Vir2Drug: Supplementary Content

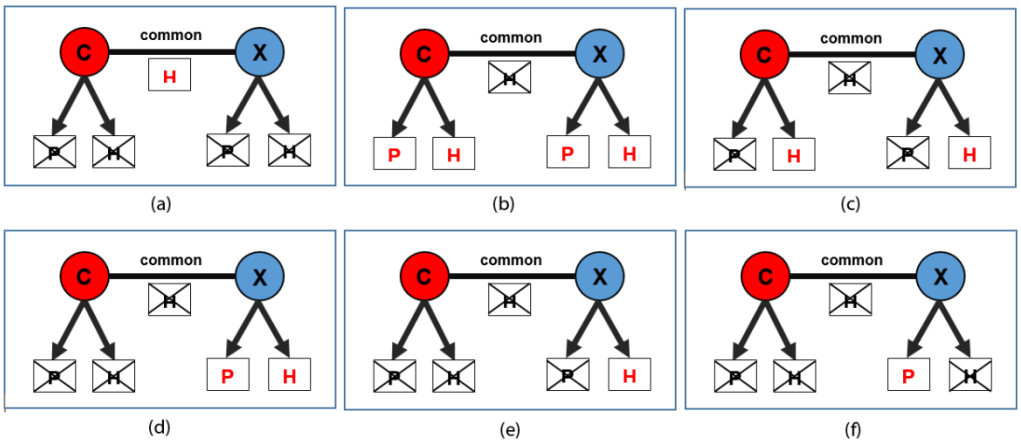
In this supplementary file we describe in details the drug screening and scoring processes mentioned in the manuscript.

THE DRUG SCREENING INTERMEDIATE PROCESS

The drug screening process is an intermediate process that aims to screen drugs based on the edge properties of the P2P network described in the manuscript, and to further provide a separate drug list for each edge in the network. Drug screening is performed by means of two diverse approaches: (1) the PPI network-based drug screening process which comes with 6 diverse methodologies that draw from the PPI information of the pathogens involved in the P2P network, and (2) the taxonomy-based drug screening process which is based on the taxonomic distance in-between two pathogens that form a specific edge. These read as follows:

PPI network-based drug screening process

This is a generalized methodology which draws from both the pathogens and the host PPI interactions of the two pathogens that form an edge. Specifically, the method examines all the edges in the P2P network and brings drugs which target specific proteins of the two connected pathogens. However, at this stage of analysis, someone may find several PPI combinations that may change significantly the outcome of the drug screening process. Specifically, there are several drugs that target directly the pathogen's proteins rather than the host-proteins a specific pathogen interacts with. For example, the *Tyrosine-protein kinase transforming protein Abl (P00521)* of the *Abelson murine leukemia virus* is targeted by four drugs according to the DrugBank repository. These include: (1) PD173955, (2) 2-(N-morpholino)ethanesulfonic acid, (3) N-[4-Methyl-3-[[4-(3-Pyridinyl)-2-Pyrimidinyl]Amino]Phenyl]-3-Pyridinecarboxamide, and (4) ACA 125. On the contrary, the host proteins the specific virus interacts with are not targeted by any specific drug. In that case, the protein commonality concept especially at host-protein level may not be an optimal approach since there are only direct candidate drugs for this virus. In order to bypass such inconsistencies and to further optimize the drug screening process, in the following subsections we propose six diverse methodologies for drug screening, focusing on: (1) either host proteins or pathogen's proteins included in the pathogen under study as well as in the X connected pathogen, and (2) either common host proteins or common pathogen's proteins shared in between two pathogens. Figure 1, depicts the way in which those methodologies draw information from the two pathogens (nodes) that form an edge.



Supplementary Figure 1: Two pathogens (nodes) that form an edge according to common proteins. The letter P, refers to the proteome. The letter H: refers to the host-proteins of the specific pathogen. The C refers to the candidate pathogen under study. The letter X refers to the X-pathogen that shares its proteins with the C pathogen. The red node refers to the pathogen under study. The red marked letters refer to the protein set where the specific methodology draws from.

