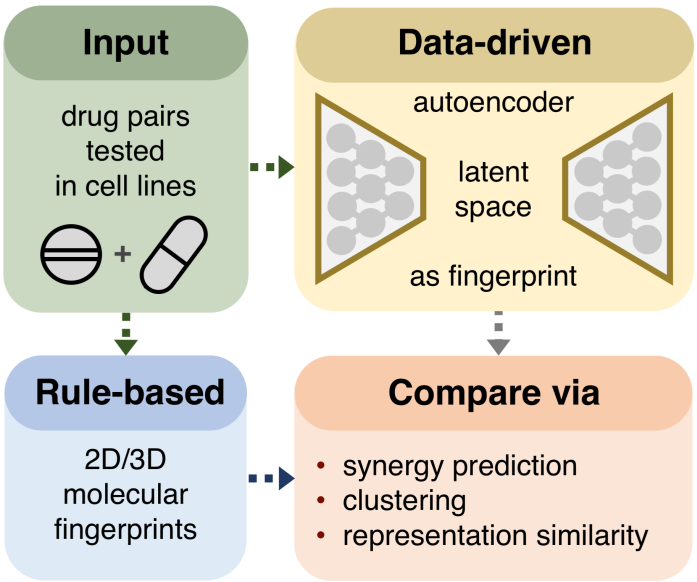
COMPARATIVE ANALYSIS OF MOLECULAR REPRESENTATIONS IN PREDICTION OF CANCER DRUG COMBINATION SYNERGY AND SENSITIVITY

B. Zagidullin[1,*], Z. Wang[2], Y. Guan[3], E. Pitkänen[4], and J. Tang[1,*]

[1]*Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Finland*
[2]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA*
[3]*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, USA*
[4]*Institute for Molecular Medicine Finland (FIMM) and Applied Tumor Genomics Research Program, University of Helsinki, Finland*
[*]{bulat.zagidullin,jing.tang}@helsinki.fi

ABSTRACT

Application of machine and deep learning (ML/DL) 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 DL 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 4 153 molecules tested in 112 cancer cell lines. We evaluate the clustering performance of molecular fingerprints and quantify their similarity by adapting Centered Kernel Alignment metric. Our work demonstrates that in order to identify an optimal 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.



Keywords drug combinations · drug synergy · machine learning · molecular fingerprints

1 Introduction

In the past years machine and deep learning (ML/DL) methods have received a considerable amount of attention in biomedicine, drug discovery and computational chemistry [1–3]. Deep neural networks achieve state-of-the-art performance in medical computer vision tasks and protein structural modelling, enabling de novo generation of drug candidates and development of prognostic clinical models [4–8]. However, such outstanding performance of DL models is context-dependent [9–12]. While quantitative metrics are routinely and effectively used to compare ML/DL methods, overreliance on them is a well-known issue in machine learning, economics, social science, clinical practice and other complex systems [13–17]. It is beneficial to supplement rankings on benchmark datasets with estimates of model uncertainty and robustness, as well as ability to generalize on unseen data [18–20]. Such qualitative aspects are particularly important in biomedical research and individualized medicine, where experimental design choices are guided by the *in silico* model predictions, as exhaustively testing all combinations of studied variables is usually not viable due to the combinatorial explosion [21,22].

Advances in high throughput screening of bioactive compounds in cancer cell lines promote development of personalized and effective cancer treatments [23]. A major goal in such drug sensitivity and resistance testing studies is to prioritize promising treatment regimens for further experimental validation using *in silico* rational design of drug combinations [24]. 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 [25,26]. 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 [27]. It is known that the choice of an appropriate null model of no interaction is crucial for an accurate assessment of drug synergy. In the current work we utilize four such baseline models, namely Bliss independence (Bliss), Highest single agent (HSA), Loewe additivity (Loewe), and Zero interaction potency (ZIP). Bliss model is based on probabilistic independence of drugs effects, such that single agents are considered to be competing, but independent perturbations each contributing to a total effect [28]. HSA assumes that an expected drug combination effect is the higher of the two single-agent effects at corresponding concentrations [29]. Degree of non-interaction according to Loewe is equal to the outcome of a sham experiment, that is combining a drug with itself [30,31]. Zero Interaction Potency (ZIP) assumes that non-interacting drugs minimally impact each other's dose-response curves [32]. While synergy measures how compounds in question modulate each other's biological activity, combination sensitivity score (CSS) is designed to inform about drug combination clinical efficacy. It is defined as an average of areas under two dose-response curves, whereby each curve is generated by fixing the concentration of one of the drugs at its IC_{50} value and varying the second one over the full range of tested doses [33]. This makes CSS dose-independent and numerically equivalent to percent growth inhibition.

Application of drug structural information in rational design of drug combinations is related to Quantitative Structure Activity Relationship (QSAR) modelling and Virtual Screening (VS) [34,35]. The QSAR captures mathematical associations between chemical structures and assay endpoints based on the assumption that structurally similar compounds have similar (bio)activity properties, while in the VS studies candidate molecules are prioritized for subsequent experimental validation according to *in silico* screening results [36]. Many types of molecular fingerprints, representing drugs as numeric arrays, have been developed by integrating empirical data from high throughput screening studies with aspects of molecular geometry [37,38]. One of the most well-known examples of such representations is MACCS structural keys that are path-based topological fingerprints combining information about atom properties, bond properties and 2D chemical structures [39,40]. Another type of commonly used rule-based molecular representations are circular topological fingerprints that describe combinations of non-hydrogen atom types and paths between them within a predefined atom neighborhood [41,42]. Recently, data-driven molecular fingerprints generated by means of DL algorithms have been shown to perform well in various QSAR/VS tasks [43]. Most such DL models 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 representation ($\mathbf{z}$), which is then used to reconstruct the original input [44]. When an encoder-decoder neural network is trained on chemical structures, its latent space is interpreted as a data-driven fingerprint. Examples of the early DL fingerprinting models include a Convolutional Neural Network-based (CNN) Chemception and a Recurrent Neural Network-based (RNN) SMILES2Vec, as well as a Variational Autoencoder (VAE) model with a CNN encoder and Gated Recurrent Unit-based (GRU) decoder [45–49]. Development of attention methods for sequence data further reinforced the importance of DL in the creation of data-driven fingerprints, whereas generative neural networks enabled de novo generation of chemical libraries by sampling from trained models' latent spaces [50–54]. Most aforementioned

DL solutions operate on images of molecules or their sequence-based representations, such as SMARTS or SMILES [55–57]. An alternative type of neural fingerprints based on Geometric Deep Learning (GDL) models, such as Graph Convolutional Networks (GNN) and Message Passing Neural Networks, utilizes natural representations of molecules in form of graphs with atoms as edges and bonds as vertices [58–60]. Any data-driven molecular representations may be supplemented with global molecular properties by concatenating them to the model's latent space or through the application of multitask learning methods [61]. Graph-based DL models extend that functionality further and make use of positional information pertinent to single atom and bond properties [62–64].

There exists a number of extensive benchmark datasets for ranking DL models in QSAR/VS-related tasks, such as MoleculeNet designed for molecular machine learning, or GDL-focused Open Graph Benchmark and Benchmarking GNNs [65–67]. We are not aware of any systematic evaluation of molecular fingerprints' performance in prediction of drug combination effects. To address the gap, we compare 11 molecular representations, comprising seven DL and four rule-based variants in prediction of cancer drug combination synergy and sensitivity. Owing to the lack of consensus on the optimal formulation of drug synergy, we consider four competing non-interaction null models based on distinct valid assumptions and identify a preferred model for use in cancer drug combinations research based on 17 271 848 data points from 14 cancer drug screening studies [68,69]. To compare fingerprints of different types and sizes, we use Linear Discriminant Analysis to cluster compounds into 10 WHO ATC classes and measure fingerprint similarity by adapting Centered Kernel Alignment as a distance metric [70–72]. Even though benchmark performance results may not translate well due to a gap between quantitative improvements on standardized datasets and impact of such gains outside of academic research, we hope that our work will contribute to the rational design of effective drug combination and promote a more effective use of novel Deep Learning methods in the biomedical research [73].

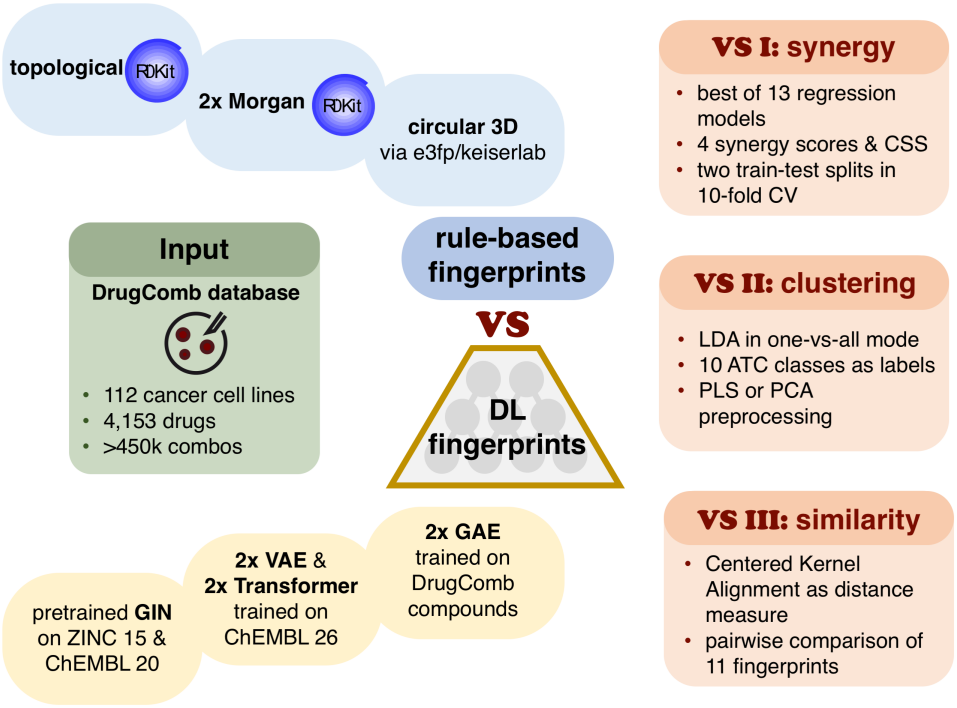


Figure 1: Study workflow. Compounds found in combinations in the DrugComb database are represented using four rule-based (blue) and seven data-driven (yellow) fingerprint types. Rule-based fingerprints include topological and 2D/3D extended connectivity variants. Data-driven fingerprints are generated using two Variational Autoencoder and two Transformer models trained on ChEMBL 26, Graph Autoencoder trained on DrugComb compounds and a pre-trained Graph Infomax model (GIN). Prediction of drug combination sensitivity and four synergy scores is used to identify the best-performing synergy model (VS I). One-vs-all fingerprint classification using Linear Discriminant Analysis (VS II), molecular representation similarity based on the Centered Kernel Alignment (VS III) together with the VS I result are used to compare molecular representations between each other.

