Bioinformatics, 37, 2021, i42–i50
doi: 10.1093/bioinformatics/btab336
ISMB/ECCB 2021

Investigation of REFINED CNN ensemble learning for anti-cancer drug sensitivity prediction

Omid Bazgir[1], Souparno Ghosh[2] and Ranadip Pal[1,*]

[1]Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX 79409, USA and [2]Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Anti-cancer drug sensitivity prediction using deep learning models for individual cell line is a significant challenge in personalized medicine. Recently developed REFINED (REpresentation of Features as Images with NEighborhood Dependencies) CNN (Convolutional Neural Network)-based models have shown promising results in improving drug sensitivity prediction. The primary idea behind REFINED-CNN is representing high dimensional vectors as compact images with spatial correlations that can benefit from CNN architectures. However, the mapping from a high dimensional vector to a compact 2D image depends on the a priori choice of the distance metric and projection scheme with limited empirical procedures guiding these choices.

Results: In this article, we consider an ensemble of REFINED-CNN built under different choices of distance metrics and/or projection schemes that can improve upon a single projection based REFINED-CNN model. Results, illustrated using NCI60 and NCI-ALMANAC databases, demonstrate that the ensemble approaches can provide significant improvement in prediction performance as compared to individual models. We also develop the theoretical framework for combining different distance metrics to arrive at a single 2D mapping. Results demonstrated that distance-averaged REFINED-CNN produced comparable performance as obtained from stacking REFINED-CNN ensemble but with significantly lower computational cost.

Availability and implementation: The source code, scripts, and data used in the paper have been deposited in GitHub (https://github.com/omidbazgirTTU/IntegratedREFINED).

Contact: ranadip.pal@ttu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A primary objective of precision medicine for cancer is the selection of an anti-cancer drug or a drug combination that is most effective for the individual patient (Garnett, 2012). A multitude of methods have been proposed to address the issue of anti-cancer drug sensitivity prediction using high-dimensional genomics or chemical drug descriptors data, but there exists room for achieving significant improvement (Barretina, 2012; Chiu *et al.*, 2020; Costello, 2014; Romm and Tsigelny, 2020; Wan and Pal, 2014). To offer enhanced predictive performance, numerous deep learning based models have been introduced recently (Chang, 2018; Chiu *et al.*, 2020; Keshavarzi Arshadi, 2019; Liu, 2019; Xia, 2018; Yu, 2019), that are primarily either deep neural network (DNN) or 1D convolutional neural network (CNN) based approaches. These methods take the input data as a 1-D vector (Mostavi, 2020a), whereas the 2D CNN based method reshape the 1-D vector into a 2D matrix, using some form of lexicographic ordering, which does not preserve the embedded pattern of the data (Mostavi, 2020b).

We developed the REFINED (REpresentation of Features as Images with NEighborhood Dependencies) (Bazgir, 2020) procedure as a general unsupervised isometric mapping to convert high-dimensional vectors into images for training CNN models. We considered a collection of chemical descriptors associated with a drug (or the set of gene expressions associated with a cell line) as a d– dimensional vector of features predicting the efficacy of the drug on a cell line. Thus for n independent drugs (or cell lines), it is a standard univariate high dimensional regression problem. The novelty of our REFINED projection, however, is to represent the foregoing p-dimensional feature vectors (chemical decriptors or gene expressions) as compact images where locally adjusted Bayesian Multidimensional Scaling, (MDS) solution is used to infer the location of each coordinate of the original high dimensional vector on a bounded subspace of $\mathbb{R}^2$. The dependence among the coordinates of the high dimensional vector induces spatial association in 2D images that is then exploited by the CNN based architecture of the predictive model. We note that REFINED is a general framework that can be applied to any prediction problem involving scalar responses and high dimensional correlated regressors.

For illustrative purpose, we demonstrated that REFINED-CNN model provided better predictive performance as compared to DNN

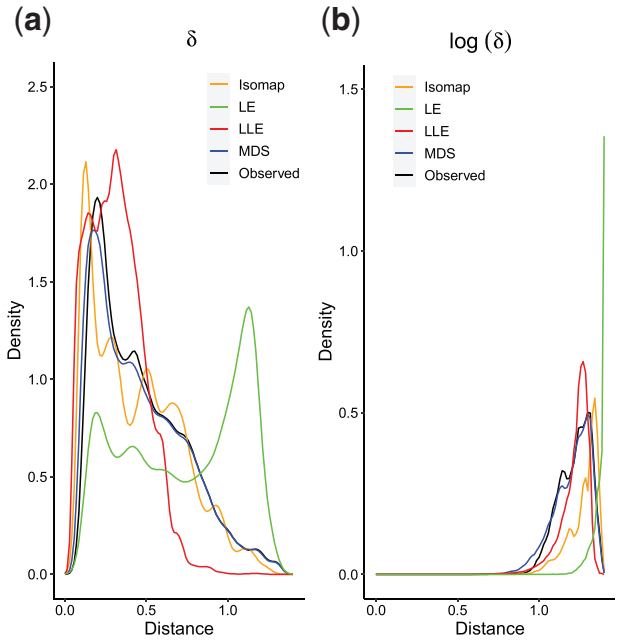

Fig. 1. Density of distances. Kernel density estimate of between features'observed (Euclidean) distance versus distances of projection in 2D space by the 4 DR techniques in regular and log scale

or 2D random projection based CNN models in publicly available pharmacogenomics data- the NCI60 and GDSC datasets. In the NCI60, we used the chemical descriptors of each drug as input features. For GDSC dataset, both gene expressions and drug descriptors were used input features (Bazgir, 2020).

However, in the original form of REFINED, we need to choose a distance metric a priori to define the 'observed' distances among the coordinates of the original high dimensional vector and, based on that choice, choose an appropriate projection scheme. For instance, if Euclidean (geodesic) distance is chosen to measure the distances in ambient dimension, MDS (Isomap) is usually chosen to initialize the dimension reduction process. In Figure 1a, we show the distribution of Euclidean distances among chemical descriptors of drugs in ambient dimension and distribution of distances in 2D under various choices of projection schemes for NCI-60 dataset. Observe that, distribution of projected distances obtained under local non-linear dimension reduction approach (LE and LLE) are very different when Euclidean distance is chosen to measure the distance in ambient dimension. If a natural distance metric is not available for the problem at hand, then, ideally, we need to obtain REFINED projections for different distance measures (local versus global, Euclidean versus Geodesic, etc.); obtain the predictions for each candidate distance measure, and choose the one that produces the best cross-validated prediction performance. Even if an a priori dissimilarity measure among the coordinates are supplied, we need to choose an appropriate projection scheme (MDS, Isomap, etc.) to begin the process of REFINED projection and choose the best initial projection scheme via cross-validation. Evidently, the predictive CNN needs to be fitted for each candidate distance metric/initial projection schemes, resulting in high computational cost.

