P16

GENerateZ: Automatic De Novo Design of Anticancer Drugs using Transcriptomic Data, Genetic Algorithms and Variational Autoencoders

Hans W. A. Hanley, Garrett M. Morris

UNIVERSITY OF OXFORD

DEPARTMENT OF STATISTICS

24-29 St. Giles Department of Statistics, OX1 3LB, University of Oxford, Oxford, UK

Introduction

We propose a novel machine learning architecture and technique for de novo drug discovery of anti-cancer drugs by using discrete representations of drugs' chemical compositions and the transcriptomics of targets. In particular, we generate novel compounds optimized for high efficacy against specific types of cancerous cells.

Important Chemical Properties

QED- Quantitative Estimation of Drug-likeness or QED is a score used to evaluate a drug-like compound's favourability.

SAS- Synthetic Accessibility Score measures the ease of synthesizing a particular compound.

IC50- The half-maximal inhibitory concentration or IC50 is a measure of a drug's efficacy. It measures how much of a drug is needed to inhibit a biological process by 50%

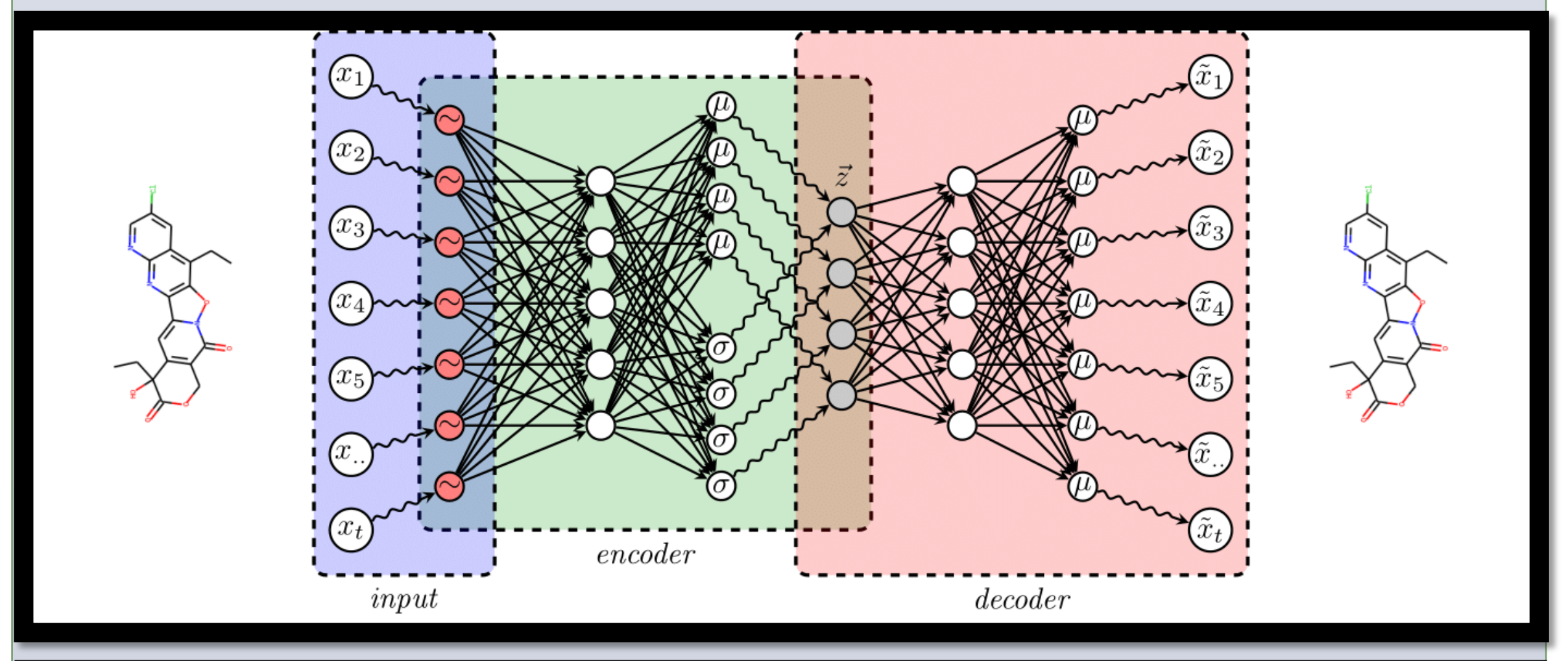
VAE: Modelling Compounds

We use **SMILES [1]**, **DeepSMILES [2]**, and **SELFIES [3]** to represent compounds