2 Methods

2.1 Data provenance

The DrugComb data portal, one of the largest public drug combination databases, is used to access combination synergy and sensitivity data [74]. Its October 2019 release contains standardized and harmonized results of 14 drug sensitivity and resistant studies on 4 153 drug-like compounds screened in 112 cell lines for a total of 447 993 drug-combination-cell-line tuples. Each data point is characterized by a combination sensitivity score (CSS) and four drug combination synergy scores, namely Bliss, HSA, Loewe and ZIP. Chemical structures of molecules are obtained from ChEMBL (release 26) as SMILES strings [75]. SMILES are standardized by constructing molecular graphs, stripping salt residues and solvent molecules, and regenerating the original input [76]. Strings 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 compound pre-training dataset consists of 1 795 483 unique SMILES with a median length of 48 and a median absolute deviation of 10.

2.2 Molecular representations

Fingerprints are represented as arrays of n elements (bits) long, where n ranges between 16 and 1 024 depending on a fingerprint type. Even though n values up to 16 384 have been tested in literature and were found to positively correlate with better prediction performance, not all the studies support these findings [39,77]. All individual values of data-driven fingerprints are in the double-precision floating-point format, rule-based fingerprints are in the binary format.

2.2.1 Rule-based fingerprints

SMILES are used to generate four types of topological fingerprints: a 1 024 bits long path-based fingerprints (Topological), two circular variants of different lengths (Morgan 300 and Morgan 1 024 and a 3D circular 1 024 bits long fingerprint (E3FP). 2D Topological and Morgan fingerprints are selected due to their good performance in a variety of small molecule virtual screening experiments [39,42]. E3FP fingerprints based on 3D molecular topology are generated following the *no_Stereo* model, such that a single lowest energy conformer, selected from 20 candidate representations, undergoes five rounds of neighborhood sampling performed within an iteratively expanding by a factor of 1.7 radius [78].

2.2.2 Deep learning-based fingerprints

Seven data-driven molecular fingerprints of different lengths are generated from four types of unsupervised encoder-decoder DL models, namely a Graph Convolutional Network [79] (GAE), a Variational Autoencoder [47] (VAE), a Transformer [50] and a Graph Isomorphism Network (Infomax) [80–82]. GAE is trained on 4 153 DrugComb compounds, VAE and Transformer models - on 1 795 483 molecules from DrugComb and ChEMBL 26 databases. A pre-trained Infomax model employed 465 000 molecules from ChEMBL 20 and 2 million structures from ZINC 15 [83].

- 16 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, $\mathbf{z}_{GAE}$. 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 GAE fingerprint [84]. If small molecules result in Σ diagonal shorter than 16 bits, then zero-padding is applied.
- 64 bits long GAE are generated by concatenating average, min- and max-pooled representations of $\mathbf{z}_{GAE}$ to a 16 bits long GAE fingerprint.
- VAE fingerprints are 16 and 256 bits long defined using latent spaces, $\mathbf{z}_{VAE\ 16}$ and $\mathbf{z}_{VAE\ 256}$ of two VAE models.

- 64 bits long Transformer fingerprints are constructed by concatenating average- and max-pooled $\mathbf{z}_{TRANSFORMER\ 16}$ with the first output of its last and second last recurrent layers. Transformer 1 024 bits variant is generated from the $\mathbf{z}_{TRANSFORMER\ 256}$ in a similar manner.
- Infomax fingerprints are 300 bits long.

2.3 DL models used for fingerprint generation

2.3.1 Graph Autoencoder (GAE)

Graph Autoencoder model is a Graph Convolutional Network [58] that uses a graph $\mathcal{G}$, with $\mathcal{V}$ nodes and $\mathcal{E}$ edges as input, where $\mathcal{V}$correspond to atoms and $\mathcal{E}$ to atomic bonds. Additional numeric features may be incorporated via node or bond feature matrices [85]. A graph $\mathcal{G}$ is represented with an adjacency matrix, A $\epsilon\ \mathbb{R}^{|\mathsf{V}|\times|\mathsf{V}|}$, where $|\mathsf{V}|$ are node indices, such that non-zero A elements correspond to existing molecular bonds [86]. A is normalized to be symmetric and contain self-loops following [87]:

$$A_{self-loop} = A + I \tag{1}$$

$$A_{norm} = D^{-1/2} A_{self-loop} D^{-1/2}, \tag{2}$$

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 $A_{self-loop}$. To initialize a GAE model a node matrix of 54 atom features is used, 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. In order to make the model compatible with previously unseen atoms a placeholder for an unknown atom type is added. The encoder part of the GAE model consists of seven convolutional layers with sum pooling followed by ReLu activation [88]. The decoder part is a dot product of $\mathbf{z}_{GAE}$ 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.

2.3.2 Variational Autoencoder (VAE)

Two VAE models are trained with the embedding sizes of 16 or 256 bits. Both models have a 54 characters 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 [89]. 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 [90].

2.3.3 Transformer

Two Transformer models are trained with the embedding sizes of 16 or 256 neurons. 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 [52]. Loss function is negative log likelihood. Network weights are initialized with Xavier uniform.

2.3.4 Graph Isomorphism Network (Infomax)

Deep Graph Infomax is based on one of the most expressive GNN architectures [91]. It can also be effectively pre-trained [80]. Unlike deterministic graph models like GAE that sample only local atom neighborhoods of size equal to the number of convolutional layers, Infomax architecture is designed to maximize mutual information between local and global graph representation [81].

2.3.5 Model training

5-fold cross-validation is used for training all Deep Learning models. Transformer and VAE 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 $1 \cdot 10^{-3}$ [92]. Training of each model is halted once the learning rate reaches $1 \cdot 10^{-6}$ or loss reaches zero [67,93]. GAE hyperparameters are optimized using tree-structured parzen estimators with a budget of 1 000 iterations, other models employ random search [94]. Further details can be found in the Supplementary section.

2.3.6 Model input

Drug fingerprints and one-hot encoded cell line labels are used as inputs to regression models. A drug combination fingerprint is generated by concatenating single drug molecular embeddings. Topological 1 024 bits long fingerprints are bit-averaged, since path-based fingerprints deterministically map molecular moieties to an array index [95]. Predictions are done on 362 635 cell line-drug combination tuples of 3 421 compounds filtered by the SMILES strings (SMILES-filtered), and on 447 993 tuples of 4 153 molecules when filtered by the CIDs (CID-filtered). Four dose-independent synergy scores are obtained by averaging experimental replicates across 14 studies, followed by averaging dose-dependent synergy scores for each cell line-drug combination tuple. CSS score is dose-independent.

2.3.7 Cross-validation and model selection

Model selection is split into three steps. In the first step 13 Regression models are tested thrice on the 10% of the full dataset, generated by sampling without replacement. 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 model using all available data in 10-fold cross-validation. Lastly, the best regression model with optimized hyperparameters is trained in 10-fold repeated cross-validation on SMILES-filtered and CID-filtered datasets with 90:10 and 60:40 train:test splits [96].

2.3.8 Regression performance metrics

Pearson correlation coefficient (PCC) and Root Mean Squared Error (RMSE) are used to assess the regression models' performance. PCC and RMSE 95% confidence intervals are calculated via Student's t-distribution estimate of Fisher's z transformed PCC values and the empirical bootstrap with 1 000 iterations and symmetric confidence intervals [97–101]. RMSE values are normalized by targets' standard deviations making them directly comparable between each other, regardless of the distributions and ranges of the target metrics. Shapiro-Wilk test is used to test the normality assumption with the null hypothesis of samples following a normal distribution [102].

2.4 Fingerprint Similarity

2.4.1 Fingerprint matrix

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

$$X = [x_1, x_2, ..., x_m]^\mathsf{T}, x_i \in \mathbb{R}^\mathsf{x} \tag{3}$$

$$Y = [y_1, y_2, ..., y_m]^\mathsf{T}, y_i \in \mathbb{R}^\mathsf{y} \tag{4}$$

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

2.4.2 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, $\mathsf{k} = x_i^T y_i$, 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:

$$K_{X,Y}^{feature} \epsilon \mathbb{R}^{\mathsf{x} \times \mathsf{y}} \tag{5}$$

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

$$K_{X,Y}^{sample} \epsilon \mathbb{R}^{\mathsf{m} \times \mathsf{m}} \tag{6}$$

2.4.3 Hilbert-Schmidt Independence Criterion

Hilbert-Schmidt Independence Criterion (HSIC) is a test statistic equal to 0 when X and Y are independent [103]. Unnormalized HSIC is without an upper bound and equal to:

$$HSIC(K_{X,Y}^{feature}) = ||Y^{\mathsf{T}}X||_{\mathsf{F}}^2 \tag{7}$$

$$HSIC(K_{X,Y}^{sample}) = trace(XX^{\mathsf{T}}YY^{\mathsf{T}}) \tag{8}$$

where $Y^{\mathsf{T}}X$ is a dot product of feature vectors and $|| \cdot ||_{\mathsf{F}}^2$ is a squared Frobenius norm and XX^{T} and YY^{T} are left Gram matrices. Notice that:

$$K_{X,Y}^{feature} = K_{X,Y}^{sample} \tag{9}$$

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

$$||Y^{\mathsf{T}}X||_{\mathsf{F}}^2 = trace(XX^{\mathsf{T}}YY^{\mathsf{T}}) = cov(X, Y) = ||X^{\mathsf{T}}Y||_{\mathsf{F}}^2 \tag{10}$$

where $cov(X, Y)$ is a cross-covariance matrix of X and Y [104].

2.4.4 Centered Kernel Alignment

HSIC is an empirical statistic that converges to its unbiased value at a rate of $\frac{1}{\sqrt{number\ of\ samples}}$ [105]. Unbiased HSIC values are used to define Centered Kernel Alignment (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:

$$CKA = \frac{HSIC(K_{X,Y}^{feature})}{\sqrt{HSIC(K_{X,X}^{feature}) \times HSIC(K_{Y,Y}^{feature})}} \tag{11}$$

2.5 Fingerprint Clustering

The WHO Anatomical Therapeutic Chemical (ATC) Classification System is used to annotate drugs according to biological systems on which they act, and their therapeutic, pharmacological and chemical properties [70]. 2 228 DrugComb compounds found in the WHO ATC database are assigned to 10 ATC classes. All, but GAE 16 bits and Morgan 1 024 bits fingerprints are included in the analysis, the GAE 64 bits variant is preferred as it is an extension of its shorter version, while the performance of Morgan 300 bits fingerprints closely resembles that of Morgan 1 024 bits in other experiments. All used fingerprints matrices are Z-score normalized, followed by two dimensionality reduction steps: PCA with a number of loadings explaining $\geq$ 0.95 variance and PLS regression with class labels as targets. Linear Discriminant Analysis (LDA) is used for one-vs-all clustering with averaged Silhouette score and Variance Ratio Criterion as performance metrics [106,107]. Silhouette score for a single point is defined as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{12}$$

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 $\neq C_i$. Variance Ratio Criterion (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 [108]. Both scores are min-max scaled to be in $[0, 1]$.

