OXFORD

Comparative analysis of molecular fingerprints in prediction of drug combination effects

B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang

Corresponding authors: Bulat Zagidullin, Research Program in Systems Oncology, Faculty of Medicine and Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. E-mail: bulat.zagidullin@helsinki.fi; Jing Tang, Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. Tel: +35 845 868 9708; E-mail: jing.tang@helsinki.fi

Abstract

Application of machine and deep learning methods in drug discovery and cancer research has gained a considerable amount of attention in the past years. As the field grows, it becomes crucial to systematically evaluate the performance of novel computational solutions in relation to established techniques. To this end, we compare rule-based and data-driven molecular representations in prediction of drug combination sensitivity and drug synergy scores using standardized results of 14 high-throughput screening studies, comprising 64 200 unique combinations of 4153 molecules tested in 112 cancer cell lines. We evaluate the clustering performance of molecular representations and quantify their similarity by adapting the Centered Kernel Alignment metric. Our work demonstrates that to identify an optimal molecular representation type, it is necessary to supplement quantitative benchmark results with qualitative considerations, such as model interpretability and robustness, which may vary between and throughout preclinical drug development projects.

Bulat Zagidullin is a PhD candidate at the University of Helsinki. He obtained a BSc degree in Biochemical Engineering from Jacobs University Bremen and a MSc degree in Pharmaceutical Biotechnology from Martin Luther University Halle-Wittenberg.
Ziyan Wang is a Graduate student at the University of Michigan, Ann Arbor. He obtained a BSE in Computer Science from the University of Michigan.
Yuanfang Guan is an Associate Professor at the University of Michigan, Ann Arbor. She obtained a BSc in Biology in University of Hong Kong and a PhD in Molecular Biology at Princeton University.
Esa Pitkänen is a FIMM-EMBL Group Leader and an Academy of Finland Research Fellow at the Institute for Molecular Medicine Finland (FIMM), Helsinki. He received his PhD in Computer Science at the University of Helsinki.
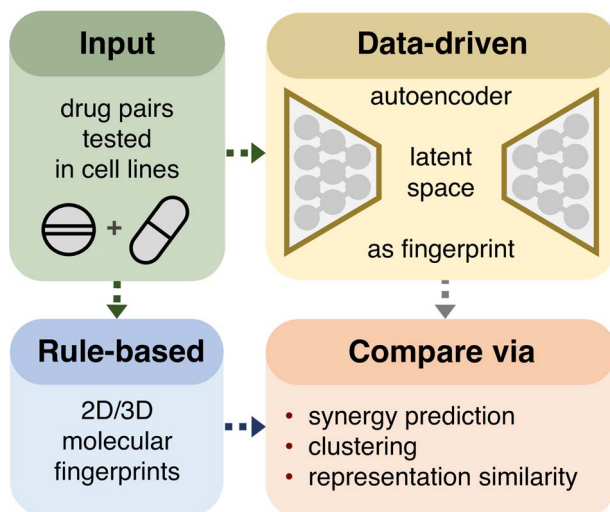**Jing Tang** is an Assistant Professor at the Faculty of Medicine, University of Helsinki. He obtained his PhD in Statistics from the University of Helsinki.
Submitted: 12 April 2021; **Received (in revised form):** 1 June 2021

Graphical Abstract

Key words: drug combinations; drug synergy; molecular fingerprints; machine learning; precision medicine

Introduction

In the past years, deep learning (DL) methods have been successfully applied to a variety of research topics in biomedicine and drug discovery [1–3]. Deep neural networks achieve state-of-the-art performance in medical computer vision tasks and protein structural modeling, enabling *de novo* generation of drug candidates and development of prognostic clinical models [4–8]. However, such performance of DL models is context-dependent [9–12]. While quantitative metrics are routinely and effectively used to compare various computational methods, overreliance on them is a well-known issue [13–18]. It is beneficial to supplement performance results on benchmark datasets with estimates of model uncertainty and robustness, as well as ability to generalize on unseen data [19–21]. These aspects are particularly important in the biomedical research, where *in silico* model predictions direct experimental design choices, as exhaustively testing all combinations of relevant factors is usually unfeasible due to the combinatorial explosion [22, 23].

Advances in high-throughput screening of bioactive compounds in cancer cell lines promote the development of personalized cancer treatments [24]. A major goal in such drug sensitivity and resistance testing studies is to prioritize promising combinatorial therapies that involve coadministration of multiple drugs [25]. By combining synergistic compounds, often with distinct mechanisms of action, it is possible to overcome single-drug resistance, produce sustained clinical remissions and diminish adverse reactions [26, 27]. Drug synergy refers to a degree of drug–drug interaction quantified as the difference between expected and observed dose–response profiles measured by a biological endpoint, such as cell viability or cell toxicity [28]. While synergy characterizes how compounds modulate each other's biological activity, combination sensitivity score (CSS) quantifies drug combination efficacy [29]. In addition to the CSS, we use four synergy scores based on distinct null models, namely Bliss independence, highest single agent (HSA), Loewe additivity and zero interaction potency (ZIP) in the regression analysis of molecular fingerprints [30–34]. Predicting drug combination synergy and sensitivity is related to quantitative structure activity relationship (QSAR) modeling and virtual screening [35, 36]. The QSAR captures mathematical associations between drug descriptors and assay endpoints based on the assumption that structurally similar compounds have similar bioactivity properties, while in the virtual screening studies candidate molecules are prioritized for subsequent experimental validation according to *in silico* prediction results [37]. Rule-based molecular fingerprints are commonly used as drug descriptors in QSAR/Virtual Screening, and MACCS structural keys based on molecular topology are arguably the most popular type of rule-based fingerprints [38–41]. Other types include circular topological fingerprints that describe combinations of non-hydrogen atom types and paths between them within a predefined atom neighborhood, and pharmacophore fingerprints that incorporate local features related to molecular recognition [42–44].

More recently, data-driven fingerprints generated by DL models have been shown to perform well in various research projects [45]. Majority of such DL fingerprints are based on the encoder–decoder architecture, whereby an approximate identity function is learned to translate high-dimensional input into a low-dimensional, fixed-size latent manifold, which is then used to reconstruct the original input [46]. When an encoder–decoder DL model is trained on chemical structures, its latent manifold is interpreted as a data-driven fingerprint. Examples of early DL fingerprinting models include a convolutional neural network (CNN), Chemception and a recurrent neural network, SMILES2Vec, as well as a variational autoencoder (VAE) model with a CNN encoder and a gated recurrent unit-based decoder [47–51]. Development of attention methods for sequence modeling further contributed to the popularity of data-driven DL fingerprints, whereas evolution of generative models enabled *de novo* molecular design through latent space sampling [52–56]. These DL solutions operate on images of molecules or SMARTS/SMILES sequences to create drug structural representations [57–59]. Further, DL fingerprints may be enriched with numerical drug descriptors through multitask DL learning methods or simply by concatenating to latent space [60]. Unlike
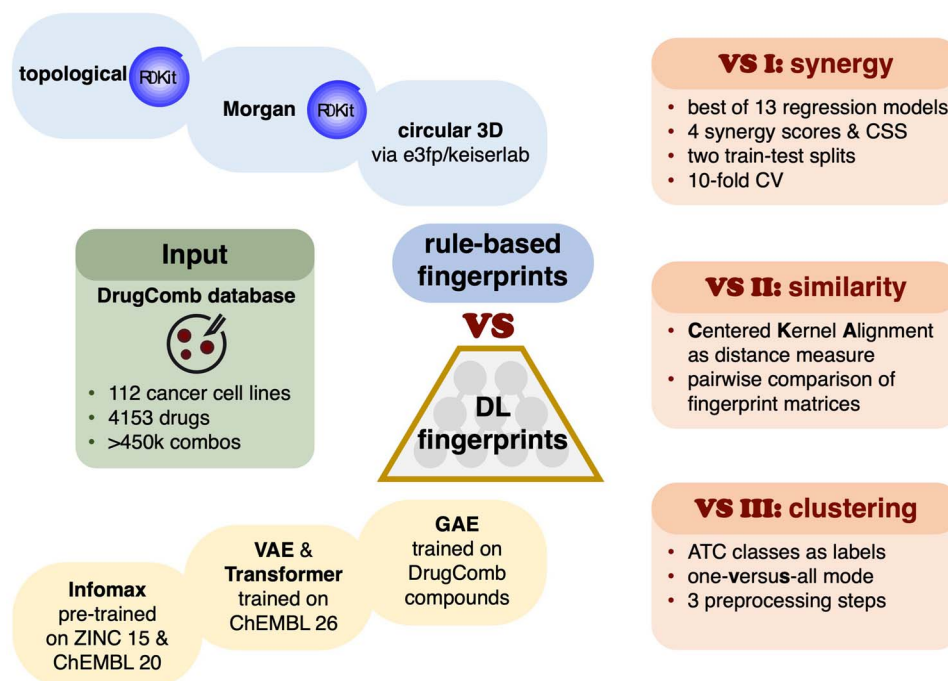
Figure 1. Study workflow. Compounds found in combinations in the DrugComb database are represented using four rule-based (blue) and seven data-driven (yellow) fingerprint types. Rule-based fingerprints include topological, and 2D/3D extended connectivity variants. Data-driven fingerprints are generated using two VAE and two transformer models trained on ChEMBL 26, GAE trained on DrugComb compounds, and a pre-trained Deep Graph Infomax model (Infomax). The fingerprints are compared in three tasks: predictions of drug combination sensitivity and four synergy scores (VS I); representation similarity based on CKA (VS II); one-versus-all fingerprint clustering based on ATC drug classes (VS III). VS I results are also used to identify the most predictive synergy model.

sequence-based versions, geometric DL fingerprints are derived from molecular graphs, and in addition to global molecular descriptors enable position-aware encoding of individual atom and bond features [61–67].

