**INFERENCE OF PERSONALIZED DRUG TARGETS VIA NETWORK PROPAGATION**

Ortal SHNaps[[1]](#footnote-1)

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: ortalits@gmail.com*

Eyal perry­

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: eyal.perry88@gmail.com*

DANA SILVERBUSH

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: dana.silverbush@gmail.com*

RODED SHARAN

*School of Computer Science, Tel Aviv University  
Tel Aviv 69978, Israel  
Email: roded@post.tau.ac.il*

This is where the abstract should be placed.

# Introduction

Precision medicine, an approach where medical treatment is tailored for a specific group of patients, is an arising paradigm in medical research and practice. Indeed, it is well known that some drugs affect only a specific sub-group of patients, while even harming other patients with the same disease [1]. To date, only few methods have been developed for personalized predictions of drugs [cite the most representative ones including our paper PREDICT].

The development of new drugs is an expensive and time consuming process; it takes about 15 years and up to 800 million dollars to convert a promising new compound into a marketed drug [2]. Therefore, there is increasing interest in computational prediction of drug targets. Previous methods for drug target prediction tackle the problem in various ways, such as literature text mining [3, 4] and protein three-dimensional structures [5-9] [The references you picked are not so good and miss many high profile ones – pls. get Dana’s help on that, also see refs and introduction in PREDICT and make sure that what you site is related – e.g. druggability doesn’t mean that it’s a drug target]. In particular, the accumulation of various types of omics data inspired methods that take advantage of the data to discover novel drug targets [10]. Li et al. [13] described a workflow that leverages information of network topology to predict novel drug targets using machine learning algorithm. The methods in [14, 15] are based on in-silico knockouts on biological derived networks.

In this work we present a novel approach to tackle the drug target inference problem from a personalized perspective using in-silico knockouts in a PPI network. Figure 1 provides an overview of the method: we start from a general PPI network and personal disease-related data. We rely on the framework described by Vanunu et al. [12] to infer gene prioritization by network propagation. We perform multiple network propagations in order to simulate the current patient state, the patient state after gene knockouts (by removing the gene's node from the network) and an estimated "healthy" state. We use these different states in order to rank the gene knockouts and retrieve a candidates list for potential novel drug targets.

The framework we presented is general and can be applied to any personalized disease-related data. However, to evaluate our performance we applied it on patients suffering from acute myeloid leukemia (AML), using data from the The Cancer Genome Atlas (TCGA) of mutated and differentially expressed genes

