Supplementary Materials for "NeoDTI: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions"

Fangping Wan[1], Lixiang Hong[1], An Xiao[1], Tao Jiang[2,3,4] and Jianyang Zeng[1,*]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China.
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
[3] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST, Tsinghua University, Beijing 100084, China.
[4] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.
[*] To whom correspondence should be addressed. Email: zengjy321@tsinghua.edu.cn.

1 Hyperparameter Selection

We used an independent validation dataset to determine the values of hyperparameters, including the dimension $d \in \{256, 512, 1024\}$ of node embedding, the dimension $k \in \{256, 512, 1024\}$ of the edge-type specific projection matrices, the repetition time $p \in \{0, 1, 2, 3\}$ for alternately repeating Steps (1) and (2), the gradient clipping norm from $\{1, 5\}$ and the number of steps for performing gradient descent. In particular, after splitting the data into training and test datasets in the cross-validation procedure, we randomly separated 0.5% of the training set as an independent validation dataset. We tuned the above hyperparameters based on this separate validation set and evaluated the performance of NeoDTI on the test dataset. In addition, we used the Adam optimizer [1] with the default learning rate 0.001 to perform gradient descent.

2 Baseline Methods

We compared the performance of NeoDTI with that of several network-based DTI prediction methods, including DT-Hybrid [2], BLMNII [3], HNM [4], MSCMF [5], NetLapRLS [6] and DTINet [7]. DT-Hybrid [2], BLMNII [3] and LPMIHN [8] have been reported as the state-of-the-art network-based prediction methods [9]. LPMIHN was not included in our comparisons because we encountered technical difficulty in implementing this method. Among the baseline methods used in our comparison tests, DTINet, MSCMF and HNM can integrate multiple heterogeneous information to predict new DTIs, while the other methods are not particularly designed to exploit multiple drug or protein network data for DTI prediction. To make a fair comparison, we followed the same strategy as in [7] to integrate multiple networks into a single network for DT-Hybrid, BLMNII and NetLapRLS in our comparison tests. In particular, for all interaction or association networks, i.e., drug-drug interaction, drug-disease association, drug-side-effect association, protein-protein interaction and protein-disease association networks, we constructed the corresponding drug-drug or protein-protein similarity networks based on the Jaccard similarities. Then, the final similarity score between drugs i and j after integrating all similarity networks was obtained by $1 - \prod_k (1 - d_{ij}^k)$, where $d_{ij}^k \in [0, 1]$ stands for the similarity between drugs i and j based on the network k. Here, k can stand for a drug-drug interaction, drug-disease association, drug-side-effect association or drug-structure-similarity network. Similarly, the final similarity score between proteins i and j after integrating all similarity networks was obtained by $1 - \prod_k (1 - p_{ij}^k)$, where $p_{ij}^k \in [0, 1]$ stands for the similarity between proteins i and j based on the network k. Here, k can stand for a protein-protein interaction, protein-disease association or protein-sequence-similarity network. Note that the edge weights in the protein-sequence-similarity network are ranged from $[0, 100]$. Here, we normalized the weights to $[0, 1]$. We used the above final similarity score as edge weights to construct the drug-drug and protein-protein similarity networks, and used them in the baseline methods DT-Hybrid, BLMNII and NetLapRLS.

For the hyperparameters used in all baseline methods, we tuned them using the same strategy as in NeoDTI. In particular, we used the following hyperparameter space for each baseline method. For DT-Hybrid, the combination parameters λ and α were both chosen