There exist several extensive benchmark datasets for ranking DL models in chemoinformatics tasks, such as MoleculeNet, Open Graph Benchmark and Benchmarking GNNs [68–70]. Despite the widespread use of molecular fingerprints, there is a lack of systematic evaluation of data-driven DL and rule-based versions. To address the gap, we study 11 types of molecular representations, comprising seven DL and four rule-based variants, in prediction of cancer drug combination synergy and sensitivity, based on 17 271 848 data points from 14 cancer drug screening studies (Figure 1, experiment VS I). By comparing four synergy scores based on distinct null models, we identify a preferred synergy formulation for use in cancer drug combinations research [71, 72]. We measure the fingerprint similarity by adapting centered kernel alignment (CKA) as a distance metric (Figure 1, experiment VS II). Lastly, we explore the downstream performance of molecular representations by clustering compounds assigned to 10 anatomical therapeutic chemical (ATC) classes in one-versus-all mode (Figure 1, experiment VS III). We believe that our work will contribute to the rational design of drug combinations, enable easier selection of molecular representations for *in silico* modeling, and promote further use of DL methods in biomedicine.

Methods

Data provenance

The DrugComb data portal, one of the largest public drug combination databases, is used to access combination sensitivity and synergy data [73]. Its October 2019 release contains standardized and harmonized results of 14 drug sensitivity and resistance studies on 4153 drug-like compounds screened in 112 cell lines for a total of 447 993 drug–drug-cell line tuples. Each pairwise drug combination is characterized by the CSS and four synergy scores, namely Bliss independence (Bliss), HSA, Loewe additivity (Loewe) and ZIP, further details are in the Supplementary Information. ChEMBL (release 26) is used to obtain SMILES strings, which are subsequently standardized by stripping salt residues and solvent molecules [74, 75]. SMILES shorter than 8 and longer than 140 characters are filtered out. PubChem identifiers (CID) are used to cross-reference compounds between the databases when necessary. The final DL training dataset consists of 1 795 483 unique SMILES with a median length of 48 and a median absolute deviation of 10.

Molecular representations

Fingerprints are numeric arrays of n elements (bits) long, where n ranges between 16 and 1024 depending on fingerprint type. Even though n values up to 16 384 have been tested in literature demonstrating a positive correlation between fingerprint size and downstream prediction performance, not all the studies support these findings [38, 76]. Fingerprints used in the current work are classified into rule-based with binary values and DL-based with continuous values. Rule-based models are further split into topological, 2D and 3D circular subtypes. DL fingerprints are split into sequence and graph subtypes. More detailed classification is found in Table 1.

Rule-based fingerprints

Four types of rule-based fingerprints used in the current work are: path-based (Topological 1024 bits long), 2D circular (Morgan

Table 1. Fingerprint taxonomy

Fingerprint	Type	Subtype	Length	Data format	Pretraining
E3FP	Rule	Circular 3D	1024	Binary	No
GAE	Data	Graph	16 and 64	Continuous	No
Infomax	Data	Graph	300	Continuous	Yes
Morgan	Rule	Circular 2D	300 and 1024	Binary	No
Topological	Rule	Path	1024	Binary	No
Transformer	Data	Sequence	64 and 1024	Continuous	Yes
VAE	Data	Sequence	16 and 256	Continuous	Yes

300 and 1024 bits long) and 3D circular (E3FP 1024 bits). Topological and Morgan variants are selected due to their good performance in Virtual Screening experiments [38, 43]. E3FP is a 3D extension of 2D extended-connectivity models, it is generated following the *no_Stereo* variant [77].

Deep learning-based fingerprints

Seven data-driven molecular fingerprints of different lengths are generated using four types of unsupervised encoder–decoder DL models, namely a graph autoencoder (GAE), a VAE, a Transformer and a pre-trained Deep Graph Infomax (Infomax).

GAE fingerprints

Sixteen bits long GAE fingerprints are defined via a diagonal semidefinite matrix of singular values Σ, obtained through the singular value decomposition of GAE embedding matrix [61, 62]. Inspired by the Ky Fan matrix k-norm, equal to the sum of k largest singular values of the matrix, the main diagonal of Σ is used as a 16 bits long fingerprint [78]. If small molecules result in Σ diagonal shorter than 16 bits, then zero-padding is applied. Sixty-four bits long GAE fingerprints are generated by concatenating average, min- and max-pooled representations of the embedding matrix to 16 bits long GAE fingerprints.

VAE fingerprints

VAE fingerprints are 16 and 256 bits long latent spaces of two independently trained VAE models [49].

Transformer fingerprints

Sixty-four bits long Transformer fingerprints are constructed by concatenating average- and max-pooled latent embeddings of the 16 bits model with the first output of its last and second last recurrent layers. Similarly, the 1024 bits variant is generated from the embedding space of the Transformer 256 bits model [52].

Infomax fingerprints

Infomax fingerprints are 300 bits long, generated using a pre-trained Deep Graph Infomax model that by design maximizes mutual information between local and global molecular graph features [79, 80].

Deep learning models used for fingerprint generation

Graph autoencoder model

GAE model uses a graph G, with V nodes and E edges as input, where V correspond to atoms and E to atomic bonds. Additional numeric features may be incorporated via node or bond feature matrices [81]. A graph G is represented with an adjacency matrix, $A \in \mathbb{R}^{|V| \times |V|}$, where $|V|$ are node indices, such that non-zero A elements correspond to existing molecular bonds [82]. A is normalized to be symmetric and contain self-loops following [83].

$$A_{self-loop} = A + I$$
$$A_{norm} = D^{-1/2} A_{self-loop} D^{-1/2}$$

where I is an identity matrix equal in size to A, D is a diagonal node degree matrix such that its main diagonal represents bond counts of $A_{self-loop}$. GAE model is initialized with a node matrix of 54 atom features, where each atom is represented by an array of one-hot encoded values denoting one of the 37 atoms types, six possible atom degrees, five atom charges, four variants of chiral configuration and an aromaticity indicator, all generated using RDKit. One-hot refers to encoding categorical variables as binary arrays. To make the GAE model compatible with previously unseen atoms, a placeholder for an unknown atom type is added. GAE encoder consists of seven convolutional layers with sum pooling followed by ReLu activation [84]. The decoder part is a dot product of the embedding matrix with itself, followed by 0.1 dropout and sigmoid activation. Cross-entropy over A_{norm} is used as a loss function. Empty nodes in A_{norm} are initialized with zeros.

Variational autoencoder model

Two VAE models are trained with the embedding sizes of 16 or 256 bits. Both models have a 54 characters in vocabulary, consisting of 53 unique alphanumeric characters found in SMILES and an additional empty character for zero-padding. Input length is 140 characters, zero-padded if necessary. VAE encoder consists of three 1D convolutional layers of 9, 9 and 10 neurons, each followed by SELU activation [85]. The decoder consists of three GRU layers with a hidden dimension of 501, followed by softmax activation. Loss function is an equally weighted combination of binary cross-entropy and Kullback–Leibler divergence. Xavier uniform initialization is used to assign the starting weights of two VAE models [86].

Transformer model

Two transformers are trained with the embedding sizes of 16 or 256 bits. The vocabulary size for both models is 58 characters including 53 unique SMILES characters and five tokens for *end-of-string*, *mask*, *zero-pad*, *unknown-character* and *initialize-decoding*. Maximum input length is 141 characters, zero-padded if necessary. Both models have four-headed attention and six transformer layers, with a dropout of 0.3 applied to the positional encoding output [54]. Loss function is negative log likelihood. Network weights are initialized with Xavier uniform.

Deep graph Infomax (Infomax) model

DGI is pre-trained on 465 000 molecules from ChEMBL 20 and on two million molecules from ZINC 15 by Hu *et al.* [80, 87].

DL model training

The GAE model is trained on 4153 DrugComb drug-like compounds, while VAE and Transformer models are trained on 1 795 483 molecules from DrugComb and ChEMBL 26 databases. Five-fold cross-validation is used for training all the DL models. Transformer and VAE models are trained for 10 epochs on each fold, GAE is trained for 40 epochs on each fold. All models use Adam optimizer with a learning rate decay and an initial learning rate of 1e-03, the training is halted once the learning rate reaches 1e-06 or loss reaches zero [70, 88–90]. GAE hyperparameters are optimized using tree-structured parzen estimators with a budget of 1000 iterations, other DL models employ random search [91]. Further training details can be found in Supplementary Information.

Regression analysis of molecular fingerprints (VS I)

Data input

One-hot encoded cell line labels and each of 11 drug fingerprints are used as inputs to regression models to predict drug combinations sensitivity and synergy. Combination fingerprints are generated by concatenating single molecular representations, topological fingerprints are bit-averaged [92]. Full dataset contains 362 635 cell line-drug combination tuples of 3421 compounds, when filtered by the SMILES strings (SMILES-filtered), and 447 993 combination tuples of 4153 molecules, when filtered by the CID (CID-filtered). For each cell line-drug combination tuple, four synergy scores and CSS sensitivity scores are obtained from DrugComb. If found, biological replicates are averaged, further, dose-dependent synergy scores are averaged inside cell line-drug combination tuple.

Cross-validation (VS I)

Model selection for the regression analysis of molecular fingerprints is split into three steps. In the first step, 13 different regression models are tested thrice in 5-fold cross-validation on the 10% of the full dataset, sampled without replacement (Supplementary Information). The goal is to identify an optimal type of a regression model for prediction of four synergy scores and the CSS. The second step concerns hyperparameter tuning of the previously selected regression model on all available data in 10-fold cross-validation. Lastly, the model is trained in 10-fold repeated cross-validation on SMILES-filtered and CID-filtered datasets with 90:10 and 60:40 train:test splits [93].

