UNIVERSITY OF CAMBRIDGE

Machine learning in Drug Discovery: Application to Cell Painting data

Srijit Seal

20 October, 2022

ss2686@cam.ac.uk

Yusuf Hamied Department of Chemistry, University of Cambridge

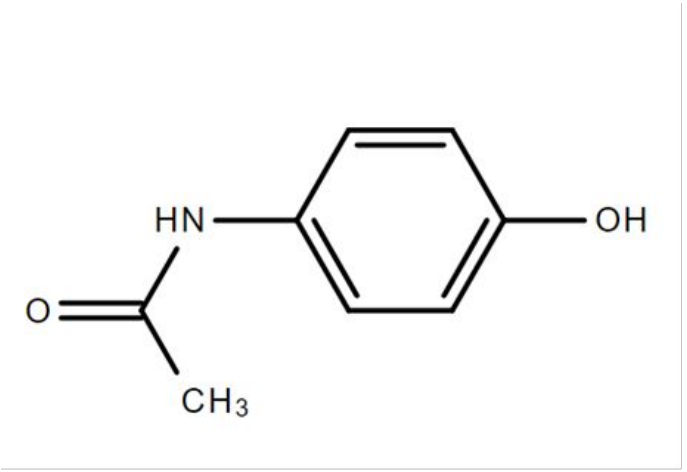
ss2686@cam.ac.uk

@srijitseal

Outline

1. Introduction: Machine Learning and Drug Discovery and Cell Painting Assay
2. Detection of Mitochondrial Toxicity
3. Combining Predictions from Cell Morphology Data and Chemical Structure
4. Interpreting Cell Morphology using Convolutional Neural Networks for Compound Toxicity
5. Conclusions

Data driven discovery



Paracetamol

Formula C8H9NO2

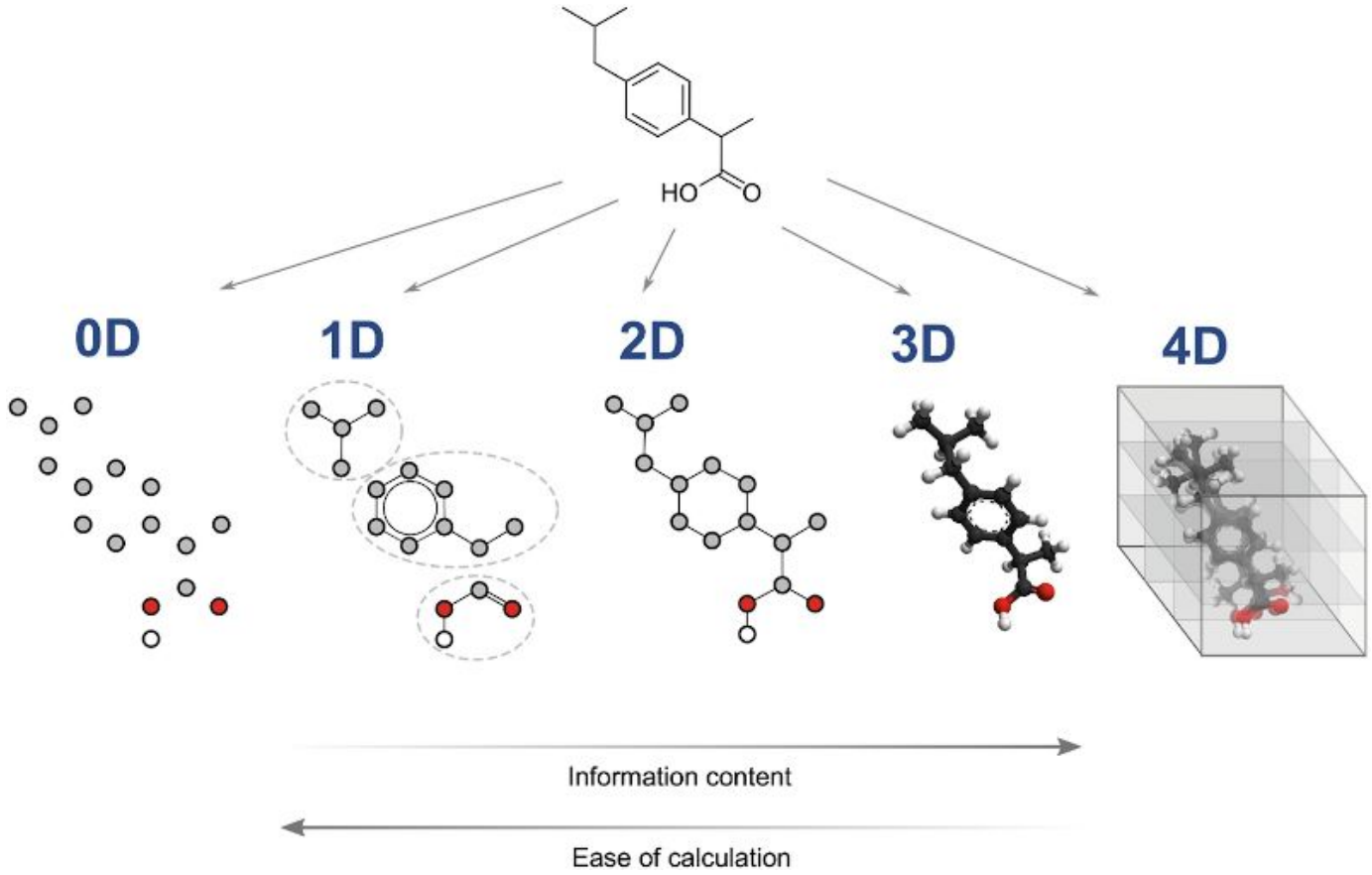
Identifiers

IUPAC name: N-(4-hydroxyphenyl)acetamide

SMILES: C1=CC(O)=CC=C1NC(=O)C

chemical structure, molecular formula, SMILES identifier of the common anti-inflammatory drug paracetamol

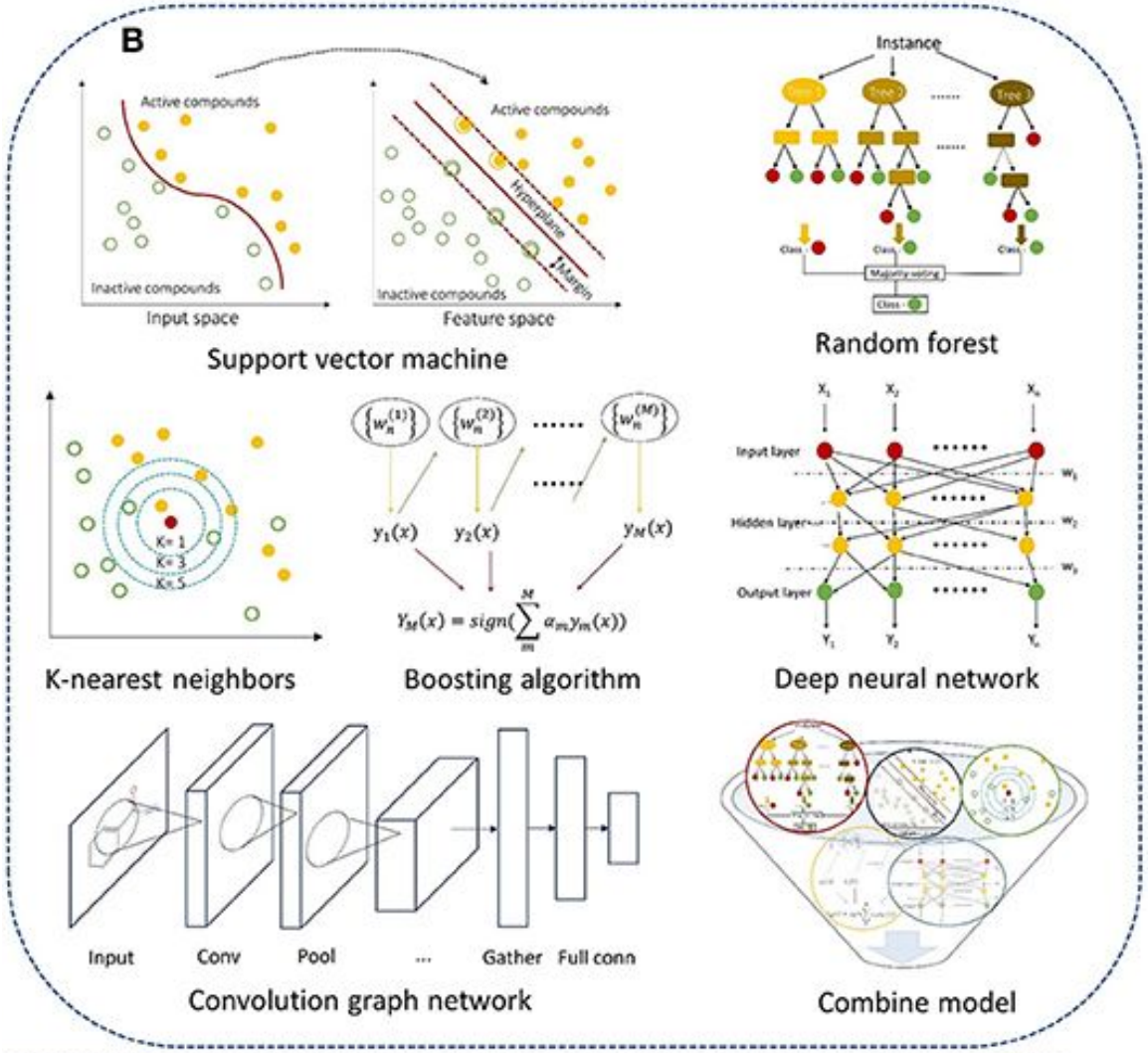
Compound Descriptors

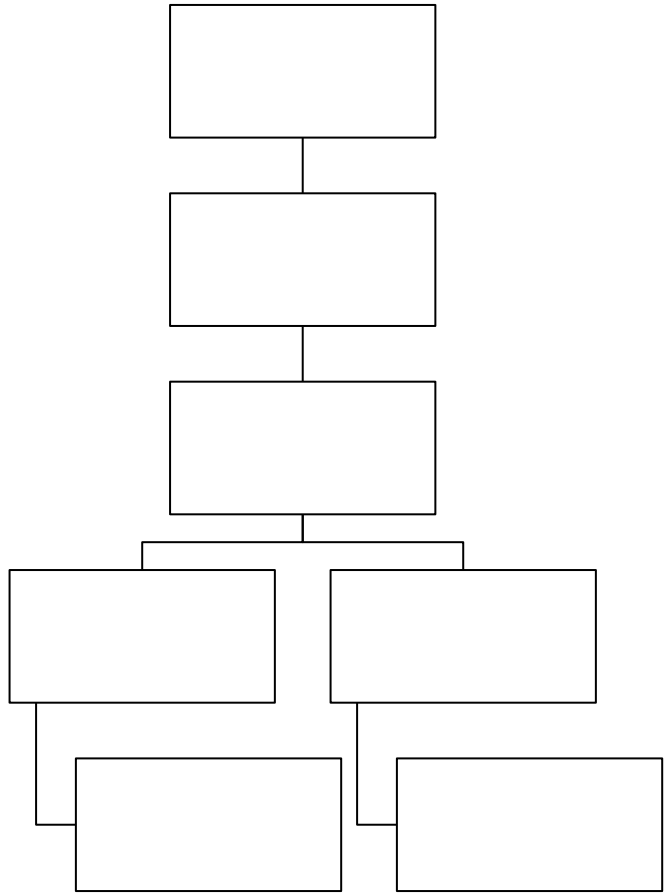


Representation of the five classes of theoretical descriptors and the relationship between their dimensionality, the information they provide and the ease of calculation.