COMMON" network scan scheme

This scheme searches for drug targets by using only the common pathogen's or host proteins between two pathogens that form an edge in the network, as shown in Figure 1a. This is a limited process that aims to restrict (and/or filter) the deriving drugs focusing only on the common subset of proteins in between two pathogens. It is recommended especially in cases where there is a large number proteins interacting with the pathogen under study, which in effect may lead to noisy and large sets of deriving drugs.

"ALL" network scan scheme

This scheme searches for drug targets by using both the pathogen's and the host proteins between two pathogens that form an edge in the network, as shown in Figure 1b. This is an unlimited/unrestricted process that aims to bring any candidate drug targeting both the pathogen's and the X-pathogen's proteins, accordingly. This method is recommended when there is a limited number of drugs that target the proteins of both connected pathogens.

"BOTH-HOST" network scan scheme

This scheme searches for drug targets by using only the host proteins of the two pathogens that form an edge in the network, as shown in Figure 1c. This is also an unlimited/unrestricted process that focuses only on the host protein sets of both pathogens.

"V2-ALL" network scan scheme

This scheme searches for drug targets by using the pathogen's proteome and the host proteins of the X-pathogen which is connected with the pathogen under study, as shown in Figure 1d. This is a subset of the BOTH-HOST scheme and it is recommended when: (a) there is not any candidate drug that targets the protein set of the pathogen under study, and (b) users want to bring drugs using only the protein sets of the X-pathogen.

"V2-HOST" network scan scheme

This scheme searches for drug targets by using only the host proteins of the X-pathogen which is connected with the pathogen under study, as shown in figure 1e. This is a subset of the V2-ALL scheme and it is recommended when: (a) there is not any candidate drug that targets the protein set of the pathogen under study, and (b) users want to bring drugs using only the host-protein set of the X-pathogen.

"V2-PROT" network scan scheme

This scheme searches for drug targets by using only the pathogen's proteins (proteome) of the X-pathogen which is connected with the pathogen under study, as shown in figure 1f.

Taxonomy based drug screening

The taxonomy of an organism is another approximation of its similarity to other organisms in terms of genetics, molecular functions and morphology. Therefore, the underlying methodology has been based on the assumption that drugs with a direct inhibitory effect against a given pathogen are more likely to have a similar effect to closely related pathogens in terms of taxonomic distance. Furthermore, drugs with pathogen protein

targets across a broad and diverse range of taxonomy distances, are expected to have a broader inhibitory effect rendering them more repurposable, as opposed to drugs which target only a specific distant group of pathogens. This broad spectrum anti-pathogenic activity can be simply captured by the maximum distance D and the entropy of the taxonomic distances between all target pathogens of a given drug. For this we extracted the NCBI taxonomy IDs of all the pathogens contained in DrugBank repository, having at least one protein that is targeted by a drug. The proposed method examines all the edges in the P2P network to bring drugs that target the proteins of the pathogen of interest.

THE DRUG SCORING INTERMEDIATE PROCESS

Each drug contained in the lists obtained from the drug screening process is further ranked by means of specific equation that draws from the proteins included in each pair of pathogens that form the specific edge. This is an initial ranking process that aims to score a deriving drug by means of the edge and the node content of the P2P network. Due to the plethora of drug information that could be obtained from the targeted proteins contained in the two pathogens that form an edge, we propose eight diverse equations for this type of ranking, each one using a different set of the following parameters:

$\mathbf{N_{vtar}}$, is the number of drug targets obtained from the pathogen's host-proteins.
$\mathbf{N_{xtar}}$, is the number of drug targets obtained from the x-pathogen host proteins.
$\mathbf{N_{vptar}}$, is the number of drug targets obtained from the pathogen's proteome.
$\mathbf{N_{xptar}}$, is the number of drug targets obtained from the x-pathogen's proteome.
$\mathbf{N_1}, \mathbf{N_2}$, are the total number of host-proteins included in pathogen and the x-pathogen accordingly.
$\mathbf{N_v}, \mathbf{N_x}$, are the total number of proteins included in the pathogen and the x-pathogen proteome accordingly.
$\mathbf{N_{dtar}}$, is the total number of drug targets that derive from DrugBank repository.
$\boldsymbol{D}$, is the taxonomic distance in between the two pathogens that form a specific edge in the P2P network.
In the following paragraphs we describe the proposed equations, providing further arguments to support this attempt.