Regression performance metrics

Pearson correlation coefficient (PCC) and root-mean-squared error (RMSE) are used to assess the regression performance. PCC and RMSE 95% confidence intervals are calculated via student's *t*-distribution estimate of Fisher's z-transformed PCC values, and via empirical bootstrap with 1000 iterations and symmetric confidence intervals [94–98]. RMSE values are normalized by standard deviations. Shapiro–Wilk test is used to test the normality assumption [99].

Related work

PCC scores of regression models, used to predict single synergy scores in three recent studies, are in Supplementary Information section.

Fingerprint similarity (VS II & VS III)

Fingerprint similarity metric

All 11 types of molecular representations vary in length and data types, making commonly used metrics, such as Jaccard–Tanimoto or cosine distances poor choices for fingerprint comparison [100–102]. Jaccard–Tanimoto is suboptimal, as it is based on bits present in one fingerprint, absent in another and shared by both [103]. Cosine distance between two vectors, defined as their inner product normalized by the corresponding L2 norms, only measures an angle between two vectors without accounting for differences in their ranges [104]. It may be possible to post-process DL fingerprints and define common distance metrics on both the binary and real-valued arrays [105]. However, we opted against it, as we are not aware of any studies that systematically assess DL fingerprint similarity or quantify downstream effects of such transformations. Recall that an inner product is an unnormalized measure of similarity allowing metrics based on the canonical correlation analysis (CCA), singular vector CCA and projection-weighted CCA to be defined on any real-valued arrays [106–108]. All these methods underperform when the number of compounds is smaller than the dimensionality of feature space, i.e., n bits [109]. It is not intuitive to use unnormalized inner product as a similarity measure, as it is unbounded and requires original data to be referenced alongside the similarity scores. Since calculation of pairwise compound distances is not a prerequisite to quantify their similarity, we compare complete fingerprint matrices using CKA, a modification of Hilbert–Schmidt independence criterion (HSIC) originally proposed to assess nonlinear dependence of two sets of variables [110].

Fingerprint matrix

Let m compounds be represented with two fingerprint matrices X and Y, where individual fingerprints x_i and y_i may be of different data types and different lengths x and y:

$$X = [x_1, x_2, ..., x_m]^T, x_i \epsilon \mathbb{R}^x$$

$$Y = [y_1, y_2, ..., y_m]^T, y_i \epsilon \mathbb{R}^y$$

X and Y are normalized by subtracting column means from the corresponding column values.

Linear kernel k

Let K be a kernel matrix, such that its entries correspond to scalar outputs of a linear kernel function k. Let k be an inner product, $k = x_i^T y_i$, where x_i and y_i are 1D vectors from two fingerprint matrices X and Y corresponding to the same compound or feature. When x_i and y_i are column vectors, K becomes a feature similarity matrix:

$$K_{X,Y}^{feature} \epsilon \mathbb{R}^{x \times y}$$

If x_i and y_i are row vectors, K is a sample similarity matrix:

$$K_{X,Y}^{sample} \epsilon \mathbb{R}^{m \times m}$$

Hilbert-Schmidt independence criterion

HSIC is a test statistic equal to 0 when *X* and *Y* are independent [110]. Unnormalized HSIC is without an upper bound and equal to:

$$\text{HSIC}\left(K_{X,Y}^{\text{feature}}\right) = \| Y^T X \|_F^2$$

$$\text{HSIC}\left(K_{X,Y}^{\text{sample}}\right) = \text{trace}\left(X X^T Y Y^T\right)$$

where $Y^T X$ is a dot product of feature vectors and $\| . \|_F^2$ is a squared Frobenius norm and $X X^T$ and $Y Y^T$ are left Gram matrices. Notice that:

$$K_{X,Y}^{\text{feature}} = K_{X,Y}^{\text{sample}}$$

Further, for centered *X* and *Y* under linear dot product kernel:

$$\| Y^T X \|_F^2 = \text{trace}\left(X X^T Y Y^T\right) = \text{cov}(X, Y) = \| X^T Y \|_F^2$$

where *cov(X, Y)* is a cross-covariance matrix of *X* and *Y* [109].

CKA (VS II)

HSIC is an empirical statistic that converges to its unbiased value at a rate of $\frac{1}{\sqrt{\text{number of samples}}}$ [111]. Unbiased HSIC values are used to define CKA, a normalized version of HSIC that ranges from 0 to 1. CKA is used to quantify the difference between two fingerprint matrices *X* and *Y*. When CKA is calculated via the feature similarity, it is defined as:

$$\text{CKA} = \frac{\text{HSIC}\left(K_{X,Y}^{\text{feature}}\right)}{\sqrt{\text{HSIC}\left(K_{X,X}^{\text{feature}}\right) \times \text{HSIC}\left(K_{Y,Y}^{\text{feature}}\right)}}$$

CKA is a non-linear extension of the CCA and does not require any assumptions about noise distributions in the datasets [112]. CKA with linear kernel is equivalent to the RV coefficient and Tucker's Congruence coefficient [109, 113–115]. If the number of samples is higher than the number of features, CKA should be calculated using feature similarities. Conversely, sample space and use of Gram matrices is preferred.

Fingerprint clustering (VS III)

The ATC classification system is used to annotate drugs according to biological systems on which they act, as well as their therapeutic, pharmacological, and chemical properties [116]. The 2228 DrugComb compounds found in the ATC database are assigned to 10 classes. All but GAE 16 bits and Morgan 1024 bits models are then used to generate nine fingerprint matrices. The generated fingerprint matrices are preprocessed 3-fold: by z-score normalization, z-score normalization followed by dimensionality reduction with PCA and z-score normalization followed by dimensionality reduction with PLS. For the PCA preprocessing, the number of loadings explaining >0.95 variance is used, PLS regression for dimensionality reduction is performed with ATC labels as targets. Linear discriminant analysis (LDA) is used for one-versus-all clustering with ATC class labels as response variables, averaged Silhouette score and variance ratio criterion (VRC) are clustering performance metrics [117, 118].

Silhouette score for a single point is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where *a* is the mean distance between point *i* and all points within its cluster C_i and *b* is the smallest mean distance between point *i* and all points in a cluster $\neq C_i$.

VRC is a ratio of between- to within-cluster variation, adjusted by the number of clusters. VRC is closely related to the F-statistic in ANOVA [119]. Both scores are min–max scaled to be in [0, 1].

Computational facilities

All models are trained on Tesla P100 PCIe 16GB GPU. VM deployment is automated with Docker 19.03.9, python-openstack 3.14.13 and Heat Orchestration Template, Newton release. All experiments are performed using: Catboost 0.24, DGL 0.5, numpy 1.19.1, mlxtend 0.17.3, Optuna 2.2.0, pandas 1.1.3, Python 3.7.6, PyTorch 1.6, RDKit 2020.03.2, scikit-learn 0.23.2, scipy 1.4.1, XGBoost 1.2.1. Figures are created in R ggplot2 3.3.2, matplotlib 3.3.2 and seaborn 0.11.

Results and discussion

Prediction of drug combination sensitivity and synergy

Regression model selection

We identify Catboost Gradient Boosting on Decision Trees (GBDT) as an optimal regression model for the prediction of drug combination sensitivity and synergy after testing 13 algorithms on the 10% of the DrugComb dataset in three replicates (Table 2). Three of the tested algorithms failed to generate any predictions and are omitted. With optimized hyperparameters GBDT models tend to reach the early stopping criterion in the last 20% of the training on all the fingerprint variants which indicates correctly tuned hyperparameters, further details are in Supplementary Information. There exist alternative dataset splitting modes that incorporate chemical similarity via Tanimoto distance or Murcko decomposition [68, 120]. While they may better mimic current drug development practices and lead to a better correlation between *in silico* predictions and prospective experimental validation, we do not expect them to produce categorically different results.

Regression performance (VS I)

Among 11 fingerprinting models, Infomax 300 and VAE 256 achieved the highest PCC in prediction of Loewe synergy score using Catboost Gradient Boosting across all the test folds, cross-validation modes and duplicate filtering methods. As seen in Figure 2 and Table 3, for the 60:40 splits on the SMILES-filtered dataset Infomax reaches a PCC of 0.6842, while VAE 256 score is 0.6813. All tested fingerprints result in the CSS prediction scores above 0.85 PCC, with Infomax 300 and VAE 256 fingerprints still ranked on top. Infomax 300 and VAE 256 have overlapping 95% confidence intervals, as such they are considered to be equally performant. E3FP is the best rule-based fingerprint and is among the top three in most experimental runs. As seen in Figure 3 and Table 4, normalized RMSE scores further corroborate that DL-based fingerprints are better than rule-based variants in regression tasks. Further regression results for 90:10 and 60:40 train:test splits using SMILES and CID-filtered datasets are in Supplementary Figures S1–S6 and Supplementary Tables S1–S6).