Since limited guidelines are available to identify an appropriate choice of distance metric/initial projection schemes in an unsupervised setting, multiple REFINED-CNNs have to be fitted regardless, resulting in the availability of an ensemble of REFINED-CNNs. Therefore, a model averaging could be performed which can improve upon the best single REFINED-CNN prediction. The goal of this study is to investigate the performance of such ensemble learners. We illustrate the advantages of three different ensemble methods: (i) model stacking, (ii) image stacking and (iii) integrated-REFINED (iREFINED) over the foregoing best single REFINED-CNN predictions. Our key contribution here is the theoretical and methodological development of iREFINED-CNN that produces predictive performance comparable to REFINED-CNN model stacking, but at a considerably lower computation cost. We apply this methodology on NCI60 and NCI-ALMANAC datasets to compare the performance of iREFINED-CNN with several competing methods. Figure 2 illustrates the framework utilized to train each deep CNN model.

2 Materials and methods

In its original form, REFINED is an unsupervised technique that projects from $\mathbb{R}^d$ to a compact subspace of $\mathbb{R}^2$, $d \gg 2$. These images are then passed on to a CNN to obtain supervised prediction. Thus, using REFINED images to train a CNN (REFINED-CNN) is, broadly, a 2-step process. This offers an opportunity to deploy ensemble learning in various different ways. In this article, we investigate three such ensembling approaches. We begin with briefly describing the process of creating REFINED images [for more details we direct the audience to Bazgir (2020)]. Next, we describe three ensemble learning approaches we used in this study. Finally, for the sake of completeness, we define the CNN architectures and Bayesian optimization framework that we utilized to select the CNNs' hyperparameters.

2.1 REFINED CNN

REFINED maps high dimensional vectors to mathematically justifiable images for training CNN models. It first uses a user-specified distance metric to obtains the initial pairwise distance matrix for the features in their original space. Then uses Bayesian multidimensional scaling (BMDS) to project the features in 2D that approximately preserve pairwise feature distances in the original space. The resulting initial feature map is then subjected to hill-climbing algorithm with the constraint that each pixel can contain at most one feature. The hill-climbing algorithm essentially provides local adjustements to arrive at a locally optimal configuration which does not produce more distortion as compared to the automorphic solution that BMDS produces. The REFINED algorithm, therefore, uses all the samples to arrive at a set of coordinates that are used to map the features into the target 2D space. Once these locations are fixed, the value of each feature, associated with a particular sample, provide the intensity at the pixel reserved for that feature. For each sample, the algorithm thus produces unique REFINED image associated with the feature vector for that sample.

By using different initial distance metric to estimate feature dissimilarity, or choosing different projection schemes to initialize REFINED procedure, different REFINED images could be obtained and consequently the REFINED-CNN's predictive performance vary across the foregoing choices. As shown in (Bazgir, 2020), REFINED CNN initialized with MDS provides better prediction error as compared to Isomap (Tenenbaum, 2000), Locally linear embedding (LLE) (Roweis and Saul, 2000) and Laplacian eigenmaps (LE) (Belkin and Niyogi, 2003) on the NCI60 dataset. Therefore, in absence of a natural measure to identify feature dissimilarities and project them to target 2D space, REFINED-CNN needs to be trained for different choices of distance metrics and initial projection schemes.

2.2 Model stacking

An immediate consequence of having REFINED-CNN being trained on different choices of distance metrics and initial projection schemes is that we have at our disposal several outputs from the CNN predictive model each associated with a different choice we made a priori. Clearly, a linear combination of these predictions, with linear weights estimated from a separate validation set, produces the REFINED-CNN model stacking. More precisely, let $\tilde{y}_a$ be the prediction of a REFINED-CNN associated with the choice of a distance metric (or projection scheme) $a = 1, 2, \ldots A$. Then, the final prediction REFINED-CNN model stacking Y_f is given by the linear regression equation

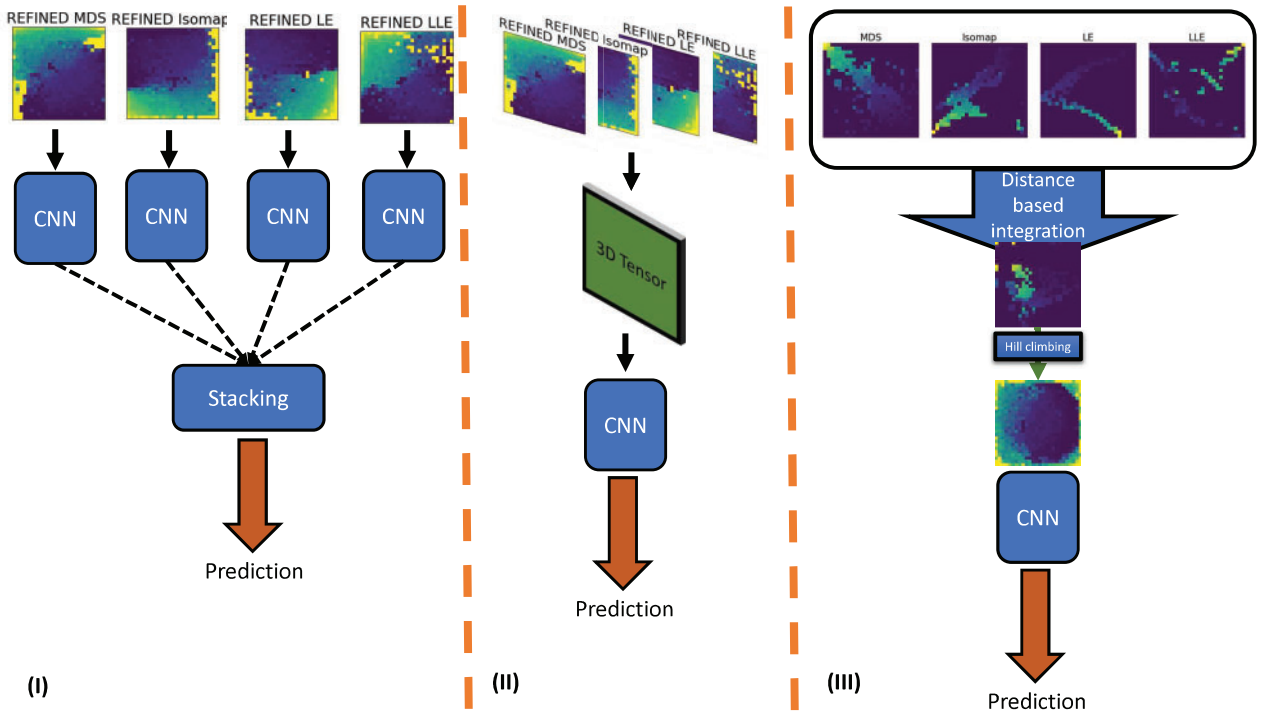

Fig. 2. Illustration of three ensemble learning approaches in this study. (I) is stacking four different REFINED CNN models to achieve the ultimate prediction. (II) is REFINED-CNN image stacking model that stack images in the z-direction prior CNN modeling and (III) is integrated REFINED CNN model that integrates all the created REFINED images into one image and then trains a CNN model

$$Y_f = \sum_{a=1}^{A} \gamma_a \tilde{y}_a + b + \epsilon \tag{1}$$

where γ_a is the linear weight associated with the choice a, b is the intercept term and ϵ is the error. MLE for regression coefficient could be estimated if the ϵ is non-Gaussian, but following (Costello, 2014; Wan and Pal, 2014) we simply use the least square solutions for the regression coeffcients. (Kondratyuk, 2020) recently showed that, in the context of CNN predictions, ensemble of models usually provide better performance than a single candidate model. We therefore use the model stacking approach to benchmark the performance of other candidate models.

