pdCSM-cancer: using graph-based signatures to identify small molecules with anticancer properties

Raghad Al-Jarf[1,2,3], Alex G.C. de Sá[1,2,3,4], Douglas E.V. Pires[1,2,3,5*], David B. Ascher[1,2,3,6*]

[1]Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne,Melbourne, Victoria, Australia

[2]Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia

[3]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

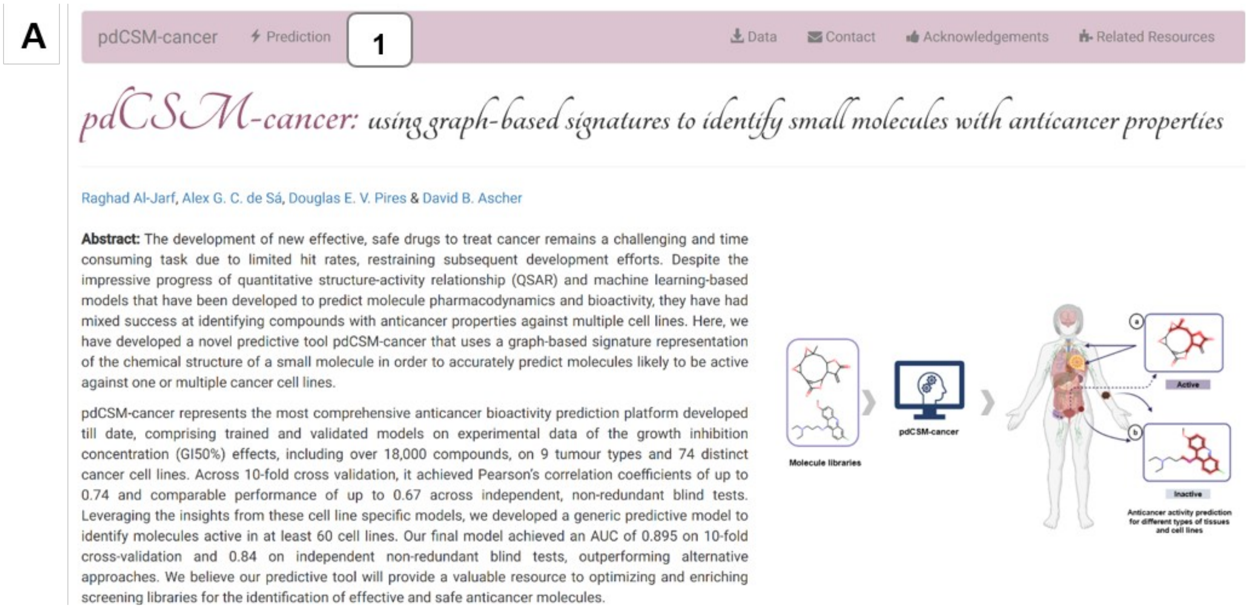
[4]Baker Department of Cardiometabolic Health, Melbourne Medical School, University of Melbourne, Melbourne, Victoria, Australia

[5]School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria, Australia

[6]Department of Biochemistry, University of Cambridge, 80 Tennis Ct Rd, Cambridge CB2 1GA

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

Main Page



pdCSM-cancer: using graph-based signatures to identify small molecules with anticancer properties

Raghad Al-Jarf, Alex G. C. de Sá, Douglas E. V. Pires & David B. Ascher

Abstract: The development of new effective, safe drugs to treat cancer remains a challenging and time consuming task due to limited hit rates, restraining subsequent development efforts. Despite the impressive progress of quantitative structure-activity relationship (QSAR) and machine learning-based models that have been developed to predict molecule pharmacodynamics and bioactivity, they have had mixed success at identifying compounds with anticancer properties against multiple cell lines. Here, we have developed a novel predictive tool pdCSM-cancer that uses a graph-based signature representation of the chemical structure of a small molecule in order to accurately predict molecules likely to be active against one or multiple cancer cell lines.