Experimental results indicate that if similarity-based clustering or identification of key molecular moieties are of

Table 2. Pearson correlation coefficients of 10 regression algorithms in prediction of synergy and sensitivity scores based on Infomax 300 and Morgan 1024 bits long fingerprints with one-hot encoded cell line labels as inputs. Models are trained in three replicates, with default hyperparameters in 5-fold CV on 10% of data. VS I

| Model	CSS	Bliss	HSA	Loewe	ZIP	Rank
sklearn GBDT	0.641	0.331	0.303	0.384	0.384	1
Random Forest	0.609	0.355	0.311	0.374	0.413	2
Catboost GBDT	0.610	0.339	0.244	0.333	0.419	3
XGBoost GBDT	0.624	0.316	0.265	0.345	0.373	4
Bayesian Ridge	0.616	0.315	0.283	0.299	0.367	5
SVR linear kernel	0.588	0.270	0.219	0.253	0.333	6
Ridge	0.599	0.251	0.219	0.287	0.311	7
Elasticnet	0.332	NaN	NaN	NaN	NaN	8
OLS	0.092	0.023	0.022	0.042	0.030	9
Lasso	0.264	NaN	NaN	NaN	NaN	10
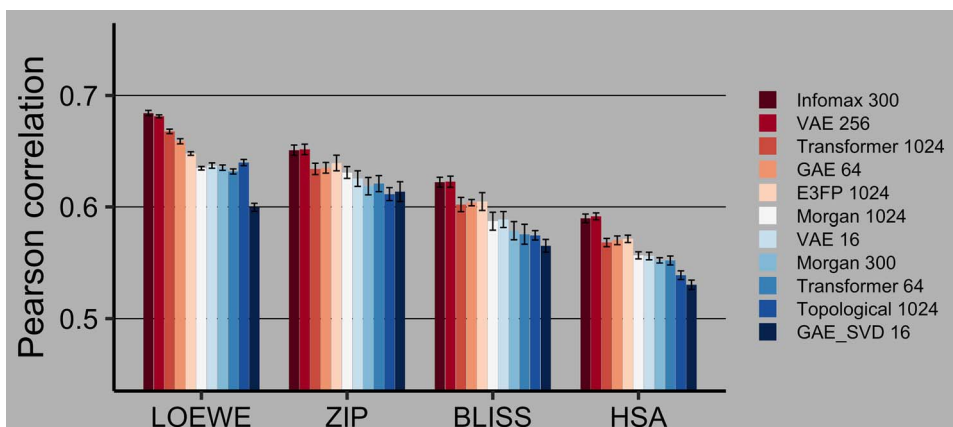
Figure 2. Drug combination synergy prediction on the SMILES-filtered dataset in 60:40 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with red. VS I task.

Table 3. Drug combination sensitivity and synergy prediction on the SMILES-filtered dataset in 60:40 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Three best models are in bold. VS I task

Fingerprint	Pearson's *r* and 95% Confidence interval				
	CSS	Bliss	HSA	Loewe	ZIP
E3FP 1024	0.8641 ± 0.0017	**0.6048 ± 0.0080**	**0.5716 ± 0.0033**	0.6479 ± 0.0016	**0.6394 ± 0.0070**
GAE 16	0.8540 ± 0.0032	0.5654 ± 0.0057	0.5304 ± 0.0042	0.5996 ± 0.0036	0.6137 ± 0.0089
GAE 64	**0.8667 ± 0.0028**	0.6038 ± 0.0028	0.5703 ± 0.0039	0.6589 ± 0.0023	0.6351 ± 0.0048
Infomax 300	**0.8761 ± 0.0019**	**0.6222 ± 0.0044**	**0.5897 ± 0.0039**	**0.6842 ± 0.0024**	**0.6509 ± 0.0046**
Morgan 300	0.8541 ± 0.0022	0.5788 ± 0.0080	0.5523 ± 0.0024	0.6352 ± 0.0024	0.6186 ± 0.0078
Morgan 1024	0.8605 ± 0.0028	0.5873 ± 0.0079	0.5568 ± 0.0032	0.6347 ± 0.0016	0.6309 ± 0.0053
Topological 1024	0.8405 ± 0.0019	0.5748 ± 0.0042	0.5390 ± 0.0039	0.6398 ± 0.0028	0.6115 ± 0.0058
Transformer 64	0.8582 ± 0.0023	0.5756 ± 0.0088	0.5522 ± 0.0040	0.6318 ± 0.0023	0.6209 ± 0.0072
Transformer 1024	0.8663 ± 0.0022	0.6021 ± 0.0064	0.5683 ± 0.0037	**0.6678 ± 0.0020**	0.6341 ± 0.0051
VAE 16	0.8616 ± 0.0018	0.5888 ± 0.0070	0.5562 ± 0.0034	0.6371 ± 0.0024	0.6254 ± 0.0070
VAE 256	**0.8759 ± 0.0022**	**0.6226 ± 0.0050**	**0.5915 ± 0.0031**	**0.6813 ± 0.0013**	**0.6516 ± 0.0047**

interest, rule-based fingerprints should be considered. Their average performance in regression is compensated by the inbuilt interpretability and robust clustering performance [121]. On the other hand, neural fingerprint models are well-suited for regression tasks, as seen in the VS I experiment. It is important to note that the differences in regression performance between rule- and DL-based fingerprints do not exceed 0.05 PCC when predicting any synergy scores or the CSS. Consistently good performance of the DL models and E3FP fingerprints may be offset by their high computational costs during model training or fingerprint generation, respectively. GAE 64 fingerprints appear to be a reasonable compromise in terms of the downstream performance and relatively short model training times.

Fingerprint similarity

CKA distance (VS II)

A heatmap of pairwise CKA similarities between 11 fingerprints, as seen in Figure 4, indicates that similar types of fingerprints
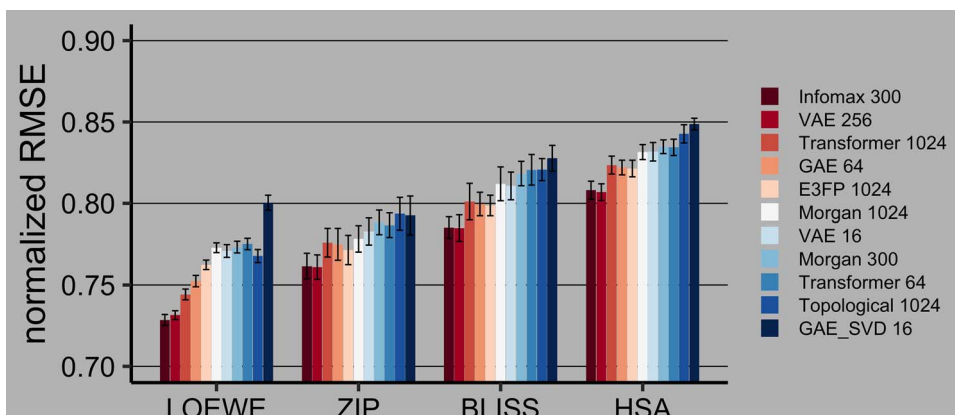
Figure 3. Drug combination synergy prediction on the SMILES-filtered dataset in 60:40 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Best models are highlighted with red. Normalized RMSE value of 1 indicates that the standard deviation of residuals is equal to the standard deviation of the target, i.e., a model that predicts mean values for all targets would have such a normalized RMSE. VS I task.

Table 4. Drug combination sensitivity and synergy prediction on the SMILES-filtered dataset in 60:40 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Three best models are in bold. VS I task

Fingerprint	Normalized root-mean-squared error and 95% confidence interval				
	CSS	Bliss	HSA	Loewe	ZIP
E3FP 1024	0.5034 ± 0.0011	**0.7987 ± 0.0063**	**0.8214 ± 0.0051**	0.7624 ± 0.0029	**0.7715 ± 0.0090**
GAE 16	0.5205 ± 0.0020	0.8277 ± 0.0079	0.8487 ± 0.0036	0.8004 ± 0.0046	0.7926 ± 0.0119
GAE 64	**0.4990 ± 0.0016**	0.7996 ± 0.0072	0.8220 ± 0.0045	0.7524 ± 0.0035	0.7750 ± 0.0098
Infomax 300	**0.4822 ± 0.0011**	**0.7852 ± 0.0066**	**0.8081 ± 0.0055**	**0.7285 ± 0.0034**	**0.7616 ± 0.0079**
Morgan 300	0.5204 ± 0.0007	0.8183 ± 0.0075	0.8348 ± 0.0042	0.7733 ± 0.0036	0.7884 ± 0.0075
Morgan 1024	0.5097 ± 0.0013	0.8120 ± 0.0104	0.8315 ± 0.0046	0.7729 ± 0.0030	0.7783 ± 0.0081
Topo^A 1024	0.5420 ± 0.0011	0.8207 ± 0.0068	0.8427 ± 0.0056	0.7678 ± 0.0040	0.7937 ± 0.0100
T^B 64	0.5134 ± 0.0014	0.8206 ± 0.0094	0.8344 ± 0.0050	0.7752 ± 0.0035	0.7867 ± 0.0075
T^B 1024	0.4998 ± 0.0012	0.8011 ± 0.0112	0.8235 ± 0.0055	**0.7442 ± 0.0033**	0.7760 ± 0.0088
VAE 16	0.5080 ± 0.0010	0.8107 ± 0.0085	0.8317 ± 0.0057	0.7709 ± 0.0039	0.7828 ± 0.0083
VAE 256	**0.4825 ± 0.0013**	**0.7849 ± 0.0081**	**0.8069 ± 0.0051**	**0.7315 ± 0.0027**	**0.7610 ± 0.0075**

cluster together. Rule-based fingerprints form two clusters corresponding to topological and circular subtypes. All the DL fingerprints generated by the trained models form the third cluster. Graph-based models appear to be far removed from all sequence and rule-based variants. GAE 64 is the most different from other trained DL fingerprints, while being co-clustered with them. Infomax 300 fingerprints, based on a pre-trained Deep Graph Infomax model, are not part of any cluster. Smaller sequence-based DL fingerprints, namely VAE 16 and Transformer 64 are at least as similar to each other, as they are to their longer in-type/subtype counterparts. We conclude that fingerprint type and subtype, as indicated in Table 1, contribute the most to the CKA similarity, followed by fingerprint pretraining status, size, and data format.

