**Εθνικό Μετσόβιο Πολυτεχνείο**
**Σχολή Μηχανολόγων Μηχανικών**
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

# *Systematic modeling of disease mechanisms and drug mode of action*

Διδακτορική Διατριβή

Χρήστος Φώτης

ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΜΗΧΑΝΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

**ΕΠΙΒΛΕΠΩΝ:**
Λεωνίδας Αλεξόπουλος, Αν. Καθηγητής, ΕΜΠ

**Εθνικό Μετσόβιο Πολυτεχνείο**
**Σχολή Μηχανολόγων Μηχανικών**
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

# *Systematic modeling of disease mechanisms and drug mode of action*

Διδακτορική Διατριβή

Χρήστος Φώτης

ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΜΗΧΑΝΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

**ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:**

Λ. Αλεξόπουλος, Αν. Καθ. ΕΜΠ (Επιβλέπων)

Κ. Κυριακόπουλος, Καθ. ΕΜΠ

Α. Χατζηιωάννου, Ερευνητής Α, ΕΙΕ

**ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:**

Λ. Αλεξόπουλος, Αν. Καθ. ΕΜΠ (Επιβλέπων)

Κ. Κυριακόπουλος, Καθ. ΕΜΠ

Α. Χατζηιωάννου, Ερευνητής Α, ΕΙΕ

Ε. Μικρός, Καθ. ΕΚΠΑ

J. Saez-Rodriguez, Καθ. RWTH Aachen

Έ. Παπαεμμανουήλ, Καθ. MSKCC

Γ. Ματσόπουλος, Καθ. ΕΜΠ

**Athens, July 2021**

# Prologue

The research for this dissertation was carried out under the supervision of the Associate professor of the Mechanical Engineering School of NTUA, Leonidas G. Alexopoulos. My research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code:T1EDK-02829).

I would like to thank the people in my life and work environment that helped me pursue my dreams and conduct my research. First of all, I would like to thank my mentor, professor Alexopoulos, for introducing me to the field of systems biology through his class in the NTUA School of mechanical engineering. From the very first lectures, I knew that this was what I wanted to do with my life. His passion and love for the field along with his entrepreneurial spirit inspired me to pursue my goals and seek a research position in his lab. His innovative thinking was transferred to me through our countless meetings and his constant guidance and mentorship made me a better researcher, as well as a better person overall. From the bottom of my heart, thank you for being a professor at NTUA, without you I would be lost. I would then like to thank my partner Nefeli, for always believing in me, supporting me and filling my life with positivity. In times that I was doubting myself, she was always there to support me emotionally and practically, by making my life beautiful. Without her support, love and motivation I wouldn't be able to do half the things I have done. I would like to thank my parents for providing me everything growing up, nourishing my passion to pursue a career in science and always believing in me, even in times when they shouldn't. Especially, I would like to thank my mother for teaching me how to study and conduct research from a very young age and for always giving me the opportunity to make my own choices. Additionally, I would like to thank my friend and colleague Nikos, for all his help during my research. I believe that we did our best work together in the lab and I am excited to see his future endeavors. Guiding him and teaching him everything I know was maybe the most valuable lesson I learned during my PhD. The lesson is that investing in the right people is the most rewarding investment that you can make. I was very lucky to have worked with him during his first steps of his research path. I would like to thank my colleague Danae for her support, which was plentiful, since we faced similar issues during our PhD thesis. Furthermore, I would like to thank all the members of the computational team that worked together with me in my research and who I can call my friends and colleagues. Your passion, thinking and work ethics motivated me to be a better leader and our success made me believe in myself. Finally, I would like to thank Teo and Asier, for teaching and mentoring me during my early days in the lab.

