Prediction of cancer drug sensitivity for drug-gene interaction discovery

Hang Su[1], Edward Yu[2], Anna Lasorella[3,4], Chris Wiggins[5,6,7], Raul Rabadan[7,8]

Affiliations: [1]Columbia College, Columbia University; [2]The Fu Foundation School of Engineering and Applied Sciences, Columbia University; [3]Department of Pathology and Cell Biology, Columbia University Medical Center; [4]Department of Pediatrics, Columbia University Medical Center; [5]Department of Applied Physics and Applied Mathematics, Columbia University; [6]Data Science Institute, Columbia University; [7]Department of Systems Biology, Columbia University; [8]Department of Biomedical Informatics, Columbia University College of Physicians & Surgeons

Glioblastomas are an aggressive form of malignant brain tumour, with high tumour heterogeneity, poor patient prognoses, and treatment options of limited effectiveness.

Leveraging genomic, and chemo-therapeutic drug response datasets in the public-domain, as well as open-source cheminformatics tools, we train modified-LASSO regression models (inspired by Netflix's recommendation system) to learn drug-gene interactions pertinent to predicting drug sensitivity. The resulting model is earmarked for eventual laboratory validation on patient-derived glioblastoma cells.

Introduction

State of the Field

Most common methods in literature include linear regression variants (LASSO, Ridge, Elastic Net, etc.), MANOVA, Random Forests, etc

Cell lines' drug sensitivity usually analysed with respect to individual drugs, in isolation of the rest

State of the art results for model predictions include mean per-drug Spearman rank correlations between 0.20-0.25 [Nguyen et al., 2016], and coefficient of determination R^2 values of 0.72 and 0.64 from 8-fold cross validation and a blind test set respectively [Menden et al., 2016]

LASSO

$$\underset{\beta}{\arg\min} \|Y - X\beta\|_2^2 + \lambda\|\beta\|$$

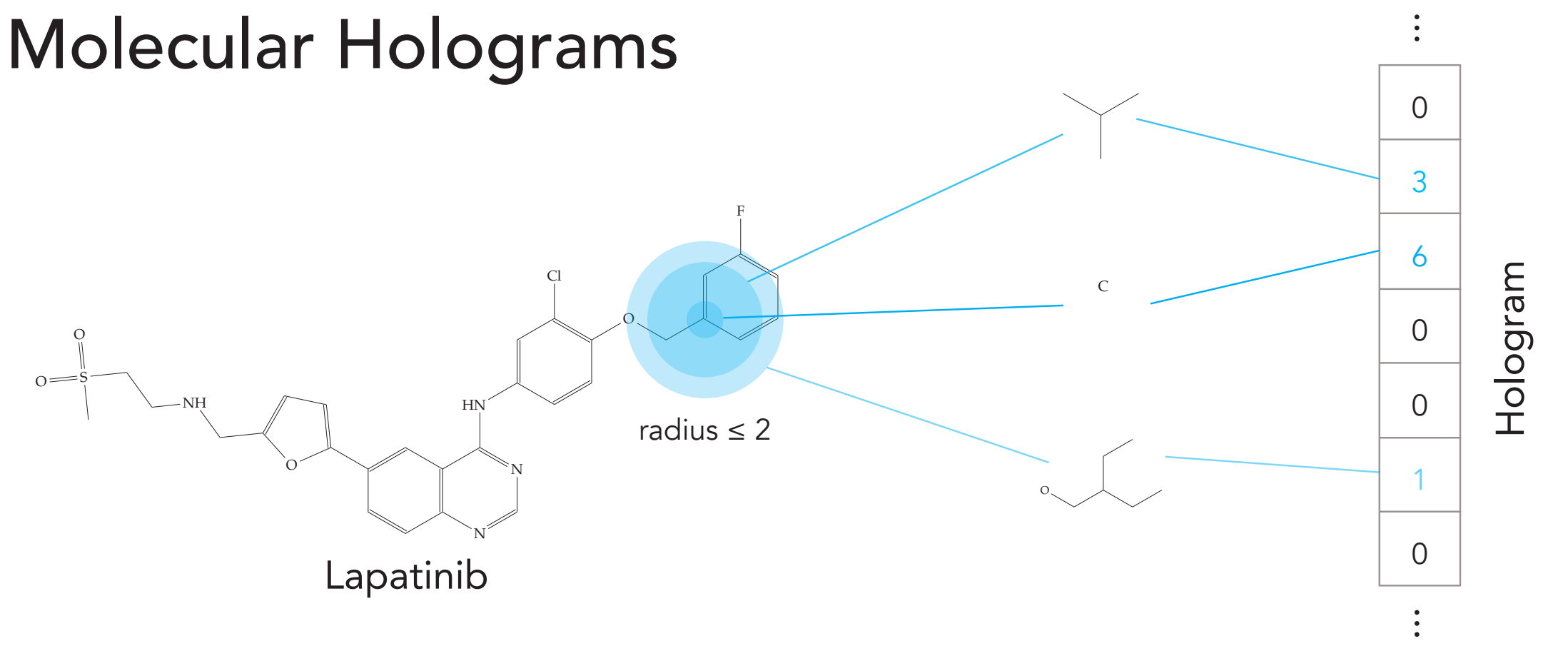
where *Y* is the response (to be approximated), *X* are the regressors, β are the coefficients of the regression (to be learned), and λ is the regularisation parameter