LDA clustering (VS III)

To further study the differences between fingerprint models, we perform one-versus-all LDA classification of 2228 compounds based on their ATC classes, using nine different fingerprinting models to represent the molecules. The GAE 16 bits fingerprints are omitted, since GAE 64 bits fingerprints extend their shorter counterparts by concatenating average, min- and max-pooled embedding spaces. Further, due to the comparable performance of Morgan 300 and 1024 bits models in VS I and VS II experiments, only Morgan 300 bits fingerprints are used in LDA clustering

experiments. VS III clustering results are in Table 5 and the overview of DrugComb compounds with the corresponding ATC classes is in Figure 5. The Infomax 300 bits model achieves the best clustering results on the z-score normalized fingerprint matrices, followed by three rule-based fingerprints. Dimensionality reduction following z-score normalization generally improves clustering performance of all rule-based fingerprints. It has the opposite effect on most DL fingerprints, with the largest reduction seen in the Infomax 300 and GAE 64 models. Longer DL sequence models, namely VAE 256 and Transformer 1024, perform better after dimensionality reduction steps, albeit with a minimal improvement in relative rankings. Such differences between graph and sequence-based DL fingerprints are supported by the CKA analysis (VS II study), indicating that the graph-based fingerprints differ the most from other DL variants.

Conclusion

Choosing an optimal fingerprint type to represent molecular features is an important step in computational drug discovery. To this end, we systematically compared 11 variants of such molecular representations in predicting drug combination sensitivity and synergy scores, and evaluated their relationships based on the clustering performance and CKA-based fingerprint similarity. We found that VAE 256 bits long and 3D circular
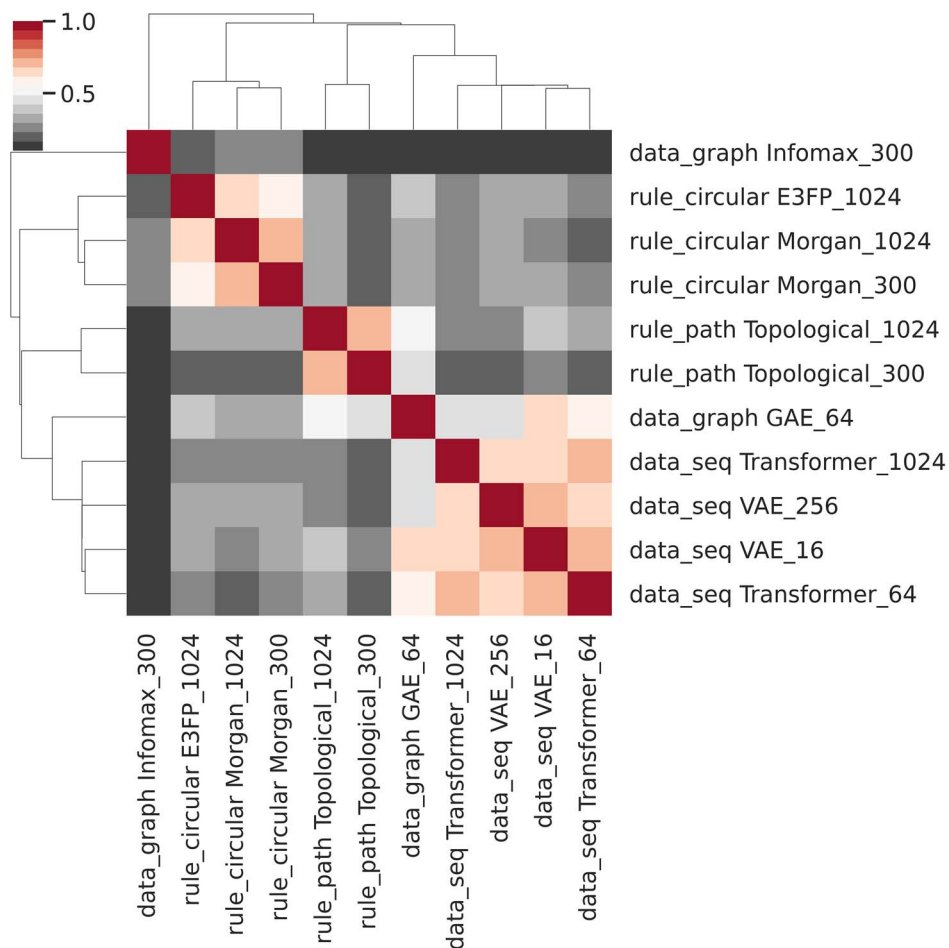
Figure 4. Heatmap of pairwise CKA distances between 11 fingerprints. Infomax and GAE 64 are DL fingerprints based on molecular graphs. VAE and Transformer are sequence-based DL fingerprints. E3FP and variants of Morgan and topological fingerprints are generated using rule-based models. VS II task.

Table 5. One-versus-all LDA clustering in 10 ATC classes of 2228 DrugComb compounds represented with nine fingerprint types. Averaged Silhouette and VRC scores are rescaled to [0,1]. Fingerprints are ranked according to scores on z-score normalized data. PCA- and PLS-based dimensionality reduction improves rule-based fingerprint (denoted by rule_ prefix) performance, most DL fingerprints (data_ prefix) decrease performance, VAE 256 and Transformer 1024 benefit from dimensionality reduction, although minimally in terms of relative ranking. VS III task

Fingerprint	z-score		z-score + PCA		z-score + PLS	
	Silhouette	VRC	Silhouette	VRC	Silhouette	VRC
data_Infomax 300	0.984	1.000	0.701	0.251	0.749	0.326
rule_E3FP 1024	1.000	0.831	0.993	0.958	0.988	0.959
rule_Topo 1024	0.979	0.771	1.000	0.997	1.000	1.000
rule_Morgan 300	0.967	0.655	0.993	1.000	0.997	0.984
data_GAE 64	0.440	0.122	0.258	0.064	0.303	0.062
data_VAE 256	0.399	0.049	0.547	0.170	0.574	0.190
data_VAE 16	0.223	0.056	0.076	0.036	0.000	0.006
data_Transformer 64	0.088	0.004	0.000	0.000	0.120	0.000
data_Transformer 1024	0.000	0.000	0.131	0.023	0.377	0.067

E3FP 1024 bits long fingerprints generated from SMILES strings, as well as Infomax 300 bits long fingerprints based on molecular graphs lead to the best regression performance. Out of the four tested synergy scores, we observe that Loewe synergy is the easiest to predict with best models reaching PCC 0.72. CSS, a measure of drug combination efficacy, can be predicted >0.85 PCC with any fingerprint type. We found that the rule-based fingerprinting methods underperform in regression tasks in comparison to the data-driven DL variants. However, the gap between the best and worst performing fingerprint models rarely exceeds PCC 0.05. Further, we adapted CKA to quantify the extent of similarity between fingerprint matrices and to demonstrate that similar types of fingerprints cluster together. An optimal similarity measure for the comparison of single rule-based and data-driven fingerprints remains an open question. Lastly, in one-versus-all compound clustering using ATC classes
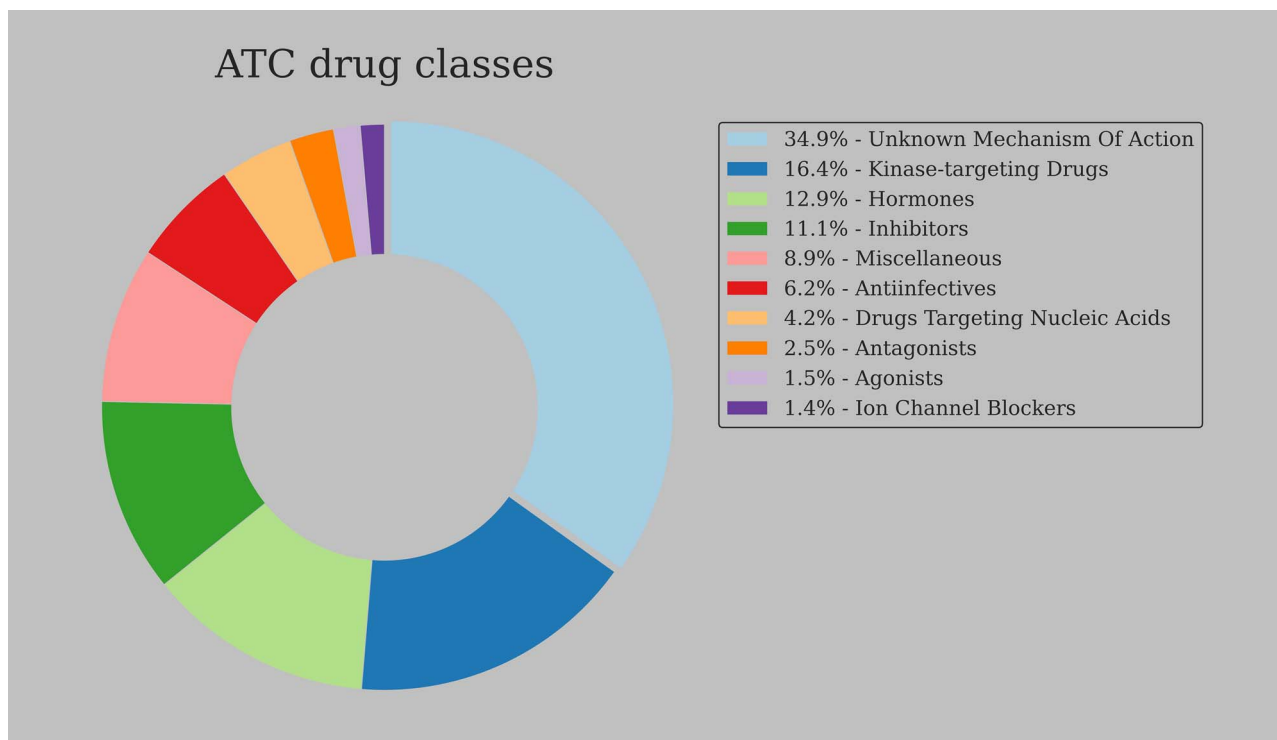
Figure 5. ATC drug classes of the DrugComb compounds ($n = 3421$). Over one third ($n = 1193$) compounds do not have a mechanism of action assigned in the ATC classification system.

as labels, rule-based fingerprints perform on par or better than the best DL representations.

