are added to the labeled set, model component updates will further propagate the erroneous behavior to subsequent predictions, resulting in an even more restricted next pseudo-labeling pool that contains more incorrectly predicted samples with increasingly high confidence (*confirmation bias*) [17]. Conversely, too low (or permissive) confidence thresholds enable the model with a large $\boldsymbol{P}_{pool}$ of samples predicted with poor class probabilities, hence more prone to being incorrectly classified. Parameter c_t therefore acts as a trade-off between the two extremes.

Alongside confirmation bias, pseudo-labeling techniques can lead to class imbalance: an over-/under-representation of classes among pseudo-labeling candidates. If a model pseudo-labels samples to a single class (e.g., due to confident predictions happening only with respect to a single label), the balance between classes in the resulting labeled set could be altered, steering the classification loss optimization towards the most represented class, and resulting in the latter influencing the component updates more than other classes. The criterion applied in lines 2 to 5 addresses class balance, by ensuring that the pseudo-labeling happens only if the candidate pool contains the same number of samples > 0 across all classes - namely the minimum amount of all classes pseudo-labeling candidates $N_{P_{min}}$. The division-by-2 factor on $N_{P_{min}}$ allows pseudo-labeling to gradually be executed over epochs, providing a logarithmic inclusion of pseudo-labeled data in the labeled set during training.

According to the confidence and balance criteria, $N_{P_{min}}$ candidate samples per class are finally pseudo-labeled. The process is described in lines 6-9: for every class, $N_{P_{min}}$ samples are randomly selected from the class-predicted candidates in $\boldsymbol{P}_{pool}$ and a *hard label* (i.e., a one-hot vector) for its predicted class is produced by updating $\boldsymbol{Y}$ accordingly [15]. The corresponding entries in $\boldsymbol{l}$ are set to 1, so that the new pseudo-labels can be considered for the computation of the classification loss $\mathcal{L}_c$.

Unlabeled regularization. In the semi-supervised learning setting, it is important to acknowledge the difference in sources between labeled cell lines and unlabeled patient tumor samples, as they may originate from distinct data distributions - the DDRd prediction generalization problem is an example of it. Despite mapping to the same target classes, samples from the labeled and unlabeled set may exhibit different distributions in the input space that can prevent the correct domain adaptation when applying pseudo-labeling after supervised training [18]. To address the issue, many pseudo-labeling techniques implement open-set approaches that treat the classes of the unlabeled set separately from the labeled set [19], [20]. A common methodology among these considers all the unlabeled data as part of an additional unlabeled data class o_u added to the output set of classes O [21].

To promote domain translation from labeled to unlabeled data, a regularization approach is proposed as an additional component for semi-supervised NMF models, named *unlabeled regularization* (*U.R.*). It is inspired by the *dustbin class* approach presented in [22], with a key novel difference. In the mentioned dustbin class approach, all unlabeled data is classified as part of an additional class o_u, and their optimization for the classification loss function is computed towards it. However, this formulation stems from the assumption that unlabeled data has a negligible probability of containing supervised samples, which is not the case when dealing with tumor patients' genomes and DDRd. With U.R., unlabeled data that is not yet pseudo-labeled contributes to the computation of the classification loss $\mathcal{L}_c$ by optimizing towards the 0-labeled class in $\boldsymbol{Y_U}$.

The 0-labeled class effectively serves as a separate additional class for unlabeled data, which guides the model to become less confident in its predictions. When optimizing towards the 0 class, we are reducing the probability of prediction for an unlabeled sample towards any of the one-hot encoded classes. For such and for our initial proposition of 0 labels for the unlabeled data target vector $\boldsymbol{Y}_U$, the classification loss under U.R. is trivialized to the traditional cross-entropy loss in eq. 8.

$$\mathcal{L}_{c_{U.R.}} = -\sum_{n=1}^{N}\sum_{o=1}^{O} y_{n,o}\log\hat{y}_{n,o} + \lambda_{L2}\sum_{w\in\boldsymbol{W}} w^2 \qquad (8)$$

After every training iteration, the unlabeled data's probability prediction for any one-hot class is actively encouraged to be smaller. At the following iteration of training, the model can further leverage the knowledge incorporated with newly pseudo-labeled samples when predicting again the unlabeled profiles.

2.1.3 Training & testing of integrated NMF models

Training. The training procedure of integrated NMF models (SNMF, SS-NMF, PSS-NMF) is split into two steps.

In the first step, 10 independent runs are executed in parallel. For each run, $\boldsymbol{S}$, $\boldsymbol{E}$ and $\boldsymbol{W}$ are initialized to a different set of random values. The model components are then iteratively updated according to the corresponding update rules (eq. 5, 6, 7). When the total loss $\mathcal{L}_{Tot}$ converges (i.e., for two consecutive epochs the difference in total loss remains under a given threshold), PSS-NMF performs the pseudo-labeling routine at the following training epoch. The model is then optimized accounting for the new labeled set, and a new pseudo-labeling procedure is executed at the following $\mathcal{L}_{Tot}$ convergence. After ensuring that a minimum amount of epochs has elapsed for all the runs, reaching a *training stoppage* checkpoint epoch, and reporting $\mathcal{L}_{Tot}$ convergence, the first training step is concluded.

In the second step of training, the set of 10 $\boldsymbol{S}$ matrices is collected and processed. Since every independent run can produce a different $\boldsymbol{S}$ due to random initialization, clustering techniques are used to ensure that a unique $\boldsymbol{S}^{clustered} \in [0,1]^{K\times M}$ is computed as a consensus of the results from all runs. The set of signature matrices of K mutational signatures each are partitioned via a variation of K-means clustering [7], where each resulting cluster contains a signature from each of the runs, and the final signatures in $\boldsymbol{S}^{clustered}$ are the centroids of the K clusters. Finally, $\boldsymbol{E}^{clustered}$ and $\boldsymbol{W}^{clustered}$ are calculated by fixing $\boldsymbol{S}^{clustered}$, setting $\lambda_c = 0$ and refitting to the input matrix $\boldsymbol{X}$.

Testing. Integrated NMF models first compute an exposure matrix $\boldsymbol{E}^{Test}$ for the new test data $\boldsymbol{X}^{Test}$, then use it together with the previously found $\boldsymbol{W}^{clustered}$ to predict the sample output labels. To obtain $\boldsymbol{E}^{Test}$, the non-negative least squares algorithm (NNLS) [23] with fixed $\boldsymbol{S}^{clustered}$ and $\boldsymbol{X}^{Test}$ is used. $\hat{\boldsymbol{Y}}^{Test}$ is computed as in Section 2.1.1: for each sample the class with the highest probability determines the final predicted label.

2.2 Evaluation

For DDRd prediction, the aim of using PSS-NMF and SS-NMF is to improve the generalization of DDRd predictions for patient tumors by incorporating mutation profiles of tumor genomes into the training. However, the absence of ground truth labels for tumors presents as a challenge to analyze model performance. Alternatively, we first evaluate the generalizability of the integrated NMF methods on the prediction of cancer types from mutational profiles. In contrast to DDRd prediction, the ground truth cancer type labels are known for both cell lines and patient tumors. PSS-NMF was benchmarked on the described experiments against S-NMF and SS-NMF models, according to signature stability, prediction accuracy and macro F1-score.

2.2.1 Cancer type prediction data

Cell line samples were sourced from the Cancer Cell Line Encyclopedia (CCLE) database [24], containing 1864 tumor-derived cell lines spanning 181 cancer types. Unlabeled patient tumor samples were collected from The Cancer Genome Atlas (TCGA) [25], containing over 20.000 sequenced genomes from oncological patients across 33 different cancer types.

To select the target cancer types, we first kept only those overlapping between the CCLE and TCGA datasets. We further refined the list by considering only cancer types with at least 15 CCLE samples and 400 TCGA samples. Next, we compared the cosine similarities for each cancer type average mutational profile across the two datasets. As 4 clusters of cosine similar tumor profiles emerged for both the CCLE and TCGA datasets, we select a cancer type per cluster according to the 4-tuple with the highest total cosine cross-dissimilarity overlapping the two datasets (for all the data selection plots and cosine similarity comparisons, see Appendix B.1).

Resultingly, we used 35 tumor cell lines from bladder (BLCA), 79 lung (LUAD), 16 skin (SKCM), and 24 uterine (UCEC) as labeled samples in the cancer type prediction task. As unlabeled samples, we considered a total of 441 BLCA, 513 LUAD, 466 SKCM, and 447 UCEC patient genomes.

2.2.2 DDRd prediction data

We used cell line data from Zou et. al [26], corresponding to a total of 173 human-induced pluripotent stem-cell samples, including replicates. Each sample had one of 42 different induced gene knockouts (KOs). Of the 42 gene KOs, 9 provided samples with mutational profiles that could be related to deficiencies in one of 3 different DDR pathways and were sufficiently distinct from 8 control cell lines: KOs of genes MSH6, MSH2, MLH1, PMS2, and PMS1 resulted in 23 samples deficient in mismatch repair (MMR); KOs of UNG and OGG1 in 8 samples deficient in base excision repair (BER); KOs of EXO1 and RNF168 resulted in 7 samples deficient in homologous recombination repair (HR).

Tumor patient samples were collected from the TCGA database. The labels, derived based on driver mutations in DNA repair genes, were originally provided by Knijnenburg et al. [27] and obtained from Volkova et al. [28]. According to the computed variant allele frequency [29], gene mutations were classified as heterozygous or homozygous, with the latter corresponding to a high degree of confidence in the inhibition of a gene [30]. Based on these labels, unlabeled tumor samples for the DDRd prediction experiment could be further categorized as homozygous, heterozygous, or wild-type (or not carrying a mutation in DDR genes) genomes. Across all cancer types, 105 tumor samples were labeled as uniquely homozygously mutated for genes involved in the MMR pathway, 36 in the BER pathway, and 96 samples in the HR pathway. The breakdown of tumor samples with gene mutation annotations is available in Appendix B.2.

The correlation between the involvement of a gene in a DDR pathway and its (non-)inhibition in a sample does not provide confirmation of the presence (or absence) of a DDRd. Therefore, Volkova's labels do not represent a ground truth for the repair status of tumor samples, but can be treated as indicative for a quantitative evaluation of predictive performance.

2.2.3 Train and test sets

In both experiments, cell line and tumor patients' samples are concatenated in unified datasets. The resulting datasets are then split into disjoint validation and test sets. The validation sets comprise 75% of the respective datasets and are further divided into 3 folds each, which are used to perform cross-fold validation on the benchmarked models for hyperparameter selection. The test sets, composed of the remaining 25% samples, are used to analyze the performance of the models on unseen data for the optimal hyperparameter configurations found during validation.

To assess the performance of PSS-NMF in different settings of unlabeled data, both experiments are conducted with varying amounts of TCGA samples. In one version, the cell line-to-patient data ratio is 2:1 (66.6% labeled cell line + 33.3% unlabeled TCGA data); in the other, the ratio is 1:1 (50% labeled cell line + 50% unlabeled TCGA data). For different ratios, the amount of labeled samples is kept the same; the amount of unlabeled samples varies according to the ratios, ensuring the same distribution of data.

Cancer type datasets. Validation folds comprise 400 cell line samples each, equally distributed across the 4 cancer type classes. Due to the lack of sufficient CCLE mutational profiles, bootstrapping as in [8] is used to ensure the same amount of samples per class. TCGA samples are then uniformly added according to the desired unlabeled data percentage, to retain balanced classes.

The test set is similarly constructed, with 400 cell lines and varying balanced amounts of TCGA profiles.

DDRd datasets. Each validation fold was composed of 400 cell line samples uniformly picked across the 3 DDRd pathways + control profiles (ensuring class balance), obtained via bootstrapping due to the limited amount of available cell line data. TCGA samples with homozygous driver mutations in DDR genes were distributed across the

3 validation folds according to their respective DDRd labels, and the remaining spots were filled with heterozygously mutated and wild-type samples until the desired labeled-unlabeled data ratio was reached. Human samples were distributed in equal proportions across validation folds, despite homozygous mutations' labels presenting the class imbalance noted in Section 2.2.2.

The test set comprised 4000 cell line samples, uniformly distributed across classes and bootstrapped to necessity. A higher count of cell lines compared to validation folds was picked for better resolution and consistency when analyzing the experimental results. For TCGA data, 57 homozygously mutated samples were included in the test set (27 MMRd-related, 21 HRd-related, and 9 BERd-related).

2.2.4 Evaluation metrics

As each training of NMF models produces a set of 10 randomly initialized solutions, it is expected that the found signatures may differ from each other. Nonetheless, a robust model should be able to consistently identify similar signatures regardless of the pre-train stochasticity. To quantify the closeness of signatures found across all runs, the metric of *stability* is introduced.

$$Stability = \frac{1}{K}\sum_{k=1}^{K} Silhouette(\text{Cluster}_K) \qquad (9)$$

Stability is defined as the average silhouette width of the K clusters of signatures found during training (eq. 9). Silhouette width (defined as in [31]) ranges from -1 to 1, with a higher value corresponding to high reproducibility (i.e., the models' ability to consistently find a similar set of signatures in each of the 10 runs). The similarity of two individual signatures is computed as cosine similarity.

Two different metrics are used to evaluate the methods' predictive performance. When the evaluation dataset is balanced (such as with both experiments' cell line data), the traditional classification metric of *accuracy* is considered.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (10)$$

Accuracy directly quantifies the rate of correct class assignments performed by a model. In eq. 10, TP and TN are respectively the true positive and true negative counts; FP and FN are respectively the false positive and false negative counts.

When dealing with class-imbalanced data (such as with human samples in DDRd experiments), accuracy is not suitable. Accuracy assigns equal weight to every prediction, hence a model could trivially predict all samples to the single dominant class in a dataset and still achieve a high score in the metric, due to the severe under-representation of other classes. To account for class imbalance, *Macro-average* (or simply *Macro*) *F1-score* is used.

$$Macro\ F1\text{-}score = \frac{1}{O}\sum_{o\in O}\frac{2\cdot Precision_o \cdot Recall_o}{Precision_o + Recall_o} \qquad (11)$$

Macro F1-score is computed as the per-class arithmetic mean of the F1-scores achieved by a model, ensuring equal weight from all classes in the final score regardless of their differences in size (eq. 11). The F1-score considers both *precision* (the ratio of TP over the sum of TP and FP) and *recall* (the ratio of TP over the sum of TP and FN) per class, penalizing models that maximize one at the expense of the other. A high macro F1-score is therefore indicative of a model that can provide good predictions across all classes.