Grisoni, F., Ballabio, D., Todeschini, R. & Consonni, V. Molecular descriptors for structure–activity applications: A hands-on approach. Methods Mol. Biol. 1800, 3–53 (2018)

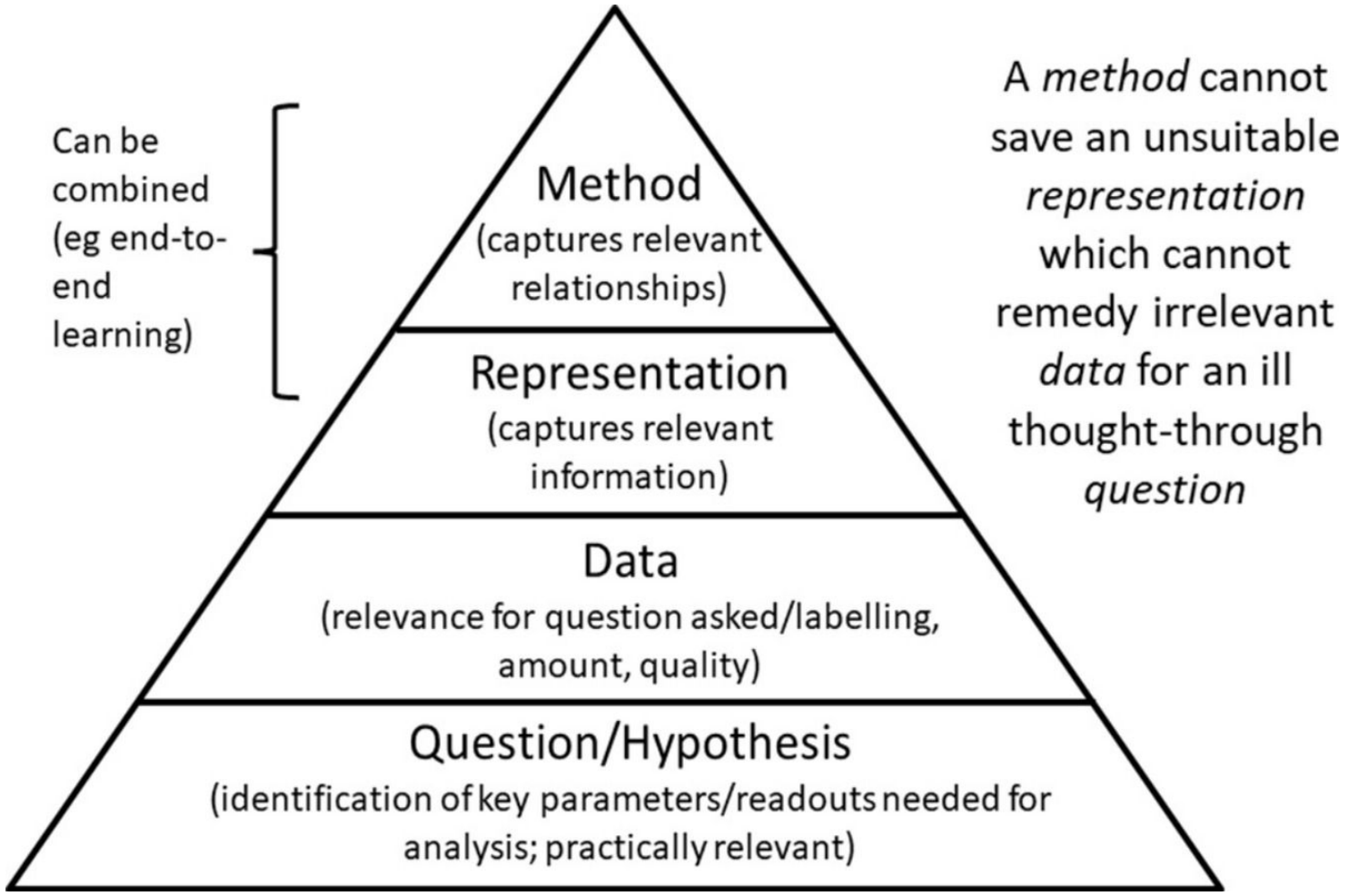
Types of Machine Learning models and standard workflow





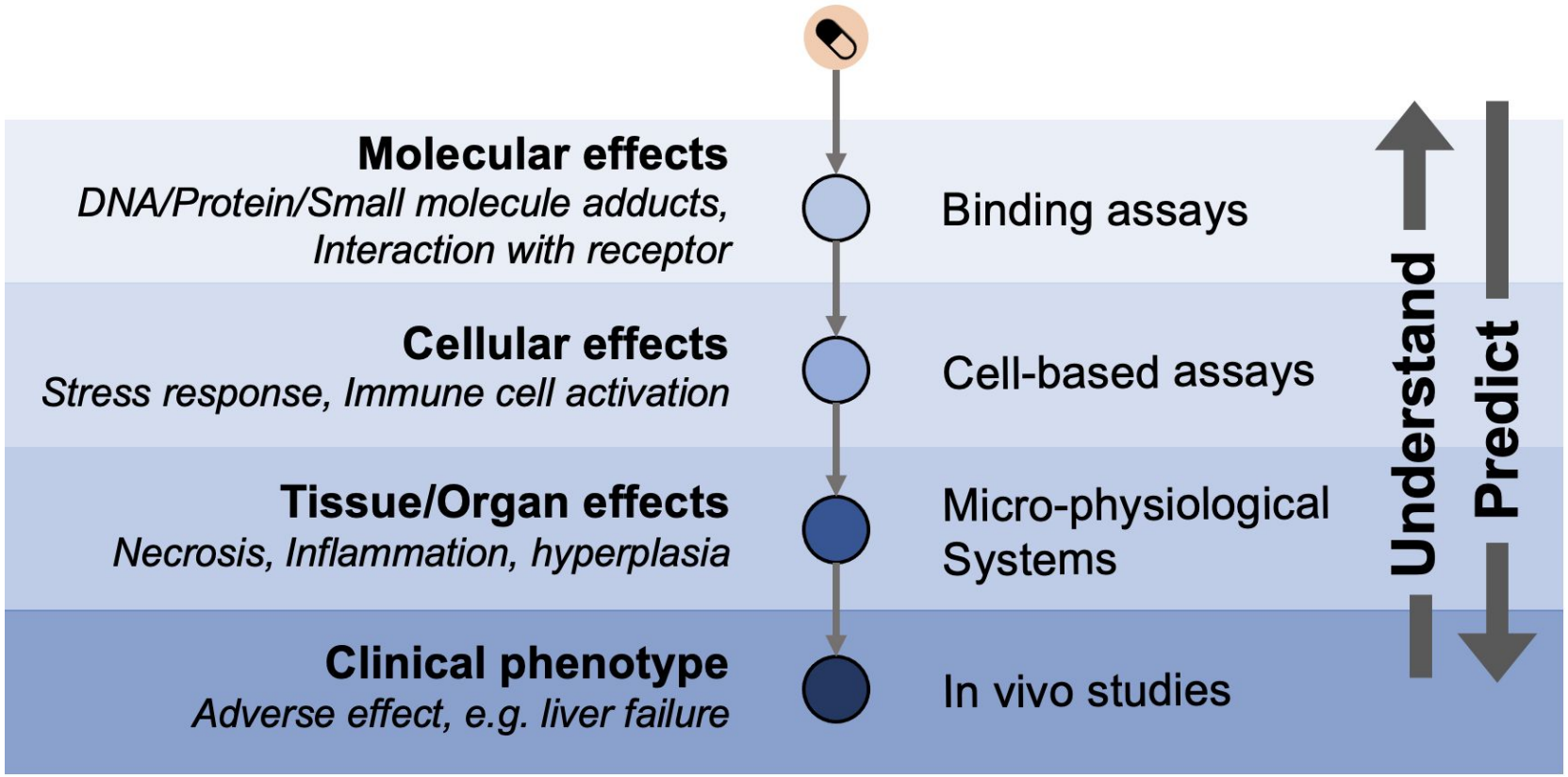
Yang, H., Sun, L., Li, W., Liu, G. & Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Front. Chem. 6, (2018)

The Pyramid of data-based discovery



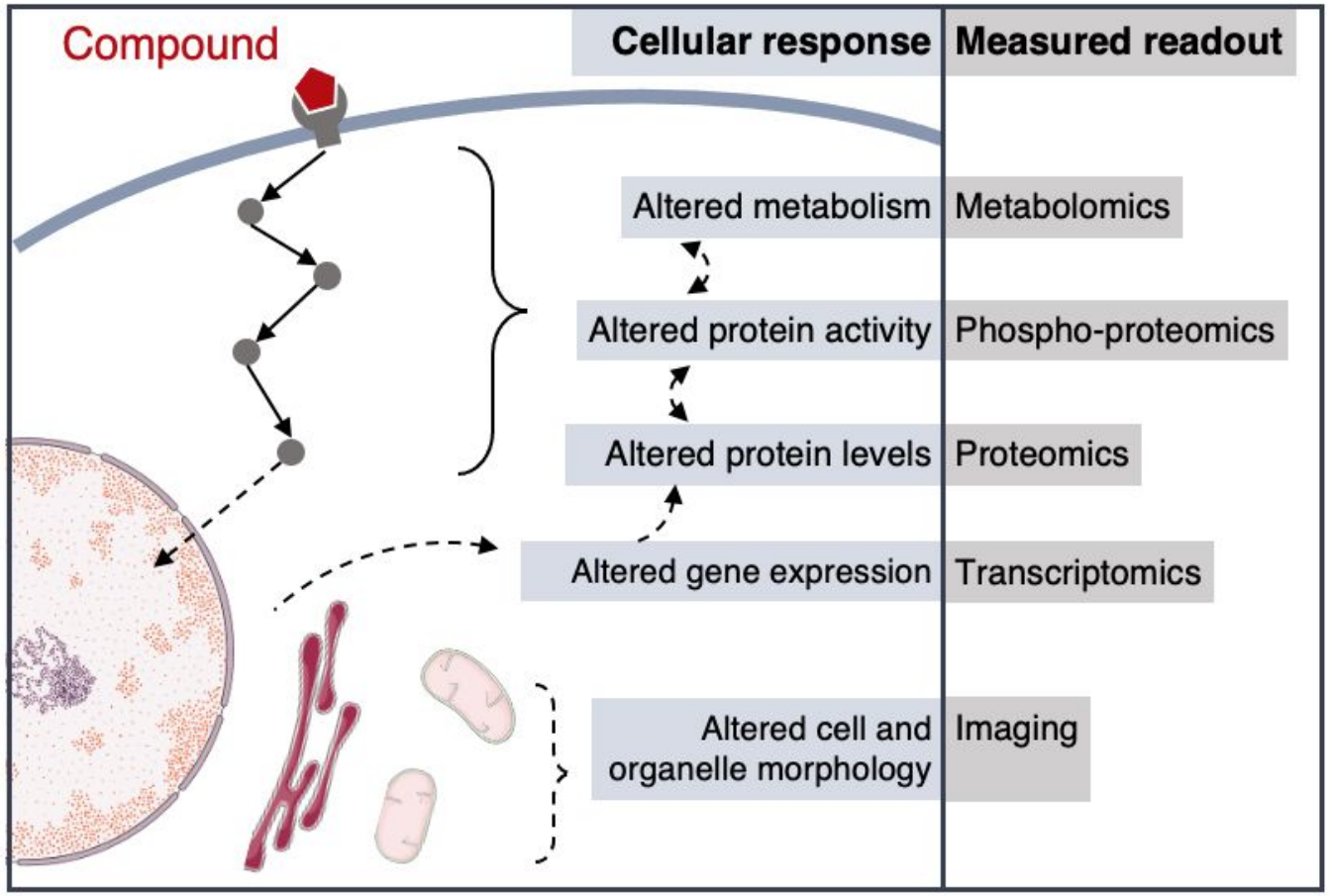
Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. Drug Discov. Today 26, 1040–1052 (2021)