We conclude that the quantitative performance differences between rule-based and DL-based fingerprints are likely to be insignificant in the context of preclinical studies of small molecule drugs [122, 123]. In order to identify an optimal fingerprint type for a given project we advise enriching quantitative performance metrics with qualitative concerns, e.g., available chemometric and DL expertise, model interpretability requirements, opinions of project stakeholders and model performance on unseen data. Fingerprints generated using the E3FP 1024, Infomax 300, Morgan 1024 and VAE 256 bits models are suggested as good starting points based on our experimental results and distinct methodologies underlying their data generating methods [124]. We recommend the Loewe synergy score for use in drug combination screening due to its best performance among four tested synergy models tested on dose–response data from 14 DSRT studies.

This work focuses on the evaluation of single fingerprint types. However, it is worth exploring the impact of combining several fingerprints together. We expect a statistically significant regression performance increase when combining molecular representations with low CKA similarity, or using models trained on multimodal data and/or key biological databases, such as Gene Ontology, Protein Data Bank and UniRef [5, 125–127]. Another line of inquiry could address high computational costs of DL and E3FP models. To this end, we suggest exploring alternative molecular representations and CPU-friendly generative models based on genetic algorithms, such as STONED on SELFIES [128]. Finally, we hope that in the future biomedical DL research will go beyond representation learning and will be used to derive novel biological knowledge by e.g., inferring synthetic and retrosynthetic chemical reactions, identifying novel

disease-associated druggable proteins and clinically actionable biomarkers [129–131].

Key Points

- To choose an optimal molecular fingerprint type, it is advised to enrich quantitative metrics of model performance with qualitative concerns related to the nature of downstream tasks, model interpretability and robustness requirements, as well as available chemometric expertise.
- Data-driven fingerprints, namely VAE 256 bits long trained on SMILES and Infomax 300 bits long-trained molecular graphs are well-suited for regression tasks. 1024 bits long 2D and 3D circular fingerprints are flexible and well-performant rule-based models fit for clustering tasks. GAE 64 bits long model may be used in any analysis scenario as a baseline option.
- Loewe synergy scores enable the highest correlation between *in silico* predictions and subsequent experimental validation of drug combination synergy in cancer cell lines.
- CKA is an effective measure of molecular representation similarity applicable to any combination of rule-based and DL fingerprints.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data and code availability

The data and code underlying this article are available in the article and in its online supplementary material.

Code: https://github.com/NetPharMedGroup/publication_fingerprint/.

Data: https://doi.org/10.5281/zenodo.4843919.

Supplementary information

Combination sensitivity and synergy scores

It is known that the choice of an appropriate null model of no interaction is crucial for an accurate assessment of drug synergy. Four such models are used in the current work. Bliss model is based on probabilistic independence of drug effects, such that single agents are competing, but independent perturbations each contributing to a total effect [30]. HSA assumes that an expected drug combination effect is the higher of the two single-agent effects at corresponding concentrations [31]. Degree of non-interaction according to Loewe is equal to the outcome of a sham experiment, that is, combining a drug with itself [32, 33]. ZIP assumes that non-interacting drugs minimally impact each other's dose–response curves [34]. CSS quantifies drug combination efficacy, which is defined as an average of areas under drug combination dose–response curves, whereby each curve is determined by fixing one of the drugs at its IC50 concentration [29].

Regression model selection

Thirteen regression models, representing a wide spectrum of ML algorithms, are compared in prediction of drug combination synergy and sensitivity. All models are tested with default hyperparameters in 5-fold cross-validation using 1024 bits long Morgan and 300 bits long Infomax fingerprints together with one-hot encoded cell line labels on 10% of randomly sampled data in three replicates. Thirteen tested regression models are: Bayesian Ridge, Catboost Gradient Boosting, ElasticNet, Gaussian Process Regression with a sum of Dot Product and White Kernels, Histogram-based Gradient Boosting, Isotonic Regression, Lasso regression, LassoLars regression, Linear regression, Ridge regression, Random Forest, Support Vector Machines with a linear kernel and XGBoost Gradient Boosting. All trees-based models are limited to a depth of six. Neural networks are not included in the comparison, since on tabular data they tend to perform on par with the previously mentioned methods, while being less interpretable and more difficult to set up [132, 133].

Top four identified regression models are bagging and boosting ensembles, followed by linear kernel Support Vector Regression and Bayesian Ridge. Gaussian Process Regression, Isotonic and Lassolars models failed to generate any predictions and their results are omitted from Table 1. Catboost implementation of GBDT is selected for further experiments due to its efficient GPU utilization and two design choices aimed at reducing overfitting: out-of-the-box categorical encoding that translates classes into numeric representations, binning them based on the expected value of target statistic, and ordered boosting whereby training data

are randomly permuted throughout tree growing process to limit unwanted target metric prediction shift, one of well-known GBDT disadvantages [134–136].

Tuned in 10-fold cross-validation best hyperparameters for Catboost GBDT are Poisson bootstrap with 0.66 subsampling ratio, L2 regularization of 9, tree depth of 10, learning rate of 0.15 with 5000 boosting iterations and 50 early stopping rounds. Overall, we conclude that ensembles are the most powerful type of tested ML algorithms in prediction of drug combination synergy and sensitivity [137].

Neural network training

For all DL models, the hyperparameter search consisted of testing activation functions (GELU, ELU, LeakyReLU, ReLU, SELU, Sigmoid, Softmax, Swish), dropout ratios (0.1–0.5 with 0.1 step size), initial learning rates (1e-01 to 1e-05, with a step size of 0.1), number of patience epochs (1–30 with a step size of 1) and the learning rate decay factor (0.9–0.1 with a step size of 0.1). Transformer models were tested with 3–6 attention heads. VAE encoders were tested with up to five convolutional layers and convolutional kernel sizes up to 10; VAE decoder is tested with up to three recurrent GRU layers of sizes up to 600.

Transformer. Two transformer models are trained in 5-fold cross-validation, for up to 10 epochs on each fold with a decay factor of 0.1 and patience of 5 epochs in batch sizes of 650 and 340 for 16 bits and 256 bits long fingerprint variants. The final cross-entropy losses are 2e-07 and 1e-08, respectively.

VAE. Two VAE models with the latent spaces of 16 and 256 neurons are trained on ChEMBL 26 using 5-fold cross-validation and 10 epochs on each fold with a decay factor of 0.2 and patience of 2 and 3 epochs, respectively. An equally weighted sum of binary cross-entropy and KL-divergence is used as a loss metric. The final losses are 0.9231 and 0.0984 for 16 bits and 256 bits models.

GAE. A single GAE model is trained on 3421 unique Drug-Comb compounds in 5-fold cross-validation mode for 200 epochs with a batch size of 340 for up to 40 epochs per fold. Learning rate decay factor is 0.1 with 30 epochs patience. Cross-entropy over the molecular graph adjacency matrix is used as a loss, with the best score of 0.8604 reached at the end of the 200th epoch.

It is likely that longer training, more extensive hyperparameter optimization, or use of alternative optimizers, such as SGD with a cyclic learning rate scheduling, may result in lower final loss values. However, we do not expect them to have a significant influence on the downstream experiments [138]. It is interesting to note that optuna-based optimization of the GAE model resulted in the encoder architecture consisting of seven convolutional layers 54-46-40-34-28-22-16 neurons wide. Such a high number of convolutional layers is somewhat unexpected, as performance of Graph Neural Networks based on spectral convolutions is expected to deteriorate with the convolutional layer count above six, most likely due to excessive feature smoothing [139, 140]. Relatively deep GAE encoder architecture may be explained by a positive correlation between the performance of DL models and the number of trainable parameters, as the seven layer GAE model with dot-product based Decoder has

circa 10 k learnable parameters, whereas large Transformer and VAE models that reach lower test losses during training have two orders of magnitude more parameters [141].

Prior work in drug combination synergy predictions

In three independent studies listed below single synergy scores are predicted. Developed models are cross-validated on single datasets.

Random forest. A Random Forest model developed by Menden *et al.* achieved a 0.3 PCC in prediction of Loewe synergy using drug and cell line labels on the AstraZeneca dataset including 910 combinations tested in 85 cancer cell lines [142].

Convolutional neural network. A CNN model introduced by Preur *et al.* achieved a 0.73 PCC in prediction of Loewe synergy from drug fingerprints, physicochemical molecular descriptors and basal expression of 3984 cancer-associated genes on the O'Neil dataset including 583 drug combinations tested in 39 cancer cell lines [143, 144].

Gradient Boosting and Random Forest. XGBoost and Random Forest models developed by Sidorov *et al.* achieved a 0.64 PCC for a modified version of Bliss synergy from drug fingerprints and their physicochemical characteristics on the NCI-ALMANAC dataset including 5232 combinations tested in 60 cancer cell lines [26, 145].