2.2.5 Benchmarked models

For the evaluation of PSS-NMF, we benchmarked the model and its variants against supervised and semi-supervised NMF for both the cancer type and DDRd prediction tasks. The following models were used:

- Supervised NMF [8]
- SS-NMF (this paper)
- SS-NMF with U.R. (this paper)
- PSS-NMF (this paper)
- PSS-NMF with U.R. (this paper)

The models were validated by performing cross-validation on the 3 folds that compose the validation sets.

The best hyperparameter configuration for each model was chosen as the combination that guaranteed pareto-optimality [32] for stability and accuracy on cell line data predictions, enforcing a minimum stability threshold of 0.9. The stability requirement was imposed to ensure that only models that could consistently find similar signatures across multiple runs were selected.

TCGA accuracy, despite being a more apt measure to quantify model generalizability than cell line accuracy, cannot be used as a metric for choosing the best-performing hyperparameter combination, as labels for unlabeled data are assumed to be absent. Generally, any metric of prediction performance on unlabeled data is impossible to determine.

Combinations of K, λ_c, and λ_{L2} parameters were cross-validated for S-NMF and SS-NMF in a grid-search manner. After finding pareto-optimal configurations for the SS-NMF models, the hyperparameter values were fixed for the respective PSS-NMFs. Multiple confidence thresholds were then validated on the same prediction and stability metrics + pseudo-labeling quality (i.e., number of pseudo-labeled samples and correctness of the predicted pseudo-labels). All pareto-optimal plots are available in Appendix C. Finally, the benchmarked models were trained with their respective optimal hyperparameter sets on the train sets and evaluated on the test sets.

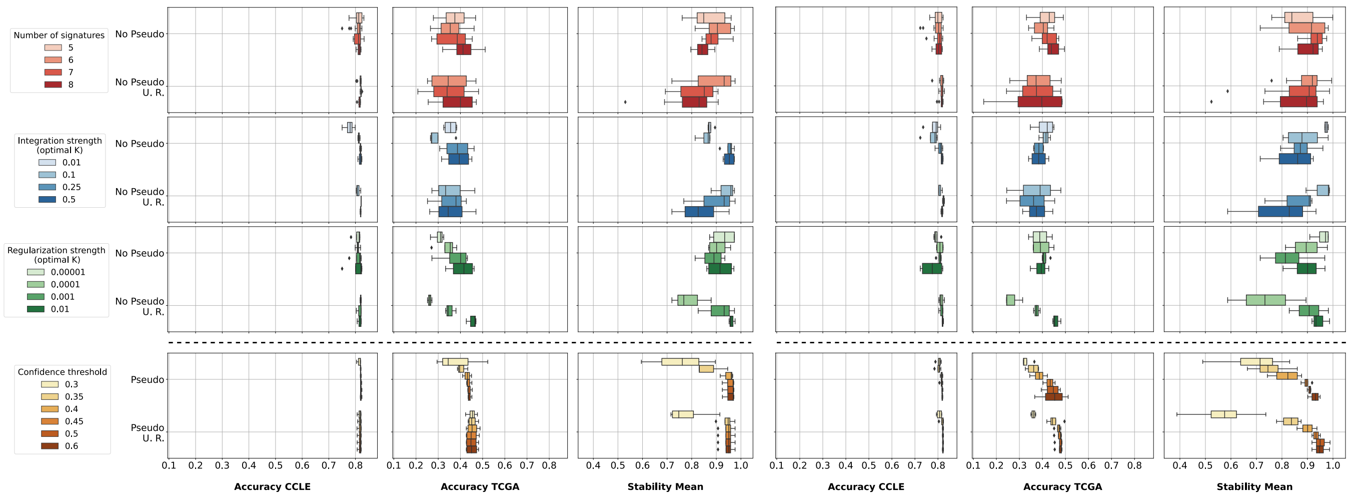
3 Results

3.1 Impact of hyperparameters on performance

To perform a quantitative hyperparameter analysis of SS-NMF and PSS-NMF, we focused on cross-validation performance for the cancer type prediction task with different proportions of unlabeled samples in the train sets (see Figure 2).

3.1.1 SS-NMF hyperparameters and performance

We assess median and interquartile range (IQR, 25th to 75th percentile) of performance obtained using each set of parameters, where a well-performing model should exhibit high median and low IQRs (variation) for both accuracy and stability (Figure 2).



(a) 33% unlabeled data

(b) 50% unlabeled data

Fig. 2: Validation results for cancer type prediction. Rows provide aggregation of results for various hyperparameters: aggregations for K are performed across all runs; aggregations for λ_c and λ_{L2} are relative to runs with the optimal value of K; aggregations for *confidence threshold* account for optimal hyperparameter configurations of the respective SS-NMF models and are separated by the others with a dashed line. Columns indicate cell line accuracy, TCGA accuracy, and mean stability reported across runs.

Unlabeled data proportion affects accuracy and stability. A higher proportion of unlabeled samples in the train set led to an average decrease in IQR for TCGA accuracies (from 0.117 with 33% unlabeled data to 0.066 with 50%) and an increase in IQR for stabilities (from 0.089 for 33% to 0.139 for 50%) when considering SS-NMF results with their optimal number of signatures (see pareto-optimal curves in Appendix C.1). When training on the dataset with 50% unlabeled data proportion, TCGA accuracies report a higher median across validation runs (0.350 for 33% vs 0.399 for 50%), but the median stability decreased (0.914 for 33% vs 0.909 for 50%).

Based on the reported results, more unlabeled samples lead to higher instability (stability columns 3 and 6 in Figure 2). It is expected, as higher proportions of unlabeled data provide a larger pool of samples that can affect the update of components initially trained only on cell line profiles, leading to a wider variety of suitable signature combinations. The TCGA mutation profiles are more convoluted than cell lines (as explained in Section 1.1), hence a larger number of TCGA samples could have a stronger impact on the decomposition of $\boldsymbol{X}$ into $\boldsymbol{E}$ and $\boldsymbol{S}$.

The decrease in median stability, however, still retains acceptable values for the higher unlabeled data proportion (above the 0.9 stability threshold discussed in Section 2.2.5). The usage of additional unlabeled data can then be accepted as it does not lead to excessively unstable models, and it also brings an increase in generalization ability with higher median and lower IQR values for TCGA accuracy.

Unlabeled regularization increases variation without improving median accuracy. The usage of U.R. leads to higher IQRs in TCGA accuracies (0.158 for U.R. models vs 0.076 for non-U.R.) and stabilities (0.131 vs 0.096 respectively) when considering aggregations of SS-NMF results for their optimal number of signatures (see pareto-optimal curves in Appendix C.1). TCGA accuracy reports a higher median value for non-U.R. models (0.368 for U.R. vs 0.381 for non-U.R.) and the same stability medians at 0.912.

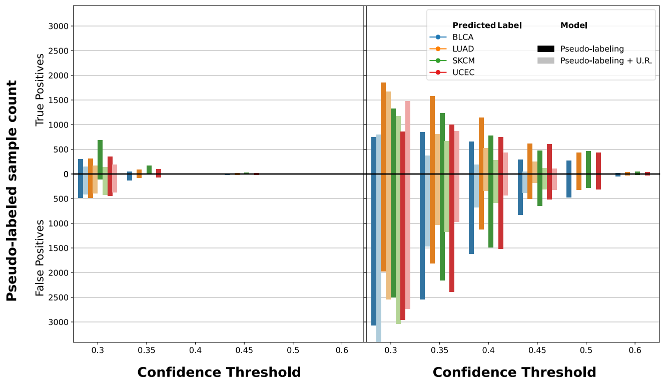
From validation runs, the unlabeled regularization component does not contribute to an improvement in model generalization ability to tumor patient data (TCGA), instead resulting in a less consistent retrieval of signatures (as indicated by U.R. models' higher stability IQRs).

CCLE accuracy is a poor indicator of generalizability. Semi-supervised models showed minimal variation in terms of CCLE accuracy. Their performance was stable at around 0.8 CCLE accuracy, with low variability (CCLE accuracy median = 0.815, IQR = 0.018). Conversely, TCGA accuracy exhibited more noticeable variation (TCGA accuracy IQR = 0.098), especially when considering the predictions across the number of signatures (see the top row in Figure 2). It suggests that CCLE accuracy does not reflect generalization ability of semi-supervised models to tumor data (TCGA), nor is indicative of the optimal hyperparameter choice in this regard. However, resorting to TCGA accuracy for choosing a model is not applicable, since unlabeled data is expected to be unannotated in the intended real-world use cases.

3.1.2 Pseudo-labeling activity analysis

PSS-NMF models were evaluated using multiple confidence thresholds coupled to the remaining pareto-optimal hyperparameter configurations of the baseline SS-NMFs (pareto plots in Appendix C.1). Pseudo-labeling activity was analyzed per confidence threshold on the quantity of pseudo-labeled samples and their pseudo-label correctness. The hereby analyzed samples account for all pseudo-labeled tumor mutation profiles (TCGA) over validation folds and runs (see Figure 3).

Higher unlabeled data proportion correlates to higher pseudo-labeling activity. When PSS-NMF is trained on the 50% unlabeled data setting, the number of pseudo-labeled



(a) 33% unlabeled data

(b) 50% unlabeled data

Fig. 3: Pseudo-labeling true and false positive counts for cancer type prediction validation folds with different unlabeled data proportions, at varying confidence thresholds. Darker bars correspond to PSS-NMF activity, while lighter bars to PSS-NMF + U.R. activity. Colors indicate the pseudo-labeled classes.

samples increases for every confidence threshold compared to the 33% unlabeled data setting. Specifically, the ratio of pseudo-labeled samples at 33% unlabeled data over pseudo-labeled samples at 50% unlabeled data grows according to a sigmoid curve starting at 5.9 for $c_t = 0.3$ and plateauing at 251 for $c_t = 0.5$ (for the complete data reporting, see Table 3 in Appendix C.3). More unlabeled profiles provided PSS-NMF with a larger pool of potential pseudo-labeling candidates, increasing the probability of predicting at least one candidate per class (crucial to satisfy the balance criterion, see Section 2.1.2). Increased pseudo-labeling could guide the model to be more confident in the predictions on other unlabeled samples, favoring additional pseudo-label candidates in the following training epochs. Supporting the claim, the average PSS-NMF (non-U.R. and U.R. results combined) prediction probability on TCGA samples for 33% unlabeled data validation runs was 0.417 (0.399 for SS-NMFs), whilst for 50% runs it was 0.475 (0.49 for SS-NMFs). Full prediction probability data is available in Table 4 in Appendix C.3.

Higher confidence thresholds decrease pseudo-labeling activity. Larger values of the confidence threshold parameter result in fewer samples being pseudo-labeled (see Table 3 in Appendix C.3). Qualitatively, higher confidence thresholds correlate with more accurate pseudo-labels; however, excessively high c_t values lead to a decrease in pseudo-labeling correctness, following the expectation presented in Section 2.1.2. $c_t = 0.4$ achieves 0.442 pseudo-labeling accuracy on average, with decreasing accuracy for both smaller and larger c_t values: the lowest $c_t = 0.3$ and the highest $c_t = 0.6$ respectively achieve 0.357 and 0.379 pseudo-labeling accuracy. For the complete results, see Table 4 in Appendix C.3.

Unlabeled regularization decreases pseudo-labeling activity. Unlabeled regularization had a strong negative impact on the pseudo-labeling process according to the validation runs, resulting in the pseudo-labeling of fewer samples. This is apparent in Figure 3 when comparing the lighter bars of U.R. models (49516 pseudo-labeled samples in total) to darker bars of non-U.R. models (31648 pseudo-labeled samples), coupled with a decrease in average pseudo-labeling precision (0.329 for U.R. models vs 0.447 for non-U.R. models).

Non-U.R. models exhibited higher pseudo-labeling activity for higher unlabeled data percentages compared to U.R. models. Moreover, a larger unlabeled data proportion in the train set enabled PSS-NMF to focus more on optimizing the loss toward tumor patient samples, instead of cell line profiles (as analyzed in Section 3.1.1).

3.2 Final (P)SS-NMF test set performance

3.2.1 (P)SS-NMF improve cell line, patient tumor prediction

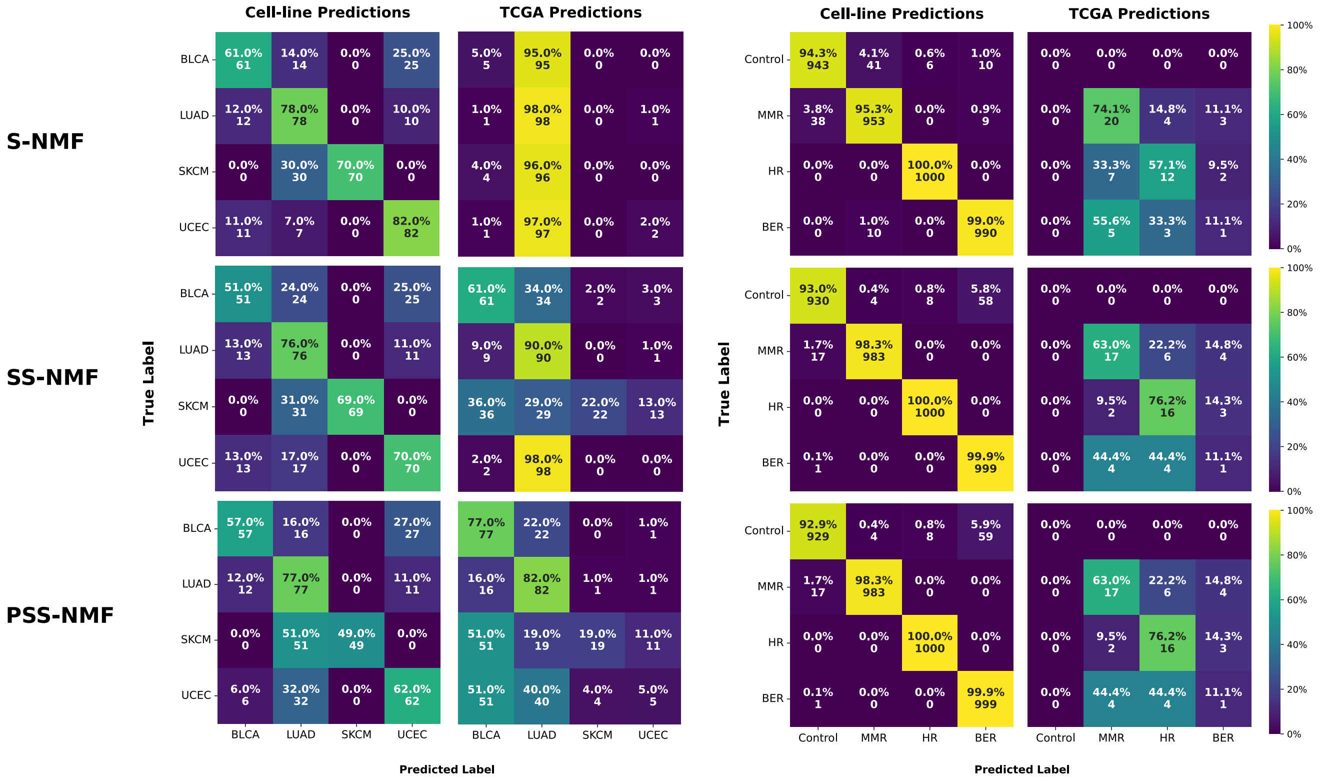
We finally assessed test set performance of PSS-NMF on the cancer type and DDRd prediction tasks using only non-U.R. models with 50% unlabeled data in the corresponding train set, given that these showed superior validation performance.