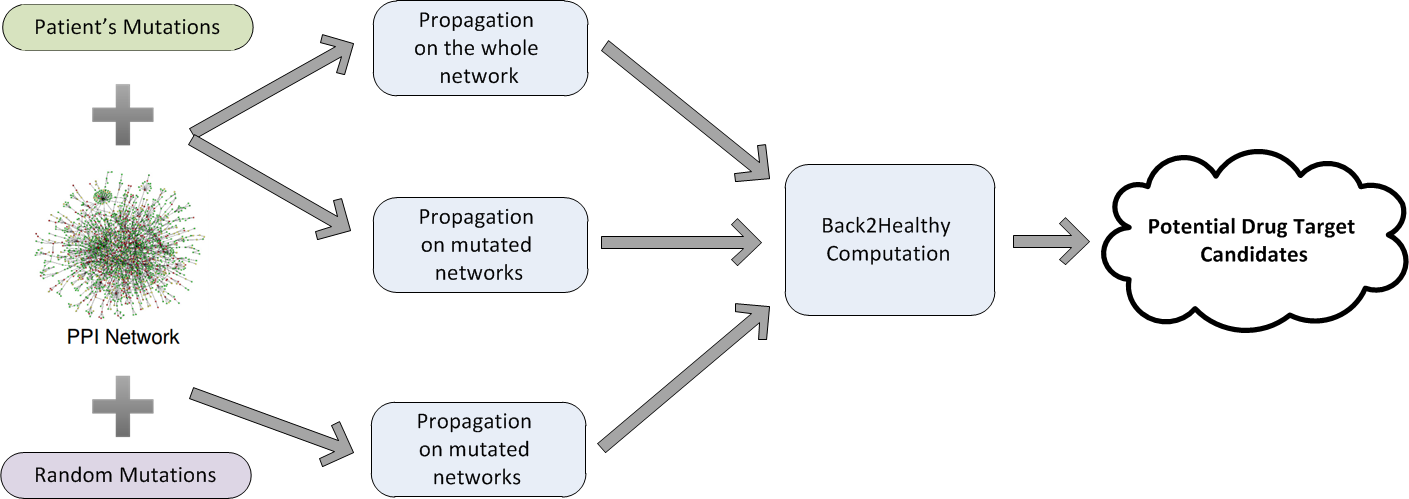


Fig. 1. An overview of the algorithmic pipeline.

# Results

We present a novel approach to tackle the drug target inference problem from a personalized perspective using in-silico knockouts in a PPI network. As described in Figure 1, we start from a general PPI network and personal disease-related data. We rely on the framework described by Vanunu et al. [12] to infer gene prioritization by network propagation. We perform multiple network propagations in order to simulate the current patient state, the patient state after gene knock-downs (by removing the gene's node from the network) and an estimated "healthy" state. We use these different states in order to rank the gene knockdowns and retrieve a candidates list for potential novel drug targets.

To evaluate our performance we applied our method to TCGA gene expression and mutation data of patients suffering from acute myeloid leukemia (AML). We gathered this dataset from the COSMIC cancer gene census [19]. First, we show that we can identify common AML causal genes by synthesizing the individual mutations set propagations and ranking according to propagation scores. Second, we show that by integrating results from a personalized knockout process we can infer potential drug targets.

We executed the algorithm using different settings for its alpha parameter (0.5, 0.75 and 0.9, see Methods) and the prior set of genes. To evaluate the results, we used three sets of known causal genes, varying in confidence and size: 10 AML causal genes from KEGG [Dana – provide the standard KEGG ref], 94 AML causal genes from COSMIC (72 of which are in our PPI network) and 533 cancer causal genes from COSMIC (363 in the network). The application of the method to each patient resulted in a propagation score for each gene. We aggregated the rank of each gene over a random sample of 100 patients to yield a gene-based score, retaining the top 10% scoring genes. We then computed the hypergeometric enrichment of this set of genes with the different sets of known causal genes. All choices of resulted in significant p-values (), though the best one out of the three is 0.9, as shown in Figure 2A. This value was used in the sequel.

Next, we wish to assess different prior knowledge (P) gene sets based on the patients’ data. Each patient holds information about each gene - whether it is mutated and/or differentially expressed. Therefore, we examined four settings – defining P based on (i) mutated genes; (ii) differentially expressed genes; (iii) both, but running them separately and averaging the propagation scores obtained; and (iv) same as (iii) but taking the maximum scores rather than the average. All variants resulted in significant p-values (). The best variant was the first – setting P to be the set of mutated genes (Figure 2B), a choice we use in the sequel.

Fig. 2. Performance evaluation under different parameter (A) and prior set (B) choices. The red line stands for a p-value of .

The previous results imply that our propagation based scores are able to infer disease-related changes in gene expression. We hypothesized that good drug targets for the disease could be genes whose knockout is predicted to reverse the disease-related effects. To identify such genes in-silico, we rerun the propagation based scoring while removing each gene in turn from the network, assessing the similarity between the obtained scores and those that characterize a “healthy” state. To this end, we use a Back2Healthy distance score (B2H; See Methods), and taking the top scoring genes as our best candidates for potential personalized drug targets.

The process above infers drug targets for each patient. As information about personalized drug targets is very scarce and hard to validate, we aggregated the results over all patients, evaluating the results using known AML drug targets derived from DrugBank [18] and COSMIC (see Methods). The top 10% scoring genes were highly enriched with known drug targets from both sources (Figure 3A [provide also p-values]). To assess the personalized approach we took, we generated a "meta-patient", using consensus (appearing in at least 5 patients) mutated and differentially expressed genes derived by aggregating all AML patients. The results were insignificant (Figure 3B), underscoring the utility of a personalized approach.