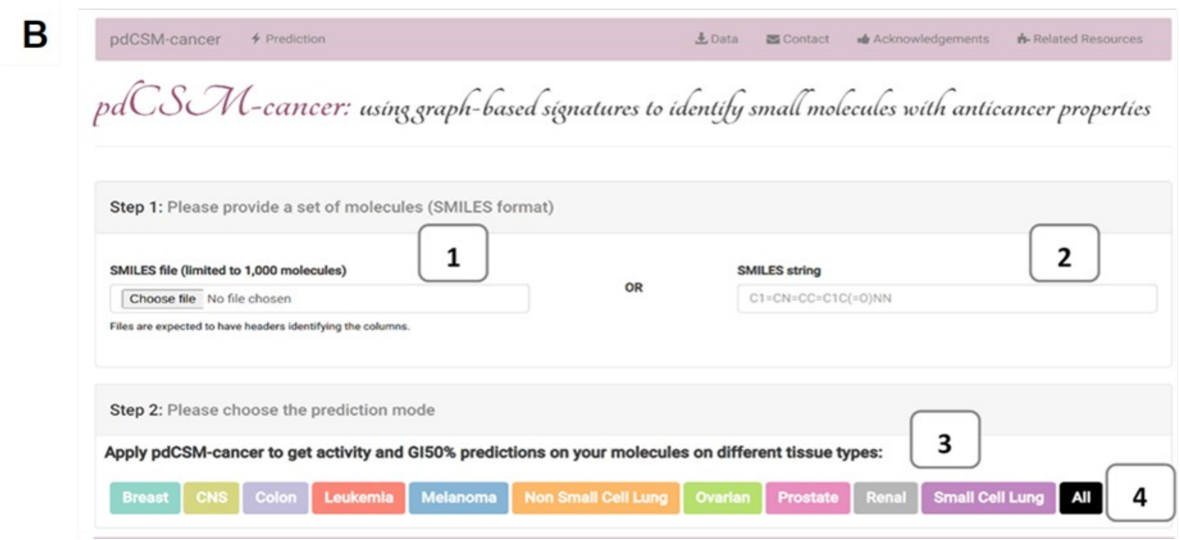
pdCSM-cancer represents the most comprehensive anticancer bioactivity prediction platform developed till date, comprising trained and validated models on experimental data of the growth inhibition concentration (GI50%) effects, including over 18,000 compounds, on 9 tumour types and 74 distinct cancer cell lines. Across 10-fold cross validation, it achieved Pearson's correlation coefficients of up to 0.74 and comparable performance of up to 0.67 across independent, non-redundant blind tests. Leveraging the insights from these cell line specific models, we developed a generic predictive model to identify molecules active in at least 60 cell lines. Our final model achieved an AUC of 0.895 on 10-fold cross-validation and 0.84 on independent non-redundant blind tests, outperforming alternative approaches. We believe our predictive tool will provide a valuable resource to optimizing and enriching screening libraries for the identification of effective and safe anticancer molecules.

About pdCSM-cancer

pdCSM-cancer is a machine learning platform that uses a graph-based signature representation of the chemical structure of a small molecule to accurately predict molecules likely to be active against one or multiple cancer cell lines, as well as pharmacodynamics properties. The platform consists of 74 regression models and a general classification model. These models were trained and tested on different experimental data sets of molecules with anticancer properties, tested against nine distinct tissue(tumor) types, including Breast, central nervous system, Colon, Leukemia, Prostate, Renal, Lung, Melanoma, and Ovarian.

(A) Depicts the main page of pdCSM-cancer. Users are directed to the submission page by clicking on "**Prediction**" at the top menu **(1).**

Submission Page



pdCSM-cancer: using graph-based signatures to identify small molecules with anticancer properties

Step 1: Please provide a set of molecules (SMILES format)

SMILES file (limited to 1,000 molecules)

Choose file No file chosen

Files are expected to have headers identifying the columns.