Cancer type prediction. SS-NMF and PSS-NMF performed better than S-NMF on TCGA data, respectively with 0.3635, 0.3842, and 0.1331 macro F1-scores (Table 1, left).

Supervised NMF (S-NMF) demonstrated a prediction bias for patient data towards the LUAD class, which SS-NMF and PSS-NMF mitigated by including unlabeled samples during training (Figure 4(a), right column). Cell line LUAD samples exhibited the total highest cosine similarity between their average mutational profile and the average per-class mutational profiles from TCGA samples, which could help explain the behavior of S-NMF (Figure 9 in Appendix D.1). Both the pseudo-labeling and non-pseudo-labeling models found optimal hyperparameter configurations for a higher number of signatures than S-NMF ($K = 5$ for S-NMF, $K = 6$ for SS-NMF, $K = 7$ for PSS-NMF). This resulted in the (P)SS-NMF models disposing of additional degrees of freedom, which could help accommodate the more diverse mutation profiles of human samples.

On cell line data, SS-NMF achieved a better macro F1-score (0.7309) than S-NMF (0.6713), whilst PSS-NMF led to worse performance (0.6218). By jointly considering the TCGA and cell line prediction results, we can notice how PSS-NMF is able to discover signatures that were more discriminative for the cancer type prediction of patient genomes, despite not retaining the entirety of patterns optimized toward CCLE data compared to SS-NMF.

Stability-wise, the semi-supervised models did not provide the same robustness in retrieving similar sets of signatures over multiple runs than S-NMF (stability of 1 for S-NMF, 0.892 for SS-NMF, and 0.947 for PSS-NMF). This could be related to the increased complexity of tumor patients' genomes compared to cell line data, which semi-supervised models incorporate in the training set for learning. A further interpretation of tumor patient and cell line data distributions is provided in Section 3.2.3 by exploring PCA decomposition plots. Nonetheless, SS-NMF and PSS-NMF resulted in sufficiently reliable models with high stability, that could produce more accurate cancer type predictions on both cell line and TCGA profiles.



(a) Cancer type predictions

(b) DDRd predictions.

Fig. 4: Confusion matrices for the test set performances of S-NMF, SS-NMF, and PSS-NMF in their optimal hyperparameter configurations across the two experiments at 50% unlabeled data. For each square in the matrices, the bottom value indicates the amount of samples belonging to the respective row's class which are predicted with the label indicated by the column; the top values indicate the percentages relative to the row totals (i.e., relative to all samples belonging to the row's class). The color gradients for each cell reflect such percentages and are interpreted with the color bars on the right.

	Cancer type prediction (50% unlabeled data)								DDRd prediction (50% unlabeled data)							
	Hyperparameters				Results				Hyperparameters				Results			
Model	K	λ_c	λ_2	c_t	Added Samples	Stability	F1-score CCLE	F1-score TCGA	K	λ_c	λ_2	c_t	Added Samples	Stability	F1-score cell line	F1-score TCGA
Supervised NMF	5	0,001	0,001	-	-	1	0,6713	0,1331	5	0,1	0,0001	-	-	1	0,9714	0,4704
Semi-Supervised NMF	6	0,5	0,01	-	-	0,892	0,7309	0,3635	6	0,1	0,0001	-	-	0,871	0,9779	0,4928
Pseudo-Labeling NMF	7	0,5	0,001	0,6	84	0,947	0,6218	0,3842	6	0,1	0,0001	0,9	12	0,871	0,9776	0,4928

TABLE 1: Test set results for benchmarked models on the two experiments at 50% unlabeled data.

DDRd predictions. Similar trends were also observed in the DDRd prediction, where the semi-supervised models outperformed S-NMF in both cell line and patient tumor prediction. However, the difference in classification performance was less prominent than what was observed for cancer type prediction: in the cancer type prediction experiment, the differences in macro F1-score between the best and worst performing models were 0.1091 for CCLE samples and 0.2511 for TCGA samples; in the DDRd prediction experiment, the reported macro F1-score differences were 0.0065 for cell line samples and 0.0224 for TCGA samples (Table 1, right).

For cell line DDRd prediction, S-NMF already achieved a high macro F1-score (0.9714), thus allowing only for marginal improvements. Nonetheless, SS-NMF and PSS-NMF were able to improve over S-NMF (macro F1-scores of 0.9779 and 0.9776 respectively), principally due to a better classification of cell line MMRd samples that compensated for the decrease in classification accuracy of control samples (Figure 4(b), left column).

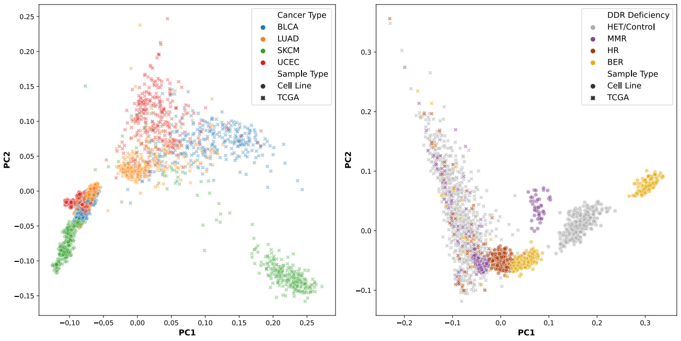
For patient tumor DDRd prediction, performances were close across all models. Supervised NMF assigned most TCGA samples to the MMRd class. This behavior was somewhat mitigated by both SS-NMF and PSS-NMF, which leveraged the finding of an additional signature to improve HRd predictions despite a smaller drop in MMRd class accuracy (Figure 4(b), right column).

Semi-supervised models were less stable than S-NMF (stability of 0.871 for (P)SS-NMF, 1 for S-NMF), as seen previously for cancer type prediction.

3.2.2 *Pseudo-labeling enforces balance in predictions*

The DDRd prediction experiment did not show any differences in predictive capabilities on patient tumor data between the two semi-supervised methods. In fact, they predicted the same deficiency labels for all TCGA samples.

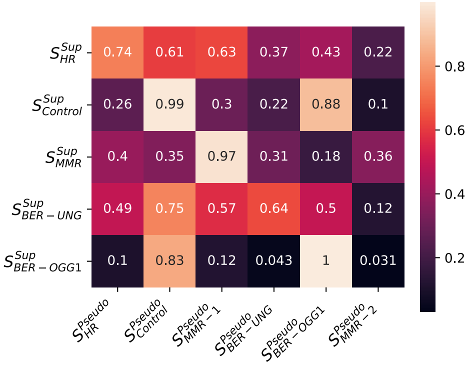
When evaluated on the cancer type prediction task, SS-NMF and PSS-NMF exhibited different behaviors. Both SS-NMF and S-NMF showed a tendency to assign most patient



(a) Cancer type.

(b) DDRd.

Fig. 5: Plots of the first two principal components (PC1, PC2) for PCA decomposition of the train sets with 50% unlabeled data proportions used in (a) cancer type prediction, and (b) DDRd prediction.



	S^{Pseudo}_{HR}	$S^{Pseudo}_{Control}$	S^{Pseudo}_{MMR-1}	$S^{Pseudo}_{BER-UNG}$	$S^{Pseudo}_{BER-OGG1}$	S^{Pseudo}_{MMR-2}
S^{Sup}_{HR}	0.74	0.61	0.63	0.37	0.43	0.22
$S^{Sup}_{Control}$	0.26	0.99	0.3	0.22	0.88	0.1
S^{Sup}_{MMR}	0.4	0.35	0.97	0.31	0.18	0.36
$S^{Sup}_{BER-UNG}$	0.49	0.75	0.57	0.64	0.5	0.12
$S^{Sup}_{BER-OGG1}$	0.1	0.83	0.12	0.043	1	0.031

Fig. 6: Cosine similarities for the matched signatures extracted by S-NMF and PSS-NMF.

tumor samples to the LUAD class, which was most prominent for S-NMF. PSS-NMF improved the predictive accuracy for the BLCA class, incurring a smaller performance drop for the classification of LUAD tumor patients' profiles.

PSS-NMF was also able to perform marginally better on UCEC recognition (the worst-predicted class), correctly classifying 5 samples compared to the 0 of SS-NMF.

3.2.3 Feature space of mutation profiles

By investigating PCA decompositions for both cancer type and DDRd data, we aimed to identify patterns that could set the two prediction tasks apart from each other in terms of model predictive performances on both TCGA and cell line samples (Figure 5).

In the cancer type prediction dataset, CCLE samples clustered completely separately from the TCGA profiles (Figure 5a); in the DDRd case, the distinction between cell line and tumor profiles presents some overlap (Figure 5b). In fact, the leftmost cluster of MMRd cell lines in Figure 5b (which involves all gene KOs other than PMS1) exhibits mutation profiles that are close to patient tumor samples along the PC1 axis. This correlates with S-NMF higher MMRd prediction rate for TCGA samples compared to the semi-supervised approaches.

DDRd cell line samples tended to form isolated clusters according to their label. Conversely, for the cancer type dataset a large amount of CCLE samples overlaid in a common area of BLCA, LUAD, and UCEC samples. This suggests that cell line mutation profiles might not be as easily classifiable with respect to their cancer type labels as they are in the DDRd prediction task. In fact, NMF approaches achieved lower classification scores on cell line data in the cancer type prediction than in the DDRd prediction.

Similarly, patient tumor TCGA samples in the cancer type dataset tended to cluster in areas with high overlap (albeit mirroring the CCLE data distribution), as a large number of TCGA samples from all cancer types were located at the center of the plot. Coupled with the lower prediction accuracy achieved on cell line data if compared to the DDRd task, it could indicate that signature decomposition approaches may not be the most informative for this type of task: other matrix factorization approaches in literature report similar performance for the cancer type prediction problem [33], [34]. Nonetheless, the DDRd tumor patient samples also exhibited such behavior, as no clear grouping emerged for any of the homozygously mutated DDRd labels. Additionally, according to the first 2 PCs, heterozygously mutated samples covered the same area spanned by homozygously mutated samples (leftmost cluster in Figure 5b). This increases the complexity of learning patterns from confidently-predicted DDRd human profiles for the semi-supervised approaches.

3.3 Signature analysis for DDRd

To provide a biological interpretation of the DDRd prediction results, PSS-NMF and S-NMF signatures were compared based on cosine similarity to the single-base substitution (SBS) tumor mutational signatures curated in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [35]. COSMIC SBS signatures were extracted using SigProfiler [7] from 2780 whole-genomes produced by the Pan-Cancer Analysis of Whole Genomes (PCAWG) Network [36]. The signatures were further validated with independent studies, to uncover their potential etiologies (or causes of emergence of such mutation patterns).

3.3.1 PSS-NMF extracts extra MMRd-related signature

Supervised NMF extracted 5 signatures from the DDRd train set profiles, while PSS-NMF identified 6 signatures. Five of the 6 PSS-NMF signatures showed high cosine similarity with one matching signature of S-NMF (from 0.64 to 1, see Figure 6). The sixth signature, labeled S^{Pseudo}_{MMR-2}, was dissimilar to all other S-NMF signatures (just 0.36 cosine similarity with the highest matching supervised signature) and represents a new mutation pattern discovered by PSS-NMF. Signature S^{Pseudo}_{MMR-2} was similar to two MMRd-related signatures in COSMIC, *SBS6* and *SBS15* (0.84 and 0.71 respective cosine similarities, see Figure 7). The other MMR signatures found by both models, S^{Sup}_{MMR} and S^{Pseudo}_{MMR-1}, were highly similar (0.97 cosine similarity) and principally related to a different set of MMRd signatures in COSMIC, namely *SBS44* (0.93 and 0.94 cosine similarities respectively) and *SBS20* (0.77 and 0.74).