- **SMILES**: a line notation for representing molecules and reactions
- **DeeepSMILES**: syntax uses only close parentheses. DeepSMILES also only uses a single symbol for ring closures.
- **SELFIES**: an alternative string-based representations of molecular graphs that are 100% robust. Namely, each SELFIES string corresponds to a valid molecule

Chemical	SMILES	DeepSMILES	SELFIES
Carbon Dioxide	O=C=O	O=C=O	[O][=C][=O]
Benzene	c1ccccc1	cccccc6	[c][c][c][c][c][c][Ring1][Branch1_1]

We use a Variational Autoencoder with cyclical annealing to model chemical compounds in a continuous latent space.



Chemistry Datasets

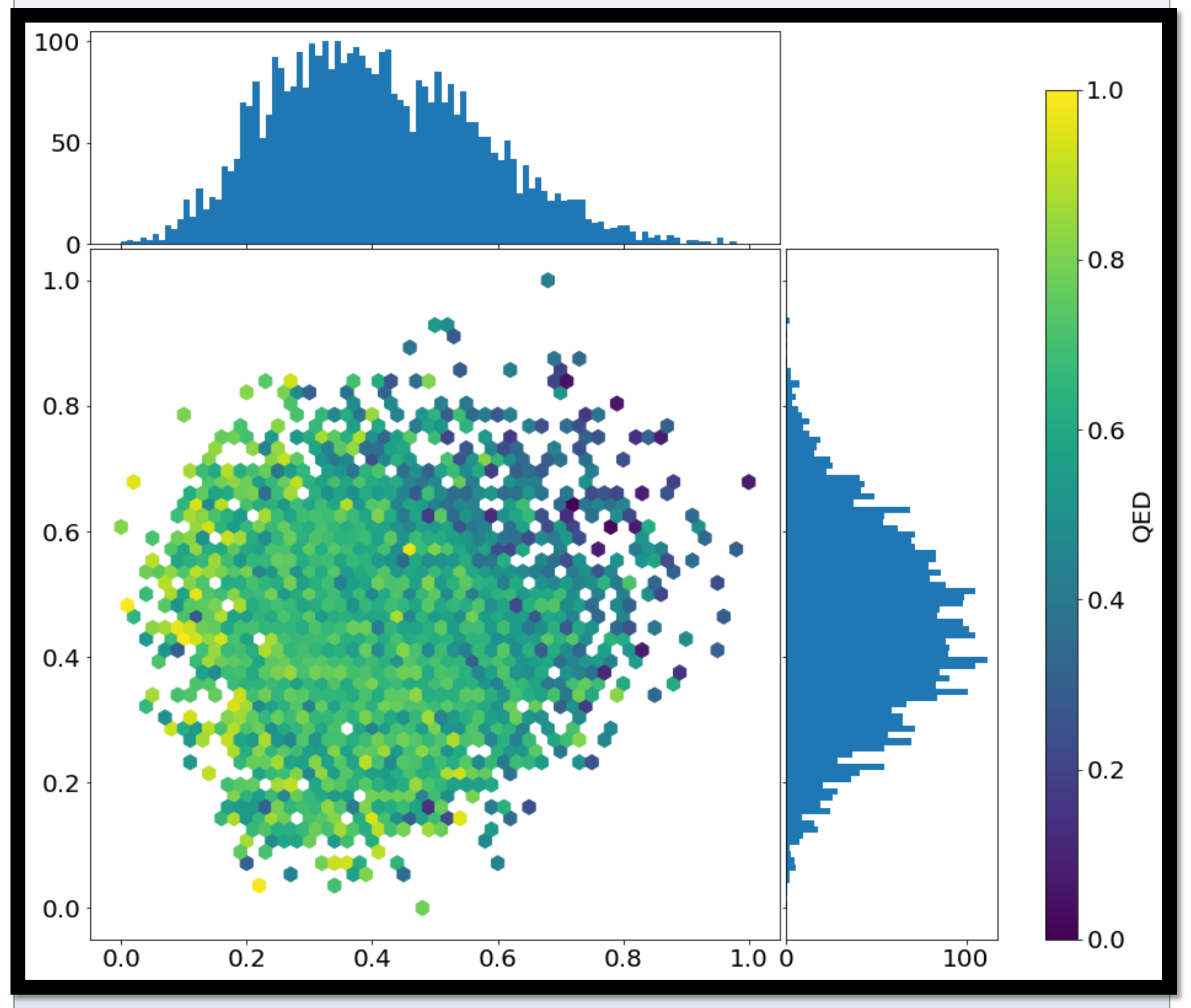
ChEMBL [4]- dataset contain 1.9M different bioactive compounds