4PH-fold drug scoring equation

This is a generalized equation that accounts for all the possible drugs targeting both the pathogen's proteome and its host-protein interactions. The equation uses four equally weighted rates (fractions) that account for both the pathogen under study and the X-connected pathogen that form the specific edge in the P2P network, as follows:

$$S_{drug} = 0.25\left(\frac{N_{vtar}}{N_1} + \frac{N_{xtar}}{N_2} + \frac{N_{vptar}}{N_v} + \frac{N_{xptar}}{N_x}\right) \quad (S1)$$

4H-fold drug scoring equation

The 4H-fold equation is an extended equation which can be used to bring drugs that target only the host-proteins of both pathogens that form an edge in the P2P network. The underlying equation can be used in cases where both pathogen proteins (proteomes) have not been targeted by any drug or have not been of crucial significance. The significance of a drug is further estimated by means of the total drug targets in the DrugBank repository. Specifically, a drug which targets a large number of proteins that are not included in the host-protein set of the two pathogens under study, is more likely to be considered as a wide spectrum drug (less significant) rather

than a candidate drug for the specific pathogen. In this line of thought, the proposed equation uses four equally weighted rates, as follows:

$$S_{drug} = 0.25\left(\frac{N_{vtar}}{N_1} + \frac{N_{vtar}}{N_{dtar}} + \frac{N_{xtar}}{N_2} + \frac{N_{xtar}}{N_{dtar}}\right) \quad (S2)$$

4P-fold drug scoring equation

The 4P-fold equation draws from the same concept described for the 4H-fold equation, but focusing only on drugs which target the pathogen's and the X-Pathogen's proteomes, accordingly. The proposed equation uses four equally weighted rates, as follows:

$$S_{drug} = 0.25\left(\frac{N_{vptar}}{N_v} + \frac{N_{vptar}}{N_{dtar}} + \frac{N_{xptar}}{N_x} + \frac{N_{xptar}}{N_{dtar}}\right) \quad (S3)$$

2PH-fold drug scoring equation

This is a subset of 4H-fold equation that aims to score drugs which target only the proteome and the host-protein set of the pathogen under study. The proposed equation uses two equally weighted rates, as follows:

$$S_{drug} = 0.5\left(\frac{N_{vtar}}{N_1} + \frac{N_{vptar}}{N_{dtar}}\right) \quad (S4)$$

2H-fold drug scoring equation

This is a subset of 4H-fold equation that aims to score drugs which target only the host-protein set of both the pathogens forming an edge in the P2P network. The proposed equation uses two equally weighted rates, as follows:

$$S_{drug} = 0.5\left(\frac{N_{vtar}}{N_1} + \frac{N_{xtar}}{N_2}\right) \quad (S5)$$

2P-fold drug scoring equation

This is a subset of 4H-fold equation that aims to score drugs which target only the pathogen's protein set of both the pathogens forming an edge in the P2P network. The proposed equation uses two equally weighted rates, as follows:

$$S_{drug} = 0.5\left(\frac{N_{vptar}}{N_v} + \frac{N_{xptar}}{N_x}\right) \quad (S6)$$

8-fold drug scoring equation

This is an extension of the generalized 4PH-fold equation that also accounts for all the possible drugs targeting both the pathogen's proteome and its host-proteins. The underlying equation is recommended in cases where there is not any a priori feedback on significant proteins and drugs targeting them. The equation uses eight equally weighted rates (fractions) that account for both the pathogen under study and the X-connected pathogen that form the specific edge in the P2P network, as follows:

$$S_{drug} = 0.125\left(\frac{N_{vtar}}{N_1} + \frac{N_{vtar}}{N_{dtar}} + \frac{N_{xtar}}{N_2} + \frac{N_{xtar}}{N_{dtar}} + \frac{N_{vptar}}{N_v} + \frac{N_{vptar}}{N_{dtar}} + \frac{N_{xptar}}{N_x} + \frac{N_{xptar}}{N_{dtar}}\right) \quad (S7)$$