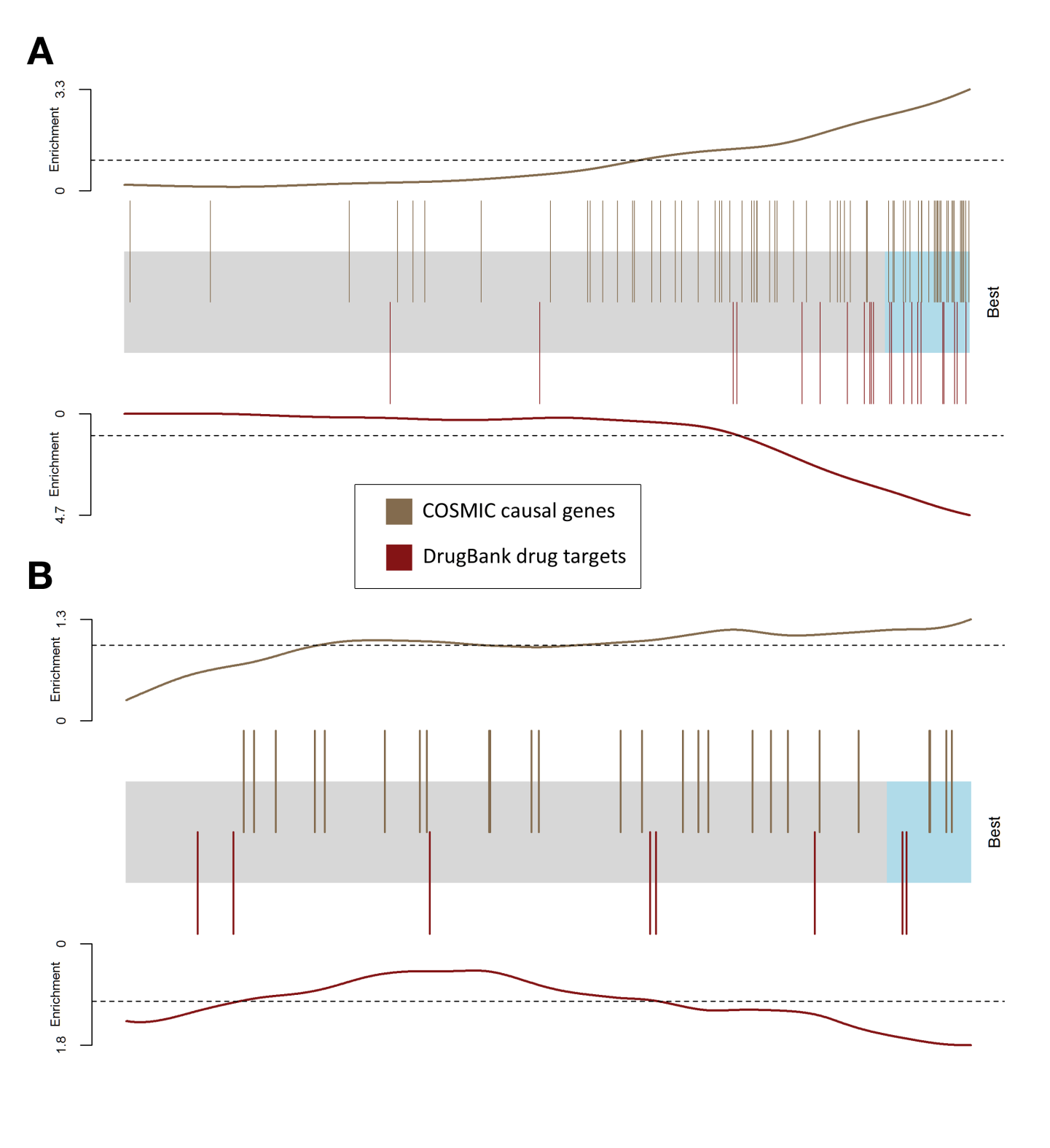


Fig. 3. Performance of drug target prediction. The candidate genes are represented by a shaded rectangle, where the top 10% are shaded cyan. Every overlaid bar stands for a single gene in a collection of known or potential drug targets. Traces above/below the bar represent relative enrichment. (A) The barcode plot was generated by running our method on each AML patient independently and aggregating the results. (B) The barcode plot was generated by running a similar single pipeline on the meta-patient.