2.3 Image stacking

Evidently, in model stacking approach (1), for each choice a, producing the REFINED images I_a, separate CNNs need to be trained. Since computational cost associated with CNN training is considerably more than producing I_a, one immediate avenue to reduce computation cost is to concatenate the REFINED images $\{I_1, I_2, \ldots, I_A\}$ to produce a 3D tensor for each sample. This 3D tensor can be passed on to the CNN architecture to train a single CNN model using all the images produced by candidate choices. The resulting 3D convolution blocks essentially learns to extract features from the tensors via the back propagation process (Ji et al., 2013; Maturana and Scherer, 2015). This approach gets rid of the linearity assumption in (1) and model averaging is done implicity. Additionally, it requires training of a single CNN thereby reducing the computation cost significantly. In the context of this study, a graphical representation of anti-cancer drug sensitivity prediction with 3D convolution blocks is shown in Figure 2.

Although this technique offers computational benefits, it still requires generation of 'A' REFINED images. More importantly, since each I_a is created independently for each choice, and because locations are not uniquely identifiable in BMDS solution, there is no guarantee that a particular feature will occupy the same coordinate in each $I_a, a = 1, 2, \ldots, A$. Consequently, when the images are concatenated, a particular coordinate often does not correspond to an unique feature across I_a, thereby severely affecting CNN's ability to extract features from the input tensors. Lack of coordinate-specific association of pixel intensities across I_a also potentially impacts the predictive performance of the CNN.

To partially address the lack of uniqueness in feature locations across REFINED images, we stack the feature maps extracted by the convolution layers instead of stacking the raw REFINED images. Toward that end, for each REFINED image, we design the convolution layer under different choices of the number and size of kernels. By allowing the kernel sizes to vary across REFINED images we can potentially capture the impact of distance metrics defined over different scale, i.e. the global versus local nature of MDS/Isomap and LE/LLE, respectively. The feature maps extracted by these convolution layers are then concatenated and passed on to the dense layers. The details of the feature map stacking is provided in Supplementary Section S4 of Supplementary Information.

Regardless, to fully alleviate this context-specific non-uniqueness problem, we need to enforce the condition that location of each feature remains same for all input REFINED images. The integrated REFINED methodology arises when this condition is enforced to infer the location of each feature.

2.4 Integrated REFINED

Consider the predictor matrix $\boldsymbol{X} = \{x_{ij}\}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$ with x_{ij} being the value of the jth feature for the ith sample. The goal of REFINED was to obtain the location of the features in a compact subset of $\mathbb{R}^2$, more specifically in $[0,1]^2$. In the following formulation, we assume each choice of initial distance metric is uniquely associated with a projection scheme leading to a total of A choice of distance metric-projection schemes pairs. Let $d_{jk,a}$ be the observed distance between the jth and the kth feature obtained using the distance metric a and δ_{jk} be the unknown Euclidean distance between these two features in unit square. Hence, $\delta_{jk} = \sqrt{\sum_l (s_{j,l} - s_{k,l})^2}$, where s is now 2D coordinate system denoting the unique location of the features j and k in unit square obtained by synthesizing $d_{jk,a}, a = 1, 2, \ldots, A$. Our goal is to

estimate $s_j \in [0,1]^2$ that remains invariant for all candidate distance metric.

Under the assumption of truncated normal distribution of $d_{jk,a}$ (Oh and Raftery, 2001), the data model associated with the distance metric a is given by $d_{jk,a} \sim N(\delta_{jk}, \sigma_a^2) I(d_{jk,a} > 0)$. For the location process, we specify a spatial Homogeneous Poisson Process (HPP) with constant intensity $\lambda = p/[0,1]^2$ which essentially distributes locations of p predictors randomly in an unit square. Since this corresponds to complete spatial randomness, an alternative specification of location process is given by $\boldsymbol{s} = \{s_1, s_2, \ldots, s_p\} \sim Uniform([0,1]^2)$ (Chandler, 2013). The advantage of this HPP specification for the location process is outlined in (Bazgir, 2020).

Let $\boldsymbol{d}_a = [d_{jk,a}],\ j,k = 1,2,\ldots,p$ be the collection of $m = \binom{p}{2}$ distances obtained under the metric a and $\boldsymbol{d} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_A]$ be the total number of distances in the dataset. Let $\boldsymbol{\delta}$ be the collection of Euclidean distances in the unit square that needs to be inferred after imposing the invariance of $\boldsymbol{s}$. Then under the assumption of conditional independence, the full data model is then given by

$$f(\boldsymbol{d}|\boldsymbol{s},\sigma^2)(\Pi\sigma_a^2)^{-\frac{m}{2}} e^{-\frac{1}{2}\sum_{j>k}\left(\sum_a \left(\frac{d_{jk,a}-\delta_{jk}}{\sigma_a}\right)^2\right)} . e^{-\sum_a \sum_{j>k} \log \Phi\left(\frac{\delta_{jk}}{\sigma_a}\right)} \quad (2)$$

where $\Phi(.)$ is the usual standard normal cdf. At the process level, we have

$$\boldsymbol{s}|p \sim Uniform([0,1]^2) \quad (3)$$

Finally, we impose the same prior for $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_A^2] \overset{iid}{\sim} \text{InverseGamma}(\alpha, \beta)$ with $a > 2,\ b > 0$.

Under this specification, the full conditional of posterior of $\boldsymbol{s}$ is given by

$$\pi(\boldsymbol{s}|\boldsymbol{d},\boldsymbol{\sigma}^2) \propto e^{\frac{-1}{2}\sum_{j>k} V\left(\delta_{jk} - \frac{\bar{d}_{jk0}}{V}\right)^2} . e^{-\sum_a \sum_{j>k} \log \Phi\left(\frac{\delta_{jk}}{\sigma_a}\right)} \quad (4)$$

where $V = \sum_{a=1}^{A} \frac{1}{\sigma_a^2}$ & $\bar{d}_{jk0} = \sum_{a=1}^{A} \frac{d_{jk,a}}{\sigma_a^2}$. They key observation is that, the location parameter of the conditional posterior of $\boldsymbol{s}$ is the weighted average of the observed distances obtained from each distance metric under consideration, with the weights being a function of the precision associated with the distribution of observed distances. The details of the derivation of (4) is relegated to Supplementary Section S1.1 of Supplementary Information.

