Supplementary data

Kernelized rank learning for personalized drug recommendation

Xiao He [1,2,†,*], Lukas Folkman [1,2,†,*], Karsten Borgwardt [1,2]

[1]Machine Learning & Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland and [2]Swiss Institute of Bioinformatics, Basel, Switzerland

[†]These authors contributed equally to this work.
[*]To whom correspondence should be addressed.

SUPPLEMENTARY METHODS

Algorithm S1: Subgradient $\partial_{\boldsymbol{U}}\mathcal{L}$

Data: $\boldsymbol{U}$, $\boldsymbol{K}$, $\boldsymbol{Y}$
Result: $\partial_{\boldsymbol{U}}\mathcal{L}$

$\boldsymbol{F} = \boldsymbol{K}\boldsymbol{U}, \partial_{\boldsymbol{F}}\mathcal{L} = 0_{n,m}$;
for i=1:n **do**
  $\boldsymbol{f} = \boldsymbol{F}_i, \boldsymbol{y} = \boldsymbol{Y}_i$;
  $\boldsymbol{\theta} =$ indices of non-missing values in $\boldsymbol{y}$;
  $\boldsymbol{f} = \boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{y} = \boldsymbol{y}_{\boldsymbol{\theta}}$;
  $\boldsymbol{\sigma} = \mathrm{argsort}(-\boldsymbol{y})$;
  $\boldsymbol{f} = \boldsymbol{f}_{\boldsymbol{\sigma}}, \boldsymbol{y} = \boldsymbol{y}_{\boldsymbol{\sigma}}$;
  $\bar{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}}{\mathrm{argmax}}[1 - \mathrm{NDCG@}k(\boldsymbol{f}, \boldsymbol{y}) + \langle \boldsymbol{c}, \boldsymbol{f}_{\boldsymbol{\pi}} - \boldsymbol{f} \rangle]$;
  $((\partial_{\boldsymbol{F}}\mathcal{L})_i)_{\boldsymbol{\theta}} = (\boldsymbol{c}_{\bar{\boldsymbol{\pi}}^{-1}} - \boldsymbol{c})_{\boldsymbol{\sigma}^{-1}}$;
end
$\partial_{\boldsymbol{U}}\mathcal{L} = \boldsymbol{K}\partial_{\boldsymbol{F}}\mathcal{L}$;

SUPPLEMENTARY TABLES

Table S1. Datasets, ranging four different molecular data types, used for the evaluation of KRL.

Data type	Cell lines	Features	Missing IC_{50}
gene expression[a]	962	17,737	18%
exome sequencing[b]	953	300	19%
copy number variation[c]	985	425	19%
DNA methylation[d]	785	378	19%

All datasets were downloaded from the GDSC website:
`http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/`
[a] continuous features encoding the RMA-normalized (robust multi-array average) basal expression of 17,737 genes
[b] binary features encoding if the given cell line carried variants in recurrently mutated sites of one of the 300 candidate cancer genes (CGs) identified in the analyses of 6,815 patient tumors
[c] binary features encoding if the given cell line carried one of the 425 recurrently aberrant copy number segments (RACSs) identified in the analyses of 8,014 patient tumors
[d] binary features encoding if the given cell line carried one of the 378 hypermethylated informative CpG islands (iCpGs) located in gene promoters identified in the analyses of 6,166 patient tumors

Table S2. Sets of hyper-parameter values used to optimize each of the compared methods using three-fold cross-validation on the training set.

Method	Parameter	Range
KRL	k	$\{10\}$
	λ	$\{1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1\}$
	γ	$\{1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1\}$
LKRL	k	$\{10\}$
	λ	$\{1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1 \times 10^{6}\}$
KBMTL	α	$\{1 \times 10^{-3}, 1 \times 10^{-2}, \ldots, 1\}$
	β	$\{1 \times 10^{-3}, 1 \times 10^{-2}, \ldots, 1\}$
	γ	$\{1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1\}$
KRR	α	$\{1 \times 10^{-3}, 1 \times 10^{-2}, \ldots, 1 \times 10^{3}\}$
	γ	$\{1 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 1\}$
RF	$n\_estimators$	$\{100, 1000\}$
EN	α	$\{1 \times 10^{-3}, 1 \times 10^{-2}, \ldots, 1 \times 10^{3}\}$
	l_1_ratio	$\{0.1, 0.3, 0.5, 0.7, 0.9\}$

We used grid search to find an optimal combination of the given parameters.