from $\{0.0, 0.1, \cdots, 1.0\}$. In BLMNII, The max function g was used to integrate the interaction scores p^d and p^t, and the combination weight α was selected from $\{0.0, 0.1, \cdots, 1.0\}$. For HNM, the value of the decay factor α was chosen from $\{0.0, 0.1, \cdots, 1.0\}$. For MSCMF, the feature dimensionality K was selected from $\{50, 100\}$ and four weight parameters, λ_l, λ_d, λ_t and λ_w were chosen from $\{2^{-2}, 2^{-1}, \cdots, 2^1\}$, $\{2^{-3}, 2^{-4}, \cdots, 2^5\}$, $\{2^{-3}, 2^{-2}, \cdots, 2^5\}$, and $\{2^1, 2^2, \cdots, 2^{10}\}$, respectively. For NetLapRLS, the ratios $\lambda_{d2}/\lambda_{d1}$ and $\lambda_{p2}/\lambda_{p1}$ were chosen from $\{10^{-5}, 10^{-4}, \cdots, 10\}$ and the parameters β_d and β_p were selected from $\{3 \cdot 10^{-4}, 3 \cdot 10^{-3}, \cdots, 3 \cdot 100\}$. In DTINet, we chose f_t from $\{400, 500, \cdots, 800\}$ and f_d, f_k from $\{100, 200, \cdots, 500\}$.

3 The effects of using different drug/protein-disease edge types on the prediction performance

The drug/protein-disease association edges used for constructing our heterogeneous network were derived from the Comparative Toxicogenomics Database (CTD) [10]. These edges can be further distinguished as marker, therapeutic and inferred types. To investigate the effects of using different drug/protein-disease edge types on the prediction performance, we further conducted the following additional tests: (1) We used only binary drug/protein-disease associations classified as markers in the CTD to construct the heterogeneous network and make prediction. (2) We used only binary drug/protein-disease associations classified as the therapeutic type in the CTD to construct the heterogeneous network and make prediction. (3) We used the drug/protein-disease associations classified as the inferred type in the CTD to construct the heterogeneous network and make prediction. Here, instead of using the binary edge weights, we used the inferred scores as edge weights. (4) We did not distinguish the edge types and used all edge types (i.e., marker, therapeutic and inferred) to create the binary drug/protein-disease associations to construct the heterogeneous network and make prediction (which corresponded to our original setting). All computational experiments were conducted using a ten-fold cross-validation procedure in which the ratios of positive versus negative samples were set to 1 : 10. The results are shown in Figure S4. Although we found that using all edge types without distinguishing them yielded the best prediction performance, it is interesting to notice that using only the therapeutic edges can already produce the comparable prediction performance.

4 The effect of using the edge weight normalization on the prediction performance

To investigate the effect of using the edge weight normalization, we conducted an addition ten-fold cross-validation test in which we removed the edge weight normalization term $M_{v,r}$ in Eq. 1 from Definition 2. The results are shown in Figure S5b-c. Compared to the original results using the edge weight normalization (i.e., our original design choice), we witnessed signification performance decreases in terms of both AUROC and AUPR

when dropping the edge weight normalization term. Therefore, we kept this normalization term in our model.

5 The effect of choice of the first dimension of W^1

When setting the repetition time of neighborhood information aggregation to one, the first dimension of the weight matrix $\boldsymbol{W}^1$ in Eq. 2 from Definition 3 may have arbitrary dimension $h \in \mathbb{R}$ instead of being fixed as d. We performed additional tests to investigate the effect of the choice of h. Specifically, the original h value used in our framework was set to 1024. In our new tests, we changed h to 512 and 2048, and conducted an additional ten-fold cross-validation procedure in which the ratios of positive versus negative samples were set to $1:10$. The results are shown in Figure S5a. When $k = 512$, 1024 and 2048, the corresponding AUROC scores were 0.946, 0.946 and 0.948, respectively, while the corresponding AUPR scores were 0.833, 0.853 and 0.859, respectively. Therefore, the change of h did not significantly affect the prediction performance. We kept $h = d$ in our framework mainly because this design choice will also be convenient for performing the multiple neighborhood integration steps when necessary.