Biological effects and their Model Systems



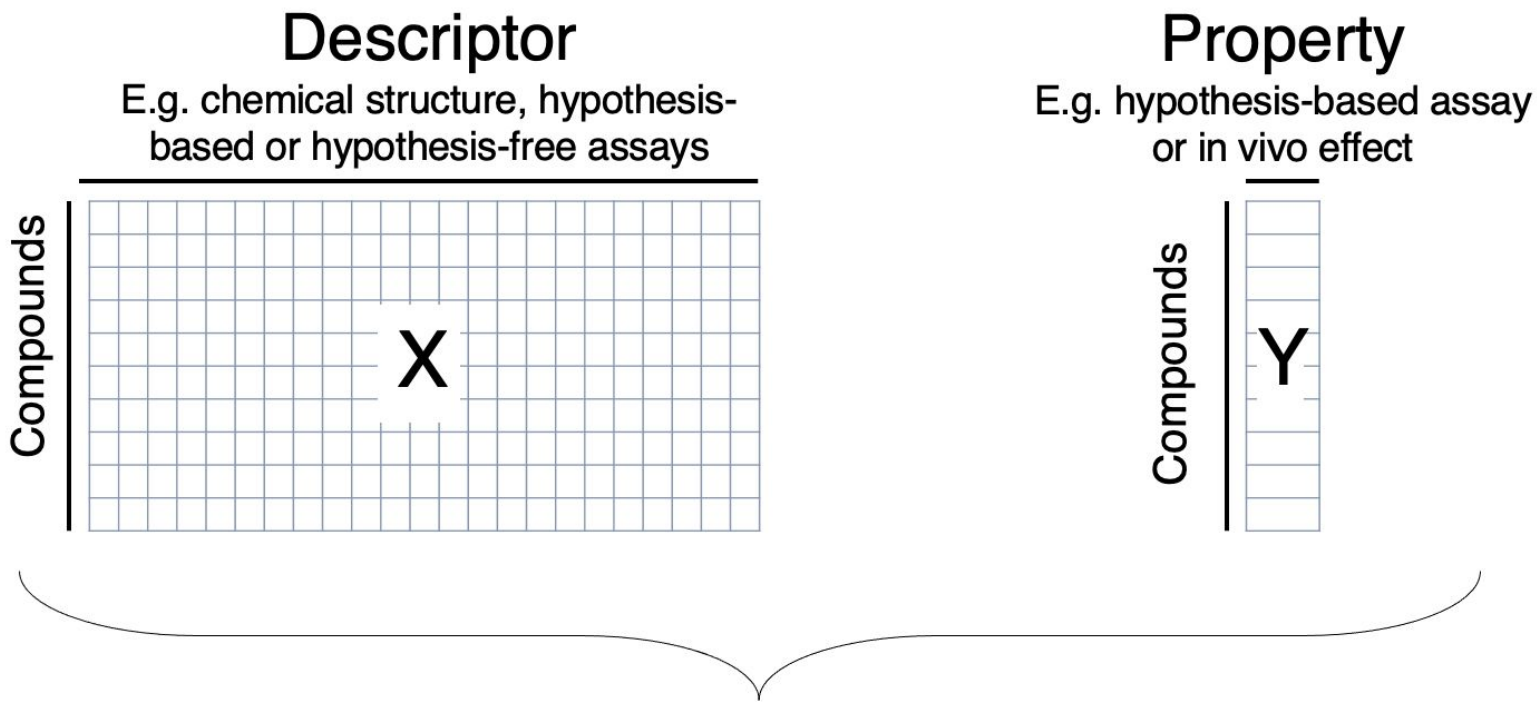
Compounds induce systems-level phenotypes through effects across multiple biological scales which can be studied using different model systems.

Biological effects and Assays to measure them…



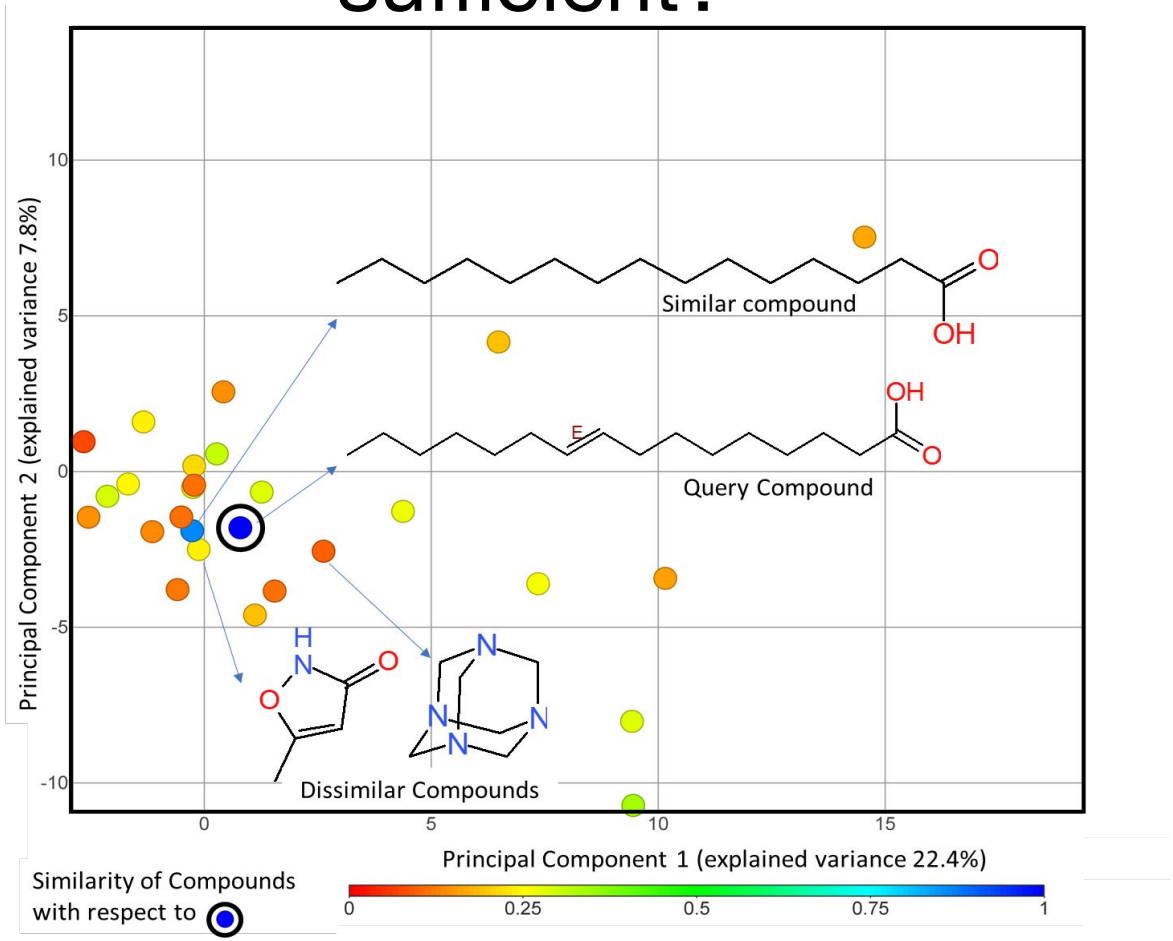
Modalities of cellular response to compound perturbation and hypothesis-free assays which measure it.

Supervised Machine Learning



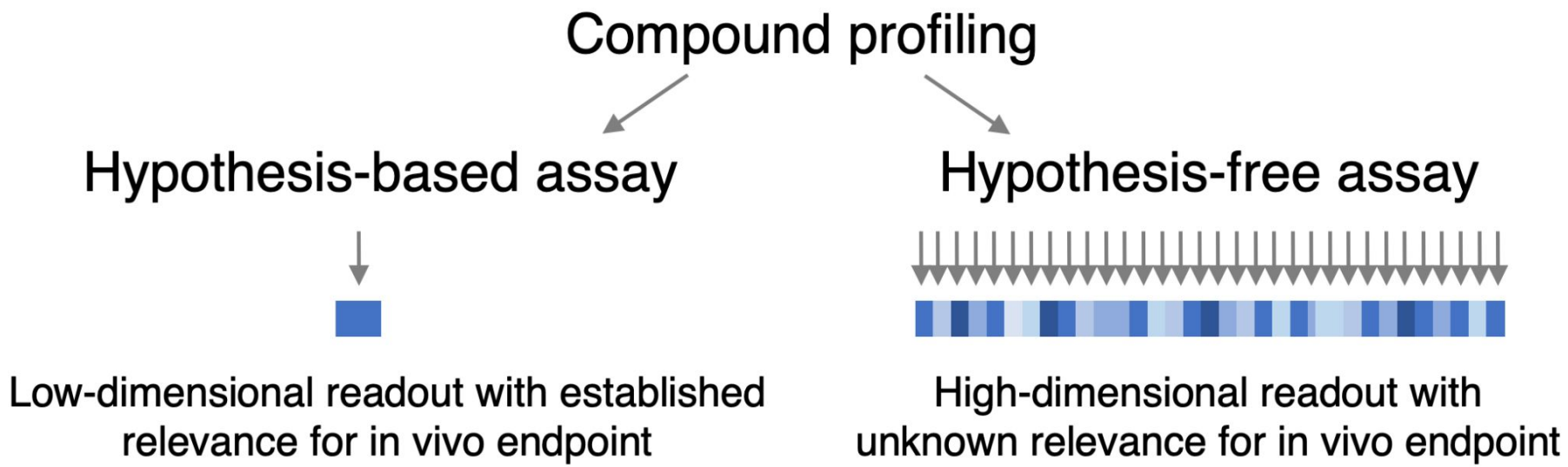
Supervised learning of properties (Y) from compound descriptors (X).

Molecular Similarity Measure: Is chemical structure sufficient?



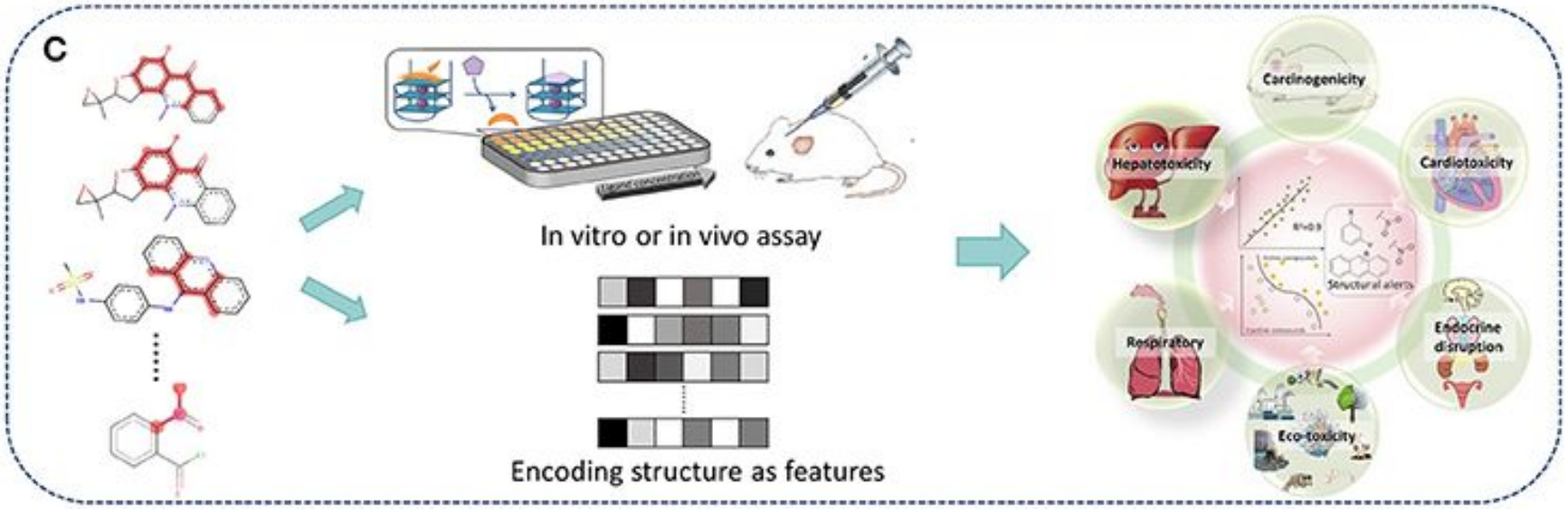
The PCA shows the global structure of the chemical space covered, while the Tanimoto similarity describes the similarity of each compound to the query compound.

Assay data: Hypothesis free and Hypothesis based



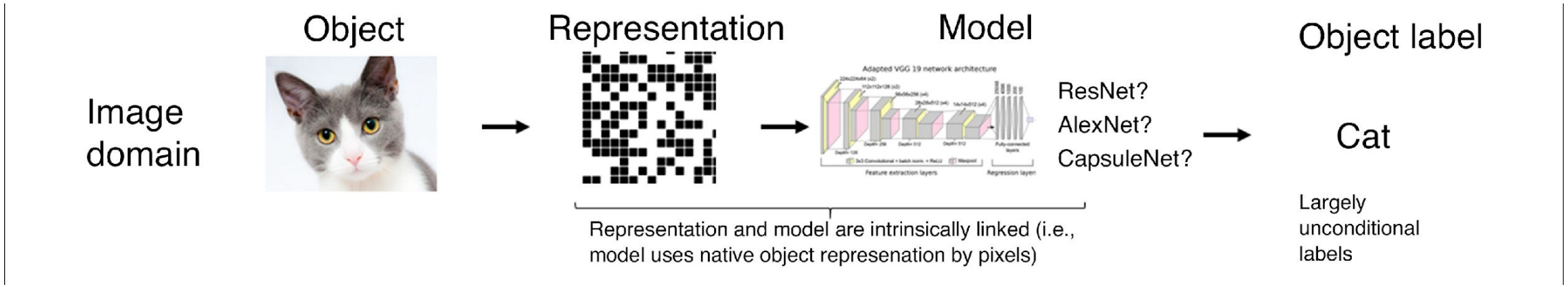
Hypothesis-based assays aim to measure a known, and often low-dimensional, readout which is mechanistically linked to an in vivo endpoint and hence generally interpretable. In contrast, hypothesis-free assays measure biological response broadly. As a consequence, their interpretation is not straightforward, however, they can potentially be informative for a wide range of endpoints.