ZINC [5]- dataset containing 254K different bioactive compounds

GDSC [6]- dataset containing data on how different anticancer compounds affect the transcriptome of different cell lines.

Information-Density of Latent Space using PCA

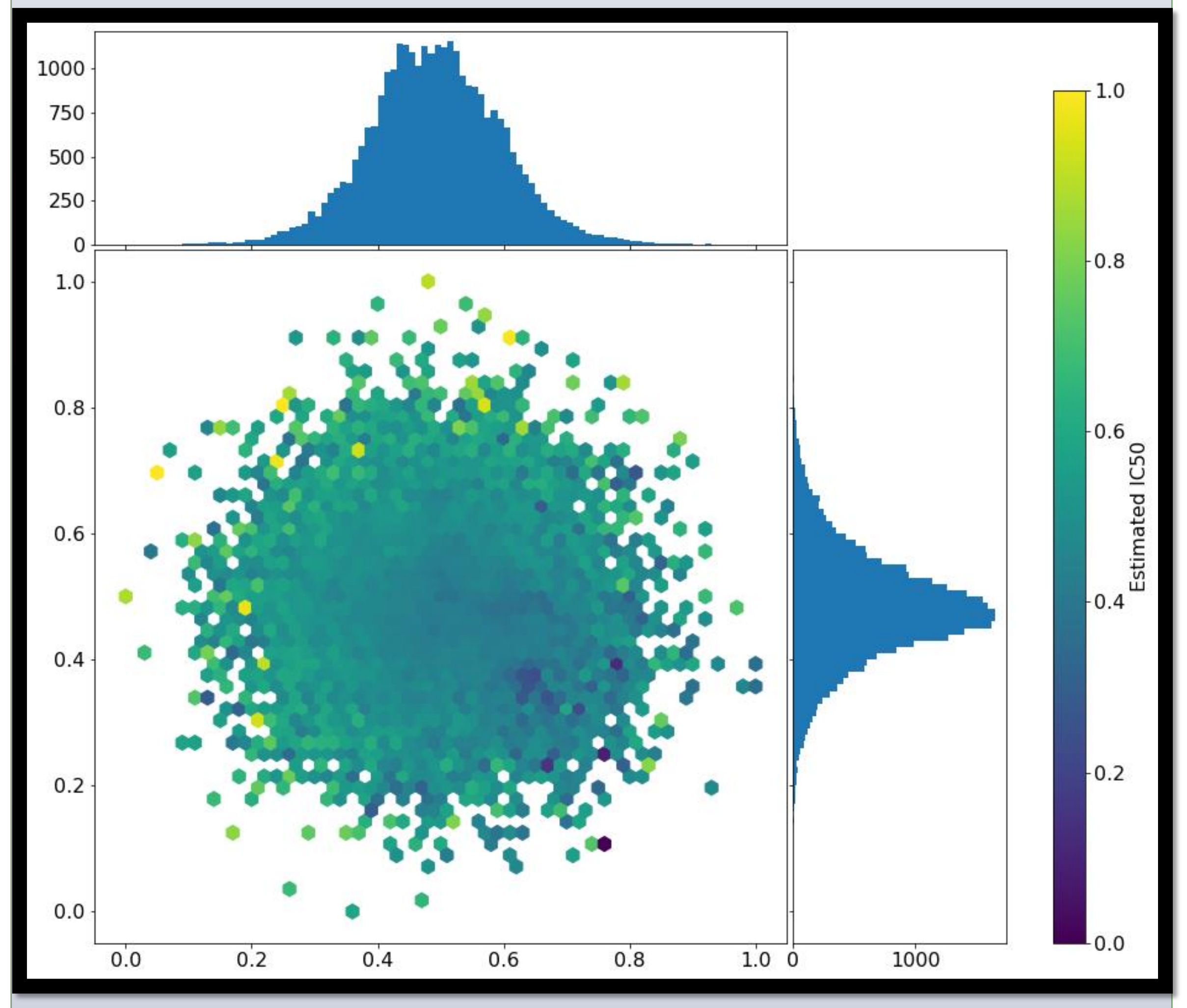
QED



Hexbin plots of mean QED of 4000 random chemical compounds from the ChEMBL test set after projecting using linear PCA the latent representations of the cyclically annealed VAE.

Latent Space Without Shaping Against Esophageal Cancer Cell Transcriptome from GDSC [6]