Collaborative Filtering via Matrix Factorisation

$$\underset{p,q,b}{\arg\min}\, C\|R - \mu - \mathbf{b}_u - \mathbf{b}_i - P^\top Q\|_2^2 + \lambda\left(\sum_u \|P_u\|_2^2 + \sum_i \|Q_i\|_2^2 + \|\mathbf{b}_u\|_2^2 - \|\mathbf{b}_i\|_2^2\right)$$

where *C* is the confidence, *R* is the ratings matrix (to be approximated), *P* is the matrix of movie attributes (to be learned), *Q* is the matrix of user attributes (to be learned), μ is the mean rating, $\boldsymbol{b}_i$ is the bias of each movie, $\boldsymbol{b}_u$ is the bias of each user, and λ is the regularisation parameter

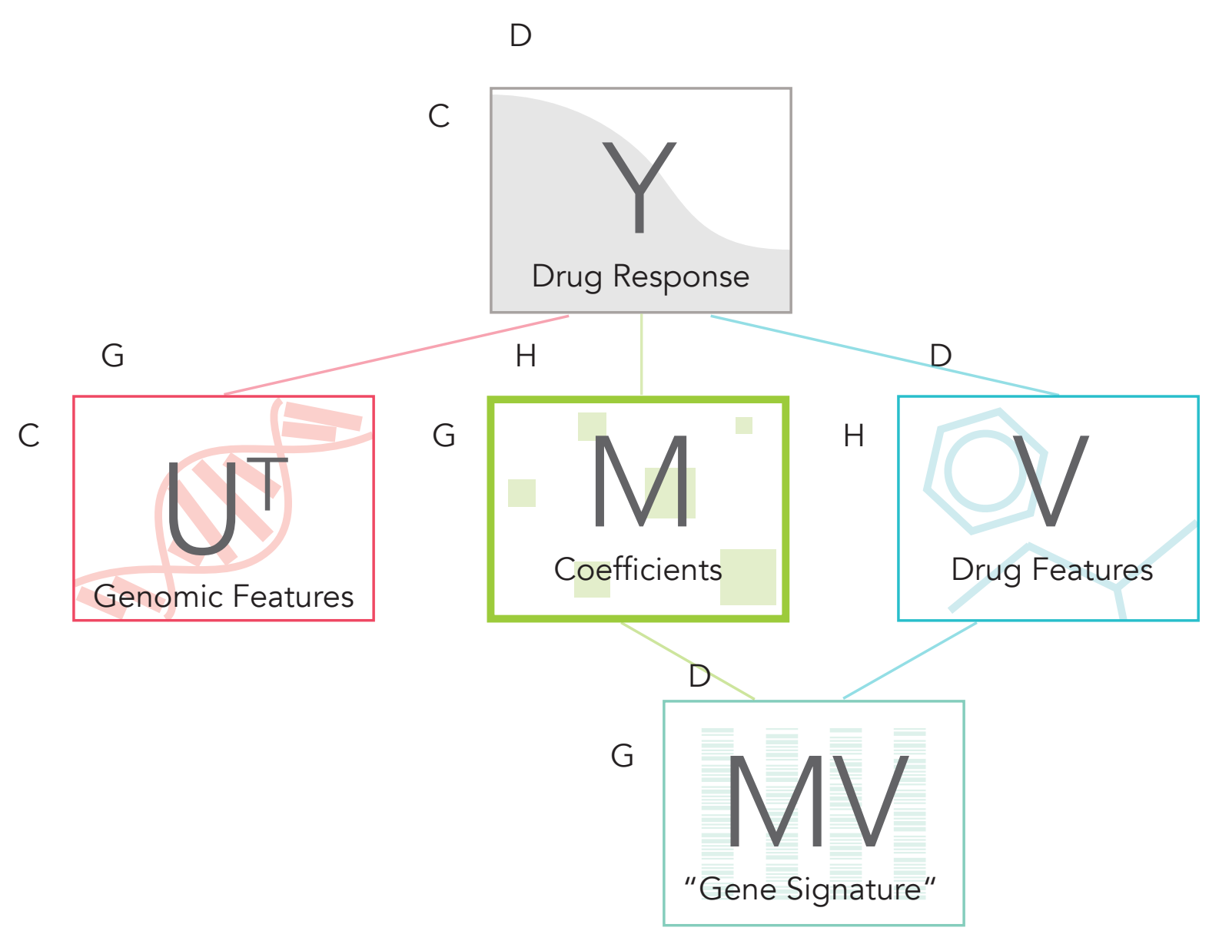
Molecular Holograms



Model

$$\underset{M}{\arg\min} \|Y - U^\top MV\|_2^2 + \lambda\|M\|_1$$

where *Y* is the drug response (to be approximated), *U* is the matrix of the genomic features of each cell line / patient (the genomic regressors), *V* is the matrix of molecular holograms (the molecular regressors), *M* are the coefficients of gene-fragment interaction (to be learned), and λ is the regularisation parameter