Schematic Representation of Workflow Compounds to Assay to Prediction



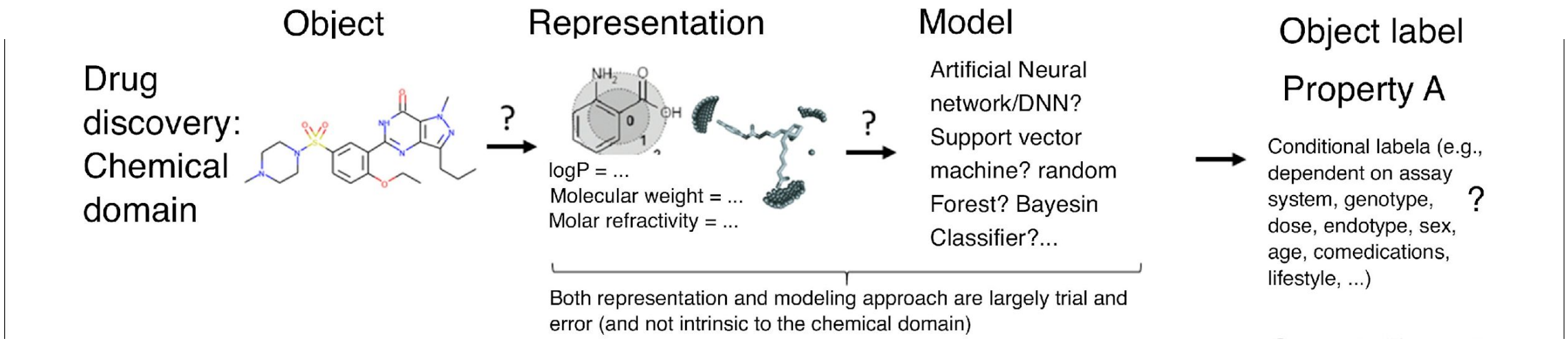
Yang, H., Sun, L., Li, W., Liu, G. & Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Front. Chem. 6, (2018)

Representing a datapoint (1 of 3)

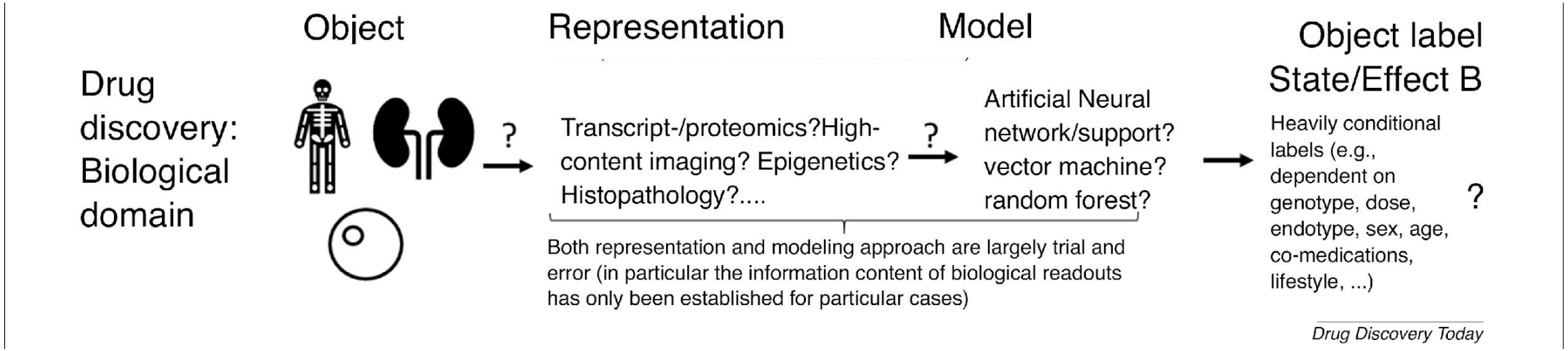


Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. Drug Discov. Today 26, 1040–1052 (2021)

Representing a datapoint (2 of 3)



Representing a datapoint (3 of 3)



Cell Painting Assay

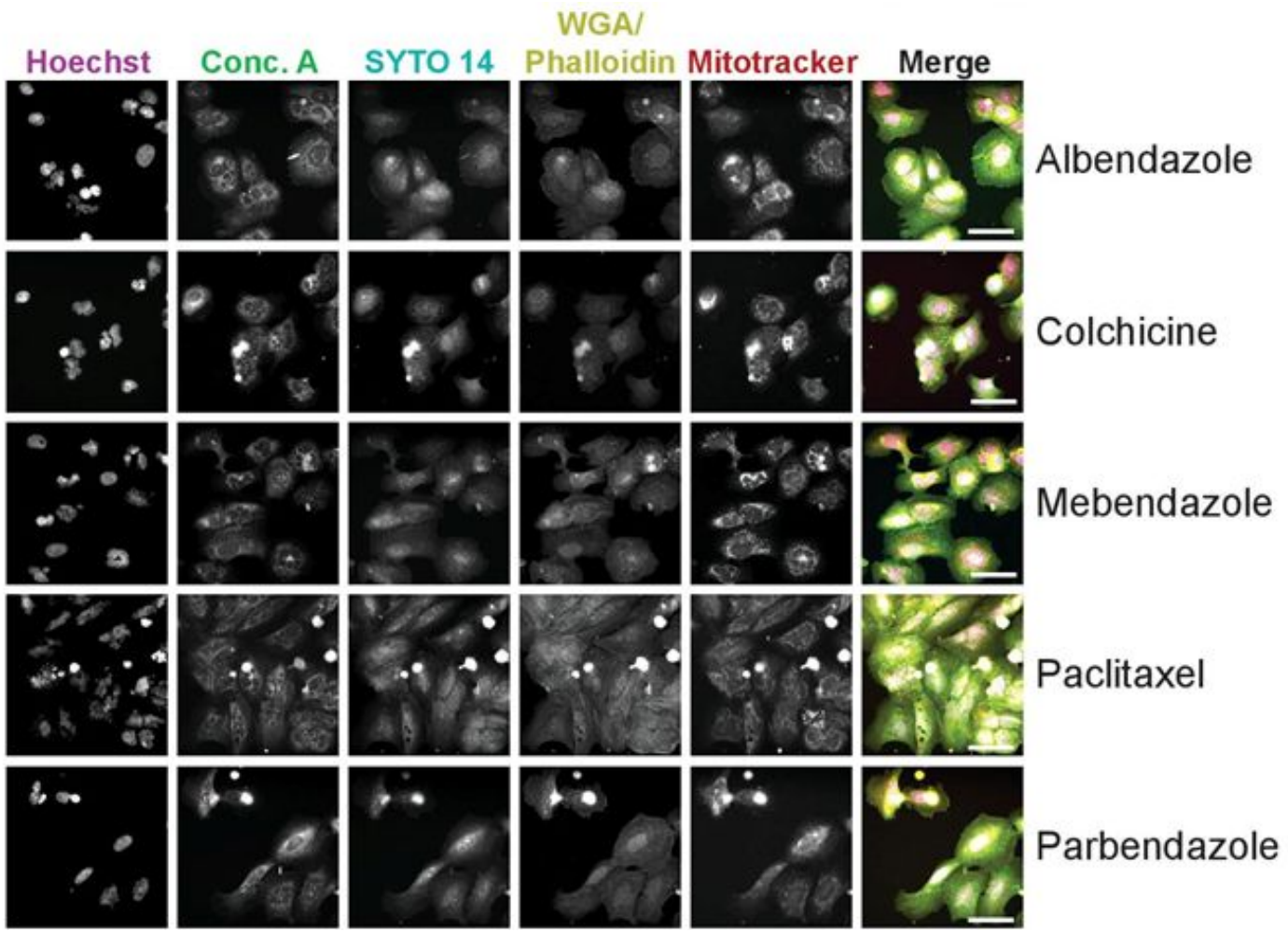
Cell Painting, human U2OS (human osteosarcoma) for more than 30,000 small molecules contains cell morphology statistics including intensity, texture and adjacency statistics.

Morphological profiling information

6 stains,

5 channels imaged,

8 constituents/organelles

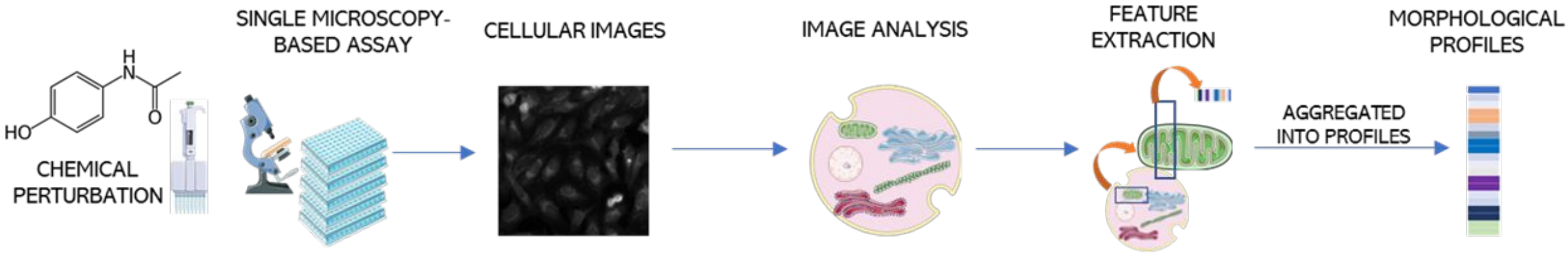


Bray, M. A.; Gustafsdottir, S. M et al., A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay. GigaScience. Oxford University Press December 1, 2017, 1–5.

Cell Painting Assay

For each identified substructures, measurements include :

- Counts: number of cells
- Size : area, volume, perimeter, diameter
- Shape
- Texture (smoothness)
- Intensity
- Spatial relationships

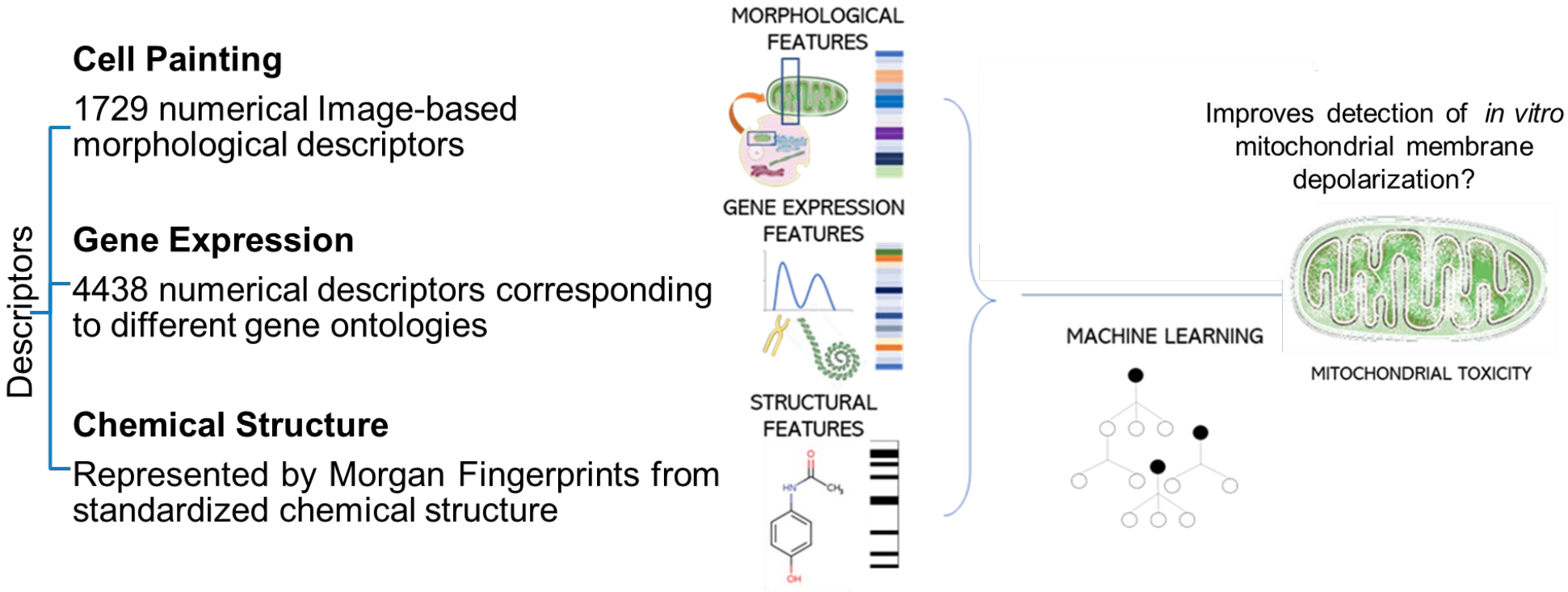


Bray, M. A.; Gustafsdottir, S. M et al., A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay. GigaScience. Oxford University Press December 1, 2017, 1–5.