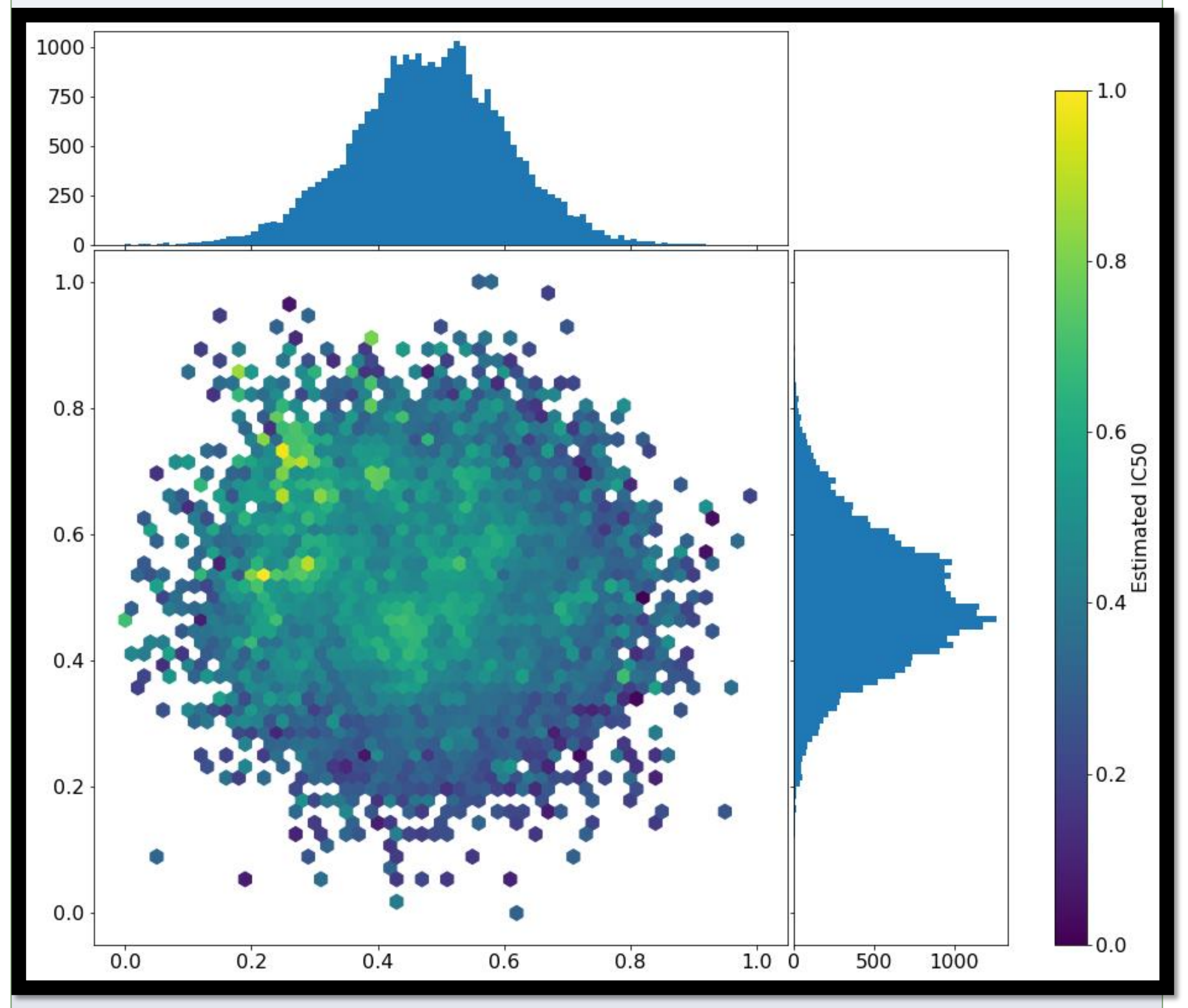
Normalized Estimate of log IC50



Normalized [0,1] log IC50 values for chemical compounds after projecting using linear PCA against the UMC-11 cell