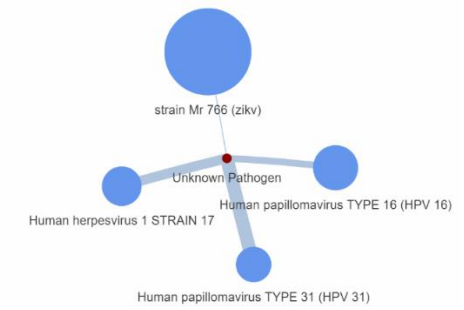
Depending on the target availability the specific equation can be transformed into one of the previous described equations since fractions with zero values do not take part in the overall score estimation.

X-TAXA drug scoring equation

Each drug in the lists obtained by the drug screening process, is ranked by means of the taxonomic distance D in between the two pathogens that form a specific edge in the P2P network.

EVALUATION

Vir2Drug is a tool that can support a plurality of biological scenarios and research questions; thus, the validation criteria are strongly dependent on the rules that clearly define a plausible and reasonable outcome. This in effect makes almost impossible to use validation criteria for any type of these scenarios especially when dealing with drugs where there is not any solid ground truth set of either ranked drugs or ranked proteins. However, in the prospect to validate the tool, in the manuscript we provided three candidate scenarios with promising results which can drive the users to work in these directions. Herein we further examine the possibility for potential outliers (drugs) that may derive using a less grounded scenario. For example, in drugbank repository there is not any drug that targets directly the BRCA1 gene which is strongly connected with types of Breast Cancer. Thus, if someone tries to find drugs for Breast cancer using the BRCA1 as input into Vir2Drug pathogen-based methodologies, this might not be an appropriate scenario. However, dealing with such an insufficient scenario step by step by firstly examine "which pathogens interact with the BRCA1 gene", we will get an interesting protein-based ranked network of pathogens as shown in supplementary figure 2. Herein the unknown pathogen refers to a hypothetical entity that has a highly significant gene, namely the BRCA1. It is observed that the assumed Breast-Cancer-related pathogen, is strongly connected with strains of the Human papillomavirus, the zikv and the Human herpesvirus. Indeed, recent studies have shown that these pathogens could play a role in the genesis of breast neoplasia and other cancer types (1-5).



Supplementary Figure 2: A network of pathogens connected through the BRCA1 gene.

Drawing from the fact that some of the top-rated drugs might be outliers and the BRCA1 is not targeted directly from any available drug, we further used the V2-HOST methodology to screen drugs that target the host proteins of the above-mentioned neighbouring pathogens. Herein potential outliers may derive, if we select methods that account the pathogen's proteome; this will bring mostly antiviral drugs that target the proteome of those viruses. Indeed, by using the [V2-PROT and 4PH-fold] screening and scoring methodologies the 5 top-rated drugs where: Acyclovir, Valaciclovir, Penciclovir, Famciclovir and Foscarnet. These drugs may be considered as outliers since the goal is to find candidate drugs for Breast Cancer and not for specific Viruses. In this line of thought, we used the [V2-HOST and 4H-fold] screening and scoring methodologies, which take into account only the host protein interactions, since we need to avoid screening antiviral drugs that target the pathogen's proteome. Keeping only the 20 top-rated ones as showing in table 1, we can clearly see that six of these drugs (depicted in red color)

are investigational for the treatment of several cancer types; a fact which approves that the underlying list is not a rare artifact but a significant list that should be considered carefully from a pharmacologist.