As displayed in Figure 8, MMRd cell line and tumor samples exhibited the largest exposure to MMRd-related

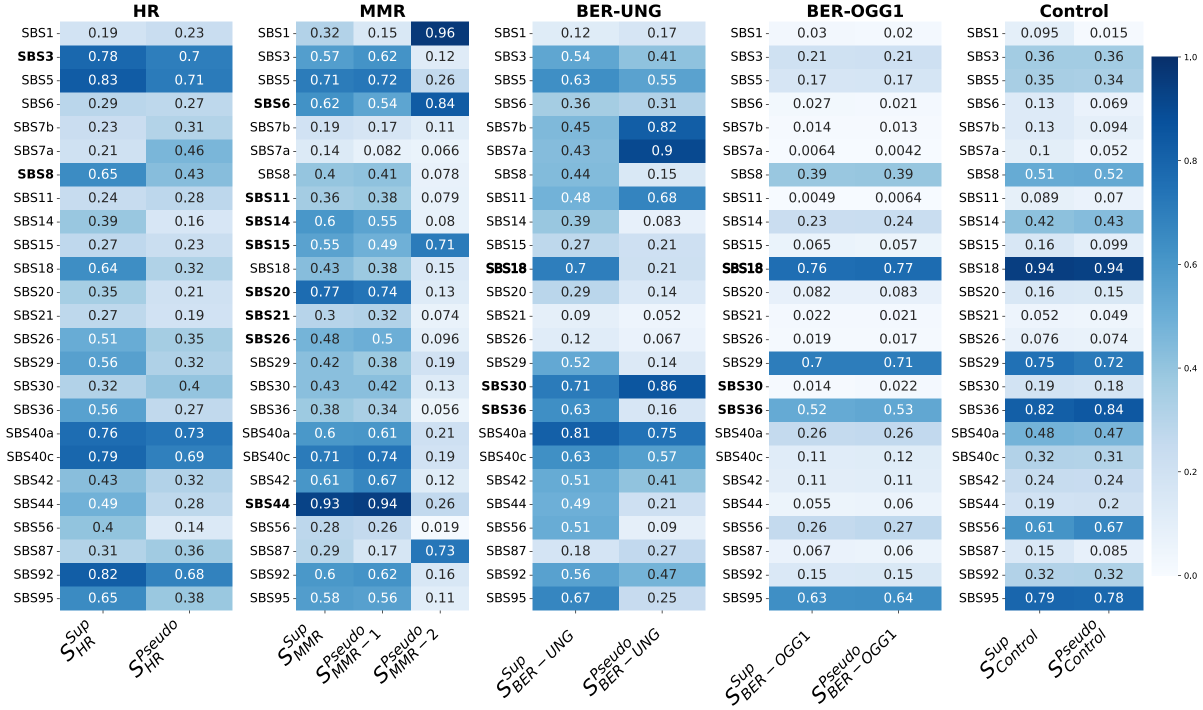


Fig. 7: Comparison of S-NMF and PSS-NMF signatures (x-axis) to mutation signatures available in COSMIC (y-axis). Signatures are grouped per DDRd: for each group, in bold are reported the COSMIC signatures known to have the corresponding DDRd etiology. Numbers represent the cosine similarities across COSMIC, S-NMF (indicated with *Sup*), and PSS-NMF (indicated with *Pseudo*) signatures. Darker colors indicate higher similarities (see the color bar on the right).

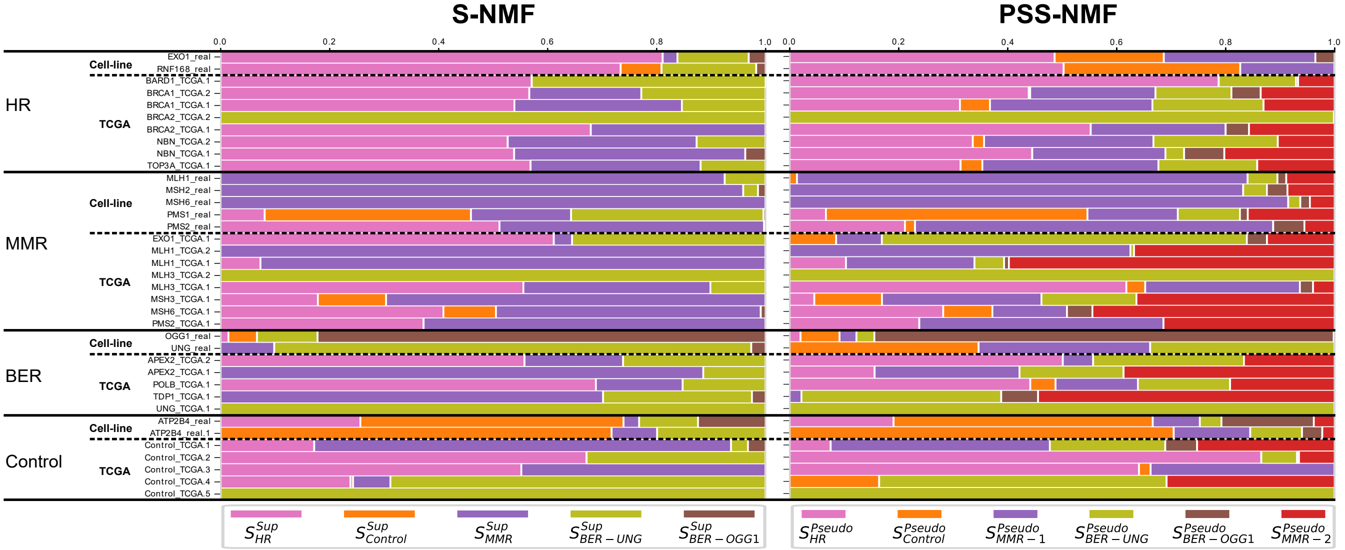


Fig. 8: Exposures for S-NMF and PSS-NMF color-matched signatures on a selection of DDRd + control samples. Profiles are grouped per DDR label, with dashed lines separating cell line from TCGA samples.

signatures for both models. On the labeled cell line set, S-NMF struggled to assign high S^{Sup}_{MMR} exposure to the PMS1-KO mutation profile, resulting in the misclassification of its bootstrapped samples. It seemed that PSS-NMF corrected its predictions by decomposing PMS1-KO profiles with S^{Pseudo}_{MMR-2}, as these samples showed the highest exposure to the newly found signature (colored red in the exposure plot) among all cell line samples.

3.3.2 Extra PSS-NMF signature present in patient tumors

The additional signature S^{Pseudo}_{MMR-2} was also highly similar (0.96 cosine similarity) to *SBS1* in COSMIC, a known clock-like signature caused by spontaneous deamination of aging cells, and also related to unrepaired $C > T$ mismatches. It could hint at a double etiology associated with S^{Pseudo}_{MMR-2} and its emergence in the semi-supervised context for the DDRd task: on one side, it can be related to the semi-supervised training on tumor patient mutation profiles; on the other, the similar *SBS1* signature is often co-occurrent in tumors exhibiting microsatellite instabilities associated with MMRd [37]. As evidence of the somewhat tumor-specific nature exhibited by the signature newly found by PSS-NMF, we investigated the TCGA profile decomposition plots in Figure 8: PSS-NMF assigned S^{Pseudo}_{MMR-2} exposure to all TCGA samples, and no other DDR-deficient cell line profile (apart from the MMRd ones) exhibited any exposure to S^{Pseudo}_{MMR-2}.

3.3.3 $S^{Pseudo}_{BER-UNG}$ more COSMIC-unique vs. $S^{Sup}_{BER-UNG}$

Both S-NMF and PSS-NMF found two signatures that could be associated with BERd, bearing different etiologies.

The first pair of signatures, $S^{Sup}_{BER-UNG}$ and $S^{Pseudo}_{BER-UNG}$, exhibited larger exposures in cell line samples with UNG knockouts (Figure 8), hence their nomenclatures. They showed the lowest similarity across the matched pairs of supervised and pseudo-labeling signatures, with a cosine similarity of 0.64 (Figure 6). In Figure 7, we can see how the signature $S^{Sup}_{BER-UNG}$ achieves 0.71 cosine similarity with *SBS30* (related to mutations that lead to inactivation of the NTHL1 gene), 0.7 with *SBS18* and 0.63 with *SBS36* (both related to MUTYH mutations). Both NTHL1 and MUTYH genes are linked to the BER pathway [38].

$S^{Pseudo}_{BER-UNG}$ was even more similar to *SBS30* (0.86 cosine similarity), but less so to *SBS18* (0.21) and *SBS36* (0.16).

3.3.4 PSS-NMF signatures generalize marginally better than S-NMF on tumor samples for MMRd and BERd

When inspecting the sample decomposition plots in Figure 8 for patient tumor mutation profiles, PSS-NMF exposures to related DDRd signatures were overall higher than those of S-NMF: for TCGA samples homozygously mutated in MMRd-driver genes, the PSS-NMF MMRd signatures had an average exposure of 0.545 vs 0.514 with S-NMF; for the BERd signatures, PSS-NMF BERd signatures had an average exposure of 0.414 vs 0.365 with S-NMF. HRd samples showed the opposite trend, as PSS-NMF related signatures exposures were lower on average than those of S-NMF (0.398 vs 0.498 respectively). Both models struggled to extract a representative signature for control samples, with an PSS-NMF average control signature exposure in related samples of 0.037 vs 0.001 with S-NMF.

4 Conclusion

Pseudo-labeling Semi-Supervised NMF was implemented as an extension of Supervised NMF, integrating unlabeled samples in the factorization optimization as Semi-Supervised NMF and in the classification optimization as pseudo-labeling self-learning training regimes. The model aims to improve predictive capabilities when applied to a dataset of samples originating from a different data source than the inputs to traditional supervised learning.

Cell line accuracy proved to be a poor indicator for the models' predictions on patients' data, resulting in a suboptimal criterion of choice for the best hyperparameter combinations when optimizing toward generalizability performance. A point of improvement for semi-supervised model selection would be finding more apt criteria, for instance by investigating metrics that consider unlabeled data clustering after predictions.

The semi-supervised models benefit from training on real patients' unlabeled data, boosting the generalization of predictions on human samples in all experimental settings. Improvements on the cancer type classification task are especially evident, whilst performance gains are more limited for DDRd predictions. In the latter experiment, Pseudo-labeling NMF shows a tendency to extract mutational signatures that are more representative of cancer patients' deficiencies, at the expense of cell line decomposition correctness.

On the DDRd prediction problem, PSS-NMF performs generally on par with traditional SS-NMF approaches, due to a low pseudo-labeling activity that prevents major modifications to the extracted signatures. Nonetheless, PSS-NMF never incurs extensive performance degradations in any of the experiments, often improving the baselines by allowing for more balanced prediction accuracy across the available classes. By pseudo-labeling samples and including them in the classification loss computation, PSS-NMF occasionally worsens its inference on cell line profiles.

Alongside pseudo-labeling, an original unlabeled regularization component was validated. Validation runs were complex to interpret, as U.R. contributes to increased entropy in the collected results when compared to the non-U.R. models. However, the performances obtained by models augmented with unlabeled regularization can often be better than non-U.R. models (see Appendix D), hence further investigation could lead to promising insights. An additional regularization coefficient can be explored to limit the influence of U.R. and allow for more consistent results, dictating the extent of its influence on $\boldsymbol{E}$ values. A coefficient value $\in [0, 1]$ could be assigned to unlabeled entries during $\mathcal{L}_{Tot}$ computation and effectively weight their regularization impact on $\boldsymbol{E}$.

Pseudo-labeling itself could be improved in terms of classification correctness for pseudo-labeled samples. Possible solutions to avoid the confirmation bias emerging from supervised training could follow the ideas developed in [39], by learning an ensemble of integrated NMF models and pseudo-labeling profiles according to the majority voting obtained from each of the learned models. Another topic worth investigating would be the usage of *soft pseudo-labels* instead of assigning hard pseudo-labels to samples, therefore scaling the classification error produced by each pseudo-labeled profile according to their prediction probabilities. Pseudo-labels could also be iteratively updated during training as the model refines its predictions, contrary to the current immutable approach upon their first assignment.

SS-NMF and PSS-NMF constitute a step forward toward interpretable DDRd prediction and problem-specific signature extraction. The findings in this research could hopefully contribute to advancements in clinical applications for DDRd detection, in combination with other DDRd-specific biomarkers (microsatellite instability for MMRd; quantification of large-scale structural variants for HRd, such as telomeric allelic imbalance, large-scale transition, or loss of heterozygosity) [40].

References

[1] T. Lindahl, "Instability and decay of the primary structure of DNA," *Nature*, vol. 362, no. 6422, pp. 709–715, Apr. 1993.

[2] N. Chatterjee and G. C. Walker, "Mechanisms of DNA damage, repair, and mutagenesis," *Environ. Mol. Mutagen.*, vol. 58, no. 5, pp. 235–263, Jun. 2017.

[3] W. G. Kaelin, "The concept of synthetic lethality in the context of anticancer therapy," *Nature Reviews Cancer*, vol. 5, no. 9, pp. 689–698, Sep. 2005.

[4] N. J. Curtin, "DNA repair dysregulation from cancer driver to therapeutic target," *Nature Reviews Cancer*, vol. 12, no. 12, pp. 801–817, Dec. 2012.

[5] H. Davies, D. Glodzik *et al.*, "HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures," *Nature Medicine*, vol. 23, no. 4, pp. 517–525, Apr. 2017.

[6] X. Zou, G. C. C. Koh *et al.*, "A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage," *Nat. Cancer*, vol. 2, no. 6, pp. 643–657, Jun. 2021.

[7] L. B. Alexandrov, S. Nik-Zainal *et al.*, "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep.*, vol. 3, no. 1, pp. 246–259, Jan. 2013.

[8] S. Goossens, Y. Tepeli, and J. Gonçalves, "Integrated learning of mutational signatures and prediction of dna repair deficiencies," Master's thesis, Delft University of Technology, 2022.

[9] A. Ertel, A. Verghese *et al.*, "Pathway-specific differences between tumor cell lines and normal and tumor tissue cells," *Molecular Cancer*, vol. 5, no. 1, p. 55, Nov. 2006.

[10] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010.

[11] J. Haddock, L. Kassab *et al.*, "Semi-supervised nmf models for topic modeling in learning tasks," 10 2020.

[12] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[13] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[14] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 78. [Online]. Available: https://doi.org/10.1145/1015330.1015435

[15] A. Galstyan and P. Cohen, "Empirical comparison of hard and soft label propagation for relational classification," 06 2007, pp. 98–111.

[16] K. Sohn, D. Berthelot *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *CoRR*, vol. abs/2001.07685, 2020. [Online]. Available: https://arxiv.org/abs/2001.07685

[17] E. Arazo, D. Ortego *et al.*, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," *CoRR*, vol. abs/1908.02983, 2019. [Online]. Available: http://arxiv.org/abs/1908.02983

[18] J. Quionero-Candela, M. Sugiyama *et al.*, *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[19] W. J. Scheirer, A. Rocha *et al.*, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1757–1772, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:12035411

[20] A. Bendale and T. E. Boult, "Towards open set deep networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:14240373

[21] L. Han, H.-J. Ye, and D. chuan Zhan, "On pseudo-labeling for class-mismatch semi-supervised learning," *ArXiv*, vol. abs/2301.06010, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:254587653

[22] X. Zhang and Y. LeCun, "Universum prescription: regularization using unlabeled data," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 2907–2913.