line, a cell of a carcinoid-endocrine tumour affecting the lung.

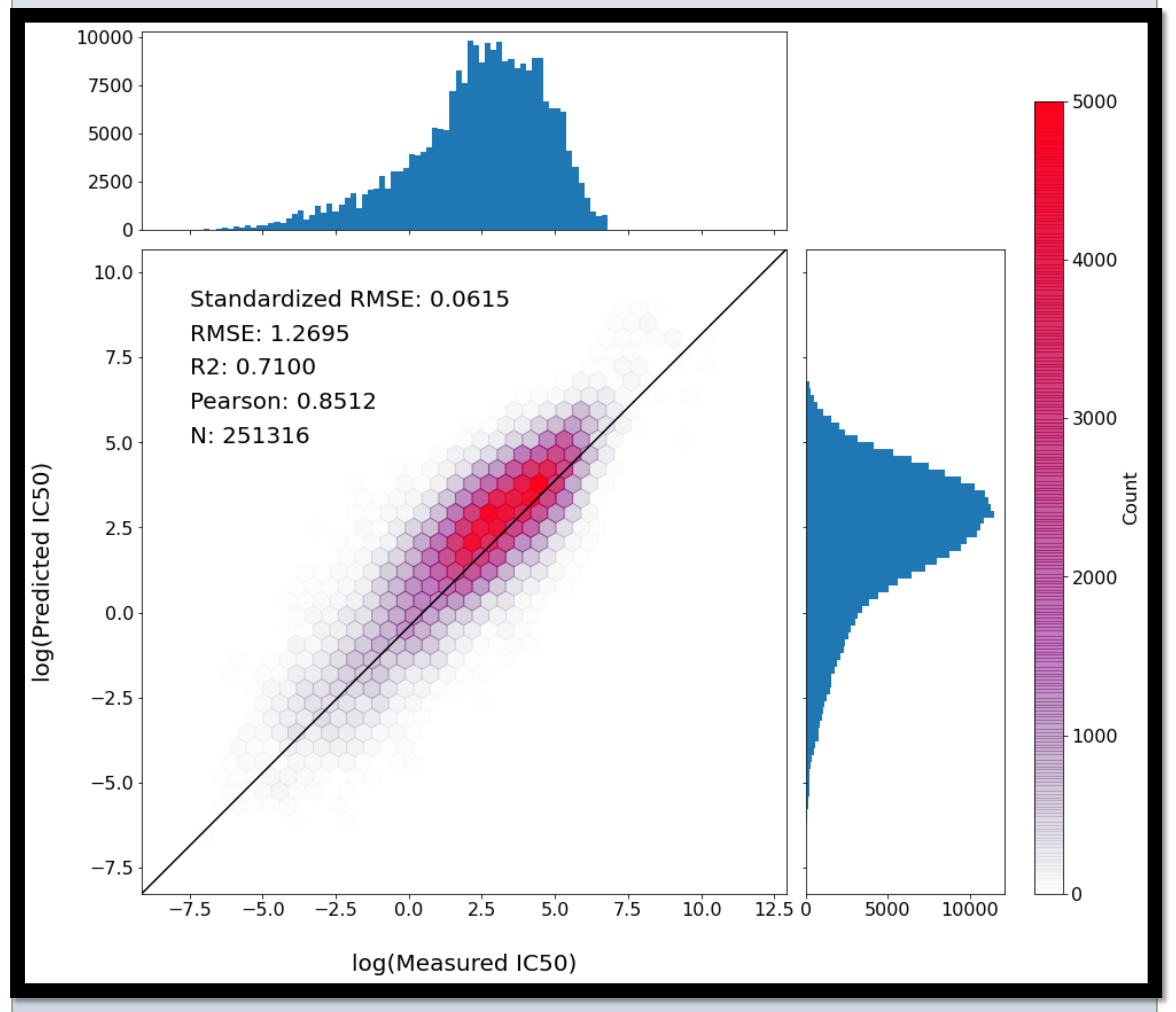
Shaping The Latent Space Against Esophageal Cancer Cell Transcriptome from GDSC [6]