If, on the other hand, we posit a log-normal distribution for $d_{jk,a}$ (Bakker and Poole, 2013), the data model associated with the distance metric a is given by $\log(d_{jk,a}) \sim N(\log(\delta_{jk}), \sigma_a^2)$. Retaining the HPP specification of location process and independent Inverse Gamma priors for $\boldsymbol{\sigma}^2$, the posterior conditional of $\boldsymbol{s}$ is given by

$$\pi(\boldsymbol{s}|\boldsymbol{d},\boldsymbol{\sigma}^2) \propto e^{-\frac{1}{2}\sum_{j>k} V(\delta^*_{jk} - \frac{\bar{d}^*_{jk0}}{V})} 2 \quad (5)$$

where $\delta^*_{jk} = \log(\delta)_{jk}$, and $\bar{d}^*_{jk0} = \sum_{a=1}^{A} \frac{\log(d)_{jk,a}}{\sigma_a^2}$. Further simplification of the location parameter in (5) yields $\frac{\bar{d}^*_{jk0}}{V} = \frac{\sum W_a \log d_{ija}}{\sum W_a}$, where $W_a = \frac{1}{\sigma_a^2}$. Clearly, the location parameter is the weighted geometric mean of $(d_{jk1}, \ldots, d_{jkA})$ with the weight being a function of precision associated with the distributional specification of $d_{..a}$. Detailed derivations of (5) is offered in Supplementary Section S1.2 of Supplementary Information.

Observe that, (4) and (5) imply that one of the ways to fix the coordinate associated with each feature across I_a, is to enforce a common $\delta(.)$ for data model associated with each distance metric. The methodological benefits of iREFINED are twofold: (i) feature-specific coordinates $\boldsymbol{s}$ can be estimated using standard BMDS solutions without explicitly specifying a composite dissimilarity measure (linear combination of initial distances either in original scale or in log scale) at the outset, and (ii) if a linear combination of the candidate distance metrics is utilized to obtain the initial dissimilarity measure

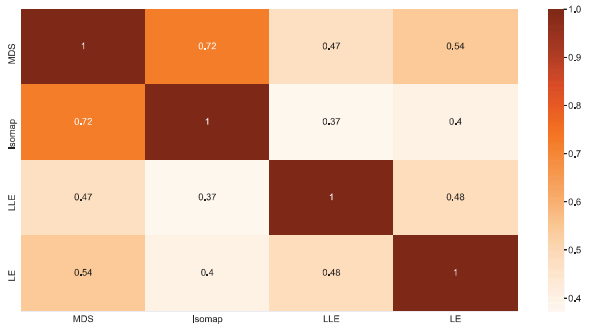

Fig. 3. Correlation between distances. Kendall's τ among the distances estimated in 2D by each DR technique and their geometric and arithmetic means

at the outset, Bayesian non-metric MDS can be performed with a suitable choice of a monotonic non-linear function $g(.)$ that connects the observed dissimilarities with δ in the following way $d_{.} \sim N(g(\delta(.)), \sigma^2) I(d(.) > 0)$ (Oh and Raftery, 2001). The fact that the composite dissimilarity measure may not be proper metric is accommodated by an explicit non-metric BMDS formulation. The computational benefit of iREFINED-CNN is obvious, it requires a single REFINED projection obtained from the estimates of $\boldsymbol{s}$ given by (4) or (5) which is subjected to the foregoing hill-climbing algorithm to arrive at single REFINED image which is then passed on to a single CNN. Consequently, regardless of the number of choice of initial distance metrics (and the associated initial projection schemes), iREFINED-CNN only requires a single full-blown training operation.

3 Application

We apply the methodologies developed in the previous section on two publicly avaiable datasets: (i) *NCI60* dataset consists of drug responses observed after application of more than 52 000 unique compounds on 60 human cancer cell lines (Shoemaker, 2006), (ii) *NCI-ALMANAC* dataset consisting over 5000 pairs of more than 100 drug responses on 60 human cancer cell lines (Holbeck, 2017). In both scenarios, we use the chemical descriptors of drugs as features to predict cell-line specific drug responses. Below we offer brief description of each dataset, outline individual REFINED projection schemes to formulate the iREFINED procedure and describe the CNN architecture.

3.1 Data description

NCI60: The US National Cancer Institute (NCI) screened more than 52 000 unique drugs on around 60 human cancer cell lines. The drug responses are reported as average growth inhibition of 50% (GI50) across the entire NCI cell panel (Gerson et al., 2018) (Shoemaker, 2006). All the chemicals have an associated unique NSC identifier number. We used the NSC identifiers to obtain the chemical descriptors associated with each drug. This information was supplied to PaDEL software (Yap, 2011) to extract relevant features for each one of the foregoing chemicals. Chemicals with more than 10% of their descriptor values being zero or missing were discarded. To ensure availability of enough data points for training deep learning models, we selected 17 cell lines with more than 10 000 drugs tested on them. Each drug was described with 672 features. To incorporate the logarithmic nature of dose administration protocol, we calculated the normalized negative-log concentration of GI50s (NORMLOGGI50). The drug response distribution for three illustrative cell lines are shown in Supplementary Figure S1 of Supplementary Information.

We considered four distance metric and associated projection schemes—MDS, Isomap, LE and LLE—to initialize the REFINED process. To investigate if these four techniques produce similar ordering of the pairwise distances between features, we calculated the following Kendall's rank correlation coefficients $\tau(R(d_{jk,a}),$

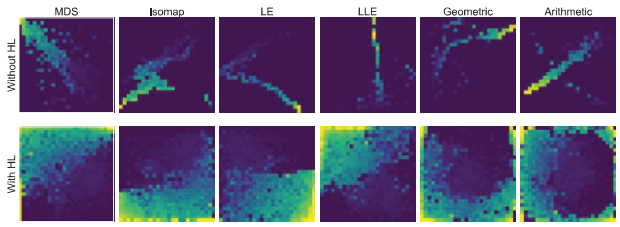

Fig. 4. Different REFINED images. REFINED images created using 4 DR technique including MDS, Isomap, LLE, LE and arithmetic and geometric average of them as initialization step at the first row before applying the hill climbing. The second Row represents the REFINED images after applying the hill climbing algorithm on each initialization step

$R(d_{jk,a'}))$, $\forall j,k = 1,2,\ldots,p$, and$a \neq a' = 1,2,3,4$ where $R(d_{jk,a})$ is the rank of the distance between features j and k obtained from the projection technique a. Figure 3 shows the heat map of the foregoing rank correlations. Evidently, there is a strong agreement between MDS and Isomap. But only moderate level of association between the global techniques and local techniques. However, the distribution of observed Euclidean distance in log-scale in Figure 1b shows better agreement with the logarithm of projected distances, across both local and global dimension reduction schemes indicating the viability of log-normal specification of the data model in the foregoing iREFINED technique. Furthermore, Supplementary Table S8 shows the Kullback–Liebler divergence between observed Euclidean distance and projected distances in both original scale and log-scale. Observe that, on an average, the KL divergence in log-scale is smaller than that in the original scale, indicating that the log-normal specification offers some protection against misspecification of the initial projection scheme.