2.6 Computational facilities

All models are trained on Tesla P100 PCIe 16GB GPU, kindly provided by the Finnish IT Center for Science. VM deployment is automated with Docker 19.03.9, python-openstack 3.14.13 and Heat Orchestration Template, Newton release [109]. 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.

3 Results and Discussion

3.1 Prediction of drug combination synergy and sensitivity

3.1.1 Baseline models

Results of three previous studies are used as baseline:

- 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 [110].
- a CNN model introduced by Preur *et al.*, achieved a 0.73 PCC in prediction of Loewe synergy from drug fingerprints, physiochemical molecular descriptors and basal expression of 3 984 cancer-associated genes on the O'Neil dataset including 583 drug combinations tested in 39 cancer cell lines [111,112].
- 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 NCI60 Almanac dataset including 5 232 combinations tested in 60 cancer cell lines [25,113].

All aforementioned models predict a single synergy score and are cross-validated on single datasets using as input one-hot encoded cell lines, one-hot encoded drugs for the Menden model and molecular fingerprints for the Preur and Sidorov models.

3.1.2 Catboost GBDT model convergence

We identify Catboost Gradient Boosting on Decision Trees (GBDT) as an optimal regression model based on 13 algorithms tested on the 10% of the DrugComb dataset in three replicates [114]. Three of the tested algorithms failed to generate any predictions and are omitted from Table 1, further details of model selection experiments are in the Supplementary section. After hyperparameter tuning, on the CID- and SMILES-filtered datasets in both 90 : 10 and 60 : 40 train-test splitting modes, GBDT models tend reach the early stopping criterion in the last 20% of the training, indicating correctly tuned hyperparameters. There exist alternative dataset splitting modes incorporating chemical similarity considerations based on Tanimoto distance or Murcko decomposition [65,115]. 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.

Table 1: Pearson correlation coefficients of 10 regression algorithms in prediction of synergy and sensitivity scores based on Infomax 300 and Morgan 1 024 bits long fingerprints with one-hot encoded cell line labels as inputs. Averaged results of three independent 5-fold CV rounds on 10% of data run with models' default hyperparameters.

model type	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

3.1.3 Prediction performance

From four tested null models of non-interaction, Loewe synergy score leads to the highest PCC using Infomax 300 and VAE 256 bits long fingerprints across all the test folds, cross-validation modes and duplicates filtering methods. As seen in Figure 2 and Table 2, for the 60 : 40 splits Infomax reaches 0.684 2 on the SMILES-filtered dataset, while VAE 256 results are 0.681 3. CSS predictions in all experiments are above 0.85 PCC, with Infomax 300 and VAE 256 still ahead of the other models. Since the top two fingerprint types have overlapping 95% confidence intervals in regression tasks, they are considered to be equally performant. E3FP is the best rule-based fingerprint and is among the top three models in a majority of experimental runs. Further, according to Figure 3 and Table 3 normalized RMSE scores corroborate that neural fingerprints outperform rule-based drug representations for all synergy models and the CSS score across all datasets and validation splits. The remaining regression results are in the Supplementary Figures and Tables section (Figures S1-S6 and Tables S1-S6).

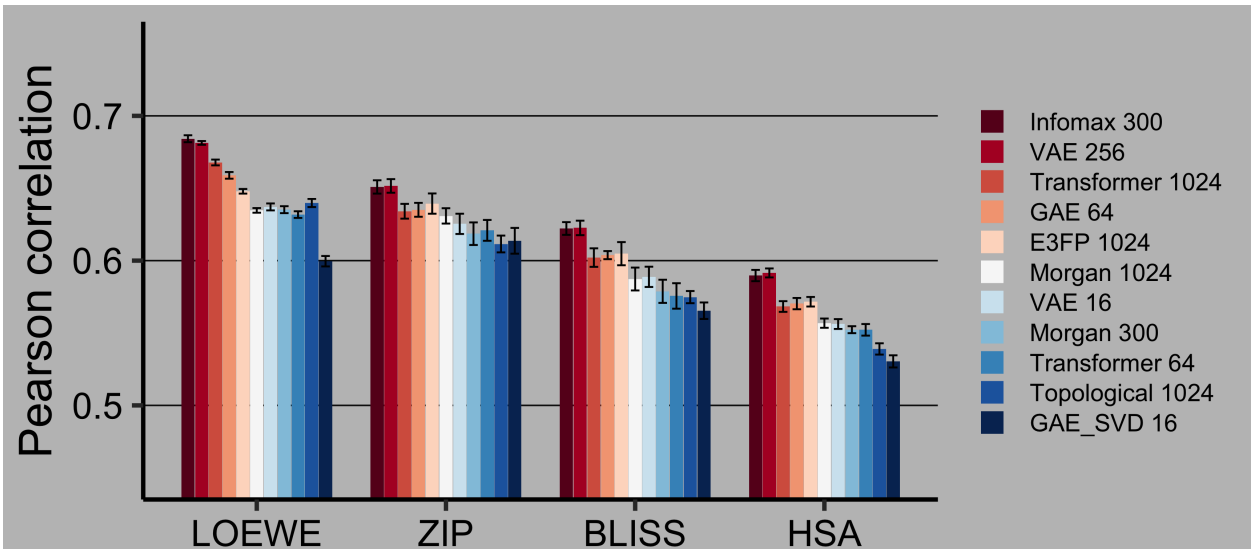


Figure 2: Drug combination synergy prediction on the SMILES-filtered dataset in 60 : 40 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with red.

Table 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 bold.

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**

If similarity-based clustering or identification of key molecular moieties is of interest, average performance of topological fingerprints in regression tasks is compensated by their inbuilt interpretability and excellent clustering performance

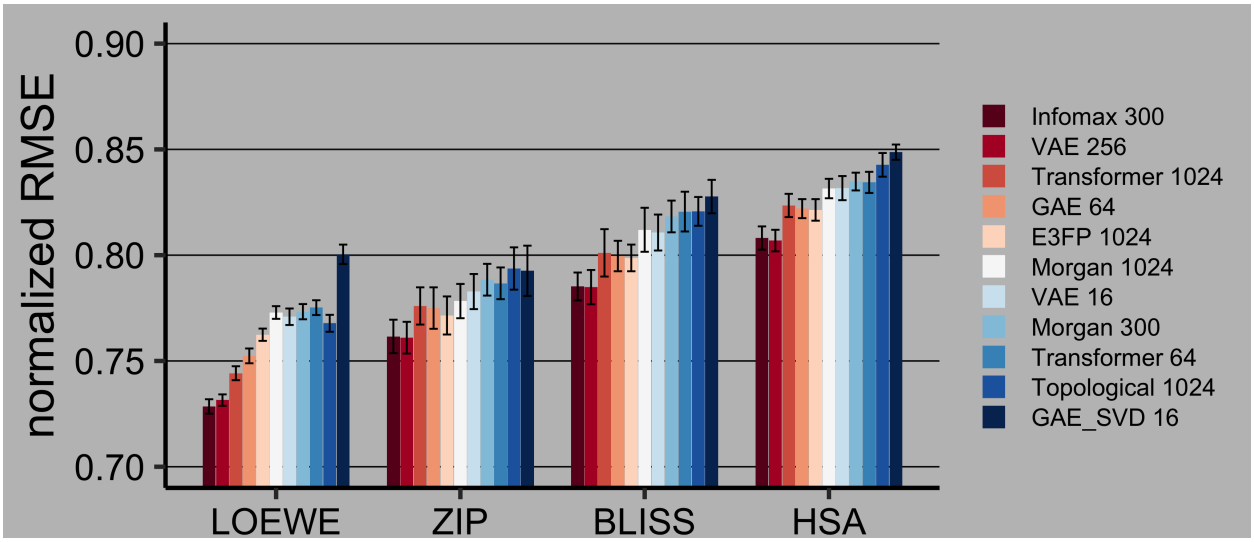


Figure 3: Drug combination synergy prediction on the SMILES-filtered dataset in 60 : 40 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Best models are highlighted with red. Normalized RMSE value of 1 indicates 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.

Table 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 bold

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
Topological 1024	0.5420 ±0.0011	0.8207 ±0.0068	0.8427 ±0.0056	0.7678 ±0.0040	0.7937 ±0.0100
Transformer 64	0.5134 ±0.0014	0.8206 ±0.0094	0.8344 ±0.0050	0.7752 ±0.0035	0.7867 ±0.0075
Transformer 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**

[116]. On the other hand, neural fingerprint models lead to the best regression results, however, generally the differences between the best and the worst performing fingerprint types do not exceed 0.05 PCC. Performance advantages of all data-driven models, as well as the E3FP fingerprints, may be offset by their high computational costs. Graph Autoencoder-based fingerprints appear to be a reasonable compromise in terms of the downstream performance and relatively fast training times.

3.2 Fingerprint similarity

3.2.1 Metric selection

All the 11 types of generated fingerprints vary in length and data types of individual bit values, further, all neural fingerprints lack any zero bits. These factors make commonly-used distance metrics, such as Jaccard–Tanimoto or cosine distance poor choices for model comparison [117–119]. Jaccard–Tanimoto is suboptimal, as it is based on bits present in one fingerprint, absent in another and shared by both [120]. Cosine distance between two vectors, defined as their inner product normalized by the corresponding L^2 norms, only measures an angle between two vectors without accounting for differences in their ranges [121]. For example, two compounds with the Tanimoto similarity of 0.13 based on the Topological 1 024 bits long fingerprints, have a cosine similarity score > 0.95 when using any of the neural fingerprints. It may be possible to post-process data-driven molecular representations and define common distance metrics on both the binary and real-valued arrays [122]. However, we opted against it, as we are not aware of any studies that systematically assess neural 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 [123–125]. All these methods underperform when a number of compounds is smaller than the dimensionality of feature space, i.e. fingerprint lengths [104]. Use of unnormalized inner product as a similarity measure on its own is non-intuitive, 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 representation similarity, we compare complete fingerprint matrices using Centered Kernel Alignment (CKA). CKA is a modification of Hilbert-Schmidt Independence Criterion (HSIC), a metric originally proposed to assess nonlinear dependence of two sets of variables using matrix multiplication [103]. It is a non-linear extension of the CCA and does not require any assumptions about noise distributions in the datasets [126]. Under linear kernel it is equivalent to the RV coefficient and Tucker's Congruence coefficient [104,127–129]. Linear kernel may be substituted with another kernel type, provided a priori compound similarities are known. If a number of samples is higher than the number of their features, CKA should be calculated using feature similarities. Conversely, sample space and use of Gram matrices is preferred.

3.2.2 CKA distance

A heatmap of pairwise CKA similarities between the fingerprints, as seen in Figure 4, indicates that similar types of fingerprints cluster together. For example, topological fingerprints of different lengths appear to be closer related to each other than to other types, 2D and 3D circular topological fingerprints display a similar trend. All the neural fingerprints generated by the trained models form the third cluster. Graph-based models appear to be far removed from the SMILES-based neural fingerprints. GAE is the least similar to other pre-trained neural models, while still being co-clustered with them. Infomax fingerprints do not cluster with any other fingerprint type. It is interesting to note that fingerprints of similar lengths, in particular VAE 16 and Transformer 64 bits, are more closely-related to each other, than to their longer in-cluster/in-class versions.