[23] V. Franc, V. Hlavac, and M. Navara, "Sequential coordinate-wise algorithm for the non-negative least squares problem," 09 2005, pp. 407–414.

[24] J. Barretina, G. Caponigro *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012.

[25] K. Chang, C. J. Creighton *et al.*, "The cancer genome atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.

[26] X. Zou, G. C. C. Koh *et al.*, "A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage," *Nature Cancer*, vol. 2, no. 6, pp. 643–657, Jun. 2021.

[27] T. A. Knijnenburg, L. Wang *et al.*, "Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas," *Cell Rep.*, vol. 23, no. 1, pp. 239–254.e6, Apr. 2018.

[28] N. V. Volkova, B. Meier *et al.*, "Mutational signatures are jointly shaped by DNA damage and repair," *Nature Communications*, vol. 11, no. 1, p. 2169, May 2020.

[29] S. P. Strom, "Current practices and guidelines for clinical next-generation sequencing oncology testing," *Cancer Biol. Med.*, vol. 13, no. 1, pp. 3–11, Mar. 2016.

[30] T. A. Knijnenburg, L. Wang *et al.*, "Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas," *Cell Rep.*, vol. 23, no. 1, pp. 239–254.e6, Apr. 2018.

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[32] A. Chinchuluun and P. Pardalos, "A survey of recent developments in multiobjective optimization," *Annals of Operations Research*, vol. 154, no. 1, pp. 29–50, October 2007. [Online]. Available: https://ideas.repec.org/a/spr/annopr/v154y2007i1p29-5010.1007-s10479-007-0186-0.html

[33] X. Lyu, J. Garret *et al.*, "Mutational signature learning with supervised negative binomial non-negative matrix factorization," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i154–i160, 07 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa473

[34] K. P. Soh, E. Szczurek *et al.*, "Predicting cancer type from tumour DNA signatures," *Genome Medicine*, vol. 9, no. 1, p. 104, Nov. 2017.

[35] Z. Sondka, N. B. Dhir *et al.*, "COSMIC: a curated database of somatic variants and clinical data for cancer," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1210–D1217, 11 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad986

[36] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, pp. 82–93, Feb. 2020.

[37] A. Farmanbar, R. Kneller, and S. Firouzi, "Mutational signatures reveal mutual exclusivity of homologous recombination and mismatch repair deficiencies in colorectal and stomach tumors," *Scientific Data*, vol. 10, no. 1, p. 423, Jul. 2023.

[38] R. D. Weren, M. J. Ligtenberg *et al.*, "Nthl1 and mutyh polyposis syndromes: two sides of the same coin?" *The Journal of Pathology*, vol. 244, no. 2, pp. 135–142, 2018. [Online]. Available: https://pathsocjournals.onlinelibrary.wiley.com/doi/abs/10.1002/path.5002

[39] J. Chavoshinejad, S. A. Seyedi *et al.*, "Self-supervised semi-supervised nonnegative matrix factorization for data clustering," *Pattern Recognition*, vol. 137, p. 109282, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322007610

[40] E. H. Stover, P. A. Konstantinopoulos *et al.*, "Biomarkers of Response and Resistance to DNA Repair Targeted Therapies," *Clinical Cancer Research*, vol. 22, no. 23, pp. 5651–5660, 11 2016. [Online]. Available: https://doi.org/10.1158/1078-0432.CCR-16-0247

Appendix A
Derivations

Components of NMF models are optimized by deriving their respective update rules via gradient descent application on $\mathcal{L}_{Tot}$. After an iteration step t during training, the models' components at $t+1$ are computed as follows:

$$\boldsymbol{S}^{t+1} = \boldsymbol{S}^{t} - \eta_S \cdot \nabla^t_{\boldsymbol{S}} \mathcal{L}^t_{Tot} \tag{12}$$

$$\boldsymbol{E}^{t+1} = \boldsymbol{E}^{t} - \eta_E \cdot \nabla^t_{\boldsymbol{E}} \mathcal{L}^t_{Tot} \tag{13}$$

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^{t} - \eta_W \cdot \nabla^t_{\boldsymbol{W}} \mathcal{L}^t_{Tot} \tag{14}$$

In the update equations, η_* indicate the respective learning rates per component, and ∇^t_* their partial derivatives of the total loss computed at iteration t. The learning rates are the same adaptive ones used in [7] and adapted in S-NMF [8] to accommodate for the presence of $\boldsymbol{W}$:

$$\eta_S = \frac{\boldsymbol{S}}{2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S}} \tag{15}$$

$$\eta_E = \frac{\boldsymbol{E}}{2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T} \tag{16}$$

$$\eta_W = \frac{\mu_W}{\lambda_c} \tag{17}$$

In the formula for η_W, μ_W is the constant learning rate for the predictor component. Gradients with respect to $\mathcal{L}_{Tot}$ are then derived. For readability purposes, we remove the superscripts to indicate the iteration of the gradient's computation.

The derivatives for $\nabla_{\boldsymbol{S}}\mathcal{L}_{Tot}$ and $\nabla_{\boldsymbol{E}}\mathcal{L}_{Tot}$ with respect to the reconstruction loss $\mathcal{L}_r$ are widely used in literature [13], hence we report the final computations as:

$$\nabla_{\boldsymbol{S}}\mathcal{L}_r = -2\boldsymbol{E}^T\boldsymbol{X} + 2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S} \tag{18}$$

$$\nabla^t_{\boldsymbol{E}}\mathcal{L}_r = -2\boldsymbol{X}\boldsymbol{S}^T + 2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T \tag{19}$$

Next, we derive $\nabla_{\boldsymbol{E}}\mathcal{L}_{Tot}$ and $\nabla_{\boldsymbol{W}}\mathcal{L}_{Tot}$ with respect to the classification loss $\mathcal{L}_c$:

$$\nabla_{\boldsymbol{W}}\mathcal{L}_c = \frac{\partial\mathcal{L}_c}{\partial\boldsymbol{W}} \tag{20}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_c = \frac{\partial\mathcal{L}_c}{\partial\boldsymbol{E}} \tag{21}$$

To complete the derivations, we first need to compute:

$$d\mathcal{L}_c = \frac{\partial\mathcal{L}_c}{\partial\boldsymbol{E}\boldsymbol{W}} : dEW \tag{22}$$

where : indicates the Frobenius inner product. We thus compute:

$$\frac{\partial\mathcal{L}_c}{\partial\boldsymbol{E}\boldsymbol{W}} = \frac{\partial\mathcal{L}_c}{\partial\boldsymbol{L}}\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{E}\boldsymbol{W}} + \frac{\partial\mathcal{L}_c}{\partial\boldsymbol{Y}\log\hat{\boldsymbol{Y}}}\frac{\partial\boldsymbol{Y}\log\hat{\boldsymbol{Y}}}{\partial\boldsymbol{E}\boldsymbol{W}} = (\boldsymbol{Y}\log\hat{\boldsymbol{Y}}) * 0 + \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) = \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) \tag{23}$$

and:

$$dEW = dE\boldsymbol{W} + \boldsymbol{E}dW \tag{24}$$

Substituting in the computation of $d\mathcal{L}_c$:

$$\begin{aligned} d\mathcal{L}_c &= \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : (dE\boldsymbol{W} + \boldsymbol{E}dW) && (25)\\ &= \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : dE\boldsymbol{W} + \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : \boldsymbol{E}dW && (26)\\ &= \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T : dE + \boldsymbol{E}^T\boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) : dW && (27) \end{aligned}$$

By combining all parts of the equation and recalling that $\boldsymbol{W}$ is constant for the gradient with respect to $\boldsymbol{E}$ (i.e., $dW = 0$) and vice versa, we obtain:

$$\nabla_{\boldsymbol{W}}\mathcal{L}_c = \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{W}} = \boldsymbol{E}^T\boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) \tag{28}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_c = \frac{\partial \mathcal{L}_c}{\partial \boldsymbol{E}} = \boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y})\boldsymbol{W}^T \tag{29}$$

Finally, the derivative of the L2 regularization term for $\boldsymbol{W}$ is standard:

$$\frac{\partial(\lambda_{L2}\sum_{w\in\boldsymbol{W}} w^2)}{\partial \boldsymbol{W}} = 2\lambda_{L2}W \tag{30}$$

To conclude, we can combine all the partial derivatives for the individual terms of the total loss function $\mathcal{L}_{Tot}$. We obtain the final gradients:

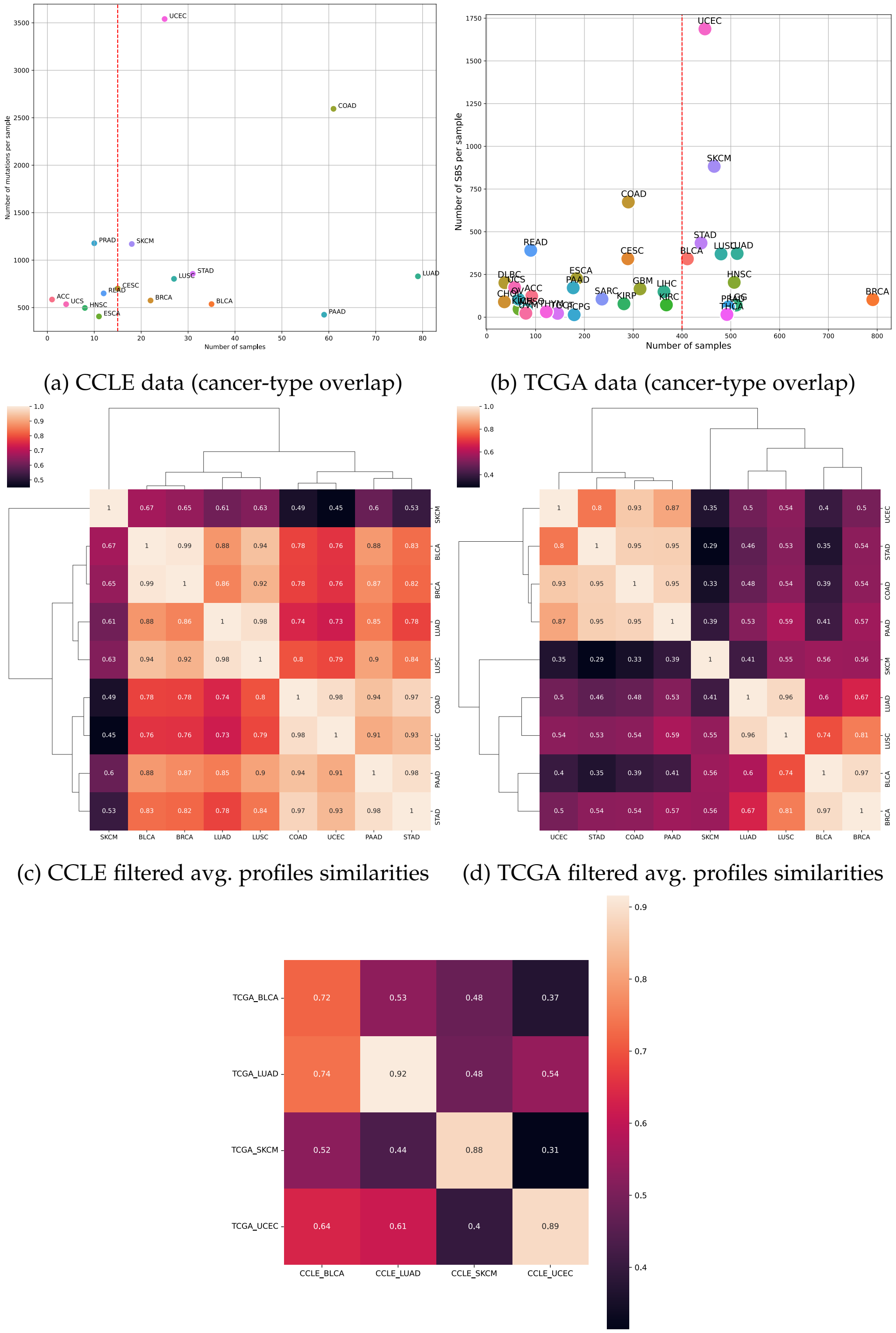
$$\nabla_{\boldsymbol{S}}\mathcal{L}_{Tot} = -2\boldsymbol{E}^T\boldsymbol{X} + 2\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{S} \tag{31}$$

$$\nabla_{\boldsymbol{E}}\mathcal{L}_{Tot} = -2\boldsymbol{X}\boldsymbol{S}^T + 2\boldsymbol{E}\boldsymbol{S}\boldsymbol{S}^T + \lambda_c\boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y} + 2\lambda_{L2}) \tag{32}$$

$$\nabla_{\boldsymbol{W}}\mathcal{L}_{Tot} = \lambda_c[\boldsymbol{E}^T\boldsymbol{L}(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + 2\lambda_{L2}\boldsymbol{W}] \tag{33}$$

Appendix B
Dataset composition

B.1 Cancer-type experiment datasets



(a) CCLE data (cancer-type overlap)

(b) TCGA data (cancer-type overlap)

(c) CCLE filtered avg. profiles similarities

(d) TCGA filtered avg. profiles similarities

(e) CCLE vs TCGA avg. profile similarities

Fig. 9: Data exploration for the CCLE and TCGA datasets. Figures (a) and (b) present scatterplots for the two datasets, with dotted red lines traced at the relative number of samples filtering values (15 for CCLE, 400 for TCGA). Figures (c) and (d) show clustermaps for the cosine similarities of the average mutational profiles for the filtered tumors in the two datasets. Figure (e) displays the cosine similarities for the chosen cancer-types' CCLE and TCGA avg. profiles.