While the first four panels of Figure 4 show the REFINED images of drug chemical descriptors created under various initializations for cell line SNB_78, last two panels show the corresponding iREFINED images under log-normal and truncated normal specifications, respectively.

NCI-ALMANAC: The NCI-ALMANAC is 'A Large Matrix of Anti-Neoplastic Agent Combinations' dataset (Holbeck, 2017) provides systematic evaluation of over 5000 pairs of 104 FDA-approved anticancer drugs were scanned against a panel of 60 human tumor cell lines (from NCI60) to discover those with enhanced growth inhibition or cytotoxicity profiles (Yang, 2020). Combination activity was reported as a 'ComboScore' that quantifies the advantage of combining two drugs (Tavakoli and Yooseph, 2019). Normalized growth percentage of ComboScore distribution for three cell lines selected randomly from NCI-ALMANAC dataset are shown in Supplementary Figure S2 of Supplementary Information. For each drug we used the same chemical descriptors obtained for NCI60 dataset using the NSC identifiers.

CNN architecture: We had two different CNN architectures; one for modeling the NCI60, and another for NCI-ALMANAC dataset. The REFINED CNN used to model NCI60 dataset, contains two convolutional and two fully connected (FC) hidden layers where each followed by a batch normalization (BN) and ReLu activation function layer. Each ReLu activation after the FC layers was followed by a dropout layer to avoid overfitting.

The REFINED CNN models of NCI-ALMANAC dataset, which predict the ComboScore of two drugs, contain two input as two different drugs in two arms. Each arm contains two convlotional layers followed by a BN and ReLu activation layer. The two arms' output then concatenated and flattened as a 1-D vector as an input of two sequential FC layers, each followed by a BN, ReLu activation function and a dropout layer.

The hyper-parameters of both these CNN models, i.e. learning rate, decay rate, decay step of the adam optimizer, number of kernels, kernel size, stride size per each convolutional layer and number of nodes per each fully connected layer, were optimized using Bayesian optimization framework (Bazgir, 2021; Bergstra, 2013) which sequentially queries a posterior model for hyper-parameter Θ derived from a sequence of surrogate models.

The hyper parameters of the CNN were optimized, for each dataset, using the training and validation set of only one cell line (HCC-2998). Then the model was trained and tested on each cell line independently. In the test phase, for each cell line, we held out a separate set of drugs in the NCI60 and separate set of drug pairs in the NCI-ALMANAC dataset.

4 Results

Several competing models were trained on the foregoing NCI60 and NCI-ALMANAC dataset. Each model was fitted separately on the drug-response data for each cell line. For each cell line, the data was randomly partitioned into training, validation and test sets. Training set consisted of 80% of the sample, 10% of the samples were used for validation and the remaining 10% formed the test set to evaluate the out-of-sample predictive performance of the competing models. To ensure direct comparability, the training, validation and test datasets remained same for all competing models.

A total of 11 models (A summary description of the baseline models are shown in Supplementary Table S1 of Supplementary Information.) were considered: (i) Ensemble REFINED-CNN model stacking, (ii) Ensemble REFINED CNN-image stacking model, (iii) iREFINED-CNN, with both weighted arithmetic mean and weighted geometric mean construction, (iv) individual REFINED CNN with MDS, Isomap, LLE and LE projections, (v) DeepSyenrgy (Preuer, 2018), (vi) (Xia, 2018) approach, (vii) Gradient Boosting Machine (Friedman, 2002), (viii) Random Forests (Ho, 1995), (ix) Support Vector Regression (Drucker, 1997), (x) Kernelized Bayesian Multitask Learning (KBMTL) (Gönen and Margolin, 2014) and (xi) Elastic Nets (Zou and Hastie, 2005). We only applied the DeepSyenrgy (Preuer, 2018) and the (Xia, 2018) approaches on the NCI-ALMANAC dataset, as they are designed for drug combination therapy modeling. We emphasize that all the competing models were independently used for prediction task. Although, GBM, SVR, RF, KBMTL could be used as non-linear/model-free stacking devices to combine the output of different individual REFINED CNNs, we did not pursue that that avenue here.

Several performance measures were used to assess the adequacy of the proposed models and compare their predictive performances. Below we describe the metrics used to evaluate the model performance:

1. Normalized root mean square error of prediction (NRMSE): The customary root mean squared error of prediction (RMSPE) of a given model was normalized by the RMSPE with sample mean as the predictor. We use NRMSE to implicitly compare all the models with respect to the baseline intercept-only model. The NRMSE formula is given by:

$$\text{NRMSEofamodel} = \sqrt{\frac{\sum_{i=1}^{n_p}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_p}(y_i - \bar{y})^2}} \tag{6}$$

where then_p is the size of test-set, y, $\bar{y}$ and $\hat{y}$ are the observed drug response, mean of the drug responses obtained from the non-test set, and predicted drug responses obtained from the model under consideration.

2. Normalized mean absolute error (NMAE): In addition to NRMSE, we use NMAE (7) so that model comparison can be performed without being severely impacted by large outliers.

$$NMAE = \frac{\sum_{i=1}^{n_p}|y_i - \hat{y}_i|}{\sum_{i=1}^{n_p}|y_i - \bar{y}|} \tag{7}$$

For both NRMSE and NMAE, smaller values indicate better predictive performance.

3. Pearson correlation coefficient (PCC) between the predicted and target values: PCC quantifies linear association between the predicted and target drug responses. Model with PCC closer to 1 would be preferred.
4. Bias reduction: We use the method described in (Bazgir, 2020; Song, 2015) to compute model bias. A simple linear regression is performed between residuals (ordinate) and predicted values (abcissa) in the test set. The angle (θ) between the best fitted regression line and abscissa is used as a measure for bias. An unbiased model is expected to produce an angle of 0°. Therefore, models with smaller value of θ is preferred.
5. Model improvement: We introduce a novel measure for model improvement that uses Gap statistics (Tibshirani, 2001) to perform a formal hypothesis test. First, we paired each model with a null model [see (Costello, 2014) for the construction of null model]. Then bootstrap samples were drawn from the drug response values of the test set along with their corresponding predicted values for each model. The null model, using the distribution of drug responses in the training set, is then used to predict drug response sampled from the test set. The process is repeated for 10 000 times and a distribution of NRMSE, NMAE, PCC and Bias, is made for each model along with the null model. For each candidate model, the bootstrapped distribution of each metric is paired with the corresponding distributions obtained from the null models.