2. Detection of Mitochondrial Toxicity

Detection of Mitochondrial Toxicity

- We explored Cell Painting features for information on mitochondrial toxicity
- Cell Painting and Gene Expression features extrapolate the applicability domain of structure-based models
- We interpret the biological significance of Cell Painting features



Dataset

Training Dataset:

- Tox21 Mitochondrial membrane potential disruption assay hit calls (summary assay)
- 382 compounds
- 62 Mitotoxic

External Test:

- Additional mitotox assays from CHEMBL, PubChem, Mitotox Database relevant to mitochondria, mitochondria potential and mitochondria complex
- 244 compounds
- 47 Mitotoxic

Hemmerich, J., Troger, F., Füzi, B. & F.Ecker, G. Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. Mol. Inform. 39, (2020)

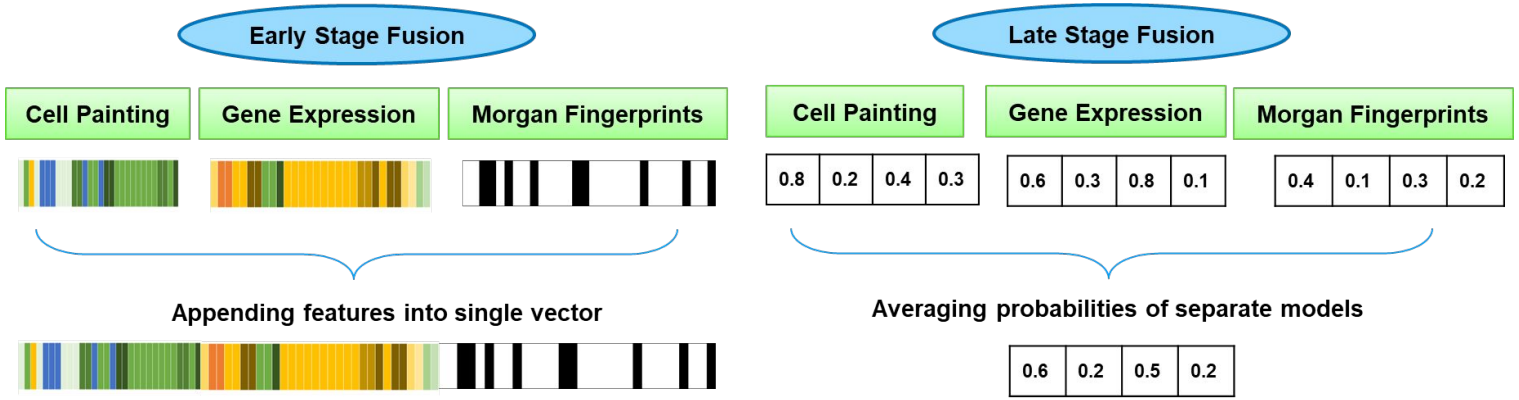
Methods

Model: Random Forest models

3 Individual Models

Cell Painting,
Gene Expression
Morgan fingerprints

2 Fusion models



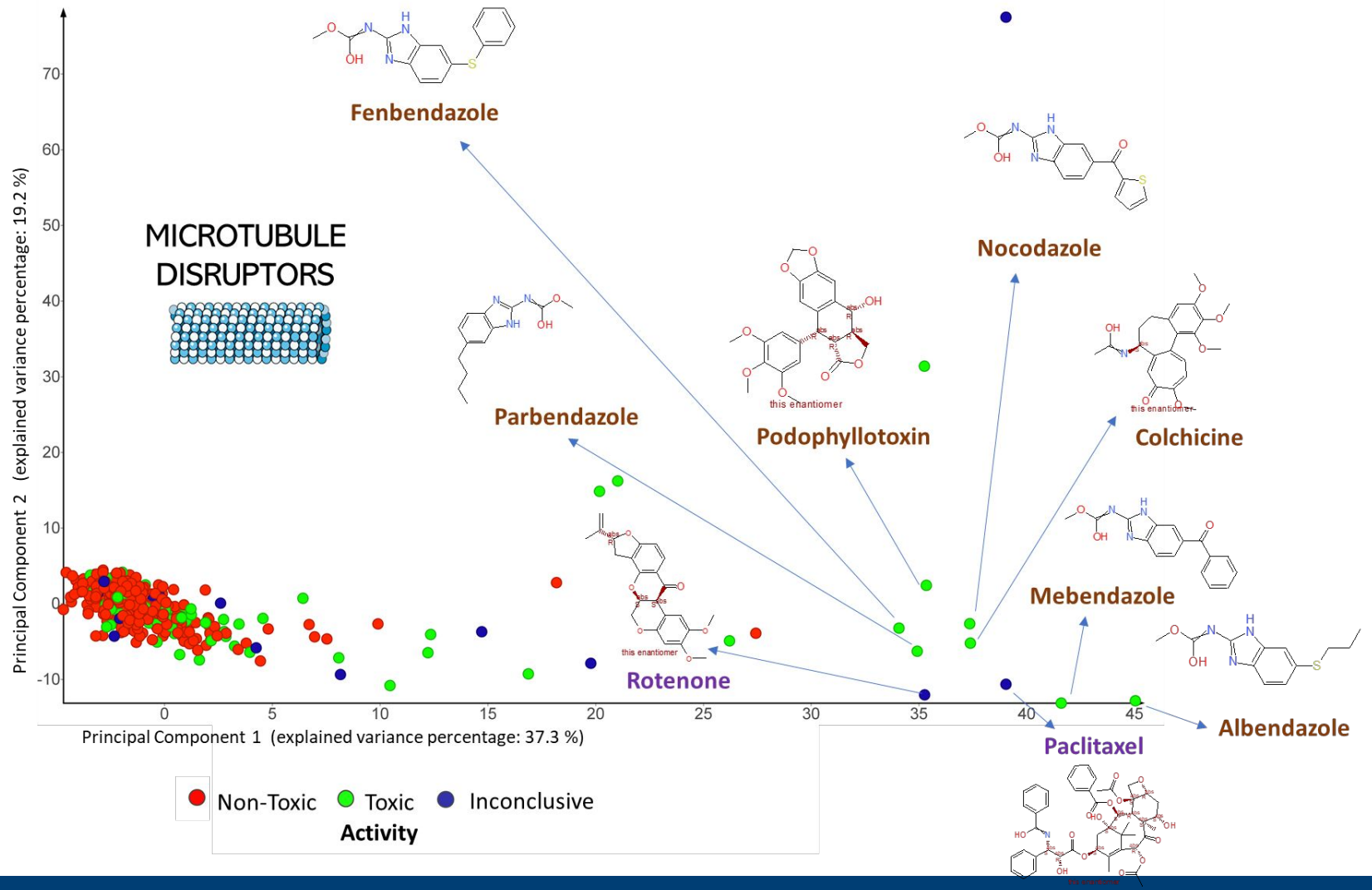
Nested Cross Validation 50 repeated 4-fold nested cross-validations on 382 compounds

Models were evaluated on an external test set of 244 compounds

Hemmerich, J., Troger, F., Füzi, B. & F.Ecker, G. Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. Mol. Inform. 39, (2020)

Morphology space clusters compounds with similar mechanisms

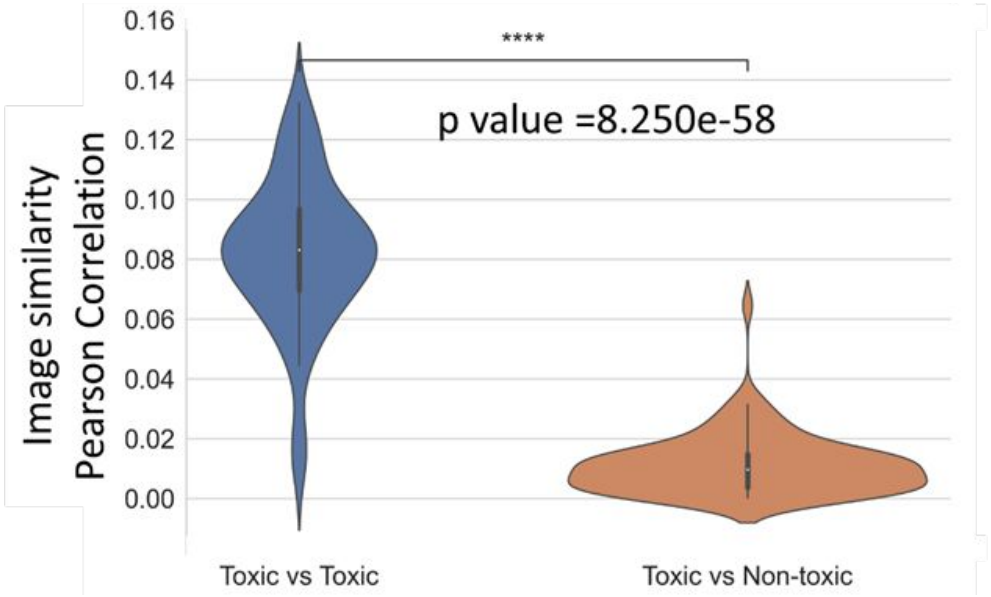
Compounds clustered further away from the distribution of majority of compounds having similar mechanisms of actions, for example, microtubule disruptors

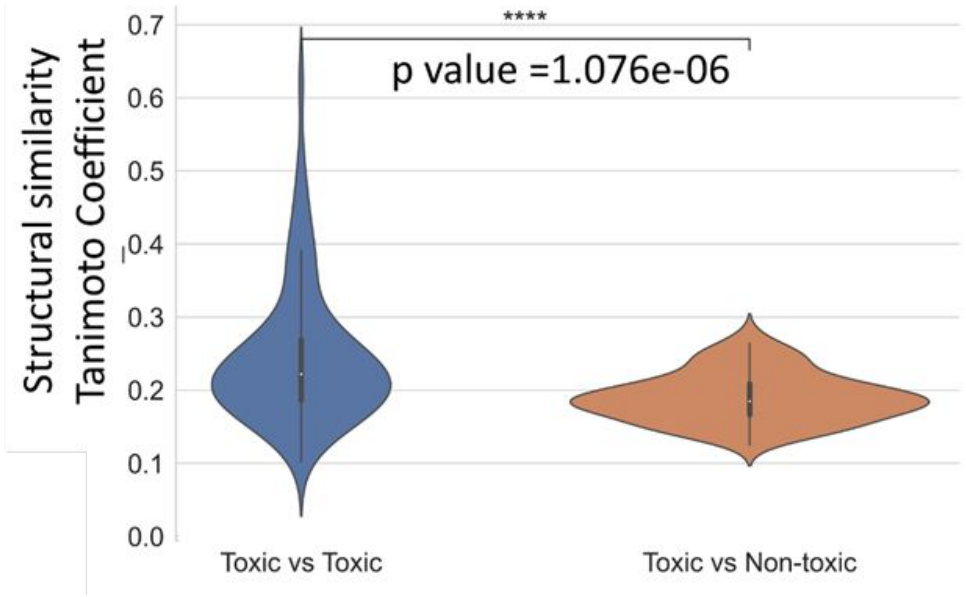


Principal Component Analysis of 542 compounds in 110-dimensional Cell Painting feature space.