Normalized Estimate of log IC50



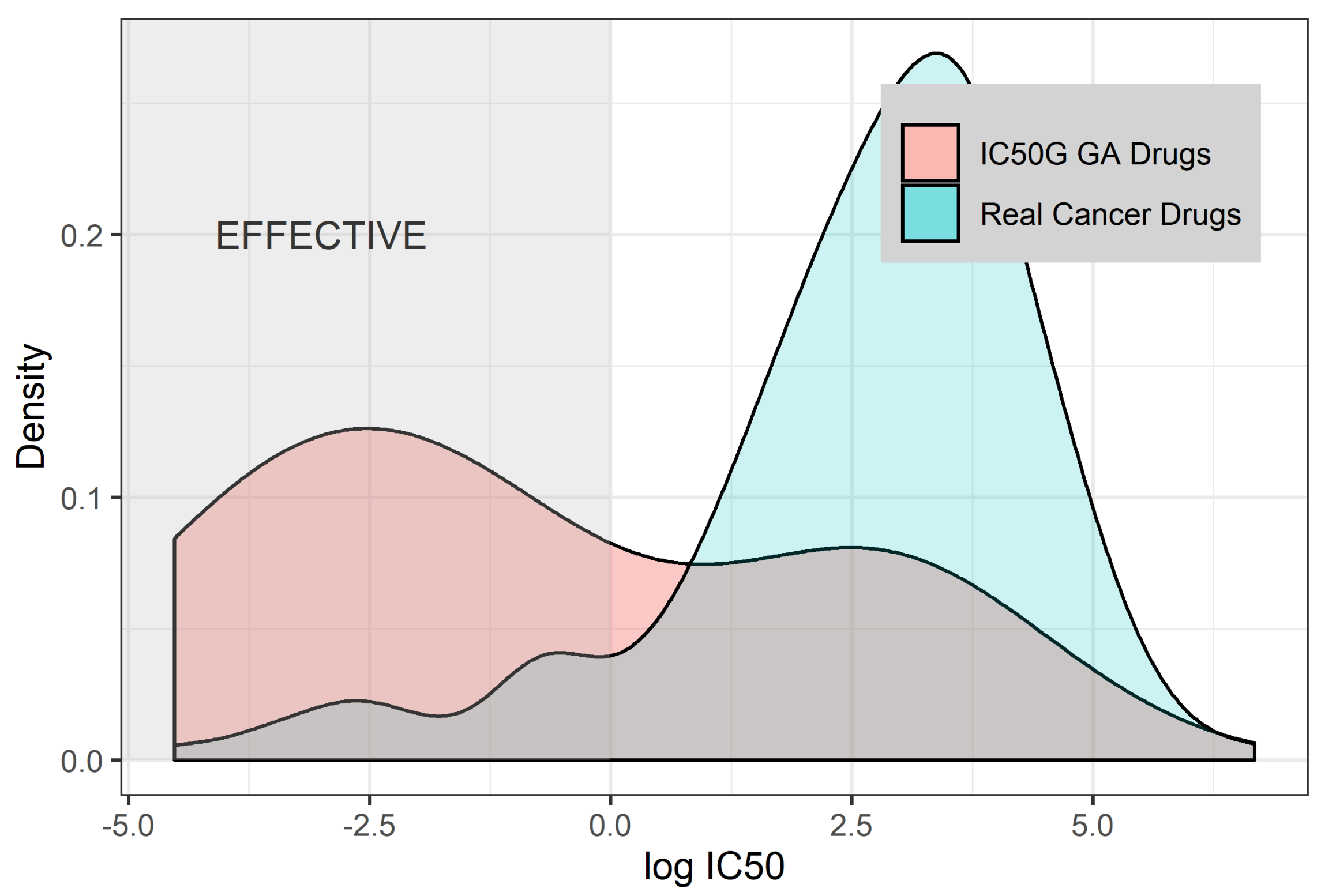
Normalized [0,1] log IC50 values for chemical compounds after projecting using linear PCA against the UMC-11 cell line, a cell of a carcinoid-endocrine tumour affecting the lung.