3.2.3 LDA clustering

To further investigate the differences between studied fingerprints, we performed a one-vs-all classification aiming to cluster compounds belonging to each of the 10 identified ATC classes. The clustering results are shown in Table 4 and the overview of DrugComb compounds with the corresponding ATC classes is in Figure 5. In the control group, i.e. without the dimensionality reduction step, the best performing method for compound clustering is based on pretrained Infomax fingerprints, followed by three rule-based fingerprints. After either PCA- or PLS-based dimensionality reduction, LDA clustering performance of all rule-based fingerprints increases, while neural fingerprints, in particular GAE 64 and Infomax 300 decrease in performance. Such differences between graph and SMILES-based molecular representations are supported by the CKA analysis, indicating that GNN-based fingerprints are the least similar to other types.

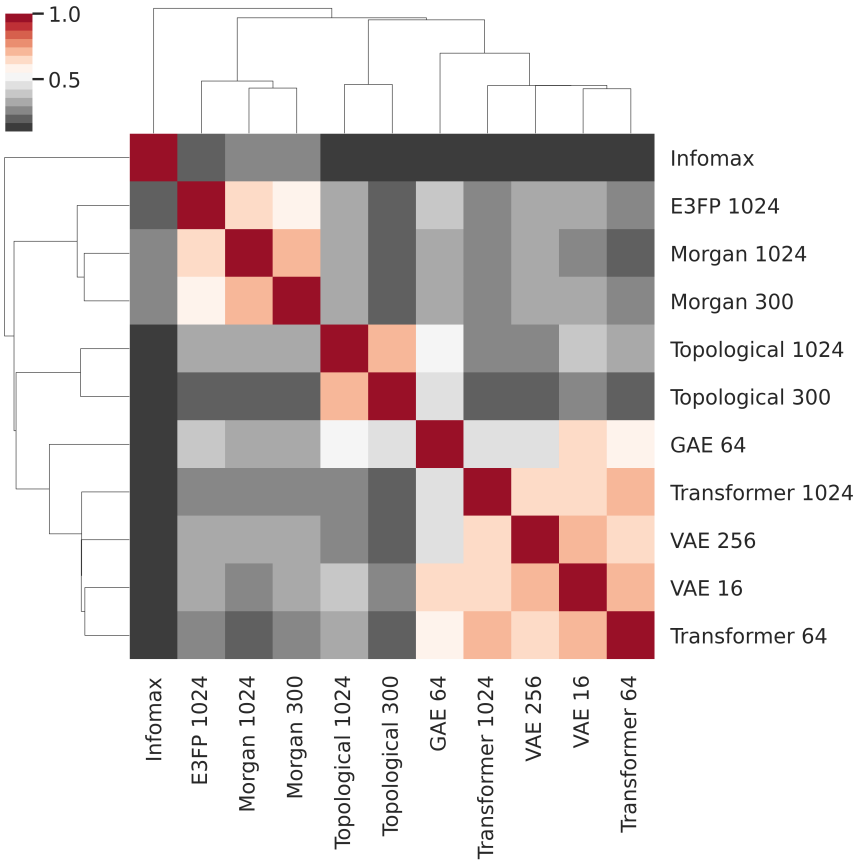


Figure 4: Heatmap with median linkage of CKA distances between 11 fingerprint types.

4 Conclusion

Choosing an optimal fingerprint type to represent molecular features is an important step in the drug discovery research. 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 and 3D circular E3FP fingerprints generated from SMILES strings, as well as Infomax 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, 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 neural variants. However, the gap between the best and worst performing fingerprints rarely exceeds PCC 0.05. Further, in one-vs-all compound clustering using ATC classes as labels rule-based fingerprints perform on par or better than the best neural representations. Lastly, we adapted Normalized Centered Kernel Alignment to quantify the extent of similarity between sets of fingerprints and identified 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.

We conclude that the quantitative performance differences between rule-based and neural fingerprints are likely to be insignificant in the context of preclinical studies of small molecular drugs [130,131]. In order to identify an optimal fingerprint type for a given project we advise enriching quantitative performance metrics with qualitative concerns, e.g. assessment of available chemometric and deep learning expertise, model interpretability requirements, opinions of project stakeholders and model performance on unseen data. Fingerprints based on the E3FP, Infomax, Morgan and VAE models are suggested as good starting points based on our experimental results and distinct methodologies

Table 4: Performance of nine fingerprint types in one-vs-all clustering of DrugComb compounds with ATC classes as labels. Average Silhouette and Variation Ratio Criterion (VRC) scores are in $[0, 1]$.

fingerprint	no dimensionality reduction		+ PCA		+ PLS	
	silhouette	VRC	silhouette	VRC	silhouette	VRC
Infomax 300	0.984	1.000	0.701	0.251	0.749	0.326
E3FP 1024	1.000	0.831	0.993	0.958	0.988	0.959
Topo 1024	0.979	0.771	1.000	0.997	1.000	1.000
Morgan 300	0.967	0.655	0.993	1.000	0.997	0.984
GAE 64	0.440	0.122	0.258	0.064	0.303	0.062
VAE 256	0.399	0.049	0.547	0.170	0.574	0.190
VAE 16	0.223	0.056	0.076	0.036	0.000	0.006
T 64	0.088	0.004	0.000	0.000	0.120	0.000
T 1024	0.000	0.000	0.131	0.023	0.377	0.067

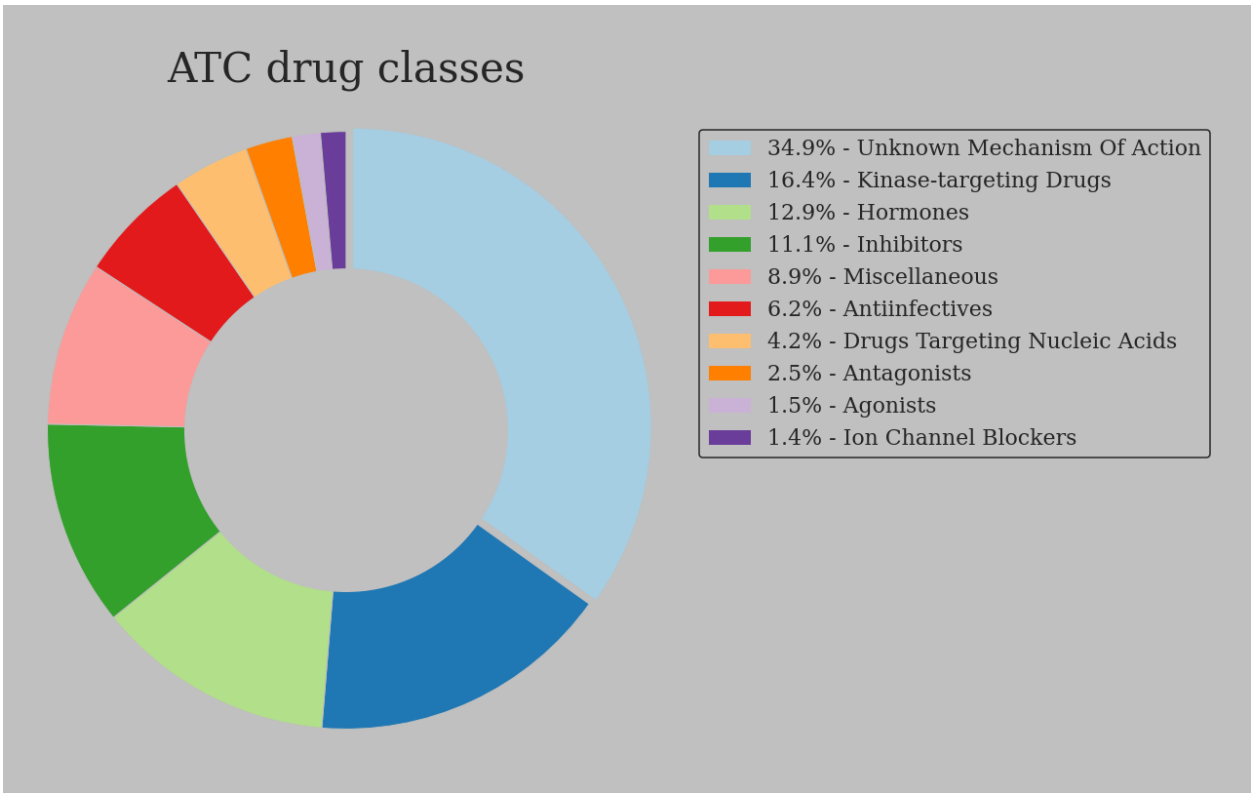


Figure 5: WHO ATC classes of the DrugComb compounds (n=3 421).

underlying their data generating methods [132]. We recommend Loewe synergy for use in drug combination screening due to its best performance among four tested synergy models based on data from 14 high throughput screening 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-based similarity, or based on models pre-trained on multimodal data and/or key biological databases, such as Gene Ontology, Protein Data Bank and UniRef [5,133–135]. Another line of inquiry could address high computational costs of data-driven and 3D rule-based models. To this end we suggest exploring alternative molecular representations and CPU-friendly generative models based on genetic algorithms, such as STONED on

SELFIES [136]. Finally, we hope that biomedical DL research will go beyond representation learning and focus on deriving novel insights by e.g. inferring synthetic and retrosynthetic chemical reactions, identifying novel disease-associated druggable proteins and clinically actionable biomarkers [137–139].

5 Key points

- To choose an optimal molecular fingerprint type it is advised to enrich quantitative metrics with qualitative concerns, such as the nature of downstream tasks, interpretability and robustness requirements and available chemometric expertise.
- Data-driven fingerprints, in particular based on Variational Autoencoders or Graph Neural Networks, as well as rule-based 2D and 3D circular topological fingerprints are flexible and well-performant molecular representation variants fit for any drug combination research project.
- Loewe formulation of non-interaction enables the highest correlation between in silico predictions and subsequent experimental validation of drug combination synergy in cancer cell lines.
- Centered Kernel Alignment is an effective measure of molecular representation similarity applicable to any rule-based and data-driven fingerprint types.

6 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 (direct download): `https://a3s.fi/swift/v1/AUTH_18b6635de4a24bdeb650a7c2bd90dadc/doses_oct2019`

7 Description of the authors

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.

8 Funding

This work was supported by the European Research Council (ERC) starting grant DrugComb (Informatics approaches for the rational selection of personalized cancer drug combinations) [No. 716 063] (to BZ and JT). BZ is also supported by the faculty funded PhD position in the Integrative Life Science Doctoral Programme, University of Helsinki. EP is supported by the Academy of Finland grant [No. 322 675].

9 Potential conflicts of interest

None declared

10 Supplementary Information

10.1 Regression model performance

13 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 1 024 bits long Morgan and 300 bits long Infomax fingerprints together with one-hot encoded cell line labels. 13 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 aforementioned methods, while being less interpretable and more difficult to set up [140,141].