A model is deemed to provide significant statistical improvement over the null model if each performance metric is stochastically *better* than its counterpart obtained from the null model. Therefore, we concatenated the bootstrap replicates of performance metrics under the candidate model and null model and formally tested for the presence of at least two clusters using gap statistics in a completely unsupervised fashion. If gap statistics identified presence of at least 2 clusters, we performed K-means clustering. Ideally, the clustering procedure should be able to distinguish replicates coming from null model and candidate model. Hence, an adequate model will produce little overlap between the clusters associated with the candidate model and those associated with the null model. Additionally, all models were subjected to a robustness analysis (Costello, 2014), where we calculated how many times each ensemble REFINED model outperforms other competing models in 10 000 repetition of bootstrap sampling process (Bazgir, 2020).

We calculated 95% confidence interval for each of the foregoing performance metrics using a pseudo Jackknife-after-Bootstrap confidence interval generation approach (Efron, 1979). Multiple bootstrap sets were drawn from the test samples and then the model performance metrics calculated resulting in a distribution for each metric which was used to calculate the confidence interval for a given cell line for NCI60 and NCI-ALMANAC datasets (Bazgir, 2020).

4.1 Results for NCI60

First, we report the performance of the nine candidate models, averaged over 17 cell lines, in Table 1. Observe that, although we expected that REFINED-CNN model stacking will perform best, it was not uniformly better in terms of all the evaluation metrics. Two variants of iREFINED-CNN produced better performance with respect to NMAE and Bias reduction. The REFINED-CNN image stacking performed uniformly worse as compared to the remaining ensemble REFINED models. One of the reasons for this worse performance could be the inability of image stacking approach to extract appropriate features across the REFINED images. However, all the ensemble variants uniformly outperformed single projection based REFINED models and other popular machine learning models considered here. The Supplementary Tables S9–S11 of Supplementary Information details the performance of each model with respect to the foregoing metrics for different cell lines. The 95% confidence interval for all the models per each cell line are provided in Supplementary Figures S3–S6 of Supplementary Information.

Table 1. NCI60 results

Model	NRMSE	NMAE	PCC	Bias
REFINED-CNN model stacking	**0.702**	0.653	**0.710**	0.489
iREFINED-CNN-AM	0.715	**0.630**	0.706	0.461
iREFINED-CNN-GM	0.722	0.635	0.705	**0.446**
REFINED-CNN image stacking	0.775	0.679	0.655	0.509
sREFINED with Isomap	0.787	0.716	0.644	0.509
sREFINED with LE	0.788	0.720	0.644	0.504
sREFINED with LLE	0.795	0.759	0.625	0.511
sREFINED with MDS	0.778	0.709	0.650	0.488
KBMTL (Gönen and Margolin, 2014)	0.856	0.768	0.547	0.733
XGBoost (Friedman, 2002)	0.842	0.806	0.513	0.781
SVR (Drucker, 1997)	0.870	0.806	0.525	0.755
RF (Ho, 1995)	0.880	0.846	0.486	0.816
EN (Zou and Hastie, 2005)	0.976	0.942	0.287	0.968

Note: Comparison of performance of proposed approaches, single projection based REFINED (sREFINED) and state-of-the-art methods on NCI60 dataset. The bold values indicate best performance.

In terms of improvements in prediction, we observe that REFINED-CNN model stacking decerased NRMSE, NMAE and bias by 7–9%, 6–9% and 1–2%, respectively, as compared to single REFINED model. The former ensemble model also increased the PCC by 6–9% as compared to the latter. Integrated REFINED decerased NRMSE, NMAE and bias by 6–8%, 7–12% and 3–4%, respectively, and increased PCC by 5–8% as compared to single REFINED model. However, REFINED-CNN image stacking merely decerased NRMSE, NMAE and bias by 1–3%, 2–7% and 0–1%, respectively, and increased PCC by 1–3% as compared to single REFINED model, indicating its inability to compete favorably with the previous two ensembling approached.

Turning to robustness analysis to compare integrated REFINED and REFINED-CNN model stacking models with other single REFINED CNN models, we observe that REFINED-CNN model stacking offers better performance in terms of (i) NRMSE between 73 and 80% of the times, (ii) NMAE 71–81% of the times, (iii) PCC 68–76%of the times and (iv) Bias 48–56% of the times (see Supplementary Tables S23–S26). The integrated REFINED, on the other hand, produced better performance in terms of (i) NRMSE between 70 and 78% of the times, (ii) NMAE 77-87% of the times, (iii) PCC between 67 and 75% of the times and (iv) Bias 55-63% of the times on average as compared to other single REFINED CNN models (see Supplementary Tables S15–S22). The Gap statistics also indicate that the out-of-sample performance metrics produced by REFINED-CNN model stacking and integrated REFINED are, on average, well distinguishable from the null model (see Supplementary Tables S27–S30 of Supplementary Information). Furthermore, higher values of the Gap statistics associated with ensemble models as compared to those associated with single REFINED-CNN models indicate higher degree of separation of the performance metrics clusters associated with ensemble models from the null model as compared to the single REFINED-CNN versions. The NRMSE, NMAE, PCC and Bias distribution of all the eight models along with the null model are plotted for three randomly chosen cell lines of the NCI60 dataset in Supplementary Figures S11 to S22 of Supplementary Information.

In addition to intra-REFINED comparisons, we compare our REFINED-based approaches with state-of-the-art models including: Kernelized Bayesian Multitask Learning (KBMTL) (Gönen and

Margolin, 2014), Gradient Boosting Machine (Friedman, 2002), Random Forests (Ho, 1995), Support Vector Regressor (Drucker, 1997) and Elastic Nets (Zou and Hastie, 2005). The average performance of all the models on NCI60 dataset are provided in Table 1. Observe that, on average, REFINED-based models significantly outperforms the competing non-REFINED models. The same trend is observed for most cell-lines as well. The detailed results including performance of each model for each cell line is provided in Supplementary Table S11 of Supplementary Information.

4.2 NCI-ALMANAC

In this section, we compare the performance of the foregoing three ensemble REFINED-CNN approaches with 4 single REFINED-CNN methods utlizing different projection schemes along with 6 non-REFINED predictive methods. Since this dataset offers information about responses for drug combinations, our predictors consist of two set of PaDel chemical descriptors representing two drugs for each cell line. The response consists of the '*ComboScore*' for each drug pair. We used the REFINED approach to generate the images corresponding to the drug descriptors for each drug compound in the NCI-ALMANAC dataset.