Acknowledgements

Funding

References

1. Wainberg M, Merico D, Delong A, *et al*. Deep learning in biomedicine. *Nat Biotechnol* 2018;**36**:829–38.
2. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;**15**:20170387. 10.1098/rsif.2017.0387.
3. Lo Y-C, Rensi SE, Torng W, *et al*. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 2018;**23**:1538–46.
4. McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;**577**:89–94.
5. Gainza P, Sverrisson F, Monti F, *et al*. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92.
6. Zhavoronkov A, Ivanenkov YA, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* 2019;**37**:1038–40.
7. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Sci Rep* 2018;**8**:11242.
8. Senior AW, Evans R, Jumper J, *et al*. Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.
9. Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;**110**:12–22.
10. Chen D, Liu S, Kingsbury P, *et al*. Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit Med* 2019;**2**:43.
11. Bhhatarai B, Walters WP, Hop CECA, *et al*. Opportunities and challenges using artificial intelligence in ADME/Tox. *Nat Mater* 2019;**18**:418–22.
12. Mayr A, Klambauer G, Unterthiner T, *et al*. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* 2018;**9**:5441–51.
13. Campbell DT. Assessing the impact of planned social change. *Eval Program Plann* 1979;**2**:67–90.
14. Goodhart CAE. Problems of monetary management: the UK experience. *Monetary Theory Practice*, 1st edn. 1984;91–121. https://link.springer.com/book/10.1007/978-1-349-17295-5#about.
15. Heath I, Hippisley-Cox J, Smeeth L. Measuring performance and missing the point? *BMJ* 2007;**335**:1075–6.
16. Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;**178**:1544–7.
17. Ha D, Schmidhuber J. *World Models*, 2018. arxiv preprint arXiv:1803.10122.
18. Wagstaff K. Machine learning that matters. In: *arXiv [cs.LG]*, 2012. arxiv preprint arXiv:1206.4656.
19. Hirschfeld L, Swanson K, Yang K, *et al*. Uncertainty quantification using neural networks for molecular property prediction. *J Chem Inf Model* 2020;**60**:3770–80.
20. Pahikkala T, Airola A, Pietilä S, *et al*. Toward more realistic drug-target interaction predictions. *Brief Bioinform* 2015;**16**:325–37.
21. Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem Sci* 2019;**10**:8154–63.
22. David L, Arús-Pous J, Karlsson J, *et al*. Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Front Pharmacol* 2019;**10**:1303.
23. Tang J, Aittokallio T. Network pharmacology strategies toward multi-target anticancer therapies: from computational models to experimental design principles. *Curr Pharm Des* 2014;**20**:23–36.
24. Tang J. Informatics approaches for predicting, understanding, and testing cancer drug combinations. *Methods Mol Biol* 2017;**1636**:485–506.
25. Pemovska T, Kontro M, Yadav B, *et al*. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov* 2013;**3**:1416–29.
26. Holbeck SL, Camalier R, Crowell JA, *et al*. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res* 2017;**77**:3564–76.
27. Borisy AA, Elliott PJ, Hurst NW, *et al*. Systematic discovery of multicomponent therapeutics. *Proc Natl Acad Sci U S A* 2003;**100**:7977–82.

28. Tallarida RJ. Quantitative methods for assessing drug synergism. *Genes Cancer* 2011;**2**:1003–8.

29. Malyutina A, Majumder MM, Wang W, *et al*. Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer. *PLoS Comput Biol* 2019;**15**:e1006752.

30. Bliss CI. The toxicity of poisons applied jointly1. *Ann Appl Biol* 1939;**26**:585–615.

31. Berenbaum MC. What is synergy? *Pharmacol Rev* 1989;**41**: 93–141.

32. Greco WR, Bravo G, Parsons JC. The search for synergy: a critical review from a response surface perspective. *Pharmacol Rev* 1995;**47**:331–85.

33. Loewe S. The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung* 1953;**3**:285–90.

34. Yadav B, Wennerberg K, Aittokallio T, *et al*. Searching for drug synergy in complex dose-response landscapes using an interaction potency model. *Comput Struct Biotechnol J* 2015;**13**:504–13.

35. Maggiora G, Vogt M, Stumpfe D, *et al*. Molecular similarity in medicinal chemistry. *J Med Chem* 2014;**57**:3186–204.

36. Cherkasov A, Muratov EN, Fourches D, *et al*. QSAR modeling: where have you been? Where are you going to? *J Med Chem* 2014;**57**:4977–5010.

37. Neves BJ, Braga RC, Melo-Filho CC, *et al*. QSAR-based virtual screening: advances and applications in drug discovery. *Front Pharmacol* 2018;**9**:1275.

38. O'Boyle NM, Sayle RA. Comparing structural fingerprints using a literature-based similarity benchmark. *J Chem* 2016;**8**:36.

39. Durant JL, Leland BA, Henry DR, *et al*. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**:1273–80.

40. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics, 2 Volume Set: Volume I: Alphabetical Listing/Volume II: Appendices, References*. Wiley–VCH, 2009.

41. Chuang KV, Gunsalus LM, Keiser MJ. Learning molecular representations for medicinal chemistry. *J Med Chem* 2020;**63**:8705–22.

42. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc* 1965;**5**:107–13.

43. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.

44. Voet A, Qing X, Lee XY, *et al*. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. *J Receptor Ligand Channel Res* 2014;**7**:81.

45. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* 2019;**20**:1878–912.

46. Zhang A, Lipton ZC, Li MU, Smola AJ. Dive into Deep Learning. *arXiv preprint arXiv:2106.11342*. 2021. https://d2l.ai/chapter_recurrent-modern/encoder-decoder.html (5 April 2021, date last accessed).

47. Goh GB, Siegel C, Vishnu A, *et al*. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. In: *arXiv [stat.ML]*, 2017. arxiv preprint arXiv:1706.06689.

48. Goh GB, Hodas NO, Siegel C, *et al*. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. In: *arXiv [stat.ML]*, 2017. arxiv preprint arXiv:1712.02034.

49. Gómez-Bombarelli R, Wei JN, Duvenaud D, *et al*. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci* 2018;**4**:268–76.

50. Cho K, van Merrienboer B, Gulcehre C, *et al*. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: *arXiv [cs.CL]*, 2014. arxiv preprint arXiv:1406.1078.

51. Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: *arXiv [stat.ML]*, 2013. arxiv preprint arXiv:1312.6114.

52. Honda S, Shi S, Ueda HR. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. In: *arXiv [cs.LG]*, 2019. arxiv preprint arXiv:1911.04738.

53. Manica M, Oskooei A, Born J, *et al*. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol Pharm* 2019;**16**:4797–806.

54. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Adv Neural Inform Process Syst* 2017;**30**:6000.

55. Wang Z, Liu M, Luo Y, *et al*. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. In: *arXiv [q-bio.QM]*, 2020. arxiv preprint arXiv:2012.01981.

56. Winter R, Montanari F, Noé F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 2019;**10**:1692–701.

57. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* 1988;**28**:31–6.

58. Daylight Theory Manual. Daylight Version 4.9. *Daylight Theory: SMARTS – A Language for Describing Molecular Patterns*. Laguna Niguel, CA: Daylight Chemical Information Systems, Inc., 2011. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (5 April 2021, date last accessed).

59. O'Boyle NM. Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. *J Chem* 2012;**4**:22.

60. Ramsundar B, Kearnes S, Riley P, *et al*. Massively Multitask Networks for Drug Discovery. In: *arXiv [stat.ML]*, 2015. arxiv preprint arXiv:1502.02072.

61. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: *arXiv [cs.LG]*, 2016. arxiv preprint arXiv:1609.02907.

62. Kipf TN, Welling M. Variational Graph Auto-Encoders. In: *arXiv [stat.ML]*, 2016. arxiv preprint arXiv:1611.07308.

63. Gilmer J, Schoenholz SS, Riley PF, *et al*. Neural Message Passing for Quantum Chemistry. In: *arXiv [cs.LG]*, 2017. arxiv preprint arXiv:1704.01212.

64. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al*. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In: *arXiv [cs.LG]*, 2015. arxiv preprint arXiv:1509.09292.

65. Yang K, Swanson K, Jin W, *et al*. Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 2019;**59**:3370–88.

66. Kearnes S, McCloskey K, Berndl M, *et al*. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016;**30**:595–608.

67. Chami I, Abu-El-Haija S, Perozzi B, *et al*. *Machine Learning on Graphs: A Model and Comprehensive Taxonomy* arXiv [cs.LG], 2020. arxiv preprint arXiv:2005.03675.

68. Wu Z, Ramsundar B, Feinberg EN, *et al*. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**:513–30.

69. Hu W, Fey M, Zitnik M, *et al*. Open Graph Benchmark: Datasets for Machine Learning on GraphsarXiv [cs.LG], 2020. arxiv preprint arXiv:2005.00687.

70. Dwivedi VP, Joshi CK, Laurent T, *et al*. Benchmarking Graph Neural Networks. In: *arXiv [cs.LG]*, 2020. arxiv preprint arXiv:2003.00982.

71. Meyer CT, Wooten DJ, Lopez CF, *et al*. Charting the fragmented landscape of drug synergy. *Trends Pharmacol Sci* 2020;**41**:266–80.

72. Tang J, Wennerberg K, Aittokallio T. What is synergy? The Saariselkä agreement revisited. *Front Pharmacol* 2015;**6**:181.

73. Zagidullin B, Aldahdooh J, Zheng S, *et al*. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res* 2019;**47**:W43–51.

74. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.

75. Bento AP, Hersey A, Félix E, *et al*. An open source chemical structure curation pipeline using RDKit. *J Chem* 2020;**12**:51.

76. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Chem* 2013;**5**:26.

77. Axen SD, Huang X-P, Cáceres EL, *et al*. A simple representation of three-dimensional molecular structure. *J Med Chem* 2017;**60**:7393–409.