SUPPLEMENTARY FIGURES

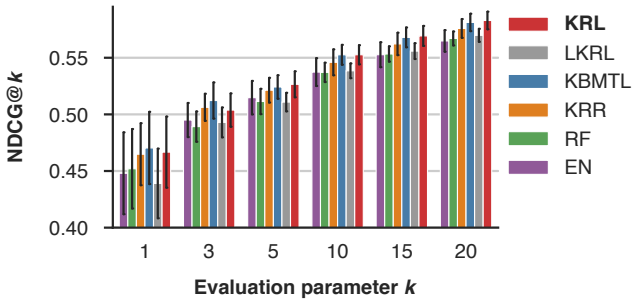

Fig. S1. Comparison of KRL with related work in terms of NDCG@k using the full training dataset for different values of the evaluation parameter k, which controls the number of predicted recommendations that are compared with the true drug ranking. The error bars show standard deviations from three cross-validation folds.

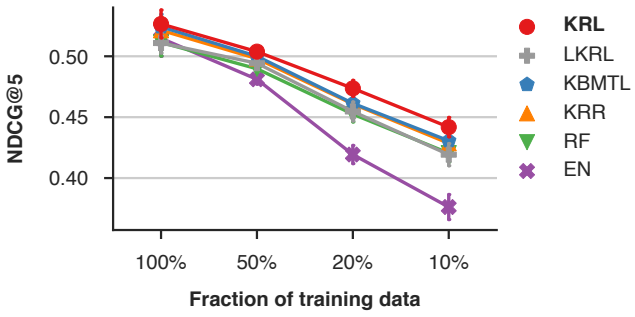

Fig. S2. Comparison of KRL with related work in terms of NDCG@5 using the subsampled training datasets. The error bars show standard deviations from ten randomly subsampled training datasets.