Top four identified 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 gradient boosting on decision trees (GBDT) is selected for further experiments thanks to its efficient GPU utilization throughout training 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 is randomly permuted throughout tree growing process to limit unwanted target metric prediction shift, one of well-known GBDT disadvantages [114,142,143].

Tuned in 10-fold cross-validation optimal Catboost GBDT hyperparameters are Poisson bootstrap with 0.66 subsampling ratio, L^2 regularization of 9, tree depth of 10, learning rate of 0.15 with 5 000 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 [144].

10.2 Neural Network Training

For all DL models the hyperparameter search consisted of testing various activation functions (GELU, ELU, LeakyReLU, ReLU, SELU, Sigmoid, Softmax, Swish), dropout ratios (0.1 to 0.5 with 0.1 step size), initial learning rates ($1 \cdot 10^{-1}$ to $1 \cdot 10^{-5}$, with a step size of 0.1), number of patience epochs ($1 - 30$ with a step size of 1) and learning rate decay factor (0.9 to 0.1 with a step size of 0.1). Both Transformer models were tested with $3 - 6$ attention heads. VAE encoders were tested with up to 5 convolutional layers and convolutional kernel sizes up to 10; VAE decoder is tested with up to 3 recurrent GRU layers of sizes up to 600.

- 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 $2 \cdot 10^{-7}$ and $1 \cdot 10^{-8}$, respectively.
- Two Variational Autoencoder 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.
- A single Graph Autoencoder model is trained on 3 421 unique DrugComb 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 or a 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 it to have a significant influence on downstream experiments [145]. 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 [146,147]. Such GAE encoder architecture may be explained by a positive correlation between the performance of Neural Networks and a number of trainable parameters, as a seven layer GAE model with dot-product based Decoder has *circa* 10k learnable parameters, whereas Transformer and VAE models have two orders of magnitude more [148].

11 Supplementary Figures and Tables

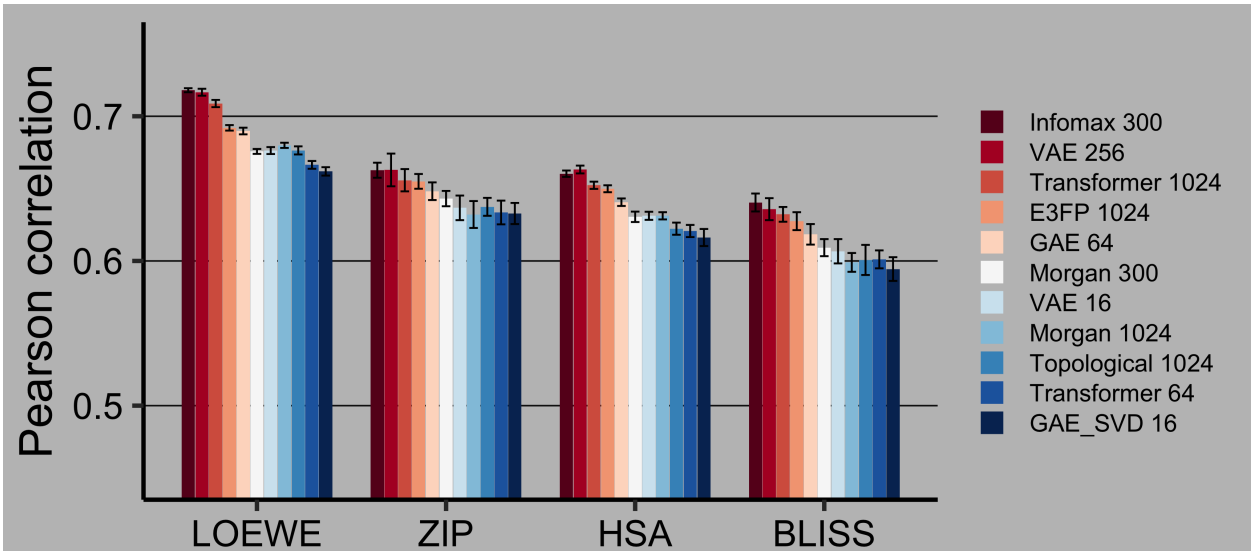


Figure S1: Drug combination synergy prediction on the CID-filtered dataset in 60 : 40 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with red.

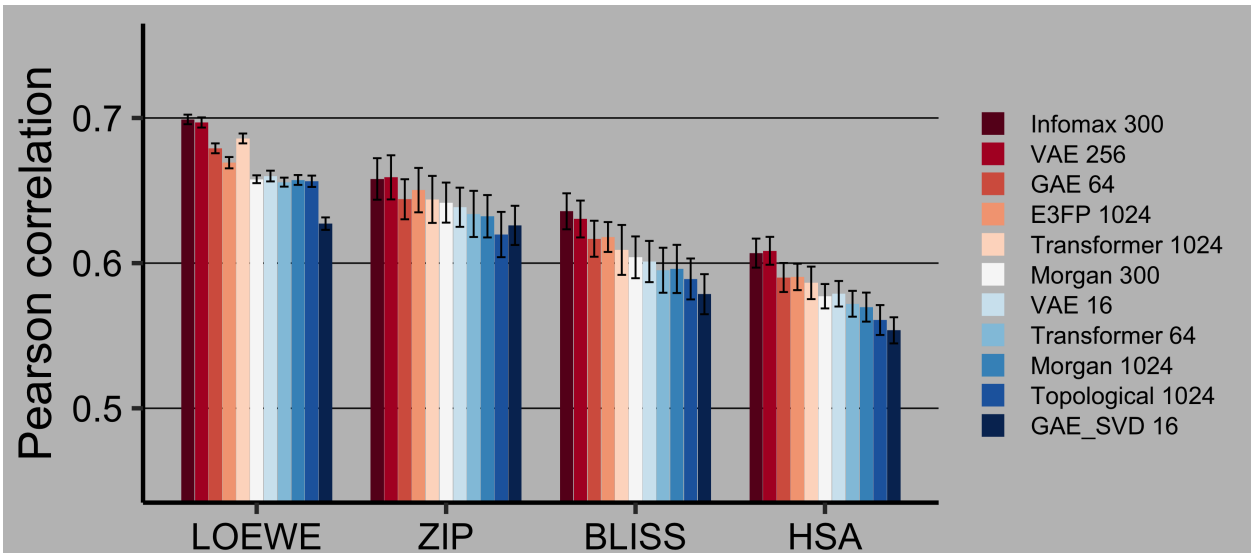


Figure S2: Drug combination synergy prediction on the SMILES-filtered dataset in 90 : 10 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with red.

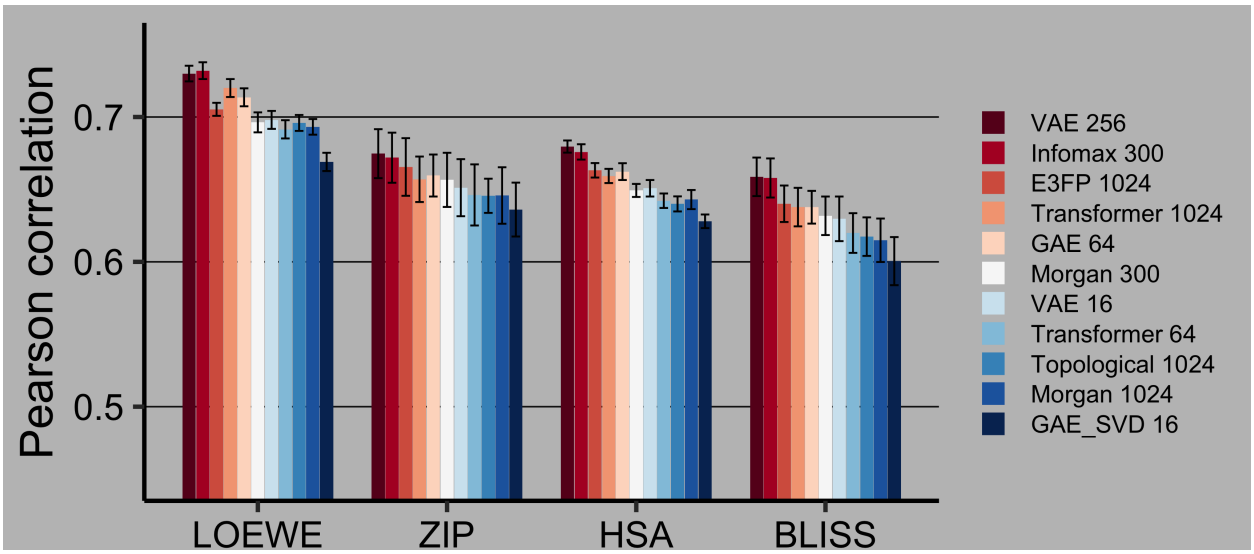


Figure S3: Drug combination synergy prediction on the CID-filtered dataset in 90 : 10 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with red.

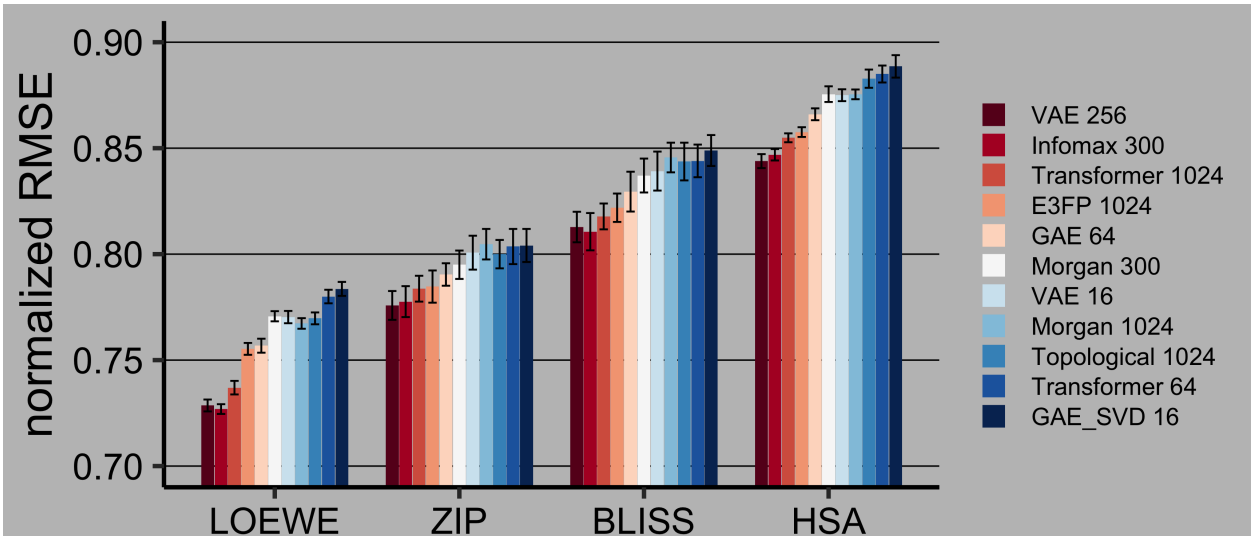


Figure S4: Drug combination synergy prediction on the CID-filtered dataset in 60 : 40 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Best models are highlighted with red.