78. Fan K. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proc Natl Acad Sci U S A* 1951;**37**:760–6.

79. Veličković P, Fedus W, Hamilton WL, *et al*. Deep Graph Infomax. In: *arXiv [stat.ML]*, 2018. arxiv preprint arXiv:1809.10341.

80. Hu W, Liu B, Gomes J, *et al*. Strategies for Pre-training Graph Neural Networks. In: *arXiv [cs.LG]*, 2019. arxiv preprint arXiv:1905.12265.

81. Schlichtkrull M, Kipf TN, Bloem P, *et al*. Modeling relational data with graph convolutional networks. *Semantic Web* 2018;593–607.

82. Hamilton WL, Ying R, Leskovec J. Representation Learning on Graphs: Methods and Applications. In: *arXiv [cs.SI]*, 2017. arxiv preprint arXiv:1709.05584.

83. von Luxburg U. A Tutorial on Spectral Clustering. In: *arXiv [cs.DS]*, 2007. arxiv preprint arXiv:0711.0189.

84. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 2011;**15**:315–23.

85. Klambauer G, Unterthiner T, Mayr A, *et al*. Self-Normalizing Neural Networks. In: *arXiv [cs.LG]*, 2017. arxiv preprint arXiv:1706.02515.

86. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 2010;**9**:249–56.

87. Sterling T, Irwin JJ. ZINC 15–ligand discovery for everyone. *J Chem Inf Model* 2015;**55**:2324–37.

88. Kingma DP, Ba JA. A Method for Stochastic Optimization. In: *arXiv [cs.LG]*, 2014. arxiv preprint arXiv:1412.6980.

89. Wang M, Zheng D, Ye Z, *et al*. *Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks*arXiv [cs.LG], 2019. arxiv preprint arXiv:1909.01315.

90. Paszke A, Gross S, Massa F, *et al*. *PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv [cs.LG]*, 2019. arxiv preprint arXiv:1912.01703.

91. Akiba T, Sano S, Yanase T, *et al*. *Optuna: A Next-Generation Hyperparameter Optimization Framework*. arXiv [cs.LG], 2019.

92. Mason DJ, Stott I, Ashenden S, *et al*. Prediction of antibiotic interactions using descriptors derived from molecular structure. *J Med Chem* 2017;**60**:3902–12.

93. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of K-fold cross-validation. *J Mach Learn Res* 2004;**5**:1089–105.

94. Efron B. *Better Bootstrap Confidence Intervals*, 1984.

95. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 1915;**10**:507.

96. Efron B, Tibshirani RJ. Introduction. *An Introduction to the Bootstrap*, 1st edn. 1993;1–9.

97. Corey DM, Dunlap WP, Burke MJ. Averaging correlations: expected values and bias in combined Pearsons and Fisher's z-transformations. *J Gen Psychol* 1998;**125**:245–61.

98. Bishara AJ, Hittner JB. Confidence intervals for correlations when data are not normal. *Behav Res Methods* 2017;**49**:294–309.

99. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965;**52**:591.

100. Bender A, Jenkins JL, Scheiber J, *et al*. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* 2009;**49**:108–19.

101. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Chem* 2015;**7**:20.

102. Todeschini R, Consonni V, Xiang H, *et al*. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *J Chem Inf Model* 2012;**52**:2884–901.

103. Todeschini R, Ballabio D, Distances CV. Similarity measures in chemometrics and chemoinformatics. *Encyclop Anal Chem* 2020;1–40.

104. Algebra, Topology, Differential Calculus, and Optimization Theory for Computer Science and Machine Learning. *Book in Progress*. In: Gallier J, Quaintance J (ed). 2020. https://www.cis.upenn.edu/~jean/gbooks/geomath.html (5 April 2021, date last accessed).

105. Szedmak S, Bach E. *On the Generalization of Tanimoto-Type Kernels to Real Valued Functions*. arXiv [cs.LG] 2020. arxiv preprint arXiv:2007.05943.

106. Schölkopf B. The kernel trick for distances. *Adv Neural Inform Process Syst* 2001;**13**:301.

107. Raghu M, Gilmer J, Yosinski J, *et al*. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Adv Neural Inform Process Syst* 2017;**30**:6078.

108. Morcos AS, Raghu M, Bengio S. *Insights on Representational Similarity in Neural Networks with Canonical Correlation*. arXiv [stat.ML] 2018. arxiv preprint arXiv:1806.05759.

109. Kornblith S, Norouzi M, Lee H, *et al*. *Similarity of Neural Network Representations Revisited*. arXiv [cs.LG] 2019. arxiv preprint arXiv:1905.00414.

110. Gretton A, Bousquet O, Smola A, *et al*. Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory*. In: *16th International Conference, ALT 2005, Singapore, October 8–11, 2005, Proceedings*. Sanjay J, Hans US, Etsuji T (eds). 2005;63–77.

111. Song L, Smola A, Gretton A, *et al*. Supervised feature selection via dependence estimation. *Proceedings of the 24th International Conference on Machine Learning* 2007; 823–30. https://doi.org/10.1145/1273496.1273600.

112. Gretton A, Herbrich R, Smola A, *et al*. Kernel methods for measuring independence. *J Mach Learn Res* 2005;**6**:2075–129.

113. Thompson JAF, Bengio Y, Schoenwiesner M. *The Effect of Task and Training on Intermediate Representations in Convolutional Neural Networks Revealed with Modified RV Similarity Analysis*. arXiv [cs.LG] 2019. arxiv preprint arXiv:1912.02260.

114. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the RV- coefficient. *Appl Stat* 1976;**25**:257.

115. Josse J, Holmes S. Measuring multivariate association and beyond. *Stat Surv* 2016;**10**:132–67.

116. PubChem. *WHO ATC Code - PubChem Data Source*. Oslo, Norway: WHO Collaborating Centre for Drug Statistics Methodology, 2018. https://www.whocc.no/atc/structure_and_principles/ (5 April 2021, date last accessed).

117. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 2009.

118. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Simul Comput* 1974;**3**:1–27.

119. Everitt BS, Dunn G. *Applied Multivariate Data Analysis*. Wiley, 2001.

120. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996;**39**:2887–93.

121. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst* 2020; 1–21. doi: 10.1109/TNNLS.2020.3027314.

122. Cortés-Ciriano I, Bender A. Reliable prediction errors for deep neural networks using test-time dropout. *J Chem Inf Model* 2019;**59**:3330–9.

123. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov Today* 2021;**26**:1040–52.

124. Gao K, Nguyen DD, Sresht V, *et al*. Are 2D fingerprints still valuable for drug discovery? *Phys Chem Chem Phys* 2020;**22**:8373–90.

125. Kuenzi BM, Park J, Fong SH, *et al*. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020;**38**:672–684.e6.

126. Elnaggar A, Heinzinger M, Dallago C, *et al*. *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing*. arXiv [cs.LG], 2020. arxiv preprint arXiv:2007.06225.

127. GüvençPaltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Brief Bioinform* 2021;**22**:346–59.

128. Nigam A, Pollice R, Krenn M, *et al*. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES. *Chem Sci* 2021;**12**:7079–90.

129. Méndez-Lucio O, Baillif B, Clevert D-A, *et al*. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat Commun* 2020;**11**:10.

130. Jin W, Barzilay R, Jaakkola T. *Discovering Synergistic Drug Combinations for COVID with Biological Bottleneck Models*. arXiv [q-bio.BM] 2020. arxiv preprint arXiv:2011.04651.

131. Bychkov D, Linder N, Tiulpin A, *et al*. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep* 2021;**11**:4037.

132. Yang Y, Morillo IG, Hospedales TM. *Deep Neural Decision Trees*. arXiv [cs.LG] 2018. arxiv preprint arXiv:1806.06988.

133. Abutbul A, Elidan G, Katzir L, *et al*. DNF-Net: A Neural Architecture for Tabular Data. 2020. arxiv preprint arXiv:2006.06465.

134. Prokhorenkova L, Gusev G, Vorobev A, *et al*. *CatBoost: unbiased boosting with categorical features*. arXiv [cs.LG] 2017. arxiv preprint arXiv:1706.09516.

135. Breiman L. Using iterated bagging to Debias regressions. *Mach Learn* 2001;**45**:261–77.

136. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev* 2021;**54**:1937–67.

137. Olson RS, Cava WL, Mustahsan Z, *et al*. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput* 2018;**23**:192–203.

138. Smith LN. *Cyclical Learning Rates for Training Neural Networks*. arXiv [cs.CV] 2015. arxiv preprint arXiv:1506.01186.

139. Brockschmidt M. *GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation*. arXiv [cs.LG] 2019. arxiv preprint arXiv:1906.12192.

140. Alon U, Yahav E. *On the Bottleneck of Graph Neural Networks and its Practical Implications*. arXiv [cs.LG] 2020. arxiv preprint arXiv:2006.05205.

141. Kaplan J, McCandlish S, Henighan T, *et al*. *Scaling Laws for Neural Language Models*. arXiv [cs.LG] 2020. arxiv preprint arXiv:2001.08361.

142. Menden MP, AstraZeneca-Sanger Drug Combination DREAM Consortium, Wang D, *et al*. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun* 2019;**10**:2674.

143. Preuer K, Lewis RPI, Hochreiter S, *et al*. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 2018;**34**:1538–46.

144. O'Neil J, Benita Y, Feldman I, *et al*. An unbiased oncology compound screen to identify novel combination strategies. *Mol Cancer Ther* 2016;**15**:1155–62.

145. Sidorov P, Naulaerts S, Ariey-Bonnet J, *et al*. Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front Chem* 2019;**7**:509.