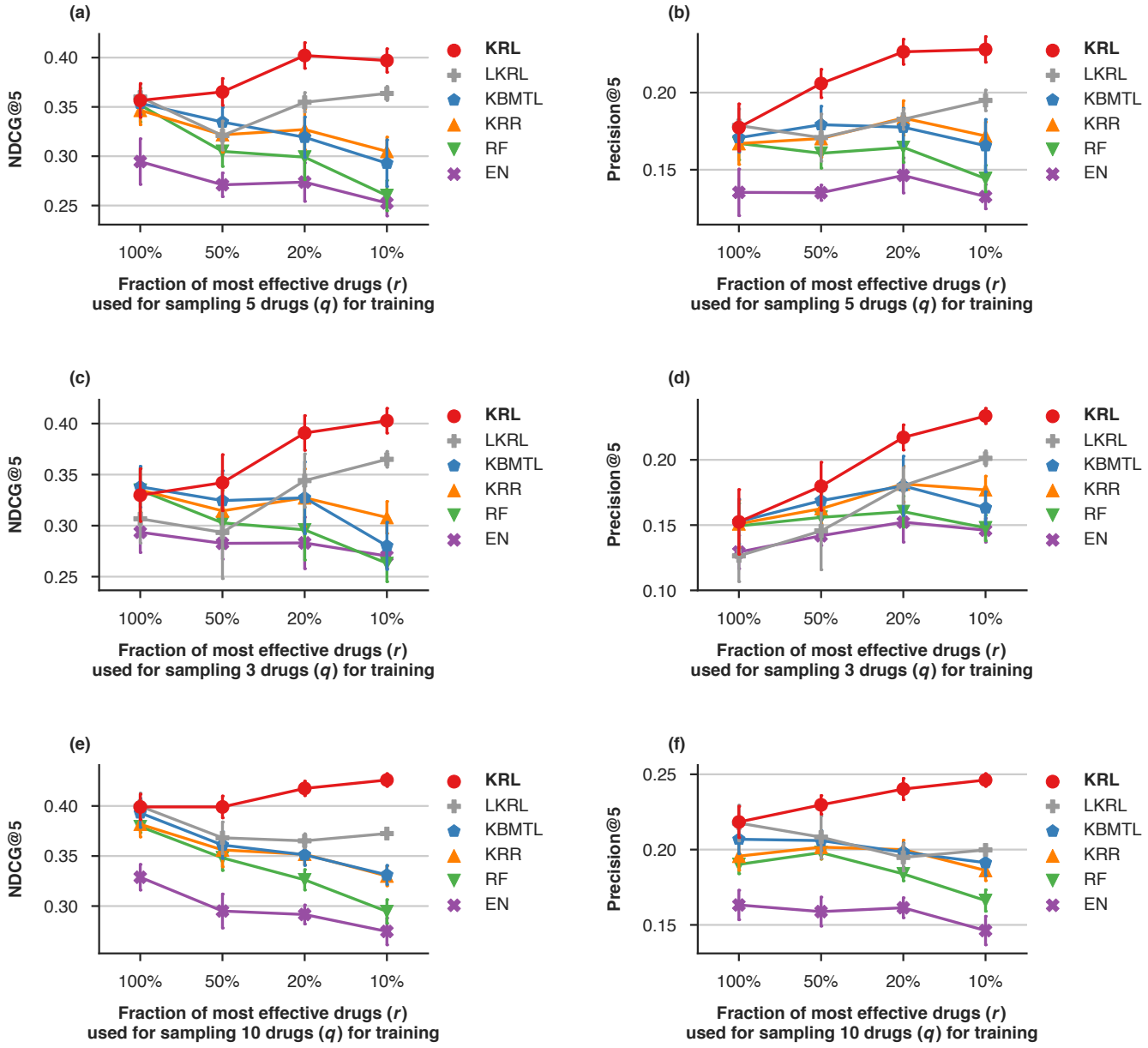

Fig. S3. Comparison of KRL with related work in terms of NDCG@5 **(a, c, e)** and Precision@5 **(b, d, f)** using the subsampled training datasets, keeping five, three, and ten drugs (q) per cell line [**(a–b)**, **(c–d)**, and **(e–f)**, respectively] sampled from a predefined fraction (r) of the cell line's most effective drugs. The error bars show standard deviations from ten randomly subsampled training datasets.