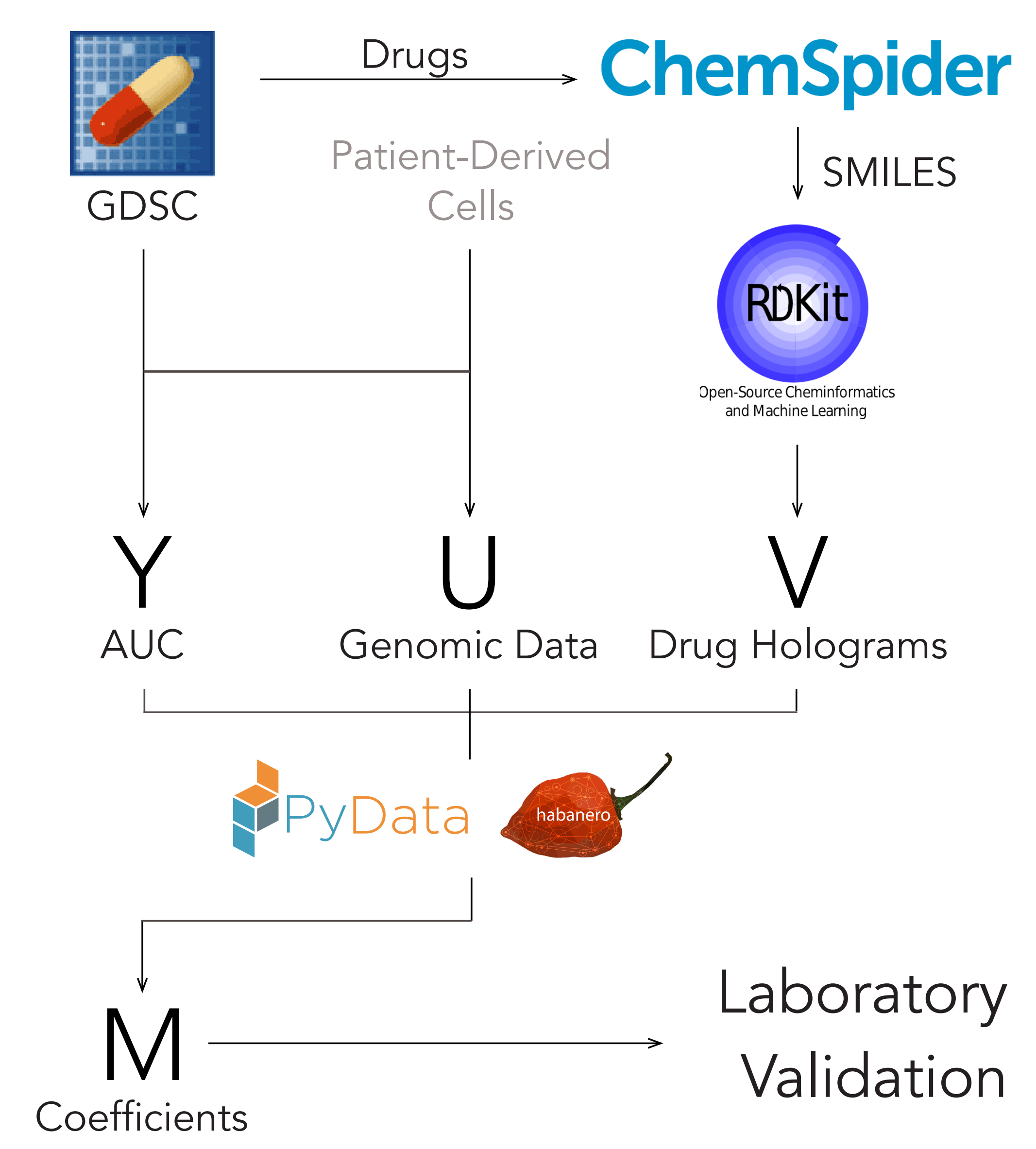
Methodology

Data

Drug response (AUC) from 1,074 cell lines (900 training, 174 eventual testing) across 265 anti-cancer drugs

Genomic data including: TCGA classification, gene expression, methylation, exon mutation, copy number

Workflow



Results

Model Validation

Gene expression dataset
900 cell lines, 265 drugs, 0.240 missing data
Training set: 0.8 split (5-fold CV), Test set: 0.2 split
R^2: 0.975 (Training), 0.974 (Test)
Spearman: 0.495 (Training; 0.216 std), 0.400 (Test; 0.197 std)