Toxic compounds are more similar in morphology space

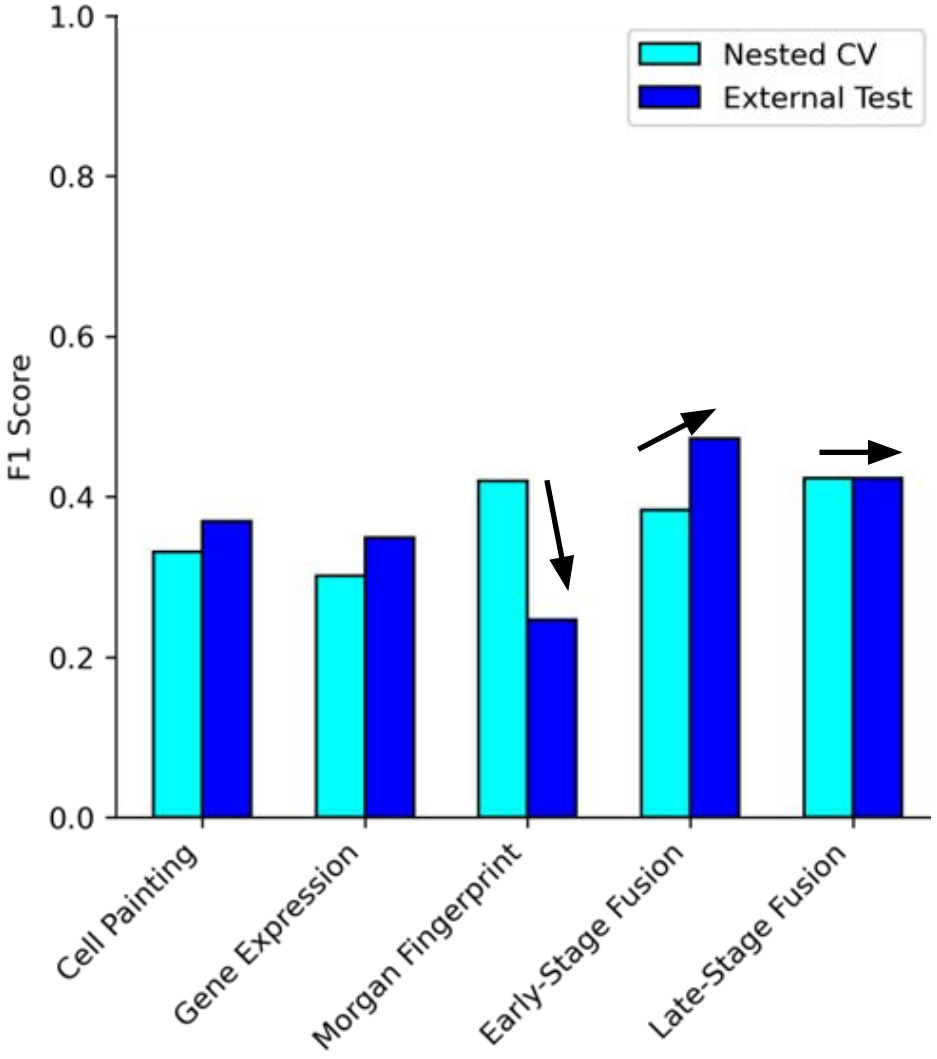
Morphological space is more able to discriminate between mitochondrial toxicants and non-toxicants than structural fingerprints.





Intra- and inter-class pairwise similarity for 486 compounds (85 mitotoxic)

Fusion models detect mitotoxicity better than structure



- External test set: F1 Score increases by 60% (0.25 to 0.42 in absolute terms) when using fusion models compared to Morgan fingerprints.
- Our method achieve higher sensitivity (0.79 in our study vs 0.37 in Apredica MitoMass[4]) with comparable balanced accuracies (0.69 in our study vs 0.65 in Apredica MitoMass).

Hallinger, D. R., Lindsay, H. B., Friedman, K. P., Suarez, D. A. & Simmons, S. O. Respirometric screening and characterization of mitochondrial toxicants within the toxcast phase i and II chemical libraries. Toxicol. Sci. 176, 175–192 (2020)

Cell Painting features related to Mitotoxicity

Biological significance of Cell Painting features with respect to Mitochondrial Toxicity :

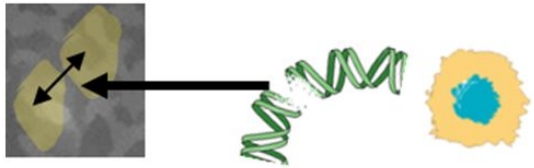
MITOCHONDRIAL FEATURES

Cells Intensity MaxIntensityEdge Mito (PPV 0.83)

Edge of segmented object potentially indicates loss of membrane integrity

FEATURES FROM OTHER IMAGE CHANNELS

Cells Correlation Costes DNA AGP (PPV 0.52)



Potentially indicates DNA fragmentation and entering apoptosis or cell death

Cell Painting predicts Mitotoxicity

- Mitochondrial toxicants significantly differ from non-toxic compounds in morphological space; clusters with similar mechanisms; granularity features are highly predictive mitochondrial toxicity

- Models combining Cell Painting, Gene Expression features and Morgan Fingerprints improved detection (by 60% from 0.25 to 0.40) compared to models using only structural features

- Models extrapolated well into new chemical space and perform with better sensitivity than some dedicated hypothesis-based experimental assays for mitochondrial toxicity

Seal, S. *et al.* Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. *Commun. Biol.* ***2022*** *51* 5, 1–15

3. Combining Predictions from Cell Morphology Data and Structural Fingerprint by Leveraging Distance to Training Data

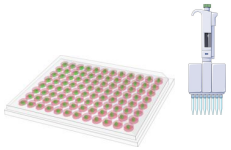
Aim

- Predictions from a model using morphology work best when compounds are similar in morphology space
- Structural models are local and work best when training compounds are similar in structure
- Instead of averaging models, can we merge morphology and structure models based on similarity to training set in both spaces?

Different regions in the morphology vs structure space
will have different weights to the individual models

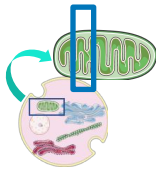
Features

Bioactivity Datasets
Assay hit calls



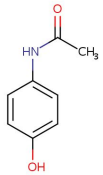
InChICode_standardised	262_740	263_741	264_742	265_743	266_744	267_745
InChI=1S/C20H22N2O3S2/c1-3-25-19(23)15-4-8-16(...	0.0	0.0	0.0	0.0	0.0	0.0
InChI=1S/C15H13N3O3/c19-13(8-12-14(20)18-15(21...	NaN	0.0	0.0	1.0	0.0	0.0
InChI=1S/C20H22N4/c1-15(16-9-3-2-4-10-16)21-19...	0.0	0.0	0.0	NaN	0.0	0.0
InChI=1S/C21H21BrN4O3S/c1-4-5-16(29-13-8-6-12(...	NaN	0.0	NaN	0.0	0.0	NaN
InChI=1S/C20H17FN2O5/c1-11-17(18(23-20(25)22-1...	0.0	0.0	0.0	1.0	0.0	0.0

Morphological descriptors
Cell Painting Data



InChICode_standardised	Cells_Correlation_Correlation_RNA_AGP	Cells_Correlation_Costes_AGP_RNA	Cells_Correlation_K_AGP_DNA
InChI=1S/C20H22N2O3S2/c1-3-25-19(23)15-4-8-16(...	-0.157794	-0.099680	-0.164429
InChI=1S/C15H13N3O3/c19-13(8-12-14(20)18-15(21...	-0.054050	-0.057673	-0.208181
InChI=1S/C20H22N4/c1-15(16-9-3-2-4-10-16)21-19...	-0.119915	-0.060212	-0.379775
InChI=1S/C21H21BrN4O3S/c1-4-5-16(29-13-8-6-12(...	-0.087385	-0.101862	-0.074942
InChI=1S/C20H17FN2O5/c1-11-17(18(23-20(25)22-1...	0.026187	-0.037733	-0.210272

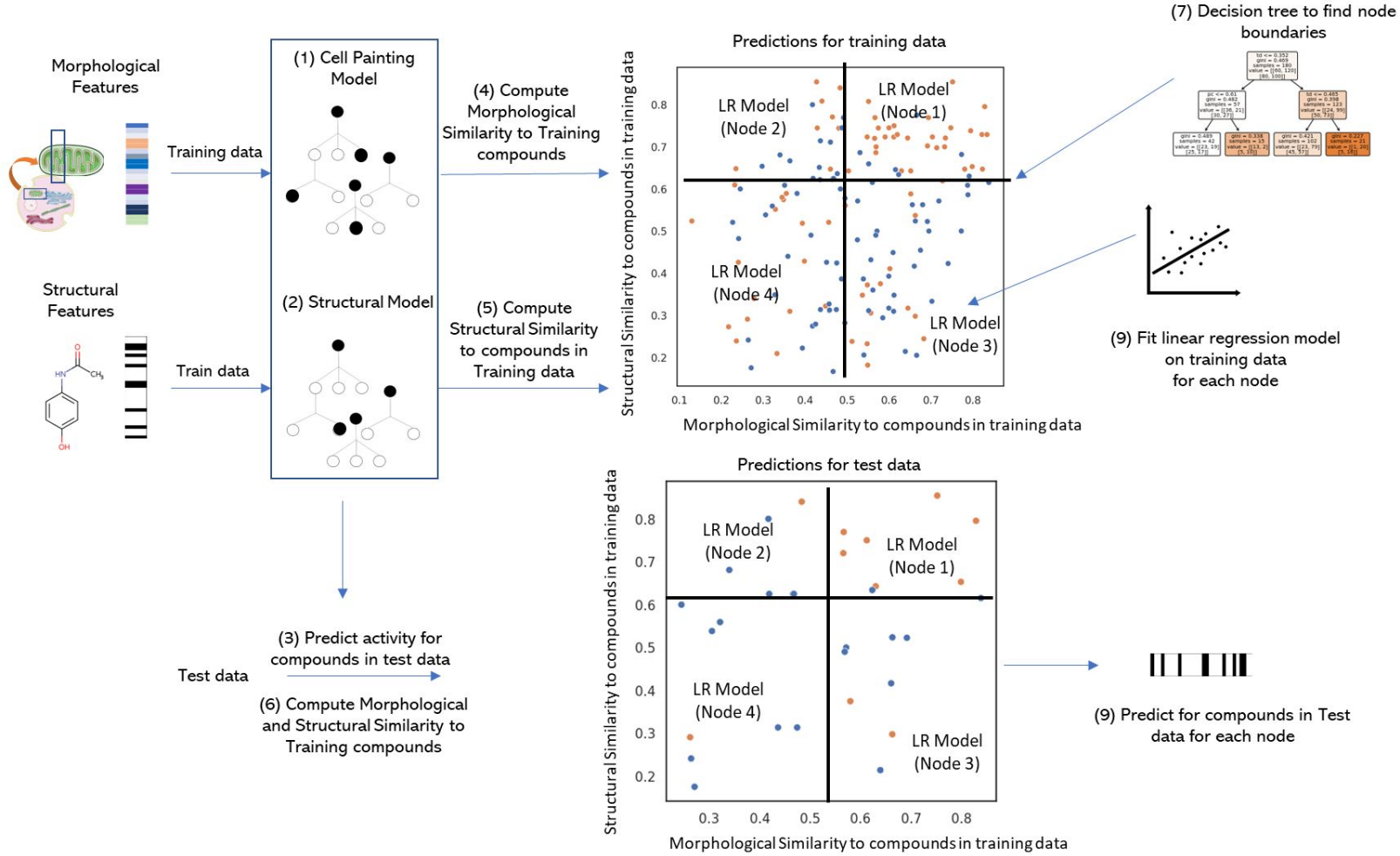
Structural features
Morgan fingerprints



InChICode_standardised	Mfp0	Mfp1	Mfp2	Mfp3	Mfp4	Mfp5	Mfp6	Mfp7	Mfp8	...	Mfp2038	Mfp2039	Mfp2040
InChI=1S/2C24H28N2O4/c2*1-3-7-17-10-11-20-21-1...	0	0	0	0	0	0	0	0	0	...	0	1	0
InChI=1S/2C28H35N3O3/c2*1-3-7-21-10-11-24-25-2...	0	1	0	0	0	0	0	0	0	...	0	1	0
InChI=1S/2C28H35N3O3/c2*1-3-7-21-10-11-24-25-2...	0	1	0	0	0	0	0	0	0	...	0	1	0
InChI=1S/2C28H29N5O4/c2*1-3-7-19-10-11-23-24-2...	0	1	0	0	0	0	0	0	0	...	0	1	0
InChI=1S/2C22H30N2O5/c2*1-3-4-15-5-6-17-20-19(...	0	0	0	0	0	0	0	0	0	...	0	1	0

Dataset and Method

92 assays and 10,402 unique compounds from ChEMBL (Hofmarcher et al)