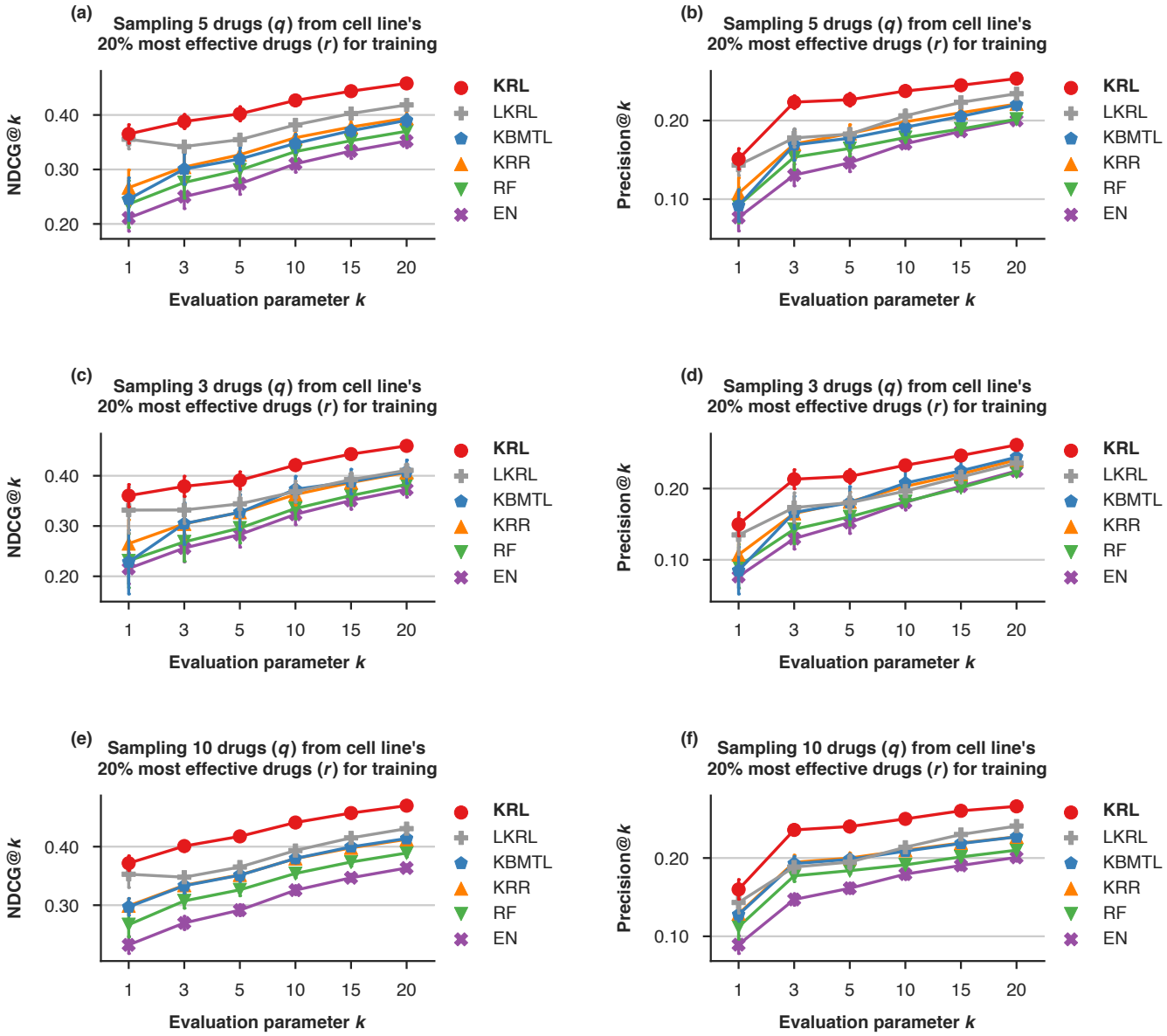

Fig. S4. Comparison of KRL with related work in terms of NDCG@k **(a, c, e)** and Precision@k **(b, d, f)**, for different values of the evaluation parameter k, using the subsampled training datasets, keeping five, three, and ten drugs (q) per cell line [**(a–b)**, **(c–d)**, and **(e–f)**, respectively] sampled from the 20% of the cell line's most effective drugs (r). The error bars show standard deviations from ten randomly subsampled training datasets.