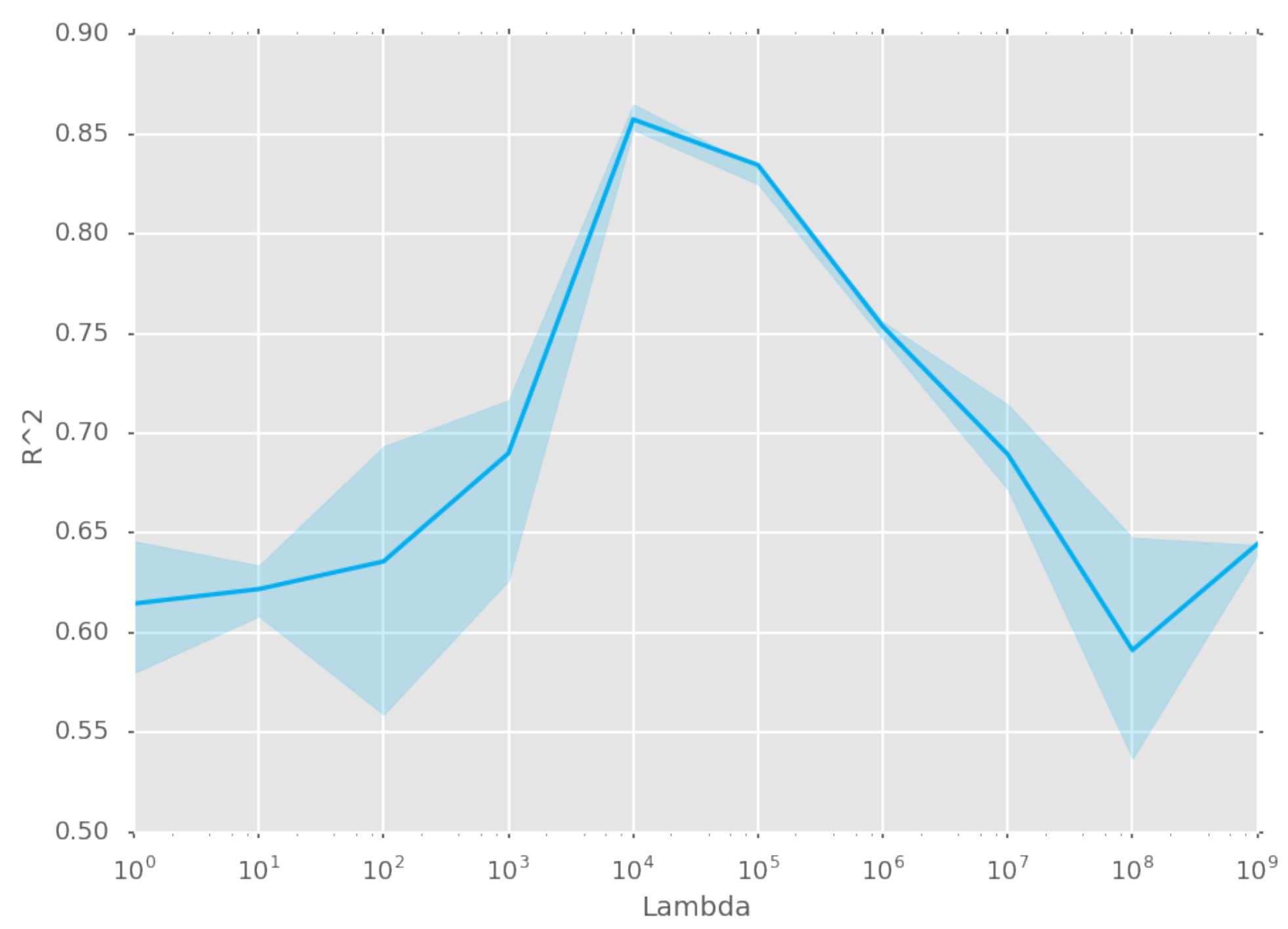
Model for IC_{50} on Copy Number Data

GENE	COEFFICIENT	DRUG NAME	FRAGMENT SMILES
MDM2	-2.336E-04	Nutlin-3a	C(c(c)c)(C(c)N)N(C)C
MDM2	-2.336E-04	Bexarotene	c(cc)(cc)C(=O)O
MDM2	-2.333E-04	Bexarotene	C(=C)(c(c)c)c(c)c
MDM2	-2.330E-04	Bexarotene	c(cc)(cc)C(c)=C
MDM2	2.327E-04	AICAR	c(nc)(C(N)=O)c(n)N