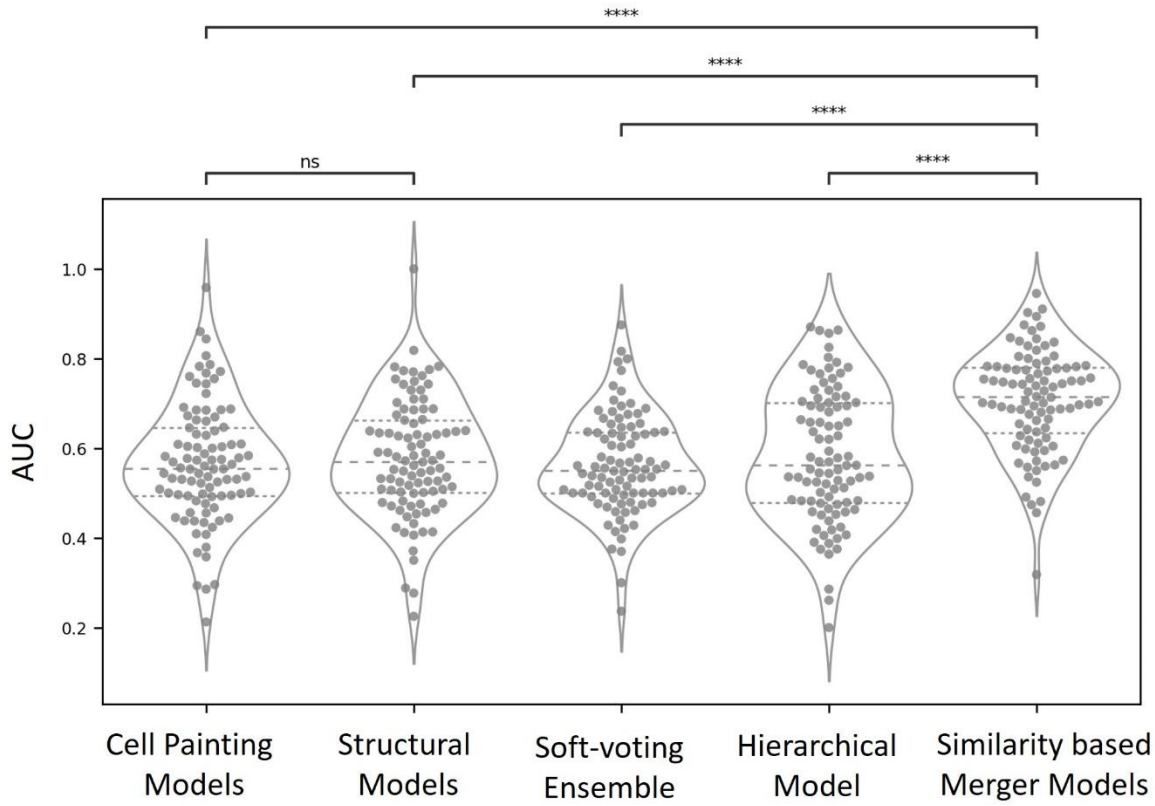


Hofmarcher et al . Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. J. Chem. Inf. Model. 2019, 59 (3), 1163–1171

Results

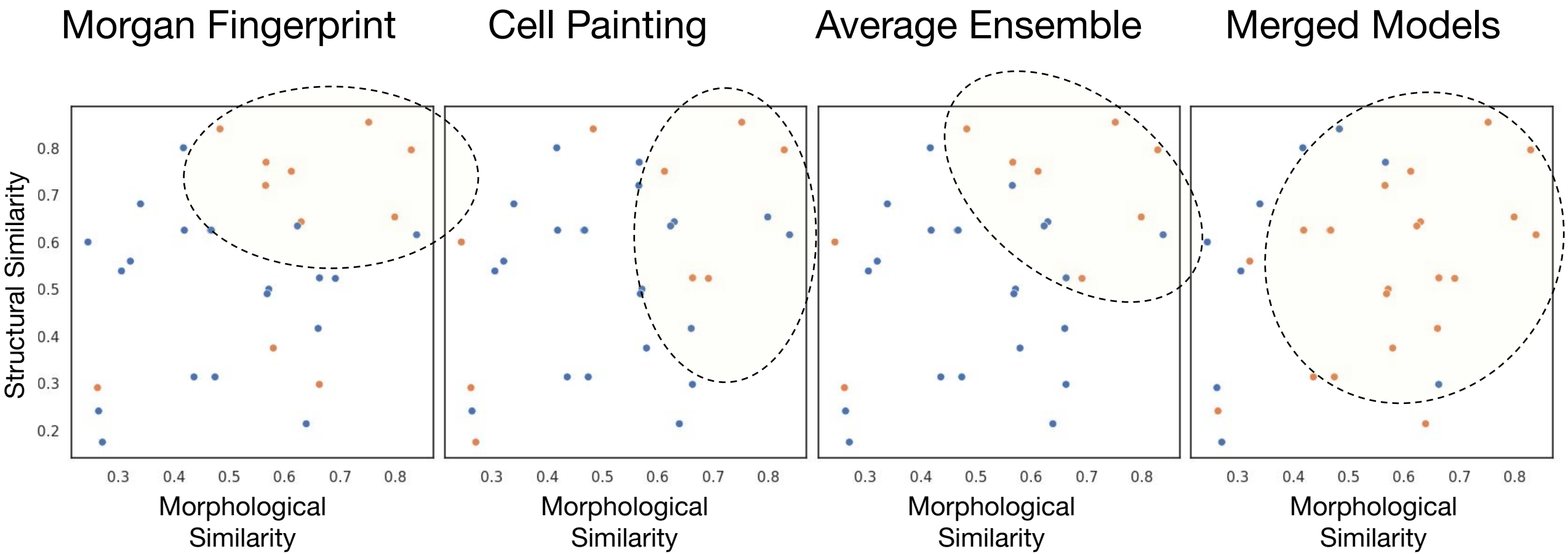
For the held-out test set, greater number of assays with AUC>0.60 higher for merged model compared to individual and average ensemble models

Significant improvement in performance when using the merged model compared to the average ensemble



t-test paired samples
p-value annotation legend:
ns: p <= 1.00e+00
*: 1.00e-02 < p <= 5.00e-02
**: 1.00e-03 < p <= 1.00e-02
***: 1.00e-04 < p <= 1.00e-03
****: p <= 1.00e-04

Merged Models can improve applicability domain



qHTS Assay for Lipid Storage Modulators in Drosophila S3 Cells

Similarity is a good measure to merge models

- Merged models improve the applicability domain
- 30 out of 90 assays exclusively show an improvement in performance AUC>0.70 in merged models which could not perform well in other models.
- Merged models can hence better combine feature spaces of structure and cell morphology

Journal of Cheminformatics BMC

CSH Cold Spring Harbor Laboratory

bioRχiv THE PREPRINT SERVER FOR BIOLOGY

Seal, S. et al. Merging Bioactivity Predictions from Cell Morphology and Chemical Fingerprint Models by Leveraging Similarity to Training Data. (accepted to Journal of Chemoinformatics)

4. Interpreting Cell Morphology using Convolutional Neural Networks for Compound Toxicity

Cell Painting Feature Space is highly correlated

- Cell morphological readouts contain information on several bioactivity endpoints
- Features are highly correlated
- Can we use interpret these features from a biological context without diluting feature importance due to correlation?
- Convolutional Neural networks leverage neighbourhood correlation
- We can obtain per endpoint and per compound importance heatmaps using Grad-CAM.

Dataset

Training Dataset: Tox21 assay (8 broad biological activities)

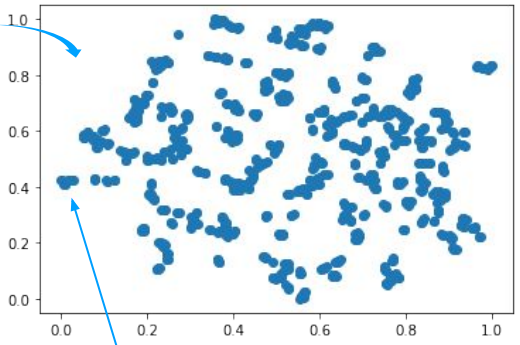
- apoptosis
- cytotoxicity BLA
- cytotoxicity SRB
- ER stress
- heat shock
- mitochondrial disruption
- oxidative stress
- proliferation decrease

Method

1. Prepare Feature Map

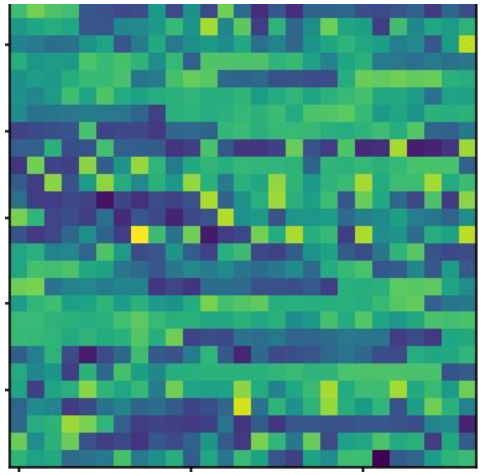
	0	1	2	3	4
Metadata_profile_id	profile_0	profile_1	profile_2	profile_3	profile_4
Metadata_cell_line	A549	A549	A549	A549	A549
Metadata_pert_name	AKT1-1	AKT1-2	ARID1B-1	ARID1B-2	ATF4-1
Cells_AreaShape_Center_Y	-0.18016	0.370572	-0.360905	0.26245	-0.110264
Cells_AreaShape_Compactness	-0.155631	-0.247842	0.79474	0.480421	-0.074895
...	...	...	...	...	...
Nuclei_Texture_SumVariance_DNA_5_0	0.923143	0.504751	-0.497296	0.063444	0.594059
Nuclei_Texture_Variance_AGP_5_0	0.944998	0.407462	-0.748232	-0.560178	0.674015
Nuclei_Texture_Variance_DNA_10_0	0.984938	0.522251	-0.51524	-0.062851	0.140325
Nuclei_Texture_Variance_DNA_20_0	1.122724	0.64437	-0.42144	0.085026	0.29123
Nuclei_Texture_Variance_DNA_5_0	0.961945	0.519441	-0.526734	0.026056	0.417465