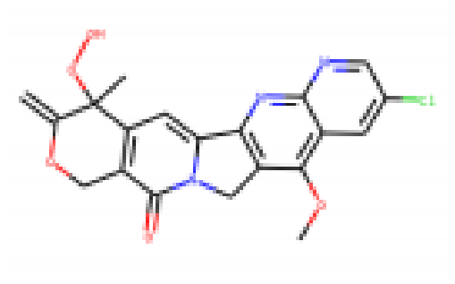
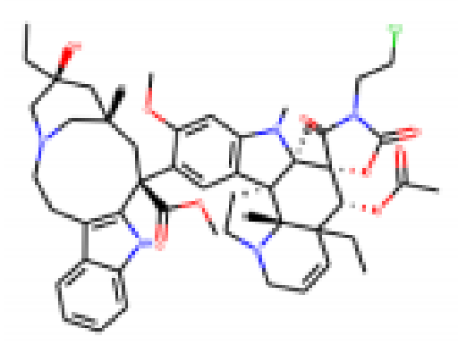
Prediction of log IC₅₀ Using Shaped Latent Embeddings



Prediction of log IC50 using transcriptomic data and SELFIES latent embeddings with IC50 latent shaping. The model was fitted in log space. RMSE was calculated after normalizing log IC50 on a [0,1] scale.

Generating New Compounds Using Genetic Algorithm



Discovered Molecule	Closest Approved Drug
 CCC1=C2C=C(C=NC2=NC3=C1ON4C3=CC5=C(C4=O)COC(=O) C5(CC)O)Cl Est $\log_{10}IC_{50}$: -4.80 (0.01) QED: 0.51 logP: 2.35 SAS: 3.40 Tox Prob: 0.11	 VINZOLIDINE c12[C@@]34[C@H]([C@]5([C@H](OC(C)=O)([C]6([C@@H]3([N] (CC=C6)CC4))CC)) Est $\log_{10}IC_{50}$: 1.99 (0.97) QED: 0.13 logP: 5.04 SAS: 7.65

Conclusions

We found that modelling compounds using VAEs was highly effective. We managed to elicit QED and SA scores implicitly in our latent space from compounds by utilizing cyclical annealing. We see we can shape our latent space using the transcriptomics of cells; we can also effectively predict IC50 using these latent vectors. We finally see that our approach can generate a host of unique compounds that are tailored for specific cell lines and types of cancer.

References

1. SMILES. Daylight Theory: SMILES. URL: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
2. Noel O'boyle and Andrew Dalke. "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures". In: (). DOI: 10.26434/chemrxiv.7097960.v1. URL: https://github.com/nextmovesoftware/deepsmiles
3. Mario Krenn et al. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. Tech. rep. arXiv: 1905.13741v2. URL: https://github.com/.
4. A. Patrícia Bento et al. "The ChEMBL bioactivity database: An update". In:Nucleic AcidsResearch42.D1 (Jan. 2014), pp. D1083–D1090.ISSN: 03051048.DOI:10.1093/nar/gkt1031.URL:https://academic.oup.com/nar/article/42/D1/D1083/1043509.
5. John J. Irwin and Brian K. Shoichet. "ZINC – A Free Database of Commercially Available Compounds for Virtual Screening". In:Journal of chemical information and modeling45.1(2005), p. 177
6. Francesco Iorio et al. "A Landscape of Pharmacogenomic Interactions in Cancer". In: Cell 166.3 (July 2016), pp. 740–754. ISSN: 10974172. DOI: 10.1016/j.cell.2016.06.017. URL: https://pubmed.ncbi.nlm.nih.gov/27397505/.