OR

SMILES string

C1=CN=CC=C1C(=O)NN

Step 2: Please choose the prediction mode

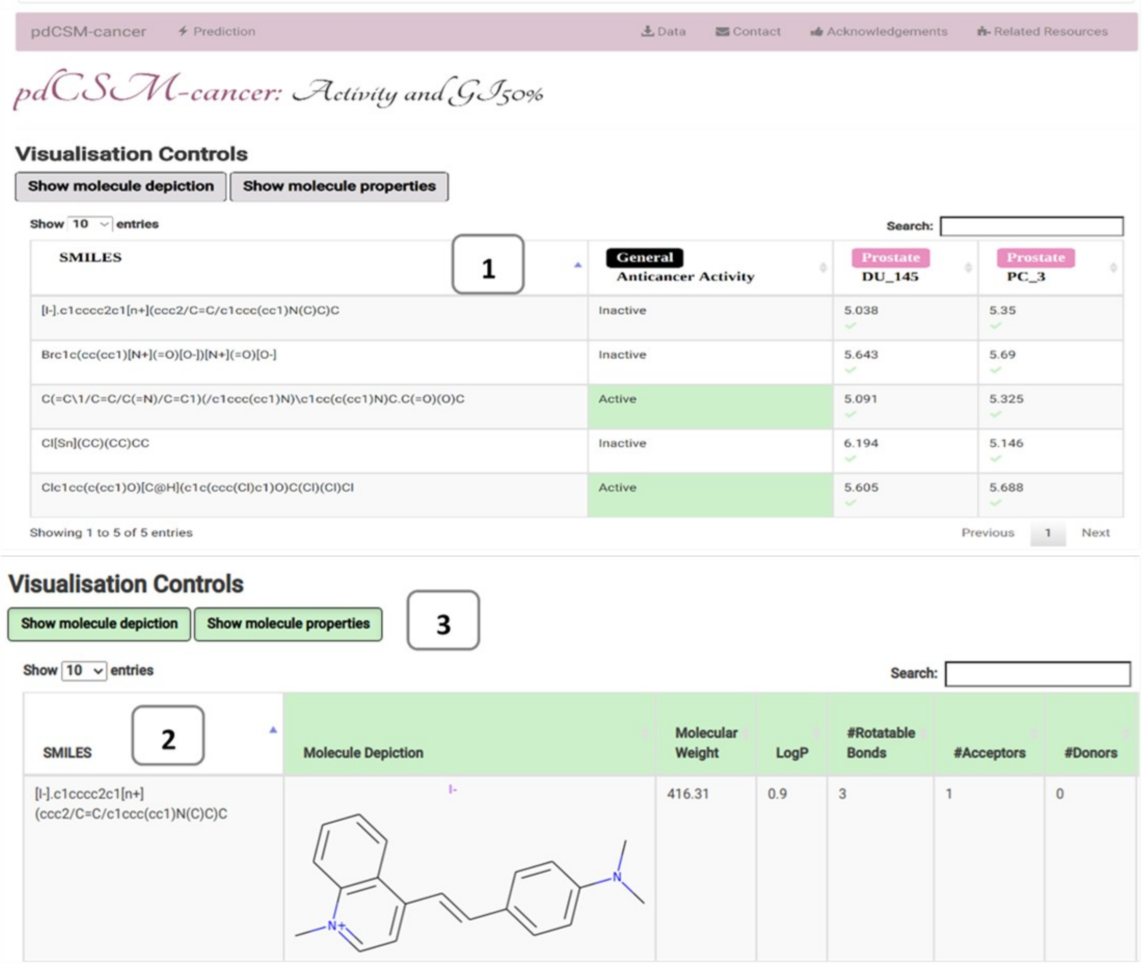
Apply pdCSM-cancer to get activity and GI50% predictions on your molecules on different tissue types:

Breast | CNS | Colon | Leukemia | Melanoma | Non Small Cell Lung | Ovarian | Prostate | Renal | Small Cell Lung | All

(B) represents the submission page. Users can either submit a set of compounds as a SMILES file **(1)** or an individual compound as a SMILES string **(2)**. Users have the options to either choose different prediction modes according to their tissue of interest **(3)** or they can choose to run all tissues **(4).**

Results Page

C



pdCSM-cancer: Activity and GI50%

Visualisation Controls

Show 10 entries

Search:

SMILES	General Anticancer Activity	Prostate DU_145	Prostate PC_3
[I-].c1cccc2c1[n+](ccc2/C=C/c1ccc(cc1)N(C)C)C	Inactive	5.038 ✓	5.35 ✓
Brc1c(cc(cc1)[N+](=O)[O-])[N+](=O)[O-]	Inactive	5.643 ✓	5.69 ✓
C(=C\1/C=C/C(=N)/C=C1)(/c1ccc(cc1)N)\c1cc(c(cc1)N)C.C(=O)(O)C	Active	5.091 ✓	5.325 ✓
Cl[Sn](CC)(CC)CC	Inactive	6.194 ✓	5.146 ✓
Clc1cc(c(cc1)O)[C@H](c1c(ccc(Cl)c1)O)C(Cl)(Cl)Cl	Active	5.605 ✓	5.688 ✓

Showing 1 to 5 of 5 entries

Previous 1 Next

Visualisation Controls

Show 10 entries

Search:

SMILES	Molecule Depiction	Molecular Weight	LogP	#Rotatable Bonds	#Acceptors	#Donors
[I-].c1cccc2c1[n+](ccc2/C=C/c1ccc(cc1)N(C)C)C		416.31	0.9	3	1	0

After choosing the prediction mode of interest, users will be redirected to a results page **(C)** where predictions for all 74 cancer cell lines (9 tissue types) specific models, anticancer activity (GI50%), general anticancer model **(1),** physiochemical properties and molecular depiction are presented in tabular format **(2)**. Users have the options to either show or hide the molecule properties and depiction **(3)**.