6 Supplementary Figures

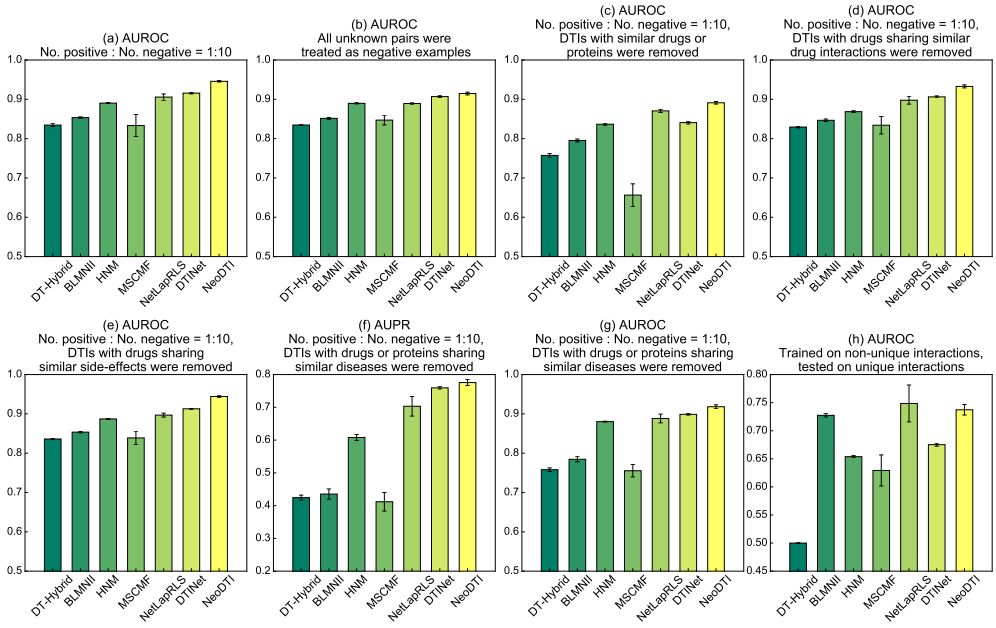

Figure S1. Supplementary results on the performance evaluation of NeoDTI on several challenging scenarios in terms of the AUPR and AUROC scores. (a) A ten-fold cross-validation test in which the ratio between positive and negative samples was set to 1 : 10. (b) A ten-fold cross-validation test in which all unknown drug-target interacting pairs were considered. (c-g) Ten-fold cross-validation with positive : negative ratios = 1 : 10 on several scenarios of removing redundancy in data: (c) DTIs with similar drugs and proteins were removed. (d) DTIs with drugs sharing similar drug interactions were removed. (e) DTIs with drugs sharing similar side-effects were removed. (f,g) DTIs with drugs and proteins sharing similar diseases were removed. (h) NeoDTI was trained on non-unique drug-target interacting pairs and tested on unique drug-target interacting pairs. More details on the baseline methods can be found in this Supplementary Materials. All results were summarized over 10 trials and expressed as mean $\pm$ standard deviation.