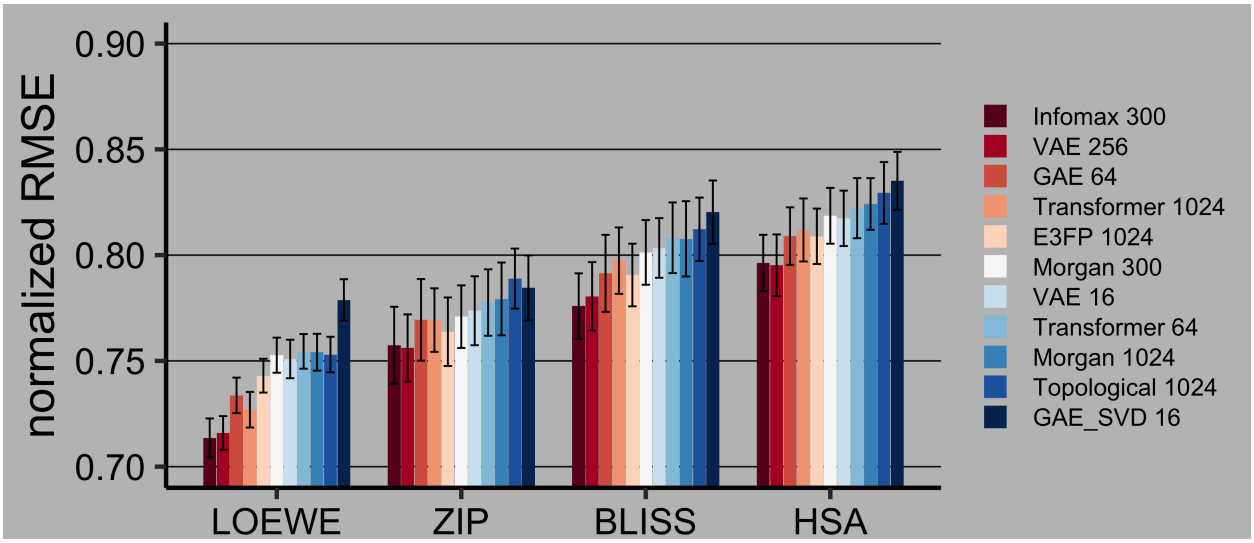


Figure S5: Drug combination synergy prediction on the SMILES-filtered dataset in 90 : 10 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Best models are highlighted with red.

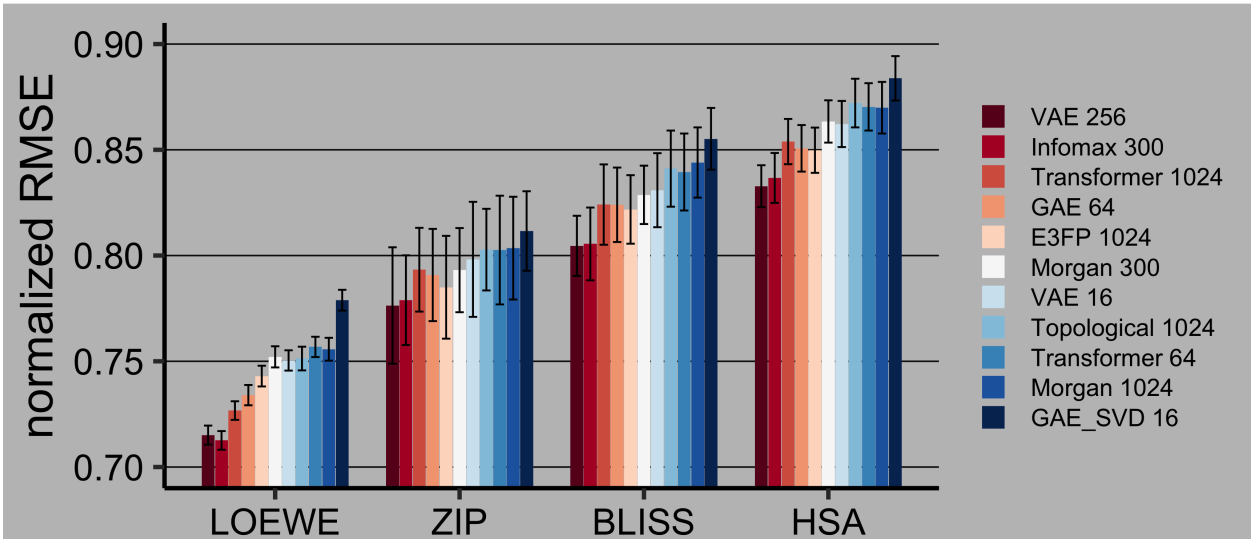


Figure S6: Drug combination synergy prediction on the CID-filtered dataset in 90 : 10 train:test split using RMSE, normalized by the target's standard deviation. 95% confidence intervals are calculated via empirical bootstrap. Best models are highlighted with red.

Table S1: Drug combination synergy prediction on the CID-filtered dataset in 60 : 40 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with bold.

fingerprint	**Pearson's *r* and 95% Confidence Interval**				
	CSS	**Bliss**	**HSA**	**Loewe**	**ZIP**
E3FP 1024	0.8819 ±0.0019	0.6275 ±0.0062	0.6499 ±0.0024	0.6920 ±0.0019	0.6550 ±0.0051
GAE 16	0.8744 ±0.0018	0.5944 ±0.0082	0.6162 ±0.0059	0.6619 ±0.0029	0.6328 ±0.0073
GAE 64	0.8800 ±0.0022	0.6184 ±0.0071	0.6406 ±0.0020	0.6898 ±0.0023	0.6482 ±0.0061
Infomax 300	**0.8907 ±0.0016**	**0.6404 ±0.0062**	**0.6603 ±0.0022**	**0.7180 ±0.0014**	**0.6627 ±0.0052**
Morgan 300	0.8771 ±0.0017	0.6092 ±0.0059	0.6305 ±0.0036	0.6757 ±0.0017	0.6431 ±0.0053
Morgan 1024	0.8712 ±0.0022	0.5990 ±0.0065	0.6312 ±0.0025	0.6799 ±0.0018	0.6321 ±0.0093
Topological 1024	0.8718 ±0.0014	0.6007 ±0.0104	0.6223 ±0.0042	0.6764 ±0.0028	0.6374 ±0.0062
Transformer 64	0.8716 ±0.0014	0.6011 ±0.0062	0.6207 ±0.0042	0.6664 ±0.0027	0.6335 ±0.0082
Transformer 1024	**0.8854 ±0.0016**	**0.6322 ±0.0051**	**0.6523 ±0.0025**	**0.7088 ±0.0025**	**0.6558 ±0.0077**
VAE 16	0.8767 ±0.0016	0.6067 ±0.0084	0.6312 ±0.0028	0.6763 ±0.0024	0.6367 ±0.0085
VAE 256	**0.8908 ±0.0023**	**0.6358 ±0.0076**	**0.6632 ±0.0027**	**0.7166 ±0.0024**	**0.6629 ±0.0113**

Table S2: Drug combination synergy prediction on the SMILES-filtered dataset in 90 : 10 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with bold.

fingerprint	**Pearson's *r* and 95% Confidence Interval**				
	CSS	**Bliss**	**HSA**	**Loewe**	**ZIP**
E3FP 1024	0.8732 ±0.0058	**0.6180 ±0.0103**	**0.5905 ±0.0091**	0.6692 ±0.0039	**0.6503 ±0.0153**
GAE 16	0.8641 ±0.0063	0.5786 ±0.0138	0.5537 ±0.0090	0.6272 ±0.0043	0.6260 ±0.0135
GAE 64	**0.8754 ±0.0064**	0.6168 ±0.0124	0.5901 ±0.0100	0.6791 ±0.0034	**0.6440 ±0.0138**
Infomax 300	**0.8826 ±0.0060**	**0.6357 ±0.0124**	**0.6069 ±0.0100**	**0.6990 ±0.0033**	**0.6580 ±0.0143**
Morgan 300	0.8702 ±0.0054	0.6040 ±0.0144	0.5772 ±0.0084	0.6578 ±0.0027	0.6417 ±0.0138
Morgan 1024	0.8640 ±0.0052	0.5960 ±0.0166	0.5697 ±0.0100	0.6573 ±0.0034	0.6323 ±0.0146
Topological 1024	0.8500 ±0.0044	0.5891 ±0.0141	0.5608 ±0.0103	0.6564 ±0.0039	0.6197 ±0.0156
Transformer 64	0.8682 ±0.0047	0.5951 ±0.0155	0.5720 ±0.0089	0.6558 ±0.0031	0.6339 ±0.0159
Transformer 1024	0.8742 ±0.0055	0.6091 ±0.0172	0.5864 ±0.0112	**0.6859 ±0.0034**	0.6439 ±0.0162
VAE 16	0.8712 ±0.0062	0.6011 ±0.0142	0.5789 ±0.0088	0.6600 ±0.0037	0.6385 ±0.0135
VAE 256	**0.8829 ±0.0063**	**0.6304 ±0.0127**	**0.6085 ±0.0096**	**0.6969 ±0.0035**	**0.6591 ±0.0152**

Table S3: Drug combination synergy prediction on the CID-filtered dataset in 90 : 10 train:test split. 95% confidence intervals are calculated via Fisher z-transformation. Best models are highlighted with bold.

fingerprint	**Pearson's *r* and 95% Confidence Interval**				
	CSS	**Bliss**	**HSA**	**Loewe**	**ZIP**
E3FP 1024	0.8849 ±0.0049	**0.6401 ±0.0126**	**0.6632 ±0.0050**	0.7053 ±0.0045	**0.6655 ±0.0199**
GAE 16	0.8761 ±0.0041	0.6005 ±0.0166	0.6280 ±0.0047	0.6690 ±0.0063	0.6361 ±0.0186
GAE 64	**0.8869 ±0.0050**	0.6377 ±0.0113	0.6623 ±0.0058	0.7136 ±0.0062	0.6596 ±0.0145
Infomax 300	**0.8942 ±0.0056**	**0.6579 ±0.0135**	**0.6759 ±0.0053**	**0.7320 ±0.0058**	**0.6719 ±0.0173**
Morgan 300	0.8824 ±0.0045	0.6318 ±0.0133	0.6493 ±0.0045	0.6963 ±0.0069	0.6566 ±0.0187
Morgan 1024	0.8748 ±0.0053	0.6149 ±0.0150	0.6430 ±0.0066	0.6932 ±0.0054	0.6458 ±0.0195
TopoA 1024	0.8647 ±0.0049	0.6174 ±0.0133	0.6400 ±0.0052	0.6959 ±0.0055	0.6456 ±0.0118
TB 64	0.8798 ±0.0059	0.6199 ±0.0137	0.6422 ±0.0050	0.6915 ±0.0063	0.6462 ±0.0211
TB 1024	**0.8853 ±0.0047**	0.6378 ±0.0133	0.6593 ±0.0049	**0.7200 ±0.0062**	0.6570 ±0.0157
VAE 16	0.8838 ±0.0053	0.6297 ±0.0154	0.6508 ±0.0057	0.6980 ±0.0062	0.6512 ±0.0197
VAE 256	0.8949 ±0.0052	**0.6587 ±0.0133**	**0.6796 ±0.0042**	**0.7300 ±0.0054**	**0.6747 ±0.0169**