Considering pairing 2 drugs with D (∼more than 100) unique NSCs for each cell line, then the total number of samples for modeling each cell line is $\binom{D}{2}$ pairs in the dataset, which is close to 5 K. For each cell line, we randomly divided the dataset into 80% training, 10% validation and 10% test sets, where each set covariates contains 672 chemical drug descriptors per each drug. REFINED-CNN model stacking and integrated REFINED CNN model outperforms all other four single REFINED CNN models whereas REFINED-CNN image stacking under-performs them in average. The REFINED-CNN model stacking and integrated REFINED CNN model achieve improvement over single REFINED CNN models in the range of: 7–10% and 2–5% for NRMSE; 8–12% and 1–5% for NMAE; 2–3% and 1–2% for PCC; 6–12% and 1–4% for Bias. The 95% confidence interval for all the models per each cell line are provided in Supplementary Figures S7–S10 of Supplementary Information.

Robustness analysis reveals REFINED-CNN model stacking offers better performance as compared to single REFINED-CNN version with respect to all performance metrics. The former produced lower NRMSE between 88 and 90% of times, lower NMAE between 93 and 95% of times, higher PCC between 83 and 86% of times, and lower Bias 78–89% of the times. Detailed results are presented in Supplementary Tables S39–S42 of Supplementary Information. Integrated REFINED also outperformed the single-REFINED variants in considerable proportion of times. The former lowered NRMSE between 53 and 68% of times, NMAE between 52 and 69% of times and Bias between 43 and 77% of the times, while increased PCC between 45 and 62% of times. The average results of the robustness analysis for each metric of the integrated REFINED are provided in Supplementary Tables S31–S38 of Supplementary Information.

Gap statistics results are provided in Supplementary Tables S43–S46 of Supplementary Information. These results follow the trend observed in the NCI60 datasets with REFINED-CNN model stacking and integrated REFINED CNNs performing considerably better as compared to the single REFINED variants. The Gap statistics distribution plots per NMRSE and NMAE metrics of each model paired with the null model along with their corresponding cluster centroids for three randomly selected cell lines are provided in Supplementary Figures S23–S34 of Supplementary Information.

We further compare the performance of our proposed approaches with state-of-the-art models including: DeepSyenrgy (Preuer, 2018), (Xia, 2018), Gradient Boosting Machine (Friedman, 2002), Random Forests (Ho, 1995), Support Vector Regressor (Drucker, 1997) and Elastic Nets (Zou and Hastie, 2005). The average performance of the models on NCI-ALMANAC dataset are provided in Table 2. The detailed results including performance of each model for each cell line is provided in Supplementary Table S14 of Supplementary Information. Once again we observe the REFINED variants are outperforming other competing non-REFINED models for most cell lines.

Table 2. NCI-ALMANAC results

Model	NRMSE	NMAE	PCC	Bias
REFINED-CNN model stacking	**0.420**	**0.361**	**0.907**	**0.168**
iREFINED-CNN-AM	0.479	0.431	0.893	0.275
iREFINED-CNN-GM	0.474	0.427	0.892	0.248
REFINED-CNN image stacking	0.561	0.524	0.856	0.362
sREFINED with Isomap	0.508	0.470	0.887	0.227
sREFINED with LE	0.489	0.443	0.884	0.238
sREFINED with LLE	0.522	0.486	0.884	0.284
sREFINED with MDS	0.514	0.474	0.877	0.259
Xie *et al.* (2018)	1.574	1.295	0.435	0.991
DeepSynergy (Preuer, 2018)	1.109	1.058	0.176	0.929
XGBoost (Friedman, 2002)	0.518	0.680	0.859	0.327
RF (Ho, 1995)	0.525	0.679	0.851	0.290
SVR (Drucker, 1997)	0.561	0.675	0.830	0.255
EN (Zou and Hastie, 2005)	0.618	0.758	0.789	0.428

Note: Comparison of performance of proposed approaches, single projection based REFINED (sREFINED) and state-of-the-art methods on NCI-ALMANAC dataset. The bold values indicate best performance.

5 Discussion

Based off (Kondratyuk, 2020; Matlock, 2018), this study developed different ensemble learning methods for REFINED-CNN predictive methodology. Our results show that standard linear stacking of multiple single REFINED-CNN improves the prediction performance as compared to the best single REFINED CNN model. To reduce the computational cost associated with linear stacking of multiple REFINED-CNN without significantly impacting its predictive accuracy, we proposed the integrated REFINED technique. Since each projection scheme captures a different embedded pattern of the data, the ensembling approach, associated with the integreated REFINED technique, provides a mathematical way to connect these patterns to reveal a more holistic picture. Robustness is achieved in the sense that model performance is no longer crucially dependent on the a priori choice of the distance metric or the initial projection scheme. Furthermore, this technique offers a way to combine metric and non-metric initial dissimilarity measures via a suitable specification of the probability model for the observed distances. The integrated REFINED also offers an heuristic advantage because we can choose the probability models for observed distances by empirically observing the observed distance histograms. Different probability models for different distance metrics could be combined by the iREFINED technique to obtain the appropriate distance averaging scheme. We proved here that weighted arithmetic and geometric means turn out to be appropriate averaging schemes for common choices of distribution of observed distances.

Through the application on both NCI60 and NCI-ALMANAC datasets, we have established the superior performance of the ensembling techniques. We benchmarked the performance of integrated REFINED with REFINED-CNN model stacking to reveal that the former produces comparable results in predicting drug sensitivity summary metrics (for example, NLOGGI50 and ComboScore) at a fraction of computation cost associated with the latter. Table 3 reveals computational time for REFINED-CNN model stacking is almost four times more than that for the integrated REFINED approach. We also observed that the integrated REFINED performed uniformly better than the REFINED image stacking model indicating the need to fix the location of feature in the set of REFINED images obtained via different projection

Table 3. Execution time comparison

Steps	iREFINED-CNN	REFINED-CNN model stacking
MDS	7 s	7 s
Isomap	21 s	21 s
LE	23 s	23 s
LLE	28 s	28 s
NMDS + DA	47 s	–
Hill climbing	8 min and 23 s	33 min and 32 s
CNN	2 h and 17 min and 36 s	9 h and 10 min and 24 s
LR	–	1 s
Total	2 h and 28 min and 25 s	9h and 45 min and 19 s

Note: Comparing execution time of each step of integrated REFINED CNN model and REFINED-CNN model stacking trained on HCC_2998 cell line data of NCI60 dataset.

schemes. We also proved that integrated REFINED emerges as a result of the constraint that requires the location of features, in the project 2D plane, must remain invariant under different projection schemes thereby offering an intuitive interpretation of the iREFINED technique.