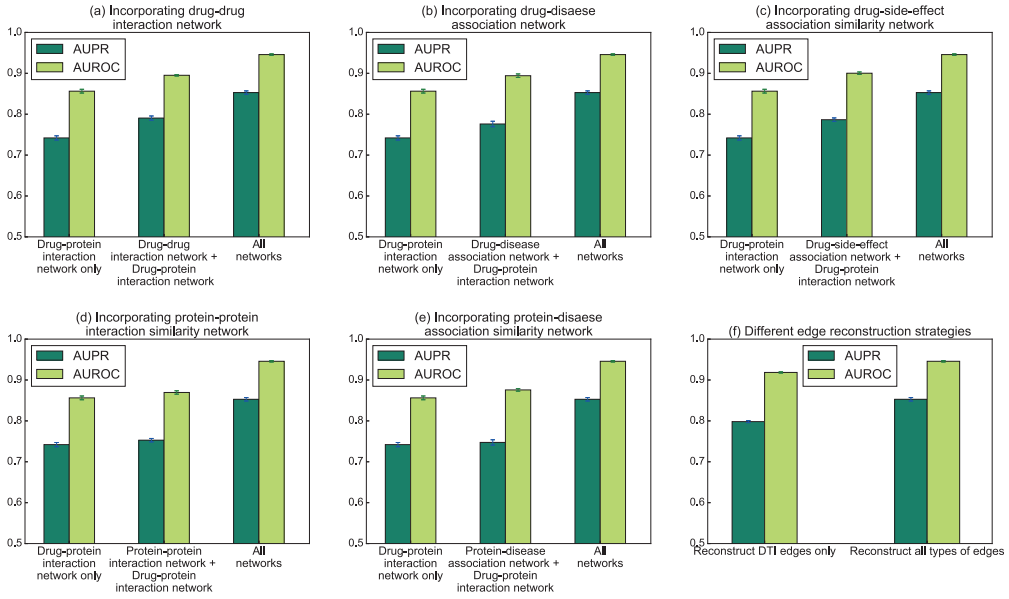

Figure S2. Supplementary results on the effects of incorporating heterogeneous information in NeoDTI. (a) Incorporating the drug-drug interaction network. (b) Incorporating the drug-disease association network. (c) Incorporating the drug-side-effect association network. (d) Incorporating the protein-protein interaction network. (e) Incorporating the protein-disease association network. (f) Performance under different edge reconstruction strategies. All results were summarized over 10 trials and expressed as mean $\pm$ standard deviation.

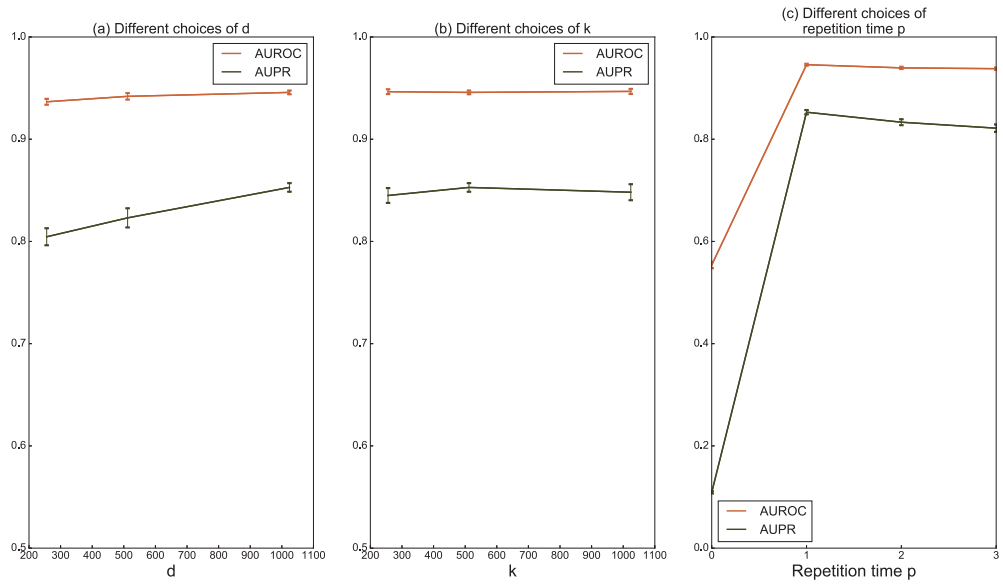

Figure S3. The robustness of NeoDTI over different choices of hyperparameters. (a) Performance of NeoDTI under different choices of the embedding dimension d. (b) Performance of NeoDTI under different choices of the dimension k of the projection matrices. (c) Performance of NeoDTI under different choices of repetition time p for performing neighborhood information integration. All results were summarized over 10 trials and expressed as mean $\pm$ standard deviation.