B.2 DDRd experiment datasets

Cancer	HR_het	MMR_het	BER_het	HR_&_MMR_het	HR_&_BER_het	BER_&_MMR_het	HR_&_MMR_&_BER_het	HR_hom	MMR_hom	BER_hom	TOT_het	TOT_hom	TOT_control	TOT_samples
ACC	12	5	2	1	2	1	9	0	12	2	32	14	46	92
BLCA	47	7	60	8	34	5	54	12	5	0	215	17	179	411
BRCA	89	14	34	25	35	6	83	11	4	0	286	15	490	791
CESC	48	10	7	50	12	0	24	6	2	0	151	8	130	289
COAD	37	17	26	14	20	15	32	3	3	2	161	8	121	290
DLBC	0	0	0	0	0	0	0	0	0	0	0	0	37	37
ESCA	15	6	32	4	12	2	38	1	2	1	109	4	71	184
GBM	17	3	57	2	50	24	38	3	1	2	191	6	117	314
HNSC	22	8	95	15	23	20	135	5	1	5	318	11	178	507
KICH	0	0	5	1	1	3	29	1	1	0	39	2	25	66
KIRC	2	39	14	11	6	126	60	2	3	4	258	9	101	368
KIRP	14	1	134	2	28	20	20	2	0	1	219	3	59	281
LGG	20	74	77	13	30	93	49	1	8	3	356	12	143	511
LIHC	23	2	86	3	62	11	46	0	1	4	233	5	125	363
LUAD	25	3	54	9	32	5	65	8	2	1	193	11	309	513
LUSC	21	8	40	11	15	5	137	7	2	2	237	11	232	480
MESO	3	2	14	0	7	9	28	1	1	0	63	2	16	81
OV	4	4	6	6	3	1	12	4	1	0	36	5	24	65
PAAD	8	0	29	5	29	7	33	1	3	1	111	5	61	177
PCPG	18	7	22	12	24	8	32	0	0	0	123	0	56	179
PRAD	6	2	198	1	104	29	49	6	3	2	389	11	95	495
READ	8	0	3	2	12	0	15	2	0	0	40	2	48	90
SARC	28	11	25	7	11	4	36	4	0	1	122	5	109	236
SKCM	33	19	55	18	54	18	91	5	13	2	288	20	158	466
STAD	23	40	74	19	27	36	63	3	8	0	282	11	146	439
TGCT	4	0	13	0	2	0	3	2	0	2	22	4	119	145
THCA	14	5	120	1	8	0	0	1	1	1	148	3	341	492
THYM	8	0	51	2	2	2	3	0	0	0	68	0	54	122
UCEC	41	69	6	40	4	3	46	4	26	0	209	30	208	447
UCS	3	2	4	3	3	1	6	1	2	0	22	3	32	57

TABLE 2: TCGA sample counts on different driver mutations (homozygous, heterozygous, or control/wild-type) for the genes involved in the DDR status considered in our experiments.

Appendix C
Hyperparameter optimization

C.1 Pareto-optimal graphs for cancer-type prediction

C.1.1 Supervised NMF optimization

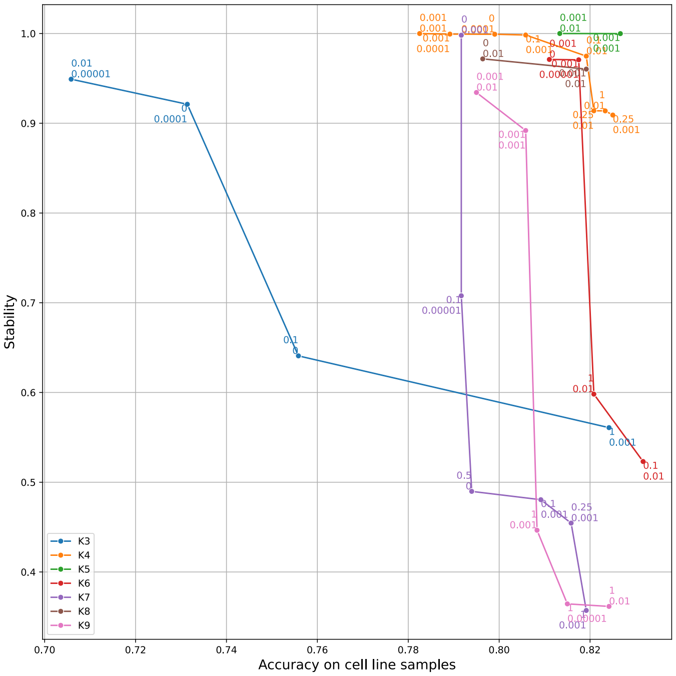
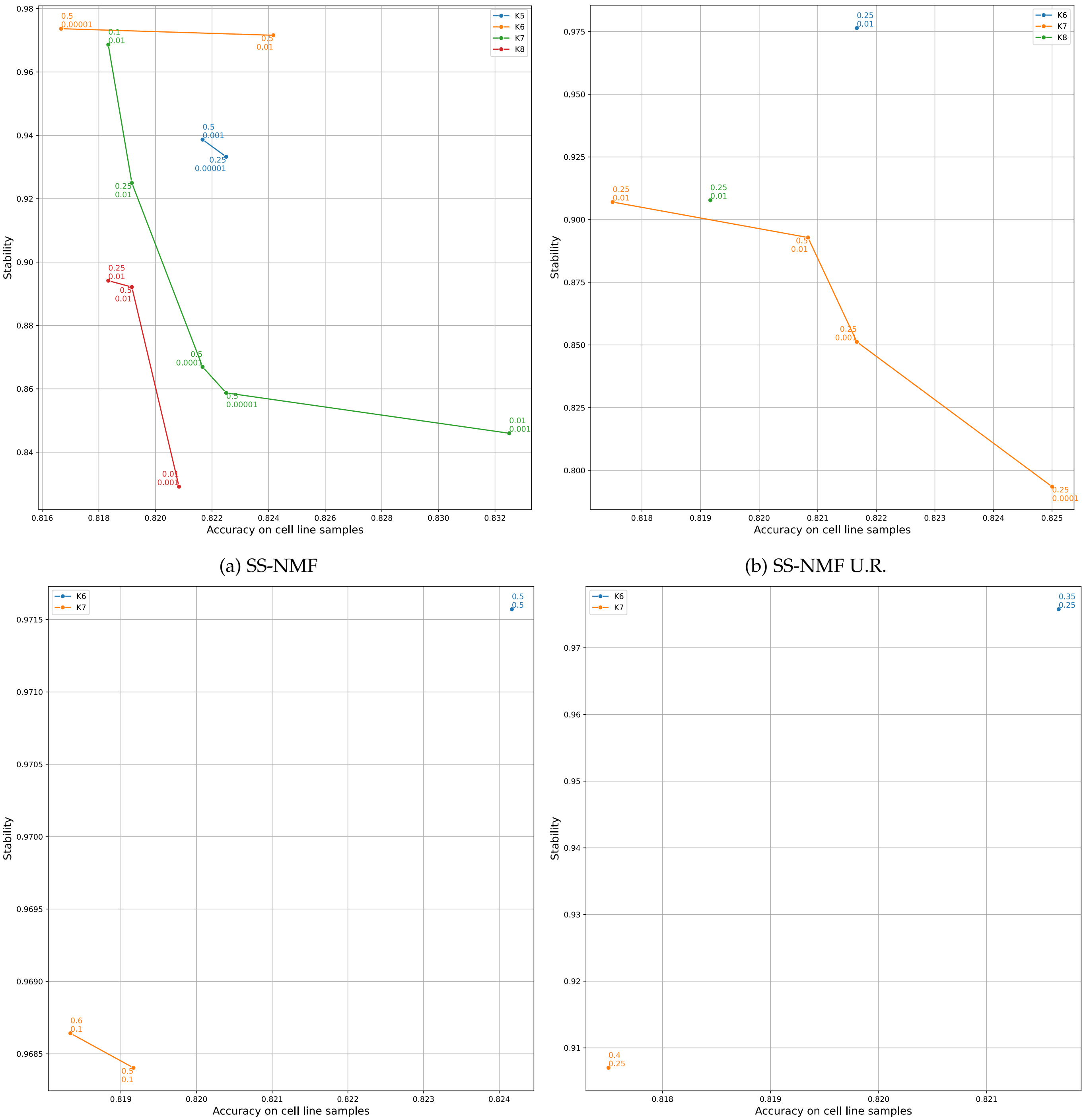


Fig. 10: S-NMF pareto-optimal hyperparameter configurations. Different lines indicate different numbers of signatures. For every point annotation, the top value indicates λ_c and the bottom value λ_{L2}.

C.1.2 33% unlabeled data proportion



(a) SS-NMF

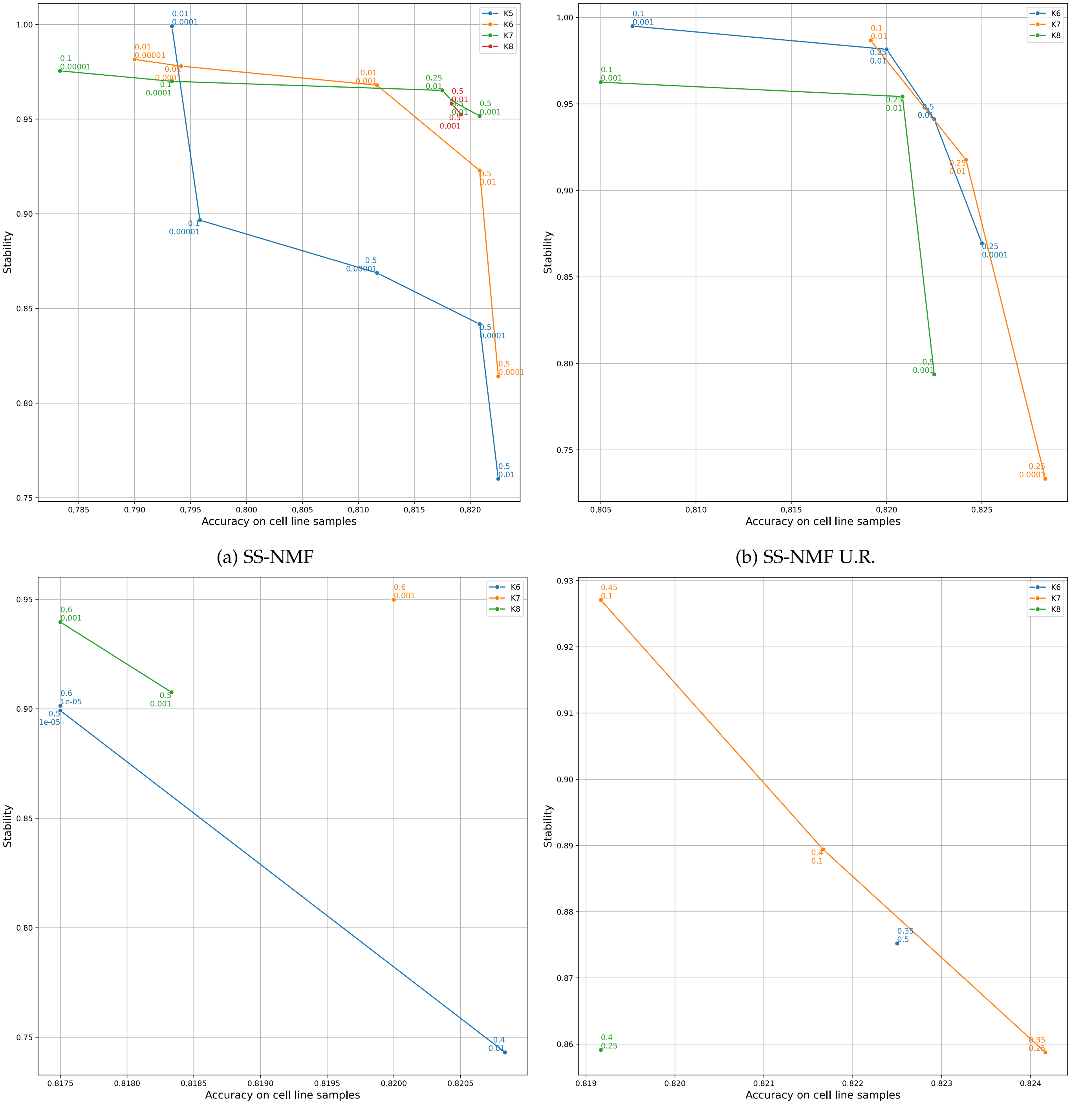
(b) SS-NMF U.R.

(c) PSS-NMF

(d) PSS-NMF U.R.

Fig. 11: (P)SS-NMF (U.R.) pareto-optimal hyperparameter configurations. Different lines indicate different numbers of signatures. For every point annotation: in Figures (a) and (b) the top value indicates λ_c and the bottom value λ_{L2}; in Figures (c) and (d) the top value indicates c_t and the bottom value λ_c for the respective SS-NMF (U.R.) hyperparameter configuration.

C.1.3 50% unlabeled data proportion



(a) SS-NMF

(b) SS-NMF U.R.

(c) PSS-NMF

(d) PSS-NMF U.R.