t-SNE of Feature Map



Cell_Texture_SumAvergage_AGP_10_0

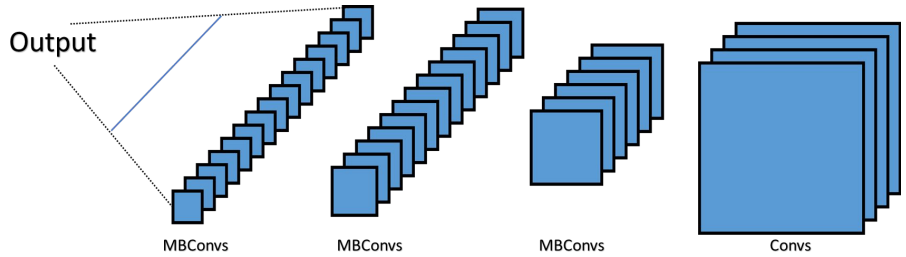
Jonker-Volgenant algorithm



Tox21 Assays

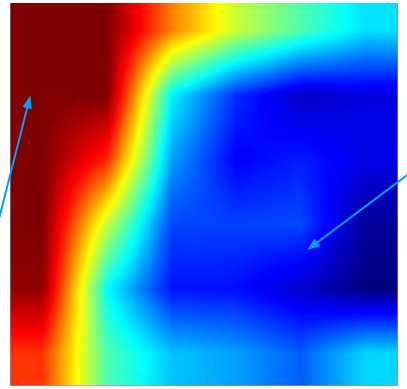
e.g. ER Stress

2. Predict Endpoint of test set



Model: EfficientNet B0

3. Interpretation using Grad-CAM



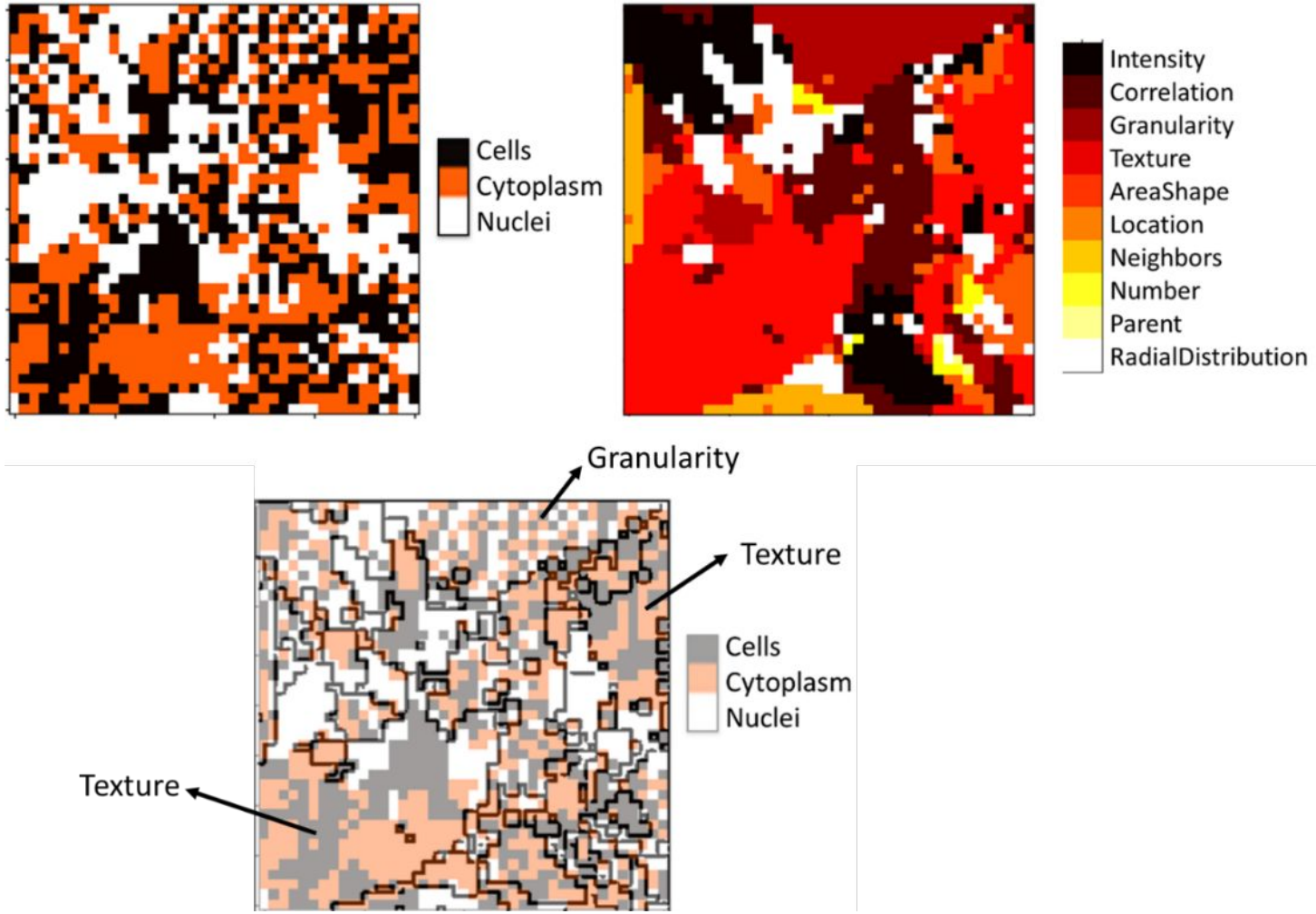
More important to model

Less contributing to model

Features are related by measurement type

- Majority of features are related by measurement function than by objects they were measured in (cells, cytoplasm, or nuclei)

- For example, granularity, features are clustered together from all compartments which means information on granularity was homogenous throughout the channels.

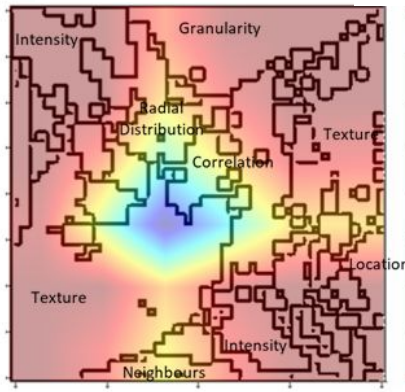


Compounds with similar MOA have similar importance regions

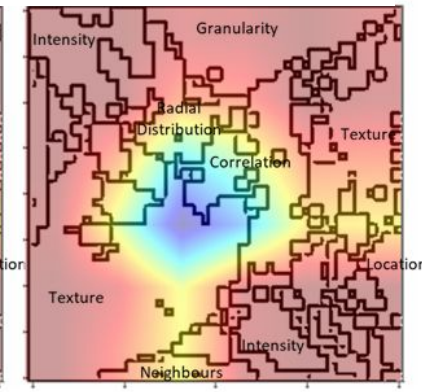
For models predicting proliferation decrease endpoint:

mitochondrial toxicity

Niclosamide

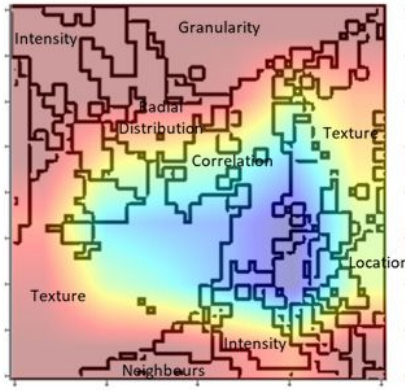


Colchicine

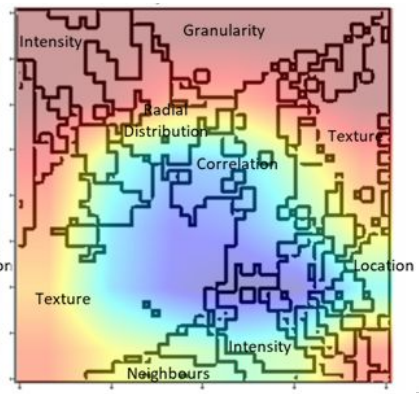


inhibits cell proliferation in the G(1) phase of the cell cycle

Lovastatin

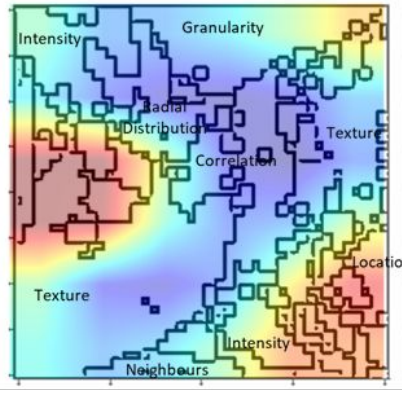


Cycloheximide

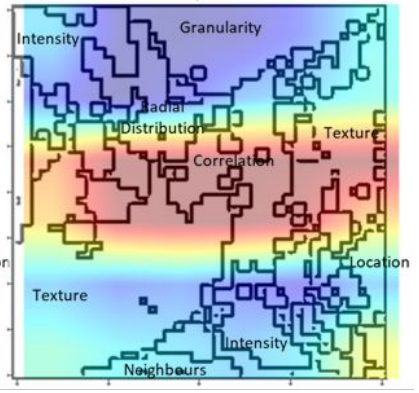


antiandrogens

Nilutamide

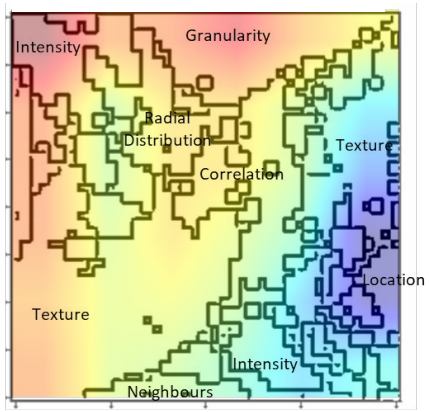


Mifepristone

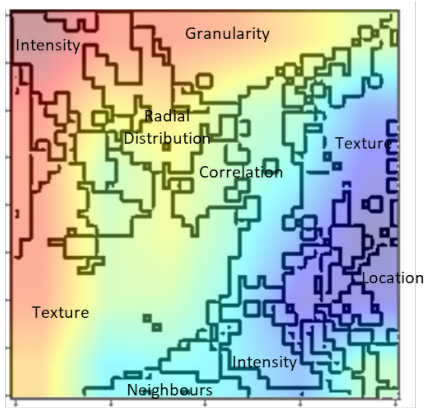


Microtubule disruptors and ER Stressors affect texture features

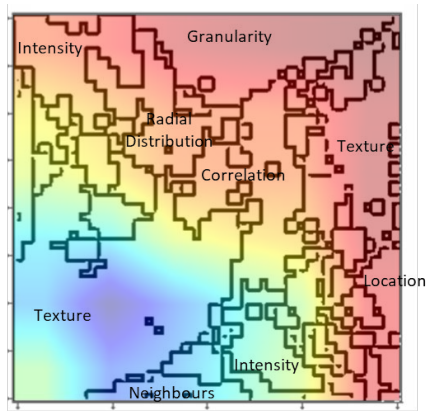
Cycloheximide



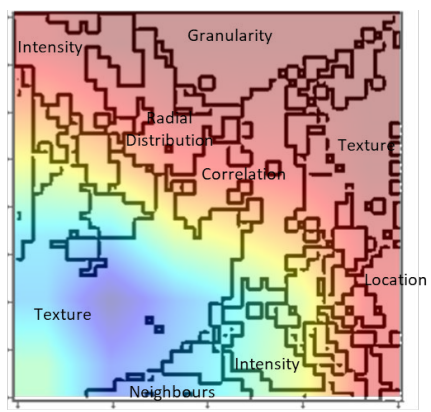
Daunorubicin



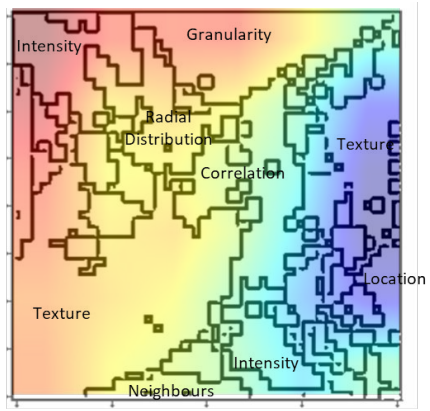
Niclosamide



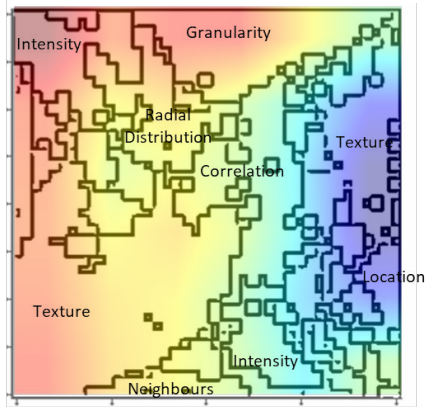
Pimozide



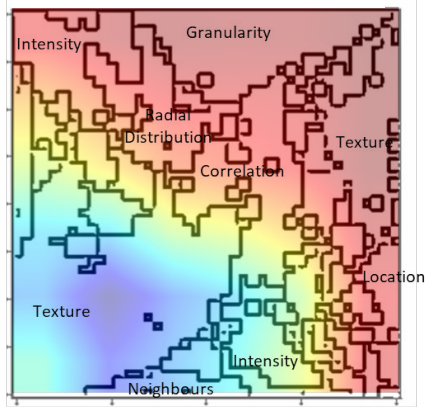
Paclitaxel



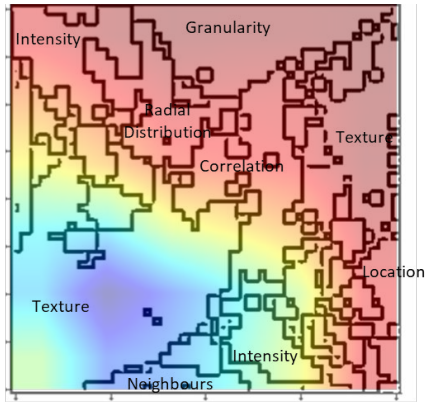
Parbendazole



Fluspirilene



2,5-Di-tert-butylhydroquinone



Microtubule disruptors causing apoptosis

Causing ER stress.

Novel Representation of Cell Painting features may reveal mechanistic understanding

- We present a new representation, modelling and interpretation technique for cell morphological
- Leverage corelated Cell Painting features in the prediction of biological processes using CNN
- Compounds with similar MOA have similar regions of interest
- ER Stressors/Microtubule disruptors exhibit importance in 2 different regions of texture features
- Understanding these morphological regions help understand Cell Painting features biologically
- We can interpret the models to guide experiments and develop new approach methodologies

Conclusions

- Similar structures causing similar effect cannot always be leveraged as it depends on the representation of the compound for ML.
- Hypothesis free data from cell morphology, like gene expression, are versatile biological descriptors of a system
- Cell morphology provides an alternative feature space to predict drug toxicity, for example mitochondrial toxicity
- Developing methods to combine morphology and structural models improve predictions further
- We could also interpret morphological readouts for insights into biological activity.
- In future, morphological data will certainly help to look at the chemical space with a biological lens

Acknowledgements

Prof Andreas Bender
Bender Group members
Prof Ola Spjuth
Dr Jordi Carreras-Puigvert
Prof Michele Vendruscolo
Prof Pietro Lio

This work was supported by:

TRINITY COLLEGE CAMBRIDGE

CAMBRIDGE COMMONWEALTH EUROPEAN AND INTERNATIONAL TRUST

CLARE HALL CAMBRIDGE

ROYAL SOCIETY OF CHEMISTRY

Researcher Development Grant

slas
Come Transform Research

Tony B. Academic Travel Award

JAWAHARLAL NEHRU MEMORIAL FUND

UPPSALA UNIVERSITET

Vetenskapsrådet
Swedish Research Council

UNIVERSITY OF CAMBRIDGE

ss2686@cam.ac.uk

@srijitseal