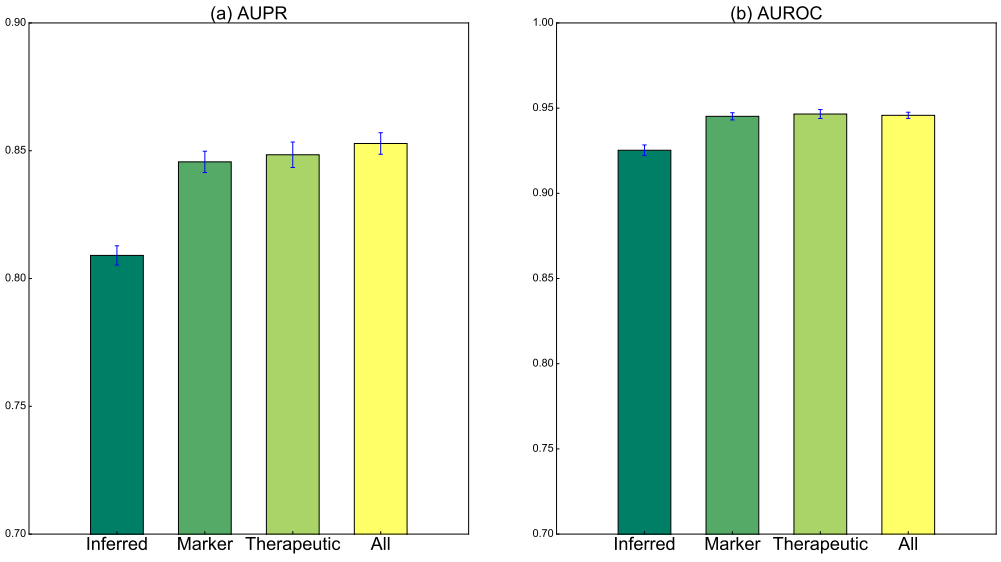

Figure S4. The effects of using different drug/protein-disease edge types derived from the Comparative Toxicogenomics Database (CTD) on the prediction performance in terms of AUPR (a) and AUROC (b).

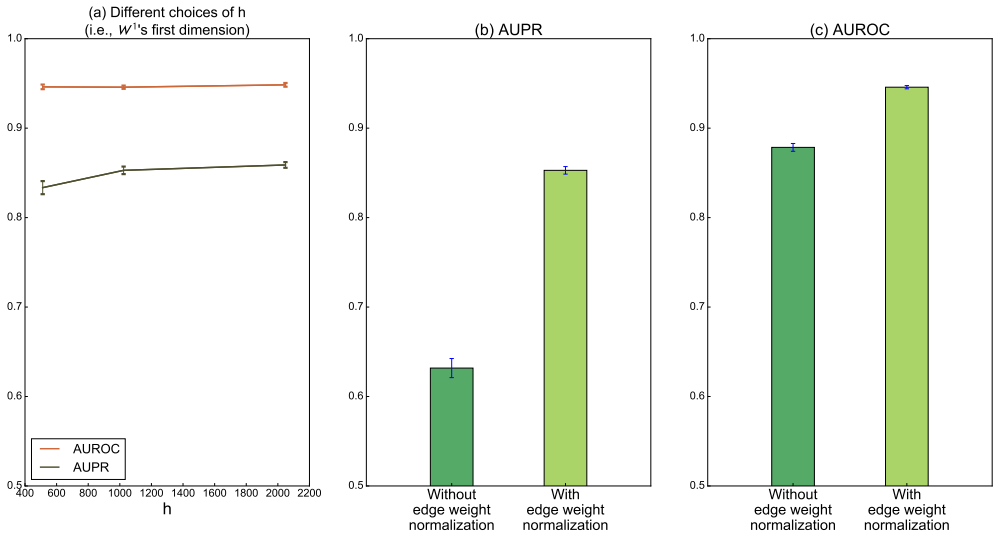

Figure S5. (a) The effect of the choice of h (i.e., the first dimension of $\boldsymbol{W}^1$ in Definition 3) on the performance. (b,c) The effects of whether using the edge weight normalization on the AUPR and AUROC scores, respectively.

7 Supplementary Tables

Node	Count
Drug	708
Protein	1,512
Disease	5,603
Side-effect	4,192
Total	12,015

(a)

Edge	Count
Drug-Protein	1,923
Drug-Drug	10,036
Drug-Disease	199,214
Drug-Side-effect	80,164
Protein-Protein	7,363
Protein-Disease	1,596,745
Total	1,895,445

(b)

Table S1. (a) The node statistics and (b) the edge statistics of the datasets used in our computational experiments. The datasets were curated in our previous study [7].