Table S4: Drug combination synergy prediction on the CID-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 bold

fingerprint	**normalized Root Mean Squared Error and 95% Confidence Interval**				
	CSS	**Bliss**	**HSA**	**Loewe**	**ZIP**
E3FP 1024	0.4925 ±0.0014	0.8219 ±0.0067	0.8576 ±0.0023	0.7553 ±0.0028	0.7847 ±0.0076
GAE 16	0.4962 ±0.0014	0.8295 ±0.0094	0.8660 ±0.0028	0.7568 ±0.0033	0.7904 ±0.0053
GAE 64	0.5070 ±0.0015	0.8489 ±0.0073	0.8489 ±0.0073	0.7836 ±0.0033	0.8041 ±0.0078
Infomax 300	**0.4748 ±0.0011**	**0.8106 ±0.0088**	**0.8469 ±0.0027**	**0.7269 ±0.0023**	**0.7776 ±0.0073**
Morgan 300	0.5018 ±0.0013	0.8371 ±0.0080	0.8755 ±0.0037	0.7707 ±0.0024	0.7950 ±0.0067
Morgan 1024	0.5130 ±0.0016	0.8456 ±0.0070	0.8754 ±0.0023	0.7673 ±0.0025	0.8047 ±0.0072
Topological 1024	0.5118 ±0.0012	0.8437 ±0.0089	0.8828 ±0.0043	0.7697 ±0.0028	0.8000 ±0.0067
Transformer 64	0.5124 ±0.0014	0.8440 ±0.0077	0.8850 ±0.0040	0.7800 ±0.0032	0.8036 ±0.0083
Transformer 1024	**0.4855 ±0.0012**	**0.8178 ±0.0061**	**0.8549 ±0.0021**	**0.7370 ±0.0032**	**0.7837 ±0.0061**
VAE 16	0.5027 ±0.0013	0.8392 ±0.0092	0.8750 ±0.0028	0.7703 ±0.0029	0.8007 ±0.0080
VAE 256	**0.4744 ±0.0019**	**0.8128 ±0.0072**	**0.8439 ±0.0033**	**0.7286 ±0.0028**	**0.7758 ±0.0068**

Table S5: Drug combination synergy prediction on the SMILES-filtered dataset in 90 : 10 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 bold

fingerprint	normalized Root Mean Squared Error and 95% Confidence Interval				
	CSS	Bliss	HSA	Loewe	ZIP
E3FP 1024	0.4881 ±0.0033	**0.7906 ±0.0148**	**0.8089 ±0.0131**	0.7430 ±0.0080	**0.7638 ±0.0162**
GAE 16	0.5041 ±0.0041	0.8203 ±0.0150	0.8351 ±0.0137	0.7788 ±0.0098	0.7845 ±0.0154
GAE 64	**0.4842 ±0.0037**	0.7914 ±0.0182	0.8090 ±0.0136	0.7337 ±0.0084	0.7694 ±0.0193
Infomax 300	**0.4708 ±0.0038**	**0.7759 ±0.0155**	**0.7963 ±0.0133**	**0.7136 ±0.0092**	**0.7574 ±0.0182**
Morgan 300	0.4935 ±0.0033	0.8013 ±0.0153	0.8186 ±0.0132	0.7527 ±0.0083	0.7709 ±0.0148
Morgan 1024	0.5045 ±0.0039	0.8077 ±0.0178	0.8242 ±0.0122	0.7541 ±0.0087	0.7793 ±0.0172
Topological 1024	0.5276 ±0.0033	0.8122 ±0.0150	0.8294 ±0.0146	0.7530 ±0.0084	0.7889 ±0.0142
Transformer 64	0.4972 ±0.0034	0.8082 ±0.0167	0.8222 ±0.0142	0.7545 ±0.0082	0.7776 ±0.0157
Transformer 1024	0.4864 ±0.0033	0.7974 ±0.0157	0.8119 ± 0.0149	**0.7269 ±0.0084**	0.7693 ±0.0150
VAE 16	0.4919 ±0.0036	0.8034 ±0.0141	0.8174 ±0.0131	0.7509 ±0.0091	0.7737 ±0.0163
VAE 256	**0.4701 ±0.0037**	**0.7805 ±0.0162**	**0.7952 ±0.0146**	**0.7159 ±0.0080**	**0.7561 ±0.0159**

Table S6: Drug combination synergy prediction on the CID-filtered dataset in 90 : 10 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 bold

fingerprint	normalized Root Mean Squared Error and 95% Confidence Interval				
	CSS	Bliss	HSA	Loewe	ZIP
E3FP 1024	0.4871 ±0.0033	**0.8218 ±0.0162**	**0.8498 ±0.0107**	0.7430 ±0.0049	**0.7850 ±0.0243**
GAE 16	0.5043 ±0.0027	0.8552 ±0.0146	0.8838 ±0.0105	0.7789 ±0.0049	0.8116 ±0.0188
GAE 64	**0.4832 ±0.0030**	0.8240 ±0.0176	0.8507 ±0.0110	0.7340 ±0.0048	0.7908 ±0.0218
Infomax 300	**0.4683 ±0.0032**	**0.8055 ±0.0172**	**0.8367 ±0.0118**	**0.7126 ±0.0044**	**0.7789 ±0.0212**
Morgan 300	0.4920 ±0.0028	0.8287 ±0.0138	0.8634 ±0.0100	0.7521 ±0.0050	0.7931 ±0.0199
Morgan 1024	0.5068 ±0.0038	0.8440 ±0.0166	0.8699 ±0.0122	0.7557 ±0.0054	0.8035 ±0.0243
Topological 1024	0.5254 ±0.0033	0.8411 ±0.0180	0.8721 ±0.0115	0.7513 ±0.0056	0.8028 ±0.0193
Transformer 64	0.4972 ±0.0036	0.8395 ±0.0182	0.8703 ±0.0112	0.7568 ±0.0048	0.8026 ±0.0257
Transformer 1024	0.4863 ±0.0029	0.8241 ±0.0190	0.8539 ±0.0107	**0.7267 ±0.0044**	0.7933 ±0.0198
VAE 16	0.4895 ±0.0030	0.8309 ±0.0175	0.8622 ±0.0109	0.7504 ±0.0048	0.7982 ±0.0272
VAE 256	**0.4665 ±0.0032**	**0.8046 ±0.0142**	**0.8328 ±0.0099**	**0.7151 ±0.0045**	**0.7764 ±0.0275**

12 References

1. Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nat. Biotechnol. 2018; 36:829–838.

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; 15:141.

3. Lo Y-C, Rensi SE, Torng W, et al. Machine learning in chemoinformatics and drug discovery. Drug Discovery Today 2018; 23:1538–1546.

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 2019; 17:184–192.
6. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577:706–710.

7. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. 2019; 37:1038–1040.

8. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci. Rep. 2018; 8:11242.

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, Patrick Walters W, Cornelis E C, et al. Opportunities and challenges using artificial intelligence in ADME/Tox. Nature Materials 2019; 18:418–422.

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; 9:5441–5451.

13. Campbell DT. Assessing the impact of planned social change. Evaluation and Program Planning 1979; 2:67–90.

14. Goodhart CAE. Problems of Monetary Management: The UK Experience. Monetary Theory and Practice 1984; 91–121.

15. Heath I, Hippisley-Cox J, Smeeth L. Measuring performance and missing the point? BMJ 2007; 335:1075–1076.

16. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern. Med. 2018; 178:1544–1547.

17. Ha D, Schmidhuber J. World Models. arXiv:1803.10122 [cs.LG] 2018; 1-21.

18. Hirschfeld L, Swanson K, Yang K, et al. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. J. Chem. Inf. Model. 2020; 60:3770–3780.

19. Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug–target interaction predictions. Brief. Bioinform. 2014; 16:325–337.

20. Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chem. Sci. 2019; 10:8154–8163.

21. 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.

22. 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.

23. Tang J. Informatics Approaches for Predicting, Understanding, and Testing Cancer Drug Combinations. Methods Mol. Biol. 2017; 1636:485–506.

24. 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; 3:1416–1429.

25. 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; 77:3564–3576.

26. Borisy AA, Elliott PJ, Hurst NW, et al. Systematic discovery of multicomponent therapeutics. Proc. Natl. Acad. Sci. U. S. A. 2003; 100:7977–7982.

27. Tallarida RJ. Quantitative methods for assessing drug synergism. Genes Cancer 2011; 2:1003–1008.

28. Bliss CI. THE TOXICITY OF POISONS APPLIED JOINTLY1. Annals of Applied Biology 1939; 26:585–615.

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

30. Greco WR, Bravo G, Parsons JC. The search for synergy: a critical review from a response surface perspective. Pharmacol. Rev. 1995; 47:331–385.

31. Loewe S. The problem of synergism and antagonism of combined drugs. Arzneimittelforschung 1953; 3:285–290.

32. Yadav B, Wennerberg K, Aittokallio T, et al. Searching for Drug Synergy in Complex Dose–Response Landscapes Using an Interaction Potency Model. Computational and Structural Biotechnology Journal 2015; 13:504–513.

33. 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(5):e1006752.

34. Maggiora G, Vogt M, Stumpfe D, et al. Molecular similarity in medicinal chemistry. J. Med. Chem. 2014; 57:3186–3204.

35. 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.

36. Neves BJ, Braga RC, Melo-Filho CC, et al. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Front. Pharmacol. 2018; 9:1275.

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

38. Chuang KV, Gunsalus LM, Keiser MJ. Learning Molecular R epresentations for Medicinal Chemistry. J. Med. Chem. 2020; 63:8705–8722.

39. Noel M. O'Boyle RAS. Comparing structural fingerprints using a literature-based similarity benchmark. J. Cheminform. 2016; 8:36.

40. Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci. 2002; 42:1273–1280.

41. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. Journal of Chemical Documentation 1965; 5:107–113.

42. Rogers D, Hahn M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 2010; 50:742–754

43. 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; 20:1878–1912.

44. Encoder-Decoder Architecture — Dive into Deep Learning 0.14.4 documentation. `https://d2l.ai/chapter_recurrent-modern/encoder-decoder.html` (5 April 2021, date last accessed).

45. 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. arXiv:1706.06689 [stat.ML] 2017; 1-38.

46. Goh GB, Hodas NO, Siegel C, et al. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. arXiv:1712.02034 [stat.ML] 2017; 1-8.

47. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci 2018; 4:268–276.

48. Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078 [cs.CL] 2014; 1-15.

49. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML] 2013; 1-14.