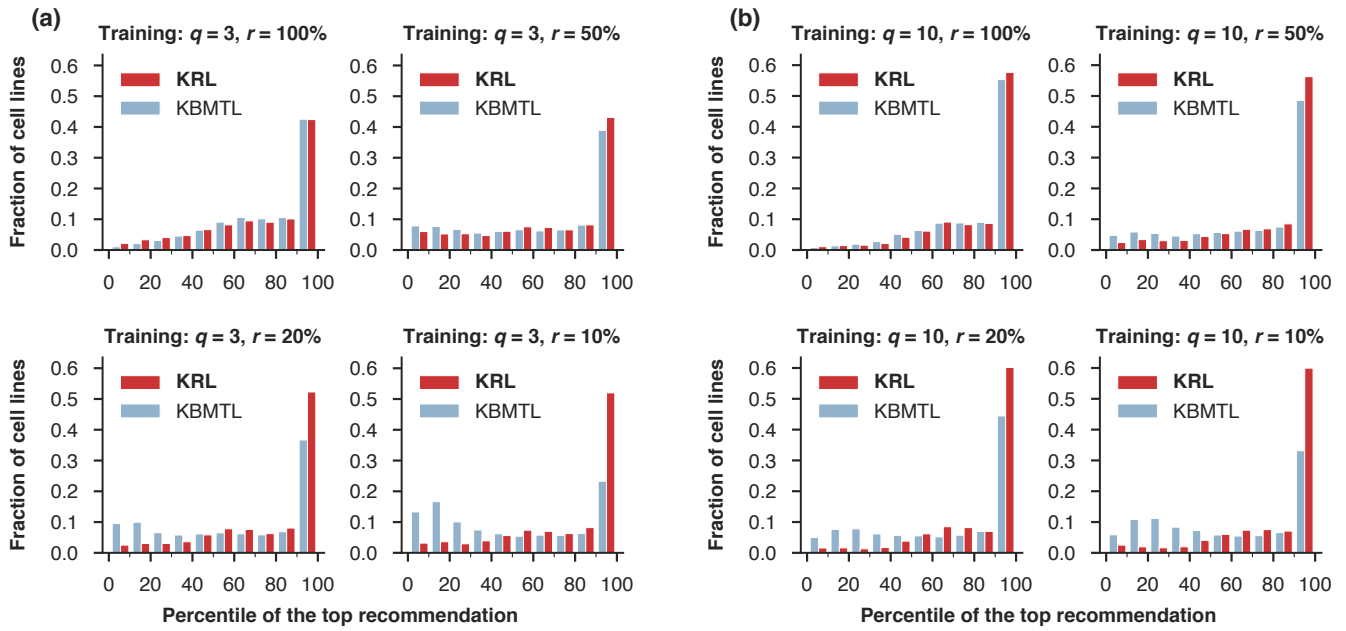

Fig. S5. Histograms comparing the distributions of percentile ranks of drugs recommended by KRL and the second best method, KBMTL, using the subsampled training datasets, keeping **(a)** three and **(b)** ten drugs (q) per cell line sampled from a predefined fraction (r) of the cell line's most effective drugs.

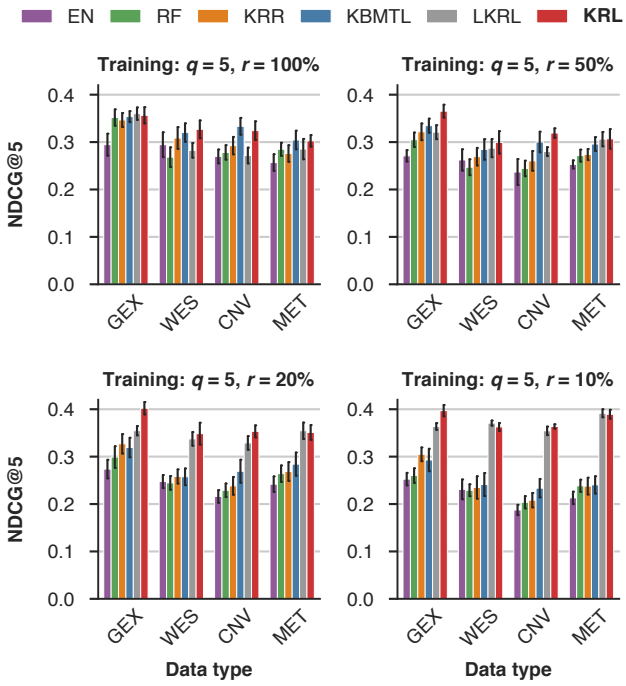

Fig. S6. Comparison of KRL with related work across the four molecular data types: gene expression (GEX), whole-exome sequencing (WES), copy number variation (CNV), and DNA methylation (MET). The six compared methods were evaluated in terms of NDCG@5 using the subsampled training datasets, keeping five drugs (q) per cell line sampled from a predefined fraction (r) of the cell line's most effective drugs. The error bars show standard deviations from ten randomly subsampled training datasets.