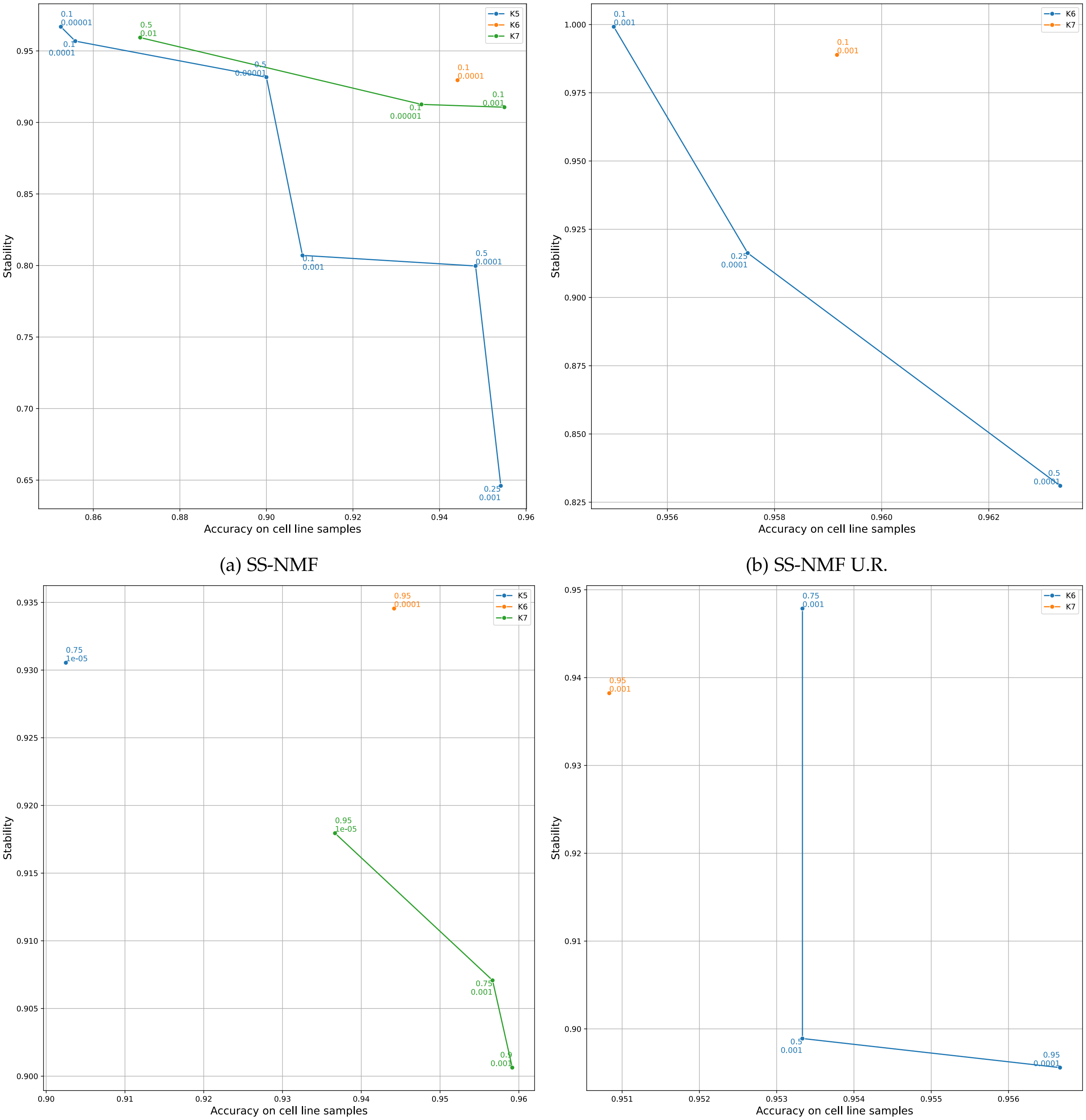
Fig. 12: (P)SS-NMF (U.R.) pareto-optimal hyperparameter configurations. Different lines indicate different numbers of signatures. For every point annotation: in Figures (a) and (b) the top value indicates λ_c and the bottom value λ_{L2}; in Figure (c) the top value indicates c_t and the bottom value λ_{L2} for the respective SS-NMF hyperparameter configuration; in Figure (d) the top value indicates c_t and the bottom value λ_c for the respective SS-NMF U.R. hyperparameter configuration.

C.2 Pareto-optimal graphs for DDRd prediction

C.2.1 Supervised NMF optimization

S-NMF optimization parameters are reported directly from the original research [8].

C.2.2 33% unlabeled data proportion



(a) SS-NMF

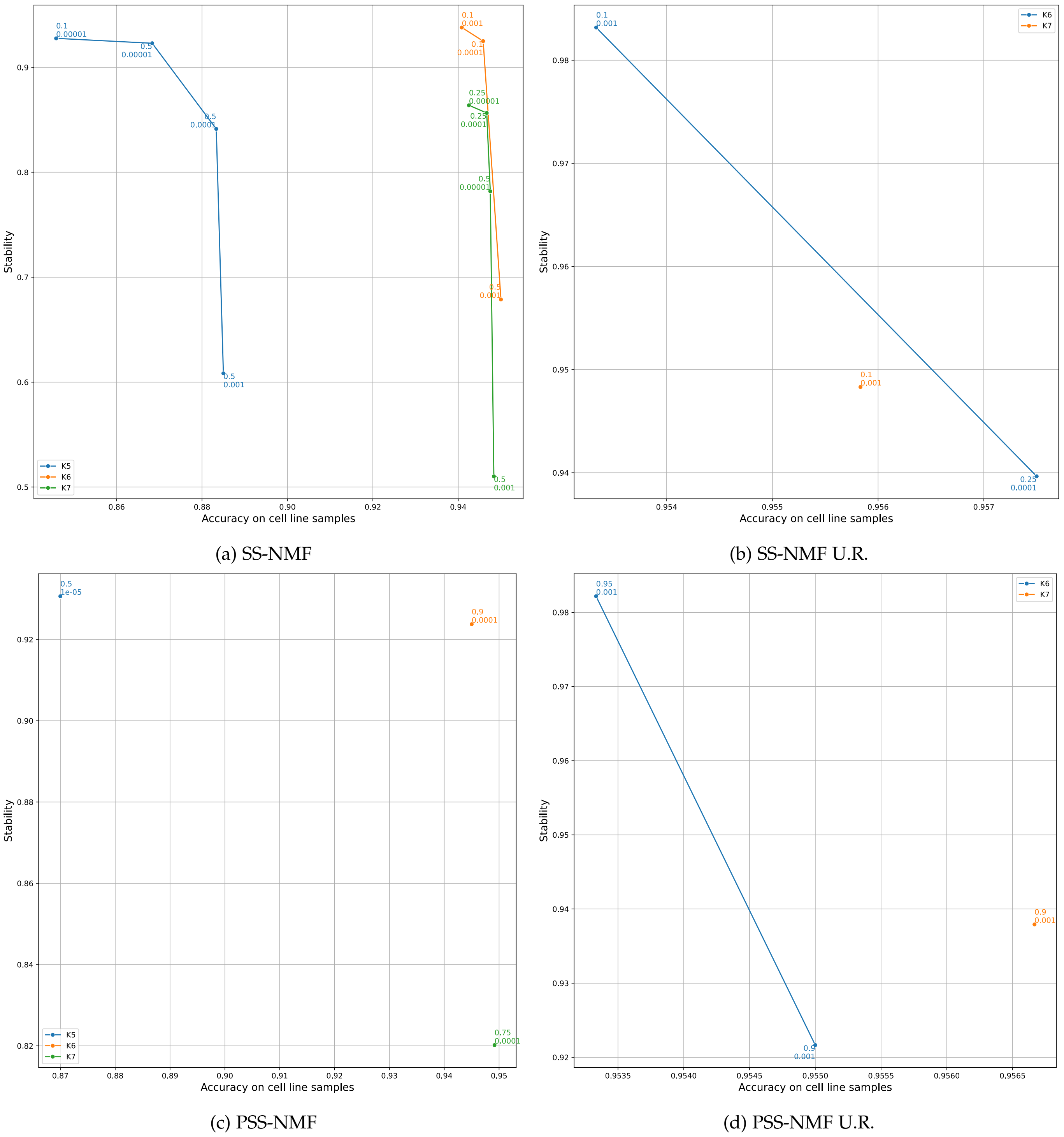
(b) SS-NMF U.R.

(c) PSS-NMF

(d) PSS-NMF U.R.

Fig. 13: (P)SS-NMF (U.R.) pareto-optimal hyperparameter configurations. Different lines indicate different numbers of signatures. For every point annotation: in Figures (a) and (b) the top value indicates λ_c and the bottom value λ_{L2}; in Figures (c) and (d) the top value indicates c_t and the bottom value λ_{L2} for the respective SS-NMF (U.R.) hyperparameter configuration.

C.2.3 50% unlabeled data proportion



(a) SS-NMF

(b) SS-NMF U.R.

(c) PSS-NMF

(d) PSS-NMF U.R.

Fig. 14: (P)SS-NMF (U.R.) pareto-optimal hyperparameter configurations. Different lines indicate different numbers of signatures. For every point annotation: in Figures (a) and (b) the top value indicates λ_c and the bottom value λ_{L2}; in Figures (c) and (d) the top value indicates c_t and the bottom value λ_{L2} for the respective SS-NMF (U.R.) hyperparameter configuration.

C.3 Pseudo-labeling activity

c_t	Pseduo-labeled samples at 33% unlabeled data	Pseduo-labeled samples at 50% unlabeled data	Ratio PL activity 50% unlabeled data over 33% unlabeled data
0.3	5408	32104	5.94
0.35	716	20884	29.17
0.4	104	12484	120.04
0.45	28	6164	220.14
0.5	12	3012	251
0.6	0	248	N.A.

TABLE 3: Number of pseudo-labeled samples when training PSS-NMF and PSS-NMF U.R. with different confidence threshold (c_t) values at different unlabeled data proportions. For completeness, the ratio of pseudo-labeled samples for the different unlabeled data proportions is also reported in the table.

Model	Unlabeled data proportion	K	λ_c	λ_{L2}	c_t	Average TCGA prediction probability	Pseudo-labeled samples	Pseudo-labeling precision
SS-NMF	33%	6	0.5	0.01	-	0.434	N.A.	N.A.
SS-NMF	50%	6	0.5	0.01	-	0.618	N.A.	N.A.
SS-NMF U.R.	33%	6	0.25	0.01	-	0.365	N.A.	N.A.
SS-NMF U.R.	50%	7	0.25	0.01	-	0.362	N.A.	N.A.
PSS-NMF	33%	6	0.5	0.01	0.3	0.564	3160	0.522
PSS-NMF	33%	6	0.5	0.01	0.35	0.459	668	0.591
PSS-NMF	33%	6	0.5	0.01	0.4	0.442	104	0.548
PSS-NMF	33%	6	0.5	0.01	0.45	0.436	28	0.500
PSS-NMF	33%	6	0.5	0.01	0.5	0.435	12	0.250
PSS-NMF	33%	6	0.5	0.01	0.6	0.434	0	N.A.
PSS-NMF	50%	7	0.5	0.001	0.3	0.595	15268	0.313
PSS-NMF	50%	7	0.5	0.001	0.35	0.611	13544	0.344
PSS-NMF	50%	7	0.5	0.001	0.4	0.573	9060	0.366
PSS-NMF	50%	7	0.5	0.001	0.45	0.523	4456	0.443
PSS-NMF	50%	7	0.5	0.001	0.5	0.507	2972	0.537
PSS-NMF	50%	7	0.5	0.001	0.6	0.485	244	0.508
PSS-NMF U.R.	33%	6	0.25	0.01	0.3	0.406	2248	0.288
PSS-NMF U.R.	33%	6	0.25	0.01	0.35	0.366	48	0.458
PSS-NMF U.R.	33%	6	0.25	0.01	0.4	0.365	0	N.A.
PSS-NMF U.R.	33%	6	0.25	0.01	0.45	0.365	0	N.A.
PSS-NMF U.R.	33%	6	0.25	0.01	0.5	0.365	0	N.A.
PSS-NMF U.R.	33%	6	0.25	0.01	0.6	0.365	0	N.A.
PSS-NMF U.R.	50%	7	0.1	0.01	0.3	0.496	16836	0.303
PSS-NMF U.R.	50%	7	0.1	0.01	0.35	0.413	7340	0.369
PSS-NMF U.R.	50%	7	0.1	0.01	0.4	0.388	3424	0.412
PSS-NMF U.R.	50%	7	0.1	0.01	0.45	0.376	1708	0.306
PSS-NMF U.R.	50%	7	0.1	0.01	0.5	0.366	40	0.250
PSS-NMF U.R.	50%	7	0.1	0.01	0.6	0.366	4	0.250

TABLE 4: Average prediction probabilities on TCGA data for SS-NMF (U.R.) and PSS-NMF (U.R.) models in the optimal hyperparameter combinations of K, λ_c, λ_{L2}. We also report the number of pseudo-labeled samples per every model.

Appendix D
Extended results

D.1 Cancer-type prediction experiment

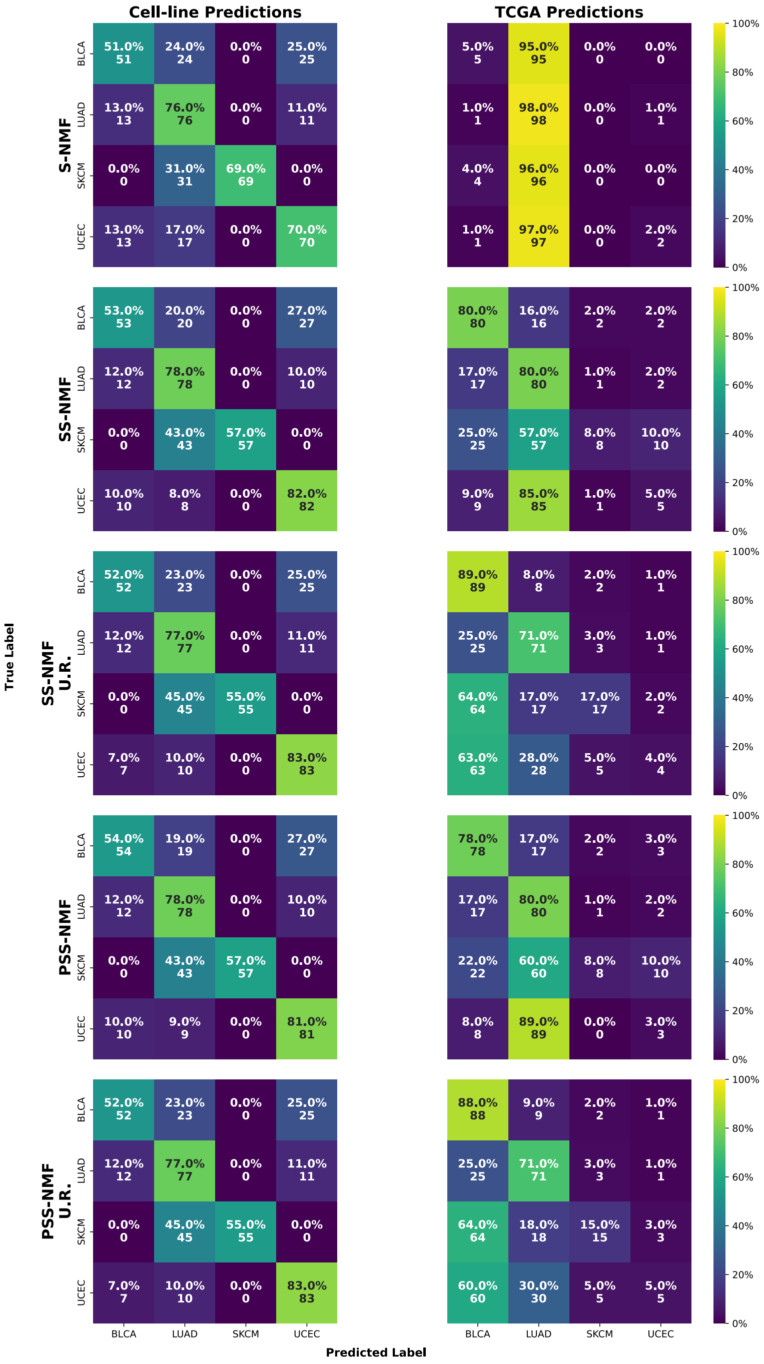


Fig. 15: Results for the cancer-type prediction experiment with 33% unlabeled data proportion, including U.R. models.