Supplementary Table 1. List of candidate drugs using V2-HOST and 4H-fold equations for screening and scoring accordingly

#	name	group	drugID	Indication
1	PRLX 93936	investigational	DB06098	Investigated for use/treatment in solid tumors.
2	(Rp)-cAMPS	experimental	DB01790	Not Available
3	Pentanal	experimental	DB01919	Not Available
4	Balanol Analog 2	experimental	DB01940	Not Available
5	3-[(3-sec-butyl-4-hydroxybenzoyl)amino]azepan-4-yl 4-(2-hydroxy-5-methoxybenzoyl)benzoate	experimental	DB02155	Not Available
6	Balanol Analog 1	experimental	DB02611	Not Available
7	Fica	experimental	DB03384	Not Available
8	3-[(5s)-1-Acetyl-3-(2-Chlorophenyl)-4,5-Dihydro-1h-Pyrazol-5-Yl]Phenol	experimental	DB03996	Not Available
9	Balanol	experimental	DB04098	Not Available
10	Monastrol	experimental	DB04331	Not Available
11	Thymectacin	investigational	DB05116	Investigated for use/treatment in colorectal cancer.
12	MCC	investigational	DB05282	Bladder cancer
13	ANX-510	investigational	DB05308	Investigated for use/treatment in breast cancer, colorectal cancer, gall bladder cancer, and pancreatic cancer.
14	OSI-7904L	investigational	DB05457	Investigated for use/treatment in gastric cancer.
15	Filanesib	investigational	DB06040	Investigated for use/treatment in cancer/tumors (unspecified).
16	3-[3-chloro-5-(5-{[(1S)-1-phenylethyl]amino}isoxazolo[5,4-c]pyridin-3-yl)phenyl]propan-1-ol	experimental	DB06897	Not Available
17	3-[3-(3-methyl-6-{[(1S)-1-phenylethyl]amino}-1H-pyrazolo[4,3-c]pyridin-1-yl)phenyl]propanamide	experimental	DB06963	Not Available
18	(2S)-1-{[5-(1H-Indazol-5-yl)-3-pyridinyl]oxy}-3-(7aH-indol-3-yl)-2-propanamine	experimental	DB06977	Not Available
19	(4R)-4-(3-HYDROXYPHENYL)-N,N,7,8-TETRAMETHYL-3,4-DIHYDROISOQUINOLINE-2(1H)-CARBOXAMIDE	experimental	DB07064	Not Available
20	(2S)-1-(6H-INDOL-3-YL)-3-{[5-(7H-PYRAZOLO[3,4-C]PYRIDIN-5-YL)PYRIDIN-3-YL]OXY}PROPAN-2-AMINE	experimental	DB07124	Not Available

REFERENCES

1. Kan, C., Iacopetta, B., Lawson, J. and Whitaker, N. (2005) Identification of human papillomavirus DNA gene sequences in human breast cancer. *British Journal of Cancer*, **93**, 946-948.
2. Wang, T., Chang, P., Wang, L., Yao, Q., Guo, W., Chen, J., Yan, T. and Cao, C. (2012) The role of human papillomavirus infection in breast cancer. *Medical Oncology*, **29**, 48-55.
3. Simões, P.W., Medeiros, L.R., Pires, P.D.S., Edelweiss, M.I., Rosa, D.D., Silva, F.R., Silva, B.R. and Rosa, M.I. (2012) Prevalence of human papillomavirus in breast cancer: a systematic review. *International Journal of Gynecologic Cancer*, **22**.
4. Fagundes, C.P., Glaser, R., Alfano, C.M., Bennett, J.M., Povoski, S.P., Lipari, A.M., Agnese, D.M., Yee, L.D., Carson III, W.E. and Farrar, W.B. (2012) Fatigue and herpesvirus latency in women newly diagnosed with breast cancer. *Brain, behavior, and immunity*, **26**, 394-400.
5. Andrade, C.B.V., Monteiro, V.R.d.S., Coelho, S.V.A., Gomes, H.R., Sousa, R.P.C., Nascimento, V.M.d.O., Bloise, F.F., Matthews, S.G., Bloise, E. and Arruda, L.B. (2021) ZIKV disrupts placental ultrastructure and drug transporter expression in mice. *Frontiers in immunology*, 1821.