Drug name	Protein name	Supporting references
Sorafenib	FLT1	[11]
Mifepristone	NR3C2	[12, 13]
Tazarotene	RXRG	[14]
Felbamate	GRIN2D	[15]
Acetazolamide	CA6	[16, 17]
Rivastigmine	CES1	[18]
Pioglitazone	PPARA	[19]
Sorafenib	CSF1R	[20]

Table S2. Eight novel drug-target interactions among the list of top 20 significant predictions derived by NeoDTI that can be supported by previous studies in the literature.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, 29(16):2004–2008, 2013.

[3] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2):238–245, 2012.

[4] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, 8(7):1970–1978, 2012.

[5] Xiaodong Zheng, Hao Ding, Hiroshi Mamitsuka, and Shanfeng Zhu. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1025–1033. ACM, 2013.

[6] Zheng Xia, Ling-Yun Wu, Xiaobo Zhou, and Stephen TC Wong. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In *BMC systems biology*, volume 4, page S6. BioMed Central, 2010.

[7] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8(1), 2017.

[8] Xiao-Ying Yan, Shao-Wu Zhang, and Song-Yao Zhang. Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network. *Molecular BioSystems*, 12(2):520–531, 2016.

[9] Maryam Lotfi Shahreza, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz, and James R Green. A review of network-based approaches to drug repositioning. *Briefings in bioinformatics*, page bbx017, 2017.

[10] Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Michael C Rosenstein, Thomas C Wiegers, et al. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, 41(D1):D1104–D1114, 2012.

[11] D Kitagawa, K Yokota, M Gouda, Y Narumi, H Ohmoto, E Nishiwaki, K Akita, and Y Kirii. Activity-based kinase profiling of approved tyrosine kinase inhibitors. *Genes to Cells*, 18(2):110–122, 2013.

[12] David Wishart, David Arndt, Allison Pon, Tanvir Sajed, An Chi Guo, Yannick Djoumbou, Craig Knox, Michael Wilson, Yongjie Liang, and Jason Grant. T3db: the toxic exposome database. *Nucleic Acids Research*, 43(Database issue):928–34, 2015.

[13] Lim Emilia, Pon Allison, Djoumbou Yannick, Knox Craig, Shrivastava Savita, Guo An Chi, Neveu Vanessa, and David S Wishart. T3db: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Research*, 38(Database issue):D781, 2010.

[14] F Alizadeh, A Bolhassani, A Khavari, S. Z. Bathaie, T Naji, and S. A. Bidgoli. Retinoids and their biological effects against cancer. *International Immunopharmacology*, 18(1):43–9, 2014.

[15] Lorraine V Kalia, Suneil K Kalia, and Michael W Salter. Nmda receptors in clinical neurology: excitatory times ahead. *The Lancet Neurology*, 7(8):742–755, 2008.

[16] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[17] Isao Nishimori, Tomoko Minakuchi, Saburo Onishi, Daniela Vullo, Andrea Scozzafava, and Claudiu T Supuran. Carbonic anhydrase inhibitors. dna cloning, characterization, and inhibition studies of the human secretory isoform vi, a new target for sulfonamide and sulfamate inhibitors. *Journal of medicinal chemistry*, 50(2):381–388, 2007.

[18] Lyudmila G Tsurkan, M Jason Hatfield, Carol C Edwards, Janice L Hyatt, and Philip M Potter. Inhibition of human carboxylesterases hce1 and hice by cholinesterase inhibitors. *Chemico-biological interactions*, 203(1):226–230, 2013.

[19] U Smith. Pioglitazone: mechanism of action. *International Journal of Clinical Practice Supplement*, 55(121):13, 2001.

[20] Katrin Ullrich, Kathrin D Wurster, Björn Lamprecht, Karl Köchert, Andreas Engert, Bernd Dörken, Martin Janz, and Stephan Mathas. Bay 43-9006/sorafenib blocks csf1r activity and induces apoptosis in various classical hodgkin lymphoma cell lines. *British journal of haematology*, 155(3):398–402, 2011.