Of course, neither the REFINED-CNN model stacking nor integrated REFINED-CNN guarantees better performance as compared to the best single REFINED-CNN in each instance. An intuitive way to decide whether these ensembling approaches should be deployed is to assess the amount of distortion induced by each individual REFINED scheme. If it appears that particular projection is producing significantly lower distortion, we recommend fitting a single REFINED-CNN associated with that projection scheme. If, on the other hand, the distortions are similar, ensembling is advised. A formal investigation into this conjecture is a future avenue for research.

Our REFINED based architecture has two major pharmacologic implications. First, if we wish to predict the efficacy of a drug on a tumor, a rich class of regressors will consist of both numerical and image variables. The set of numeric regressors consists of (i) the chemical descriptors of the drugs, and (ii) molecular characteristics of tumor that offer a genome wide profile of the tumor. The image regressors consist of histopathology images that capture the inherent heterogeneity of tumors. The REFINED technique offers a solution to this regression-on-multi-type data problem. Our technique converts all non-image regressors to legitimate images which are then processed through CNN algorithm to generate prediction. The integrated REFINED-CNN technique, developed herein, indicates that the prediction can be made robust by combining different distance metrics. Consequently, as multi-modal data collection protocols become more prevalent in the realm of pharmacogenomics, the general REFINED technique (particularly iREFINED) offers a methodology where fairly standard image based deep learning techniques can be utilized to analyze such multi-type data.

Second, we observe a high accuracy out-of-sample prediction performance of our model in NCI-ALMANAC data. This empirical predictive reliability indicates that integrated REFINED-CNN can be utilized to optimize the efficacy of a drug combination treatment regime. More specifically, given an initial choice of drug, say D_{init}, this technique can identify a set of drugs, from a given list of drugs, that are synergistic to D_{init} in the following sense. We can keep D_{init} fixed at one of the arms of the network and allow the other arm scan through the foregoing list of drugs to predict 'ComboScores'. Our bootstrapped-based inferential methodology, that enabled us to generate the intervals for NRMSE, NMAE, PCC and bias (see Supplementary Figures S7–S10 of Supplementary Information), can then be utlized to generate confidence intervals about the predicted ComboScores. This procedure can thus identify whether there exists a drug (in the list) that can be paired with D_{init} to achieve significant increase in efficacy. An exploration into this line of investigation will be conducted in future.

Funding

Research reported in this publication was supported in part by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number [R01GM122084] and by the National Science Foundation under grant number [CCF 2007903]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or National Science Foundation.

Conflict of Interest: none declared.

References

Bakker,R. and Poole,K.T. (2013) Bayesian metric multidimensional scaling. *Political Anal.*, **21**, 125–140.

Barretina,J. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bazgir,O. *et al.* (2020) Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat. Commun.*, **11**, 1–13.

Bazgir,O. *et al.* (2021) Active shooter detection in multiple-person scenario using rf based machine vision. *IEEE Sensors J.*, **21**, 3609–3622.

Belkin,M. and Niyogi,P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, **15**, 1373–1396.

Bergstra,J. *et al.* (2013) *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. International conference on machine learning*, Atlanta, Georgia, USA. pp. 115–123.

Chandler,R. *et al.* (2013) Spatially explicit models for inference about density in unmarked or partially marked populations. *Ann. Appl. Stat.*, 7, 936–954.

Chang,Y. *et al.* (2018) Cancer drug response profile scan (CDRScan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.*, **8**, 1–11.

Chiu,Y. *et al.* (2020) Deep learning of pharmacogenomics resources: moving towards precision oncology. *Brief. Bioinf.*, **21**, 2066–2083.

Costello,J. *et al.* (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1202–1212.

Drucker,H. *et al.* (1997) Support vector regression machines. In: *MIT Press, Advances in Neural Information Processing Systems, Denver, Colorado, USA*, pp. 155–161.

Efron,B. (1979) Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, pp. 1–26.

Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.

Garnett,M. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Gerson,S.L. *et al.* (2018) Chapter 57 – pharmacology and molecular mechanisms of antineoplastic agents for hematologic malignancies. In Hoffman,R. *et al.* (eds.) *Hematology*, 7th edn. Philadelphia, PA, Elsevier, pp.849–912.

Gönen,M. and Margolin,A.A. (2014) Drug susceptibility prediction against a panel of drugs using Kernelized Bayesian multitask learning. *Bioinformatics*, **30**, i556–i563.

Ho,T.K. (1995) Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Quebec, Canada **Vol. 1**. IEEE, pp. 278–282.

Holbeck,S. *et al.* (2017) The national cancer institute almanac: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.*, **77**, 3564–3576.

Ji,S. *et al.* (2013) 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 221–231.

Keshavarzi Arshadi,A. *et al.* (2019) Deepmalaria: artificial intelligence driven discovery of potent antiplasmodials. *Front. Pharmacol.*, **10**, 1526.

Kondratyuk,D. *et al.* (2020) When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570.*

Liu,P. *et al.* (2019) Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics*, **20**, 408.

Matlock,K. *et al.* (2018) Investigation of model stacking for drug sensitivity prediction. *BMC Bioinformatics*, **19**, 71.

Maturana,D. and Scherer,S. (2015) Voxnet: A 3d convolutional neural network for real-time object recognition. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 922–928.

Mostavi,M. *et al.* (2020a) Cancersiamese: one-shot learning for primary and metastatic tumor classification. *bioRxiv.*

Mostavi,M. *et al.* (2020b) Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. Genomics*, **13**, 1–13.
Oh,M.-S. and Raftery,A.E. (2001) Bayesian multidimensional scaling and choice of dimension. *J. Am. Stat. Assoc.*, **96**, 1031–1044.
Preuer,K. *et al.* (2018) Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, **34**, 1538–1546.
Romm,E.L. and Tsigelny,I.F. (2020) Artificial intelligence in drug treatment. *Annu. Rev. Pharmacol. Toxicol.*, **60**, 353–369.
Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.
Shoemaker,R.H. (2006) The nci60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
Song,J. (2015) Bias corrections for random forest in regression using residual rotation. *J. Korean Stat. Soc.*, **44**, 321–326.
Tavakoli,S. and Yooseph,S. (2019) Learning a mixture of microbial networks using minorization–maximization. *Bioinformatics*, **35**, i23–i30.
Tenenbaum,J. *et al.* (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.
Tibshirani,R. *et al.* (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **63**, 411–423.
Wan,Q. and Pal,R. (2014) An ensemble based top performing approach for NCI-dream drug sensitivity prediction challenge. *PloS One*, **9**, e101183.
Xia,F. *et al.* (2018) Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, **19**, 71–79.
Yang,M. *et al.* (2020) Stratification and prediction of drug synergy based on target functional similarity. *NPJ Syst. Biol. Appl.*, **6**, 10.
Yap,C.W. (2011) Padel-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.
Yu,H. *et al.* (2019) Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC Genomics*, **20**, 167.
Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.