# Methods

## Computing propagation score

We use the network propagation method described in Vanunu et al. [12]. The input consists of a network with V as the set of proteins, E as the set of their interactions, represents the reliability of the interaction between and , and a prior knowledge protein set . Our goal is to prioritize the proteins in with respect to . We do so by defining a function that is both smooth over the network and accounts for the prior knowledge about each node.

Regarding our framework, the set of protein-protein interactions and their reliability are taken from the HIPPIE network [16]. The prior knowledge set is derived from the patient's mutation data.

As described by Vanunu et al. we use Laplacian normalization to produce the normalized network edge weight . Briefly we construct a sparse matrix from the edge weights , and construct a diagonal matrix with . The normalized edge weight matrix is computed as .

We define a prior knowledge function such that:

With the normalized weight matrix and the prior knowledge function , we use the iterative procedure described by Zhou et al. [20] to compute . Namely, starting with , we update at iteration as follows:

The procedure is repeated iteratively until convergence; namely we stop the iterations when the following condition occurs:

The propagation score for each gene is its rank among all genes after propagating the network, where lower ranks means higher value. In case of ties the ranks of the genes is averaged.

## Back2Healthy distance score

**We need to discuss this score a bit. I assume that A is the set by which you evaluate (differentially expressed). Do you exclude from it the mutated genes? Do you exclude from it the k random genes? Think about these points.**

**E&O: A is indeed in practice the differentially expressed gene. We didn't exclude it from the mutated or the random genes. We did make sure we don't knockout any gene in A (because it would won't have a value). We're not sure why should it be excluded from the mutated genes – if one of them is differentially expressed – then the score before and after a knockout would probably be similar (since the prior knowledge would be 1), although a successful knockout might isolate the gene and leave it with a score around 1-alpha. Anyway the logic behind Back2Healthy remains true We think? And regarding the random genes, We think they should just reflect the same process we're doing for the patient but randomly, no? so if we don't exclude it from the mutated genes, then the same verdict would go for random ones.**

Let , be vectors of propagation scores for a prior gene set , where was generated by propagating on the original PPI network, while was generated by propagating on a “knockout” network, where one of the genes was removed. We define the Back2Healthy (B2H) distance between and as follows:

First, define .

For , we generate a score vector for by propagating the original PPI network and setting the prior knowledge set to be random nodes. Thus, simulating a “healthy” distribution of propagation scores for .

Next, for , define

Hence, represents the quantile of in our simulated distribution, same goes for . For our method, we used .

Finally, is defined by:

# Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer a porta elit. Etiam id turpis eu nibh porta lacinia at at metus. Proin id ullamcorper dui. Praesent eleifend ante vitae dolor euismod, faucibus rutrum mauris porttitor. Pellentesque vel nunc sed nisl elementum laoreet quis in nisl. Maecenas vulputate diam in tellus consequat convallis. Vivamus malesuada neque vel turpis dignissim vehicula. Nunc cursus diam vitae semper eleifend. Suspendisse consectetur sollicitudin molestie. Aenean congue, lorem id faucibus aliquam, nisl odio dictum nunc, in mollis ipsum urna et leo. Mauris magna nunc, elementum a est et, posuere imperdiet ex. Ut iaculis vel magna id rhoncus. Integer pretium varius eros non elementum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae;

Nullam eleifend in justo eget fermentum. Aenean id ullamcorper turpis, sit amet viverra nisi. Proin odio elit, volutpat quis volutpat quis, viverra et est. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Mauris tempus ultrices ante. Nulla volutpat neque tortor, eu faucibus augue dictum sed. Integer pretium nisi nec mollis bibendum. Donec ornare tincidunt enim, vel porttitor justo. Ut lobortis mattis fringilla.