50. Honda S, Shi S, Ueda HR. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. arXiv:1911.04738 [cs.LG] 2019; 1-9.

51. Oskooei A, Born J, Manica M, et al. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks. arXiv:1811.06802 [cs.LG] 2018; 1-10.

52. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv:1706.03762 [cs.CL] 2017; 1-15.

53. Wang Z, Liu M, Luo Y, et al. Advanced Graph and Sequence Neural Networks for Molecular Property Prediction and Drug Discovery. arXiv:2012.01981 [q-bio.QM] 2020; 1-26.

54. Winter R, Montanari F, Noé F, et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem. Sci. 2019; 10:1692–1701.

55. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Modeling 1988; 28:31–36.

56. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html|` (5 April 2021, date last accessed).

57. O'Boyle NM. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J. Cheminform. 2012; 4:22.

58. Kipf TN, Welling M. Variational Graph Auto-Encoders. arXiv:1611.07308 [stat.ML] 2016; 1-3.

59. Gilmer J, Schoenholz SS, Riley PF, et al. Neural Message Passing for Quantum Chemistry. arXiv:1704.01212 [cs.LG] 2017; 1-14.

60. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. arXiv:1509.09292 [cs.LG] 2015; 1-9.

61. Ramsundar B, Kearnes S, Riley P, et al. Massively Multitask Networks for Drug Discovery. arXiv:1502.02072 [stat.ML] 2015; 1-27.

62. Yang K, Swanson K, Jin W, et al. Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model. 2019; 59:3370–3388.

63. Chami I, Abu-El-Haija S, Perozzi B, et al. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. arXiv:2005.03675 [cs.LG] 2020; 1-44.

64. Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des. 2016 Aug;30(8):595-608.

65. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. 2018; 9:513–530.

66. Hu W, Fey M, Zitnik M, et al. Open Graph Benchmark: Datasets for Machine Learning on Graphs. arXiv:2005.00687 [cs.LG] 2020; 1-34.

67. Dwivedi VP, Joshi CK, Laurent T, et al. Benchmarking Graph Neural Networks. arXiv:2003.00982 [cs.LG] 2020; 1-30.

68. Meyer CT, Wooten DJ, Lopez CF, et al. Charting the Fragmented Landscape of Drug Synergy. Trends Pharmacol. Sci. 2020; 41:266–280.

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

70. PubChem. WHO ATC Code - PubChem Data Source. https://www.whocc.no/atc/structure_and_principles/ (5 April 2021, date last accessed).

71. Fisher RA. The use of multiple measurements in taxonomic problems. Ann. Eugen. 1936; 7:179–188.

72. Cortes C, Mohri M, Rostamizadeh A. Algorithms for Learning Kernels Based on Centered Alignment. J. Mach. Learn. Res. 2012; 13:795-828.

73. Wagstaff K. Machine Learning that Matters. arXiv:1206.4656 [cs.LG] 2012; 1-6.

74. Zagidullin B, Aldahdooh J, Zheng S, et al. DrugComb: an integrative cancer drug combination data portal. Nucleic Acids Res. 2019; 47:W43–W51.

75. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research 2012; 40:D1100–D1107.

76. Bento, A.P., Hersey, A., Félix, E. et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020; 12:51.

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

78. Axen SD, Huang X-P, Cáceres EL, et al. A Simple Representation of Three-Dimensional Molecular Structure. J. Med. Chem. 2017; 60:7393–7409.

79. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs.LG] 2016; 1-14.

80. Hu W, Liu B, Gomes J, et al. Strategies for Pre-training Graph Neural Networks. arXiv:1905.12265 [cs.LG] 2019; 1-22.

81. Veličković P, Fedus W, Hamilton WL, et al. Deep Graph Infomax. arXiv:1809.10341 [stat.ML] 2018; 1-17.

82. Ballard DH. Modular Learning in Neural Networks. Association for the Advancement of Artificial Intelligence. AAAI-97 Proceedings. 1987; 1-6.

83. Sterling T, Irwin JJ. ZINC 15–Ligand Discovery for Everyone. J. Chem. Inf. Model. 2015; 55:

84. Fan K. Maximum Properties and Inequalities for the Eigenvalues of Completely Continuous Operators. Proc. Natl. Acad. Sci. U. S. A. 1951; 37:760.

85. Schlichtkrull M, Kipf TN, Bloem P, et al. Modeling Relational Data with Graph Convolutional Networks. The Semantic Web 2018; 593–607.

86. Hamilton WL, Ying R, Leskovec J. Representation Learning on Graphs: Methods and Applications. arXiv:1709.05584 [cs.SI] 2017; 1-24.

87. von Luxburg U. A Tutorial on Spectral Clustering. arXiv:0711.0189 [cs.DS] 2007; 1-32.

88. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics 2011; 315–323.

89. Klambauer G, Unterthiner T, Mayr A, et al. Self-Normalizing Neural Networks. arXiv:1706.02515 [cs.LG] 2017; 1-102.

90. 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; 249–256.

91. Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks. arXiv:1710.10903 [stat.ML] 2017; 1-12.

92. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] 2014; 1-15.

93. Wang M, Zheng D, Ye Z, et al. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. arXiv:1909.01315 [cs.LG] 2019; 1-18.

94. Akiba T, Sano S, Yanase T, et al. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs.LG] 2019; 1-10.

95. Mason DJ, Stott I, Ashenden S, et al. Prediction of Antibiotic Interactions Using Descriptors Derived from Molecular Structure. J. Med. Chem. 2017; 60:3902–3912.

96. Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. J. Mach. Learn. Res. 2004; 5:1089–1105.

97. Efron B. Better Bootstrap Confidence Intervals. J. Am. Stat. Assoc. 1984; 82(397):171-185

98. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 1915; 10:507.

99. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CHAPMAN HALL/CRC, 1993.

100. Corey DM, Dunlap WP, Burke MJ. Averaging Correlations: Expected Values and Bias in Combined Pearsonrs and Fisher'szTransformations. The Journal of General Psychology 1998; 125:245–261.

101. Bishara AJ, Hittner JB. Confidence intervals for correlations when data are not normal. Behavior Research Methods 2017; 49:294–309.

102. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965; 52:591–611.

103. Gretton A, Bousquet O, Smola A, et al. Measuring Statistical Dependence with Hilbert-Schmidt Norms. Lecture Notes in Computer Science 2005; 63–77.

104. Kornblith S, Norouzi M, Lee H, et al. Similarity of Neural Network Representations Revisited. arXiv:1905.00414 [cs.LG] 2019; 1-20.

105. Song L, Smola A, Gretton A, et al. Supervised Feature Selection via Dependence Estimation. arXiv:0704.2668 [cs.LG] 2007; 1-9.

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

107. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun. Stat. Theory Methods 1974; 3:1–27.

108. Everitt BS, Department of Biostatistics and Computing Brian S Everitt, Dunn G. Applied Multivariate Data Analysis. Wiley, 2001.

109. OpenStack Docs: Newton. `https://docs.openstack.org/newton/` (5 April 2021, date last accessed).

110. 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. Nature Communications 2019; 10(1):2674.

111. Preuer K, Lewis RPI, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics 2018; 34:1538–1546.

112. O'Neil J, Benita Y, Feldman I, et al. An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. Mol. Cancer Ther. 2016; 15:1155–1162.

113. Sidorov P, Naulaerts S, Ariey-Bonnet J, et al. Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. Frontiers in Chemistry 2019; 7: 509.

114. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. arXiv:1706.09516 [cs.LG] 2017; 1-23.

115. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 1996; 39:2887–2893.

116. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learn Syst 2020; PP.

117. 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; 49:108–119.

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

119. 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; 52:2884–2901.

120. Todeschini R, Ballabio D, Consonni V. Distances and Similarity Measures in Chemometrics and Chemoinformatics. Encyclopedia of Analytical Chemistry 2020; 1–40.

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

122. Szedmak S, Bach E. On the generalization of Tanimoto-type kernels to real valued functions. arXiv:2007.05943 [cs.LG] 2020; 1-10.

123. Schölkopf B. The Kernel Trick for Distances. Adv. Neural Inf. Process. Syst. 2000; 13:1-7.

124. Raghu M, Gilmer J, Yosinski J, et al. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. arXiv:1706.05806 [stat.ML] 2017; 1-17.

125. Morcos AS, Raghu M, Bengio S. Insights on representational similarity in neural networks with canonical correlation. arXiv:1806.05759 [stat.ML] 2018; 1-17.

126. Gretton A, Herbrich R, Smola A, et al. Kernel Methods for Measuring Independence. J. Mach. Learn. Res. 2005; 6:2075–2129.

127. Thompson J, Bengio Y, Schoenwiesner M. The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis. 2019 Conference on Cognitive Computational Neuroscience 2019.

128. Robert P, Escoufier Y. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. Applied Statistics 1976; 25:257.

129. Julie Josse SH. Measuring multivariate association and beyond. Stat. Surv. 2016; 10:132.

130. Cortés-Ciriano I, Bender A. Reliable Prediction Errors for Deep Neural Networks Using Test-Time Dropout. J. Chem. Inf. Model. 2019; 59:3330–3339.

131. 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; S1359-6446(21)00042-8.

132. Gao K, Nguyen DD, Sresht V, et al. Are 2D fingerprints still valuable for drug discovery? Phys. Chem. Chem. Phys. 2020; 22:8373–8390.

133. 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.

134. 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:2007.06225 [cs.LG], 2020; 1-21.

135. Güvenç Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. Brief. Bioinform. 2021; 22:346–359.

136. Nigam A, Pollice R, Krenn M, et al. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules using SELFIES. https://doi.org/10.26434/chemrxiv.13383266.v2 2021; 1-17.

137. 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.

138. Jin W, Barzilay R, Jaakkola T. Discovering Synergistic Drug Combinations for COVID with Biological Bottleneck Models. arXiv:2011.04651 [q-bio.BM] 2020; 1-8.

139. 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.

140. Yang Y, Morillo IG, Hospedales TM. Deep Neural Decision Trees. arXiv:1806.06988 [cs.LG] 2018; 1-7.

141. Abutbul A, Elidan G, Katzir L, et al. DNF-Net: A Neural Architecture for Tabular Data. arXiv:2006.06465 [cs.LG] 2020; 1-17.

142. Breiman L. Using Iterated Bagging to Debias Regressions. Mach. Learn. 2001; 45:261–277.

143. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 2020; 1–31.

144. Olson RS, La Cava W, Mustahsan Z, et al. Data-driven advice for applying machine learning to bioinformatics problems. Pac. Symp. Biocomput. 2018; 23:192.

145. Smith LN. Cyclical Learning Rates for Training Neural Networks. arXiv:1506.01186 [cs.CV] 2015; 1-10.

146. Brockschmidt M. GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation. arXiv:1906.12192 [cs.LG] 2019; 1-11.

147. Alon U, Yahav E. On the Bottleneck of Graph Neural Networks and its Practical Implications. arXiv:2006.05205 [cs.LG] 2020; 1-16.

148. Kaplan J, McCandlish S, Henighan T, et al. Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs.LG] 2020; 1-19.