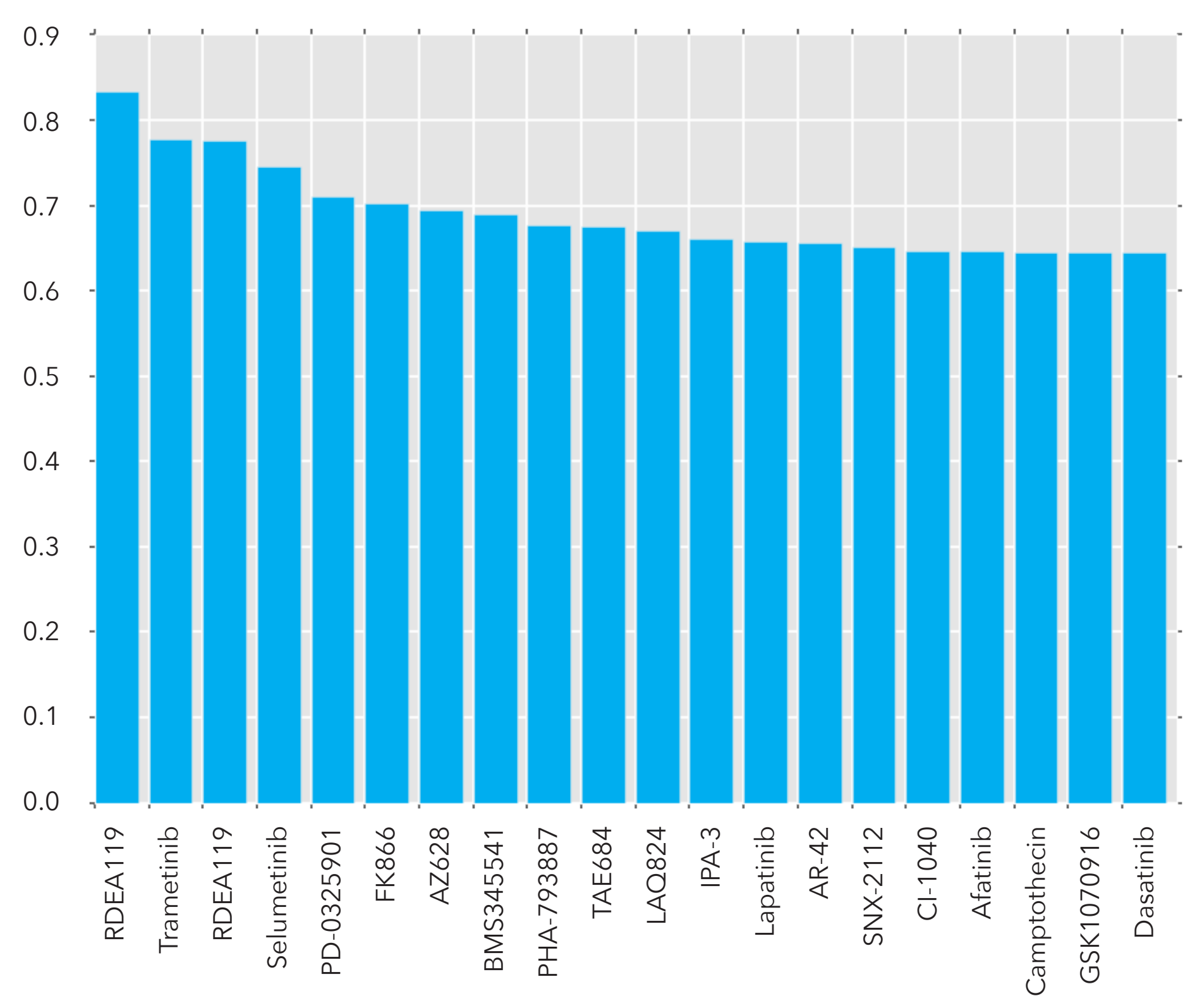
The interaction model identifies bexarotene, AICAR, and Nutlin-3A as the top scoring drugs when amplification of MDM2 is introduced as the cell feature in the model

(N.B. lower IC_{50} implies a better drug response)

Model Selection Curve for AUC on RACs Data



Spearman rank-order correlation per drug



Top 20 Spearman rank-order correlation scores on held-out data for interaction model trained on gene expression data for 900 cell lines across 265 drugs; all correlations are significant (p < 1e-8)

Discussion

More interpretable than random forest (feature importances vs. interaction coefficients)

Better able to model biological complexity than linear models

Outperforms existing models in published literature

Integration of drug molecular (QSAR) features allows for modelling of relationships between different drugs; rarely seen in literature

Future Work

Train models for datasets on copy number, methylation, etc.

Integrate models trained on different genomic data types via NLP for robustness

Integrate physical molecular descriptors (e.g. log *p*, melting point, boiling point, molar mass, etc.) into drug QSAR

Laboratory validation of newly-discovered drug-gene interactions

Acknowledgements

We would like to thank Santiago Vilar, Wang Jiguang, Mykola Bordyuh, Wu Zikai, and Iñigo Urteaga for generously lending us use of their grey matter.

Bibliography

Rdkit: Open-source cheminformatics. URL http://www.rdkit.org.

Francesco Iorio et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

Michael P Menden et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4):e61318, 2013.

Linh Nguyen, Cuong Dang, and Pedro Ballester. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *bioRxiv*, page 095224, 2016.

Harry E Pence and Antony Williams. Chemspider: an online chemical information resource, 2010.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors, volume 11*. John Wiley & Sons, 2008.

Image Credits

RDKit logo: goo.gl/zXYO6o

ChemSpider logo: http://www.chemspider.com/

GDSC logo: http://www.cancerrxgene.org/crx/gfx/cancerrxgene_logo.png

PyData logo: https://a248.e.akamai.net/secure.meetupstatic.com/photos/event/d/0/9/a/highres_447653402.jpeg

Habanero logo: https://s24.postimg.org/ekrtiru2d/site.jpg