Cras lacinia luctus mattis. Vivamus eleifend nibh eget turpis tincidunt, a tincidunt justo dignissim. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc auctor, ex vel consequat elementum, urna orci varius erat, sed lobortis sapien ante ac nunc. Nullam bibendum, ex non aliquet imperdiet, justo nisl porttitor felis, sit amet feugiat leo turpis et tellus. Aenean id nisl et nulla venenatis mattis. Donec blandit ullamcorper suscipit. In auctor nec turpis quis interdum. Nullam sapien dolor, aliquam nec sollicitudin sit amet, tincidunt quis dolor. Sed mi tellus, congue non rutrum nec, imperdiet non mauris. Cras ac dolor ipsum. Curabitur et neque non leo pellentesque lobortis. Praesent efficitur congue lacus, dapibus volutpat tellus gravida et.

# References

1. Rothwell, Peter M. "Can overall results of clinical trials be applied to all patients?." *The lancet* 345.8965 (1995): 1616-1619.
2. J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, “The price of innovation: new estimates of drug development costs,” Journal of Health Economics, vol. 22, no. 2, pp. 151–185, 2003.
3. S. Zhu, Y. Okuno, G. Tsujimoto, H. Mamitsuka, A probabilistic model for mining implicit chemical compound-gene relations from literature, Bioinformatics 21 (2005) ii245–ii251.
4. R.B. Altman, C.M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L.J. Jensen, M. Krallinger, B. Mons, S.I. O'Donoghue, M.C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay, A. Valencia, Text mining for biology-the way forward: opinions from leading scientists, Genome Biol. 9 (2008) S7.
5. A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, M. Rarey, Combining global and local measure for structure-based druggability predictions, J. Chem. Inf. Model. 52 (2012) 360–372.
6. E. Perola, L. Herman, J. Weiss, Development of a rule-based method for the assessment of protein druggability, J. Chem. Inf. Model. 52 (2012) 1027–1038.
7. P. Schmidtke, X. Barril, Understanding and predicting durggability. A highthroughput method for detection of drug binding sites, J. Med. Chem. 53 (2010) 5858–5867.
8. A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts smallmolecule druggability, Nat. Biotechnol. 25 (2007) 71–75.
9. A. Krasowski, D. Muthas, A. Sarkar, S. Schmitt, R. Brenk, DrugPred: a structurebased approach to predict protein druggability developed using an extensive nonredundant data set, J. Chem. Inf. Model. 51 (2011) 2829–2842.
10. H. Tan, X. Ge, and L. Xie, “Structural systems pharmacology: a new frontier in discovering novel drug targets,” Current Drug Targets, vol. 14, no. 9, pp. 952–958, 2013.
11. S. Erten, G. Bebek, and M. Koyut¨urk. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. Journal of computational biology, 18(11):1561–1574, Nov. 2011.
12. O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*., 6(1):e1000641, Jan 2010.
13. Li, Zhan-Chao, et al. "Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network."*Analytica chimica acta* 871 (2015): 18-27.
14. Ciaccio, Mark F., et al. "The DIONESUS algorithm provides scalable and accurate reconstruction of dynamic phosphoproteomic networks to reveal new drug targets." *Integrative Biology*, 2015.
15. Fatumo, Segun, et al. "Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico."*Infection, Genetics and Evolution* 9.3 (2009): 351-358.
16. M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2):e31826, 02 2012.
17. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328, 2004.
18. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z.T. Dame, B. Han, Y.
19. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Research, 39(suppl 1):D945–D950, 2011.
20. Zhou, D.S. Wishart, DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. 42 (2014) D1091–D1097.

1.  These authors contributed equally to this work. [↑](#footnote-ref-1)