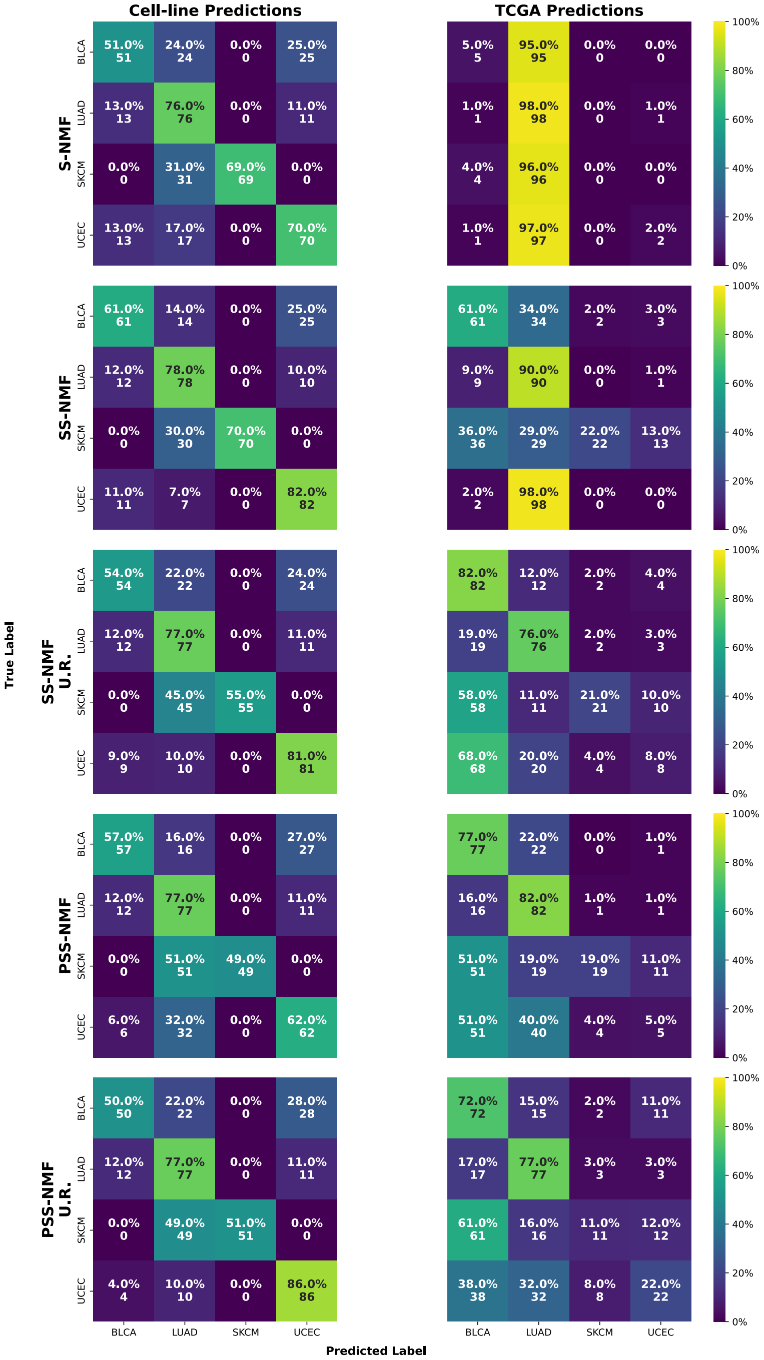


Fig. 16: Results for the cancer-type prediction experiment with 50% unlabeled data proportion, including U.R. models.

D.2 DDRd prediction experiment

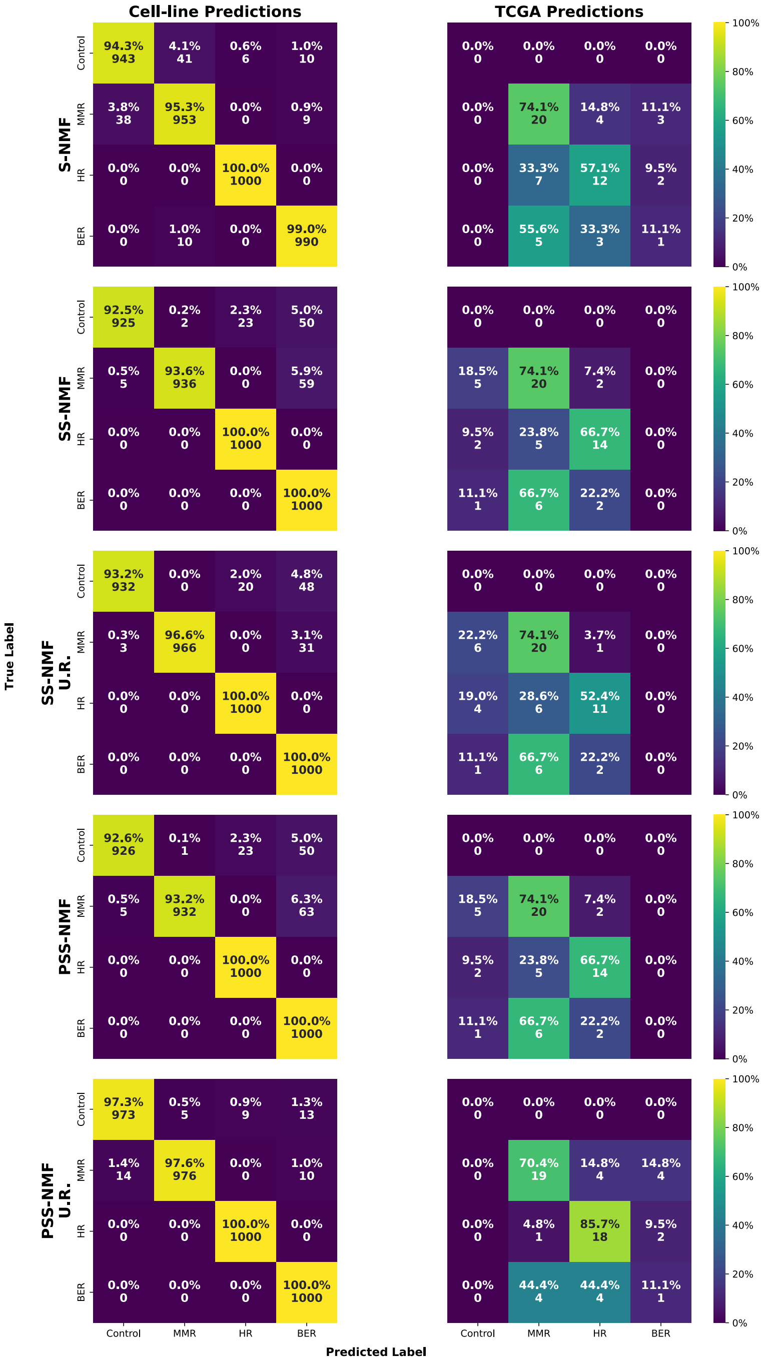
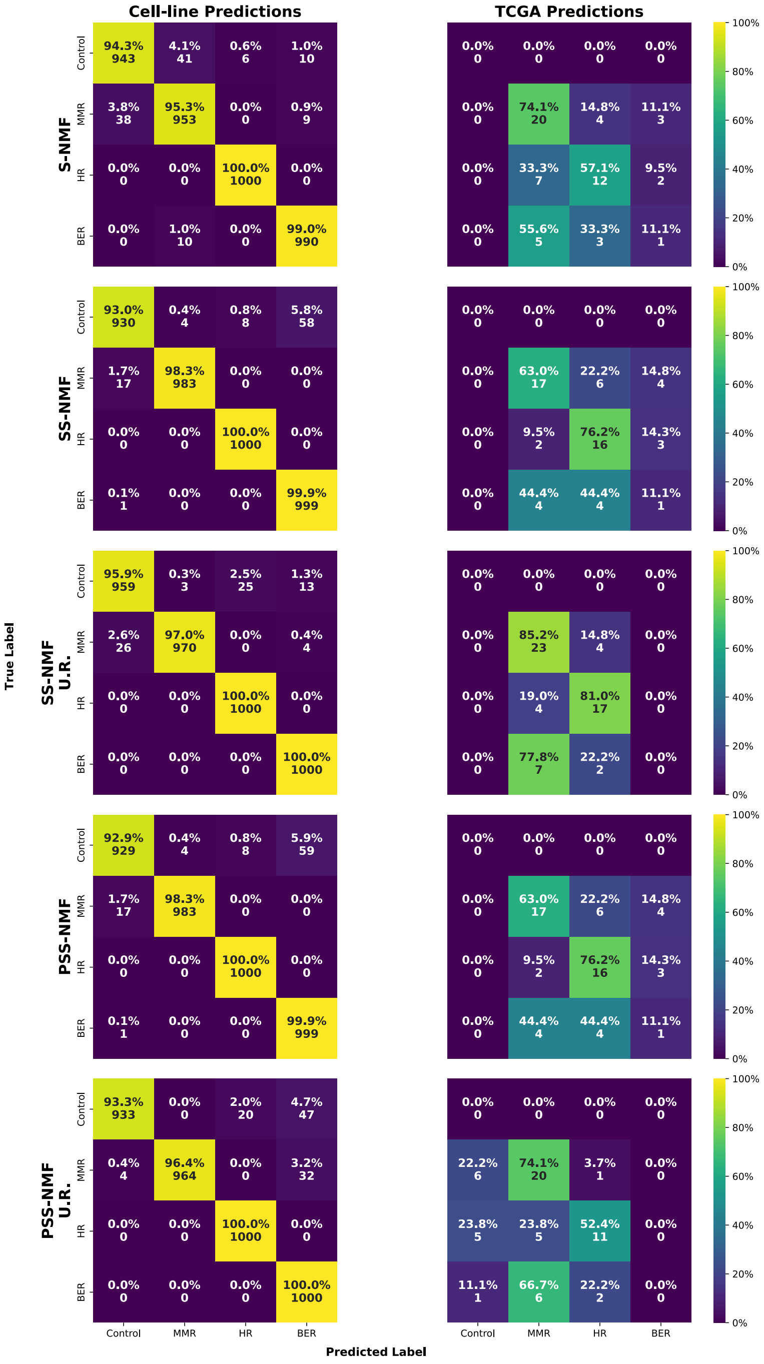


Fig. 17: Results for the DDRd prediction experiment with 33% unlabeled data proportion, including U.R. models.



S-NMF

Cell-line Predictions

	Control	MMR	HR	BER
Control	94.3% 943	4.1% 41	0.6% 6	1.0% 10
MMR	3.8% 38	95.3% 953	0.0% 0	0.9% 9
HR	0.0% 0	0.0% 0	100.0% 1000	0.0% 0
BER	0.0% 0	1.0% 10	0.0% 0	99.0% 990

TCGA Predictions

	Control	MMR	HR	BER
Control	0.0% 0	0.0% 0	0.0% 0	0.0% 0
MMR	0.0% 0	74.1% 20	14.8% 4	11.1% 3
HR	0.0% 0	33.3% 7	57.1% 12	9.5% 2
BER	0.0% 0	55.6% 5	33.3% 3	11.1% 1

SS-NMF

	Control	MMR	HR	BER
Control	93.0% 930	0.4% 4	0.8% 8	5.8% 58
MMR	1.7% 17	98.3% 983	0.0% 0	0.0% 0
HR	0.0% 0	0.0% 0	100.0% 1000	0.0% 0
BER	0.1% 1	0.0% 0	0.0% 0	99.9% 999

	Control	MMR	HR	BER
Control	0.0% 0	0.0% 0	0.0% 0	0.0% 0
MMR	0.0% 0	63.0% 17	22.2% 6	14.8% 4
HR	0.0% 0	9.5% 2	76.2% 16	14.3% 3
BER	0.0% 0	44.4% 4	44.4% 4	11.1% 1

SS-NMF U.R.

	Control	MMR	HR	BER
Control	95.9% 959	0.3% 3	2.5% 25	1.3% 13
MMR	2.6% 26	97.0% 970	0.0% 0	0.4% 4
HR	0.0% 0	0.0% 0	100.0% 1000	0.0% 0
BER	0.0% 0	0.0% 0	0.0% 0	100.0% 1000

	Control	MMR	HR	BER
Control	0.0% 0	0.0% 0	0.0% 0	0.0% 0
MMR	0.0% 0	85.2% 23	14.8% 4	0.0% 0
HR	0.0% 0	19.0% 4	81.0% 17	0.0% 0
BER	0.0% 0	77.8% 7	22.2% 2	0.0% 0

PSS-NMF

	Control	MMR	HR	BER
Control	92.9% 929	0.4% 4	0.8% 8	5.9% 59
MMR	1.7% 17	98.3% 983	0.0% 0	0.0% 0
HR	0.0% 0	0.0% 0	100.0% 1000	0.0% 0
BER	0.1% 1	0.0% 0	0.0% 0	99.9% 999

	Control	MMR	HR	BER
Control	0.0% 0	0.0% 0	0.0% 0	0.0% 0
MMR	0.0% 0	63.0% 17	22.2% 6	14.8% 4
HR	0.0% 0	9.5% 2	76.2% 16	14.3% 3
BER	0.0% 0	44.4% 4	44.4% 4	11.1% 1

PSS-NMF U.R.

	Control	MMR	HR	BER
Control	93.3% 933	0.0% 0	2.0% 20	4.7% 47
MMR	0.4% 4	96.4% 964	0.0% 0	3.2% 32
HR	0.0% 0	0.0% 0	100.0% 1000	0.0% 0
BER	0.0% 0	0.0% 0	0.0% 0	100.0% 1000

	Control	MMR	HR	BER
Control	0.0% 0	0.0% 0	0.0% 0	0.0% 0
MMR	22.2% 6	74.1% 20	3.7% 1	0.0% 0
HR	23.8% 5	23.8% 5	52.4% 11	0.0% 0
BER	11.1% 1	66.7% 6	22.2% 2	0.0% 0

Fig. 18: Results for the DDRd prediction experiment with 50% unlabeled data proportion, including U.R. models.