Christos Fotis

Athens, July 20

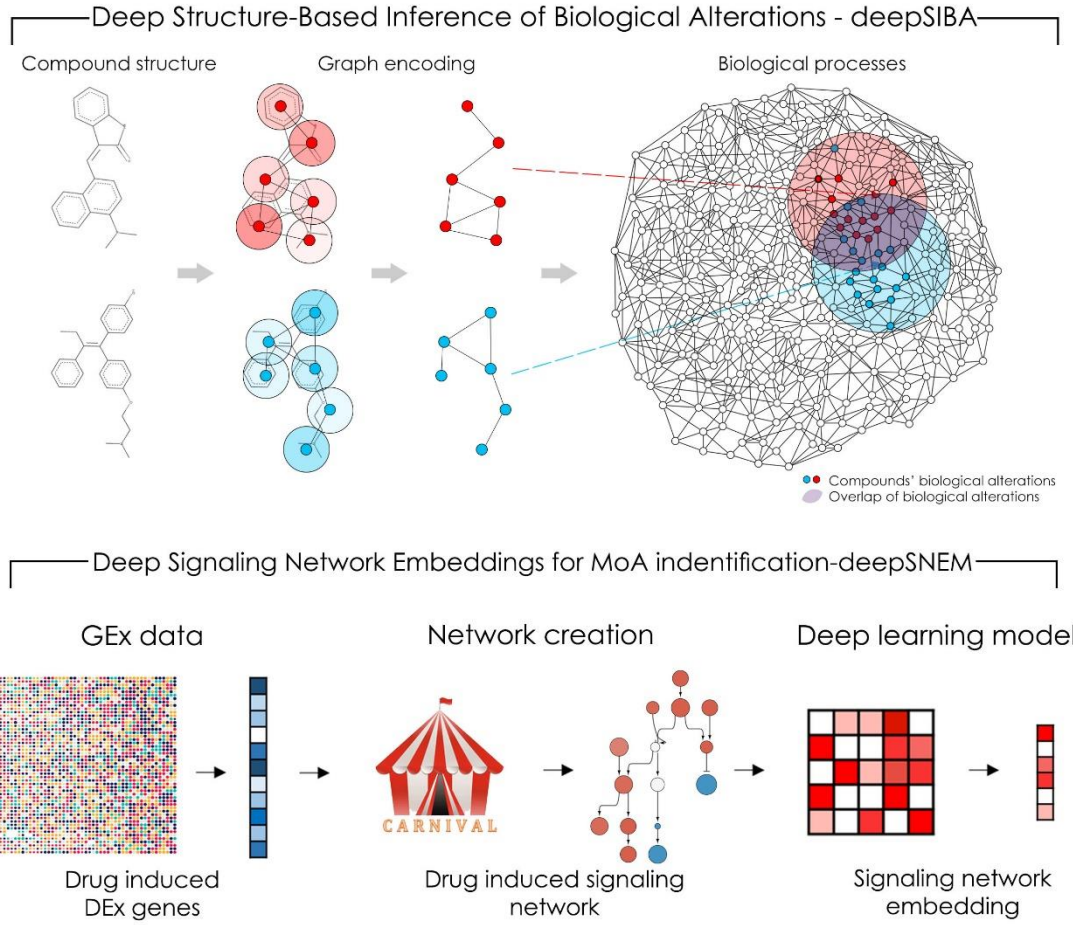
# Summary

Systems pharmacology methods aim to prioritize compounds that not only exhibit maximal binding affinity to the specified target, but also cause the desirable biological effect. One specific approach that has gained considerable attraction is modeling the cellular system as a complex network of molecular interactions, in order to identify changes in the signaling mechanism that best explain the experimental response data. In this thesis, we first present a concise review of omics repositories and knowledge bases of molecular interactions, along with network-based methods for their analysis. Furthermore, we present two novel deep learning pipelines, called deepSNEM and deepSIBA, which can be used to investigate the connection of a compound's signaling network and chemical structure to its mechanism of action (MoA) and biological effect.

DeepSNEM is a novel unsupervised graph deep learning pipeline that is trained to encode the information in the compound-induced signaling networks into fixed length high-dimensional representations. DeepSNEM is a graph transformer network, trained to maximize the mutual information between whole-graph and sub-graph representations that belong to similar perturbations. By clustering the deepSNEM representations, we were able to identify distinct clusters that are significantly enriched for specific MoAs. In order to increase the interpretability of deepSNEM, we developed a subgraph importance method to elucidate the important subgraphs that cause the signaling networks to cluster together. As a case study, deepSNEM was applied to cluster the representations of signaling networks created from gene expression profiles of various experimental platforms (MicroArrays and RNA sequencing). In order to take into account the structural attributes of compound perturbations, alongside deepSNEM, we developed the deepSIBA pipeline to investigate the connection between a compound's chemical structure and its biological effect.

DeepSIBA is a supervised Siamese graph convolutional model that is trained to predict the biological effect distance between a pair of compounds, using their molecular graphs as input. The proposed model was able to encode molecular graph pairs and identify structurally dissimilar compounds that affect similar biological processes, with high precision. Additionally, by utilizing deep ensembles to estimate uncertainty, we were able to provide reliable and accurate predictions for chemical structures that are very different from the ones used during training. Finally, we present a novel inference approach, where the trained deepSIBA models are used to estimate the signaling pathway signature of a compound perturbation, using only its chemical structure as input, and subsequently identify which substructures influenced the predicted pathways. As a use case, deepSIBA was used to infer important substructures and affected signaling pathways of FDA-approved anticancer drugs.

# Graphical abstract



# Extended summary

## Introduction

Drug discovery is a complex and time consuming process that aims to identify the right compound for the right disease and target. Despite the development of many successful drugs, the attrition rates still remain high. Recent advances in systems-pharmacology and omics technologies have led to the development of computational tools that aim to model the biological effect of the compound perturbation in the cellular system. These tools, based on biological pathways and signaling networks, offer a systematic approach to unravel a compound's Mechanism of Action (MoA) and prioritize compounds that have the desired effect for further experimental validation. In this thesis, we first provide a thorough review of omics databases and knowledge bases of molecular interactions, along with network-based modeling tools that can be applied across all stages of the drug discovery pipeline to elucidate the compound's MoA. Furthermore, we provide a concise list of studies that have successfully implemented these network-based approaches for various drug discovery projects. However, due to their complex structure, large scale datasets of compound-induced signaling networks and methods specifically tailored to their comparison are still very limited. One approach that holds promise to overcome these limitations is the use of graph deep learning models to transform signaling networks into high-dimensional representations. On this front, we proposed the deepSNEM pipeline that uses an unsupervised graph transformer network to encode a compound's signaling network and investigate its relationship with the compounds' MoA. However, systems pharmacology approaches that rely on cellular response data are limited in their application to compounds with available data and most importantly do not take into account the compounds' structural attributes that are closely related to their efficacy, effect and toxicity profiles. To this end, we aimed to use graph deep learning to match the chemical structure of compound perturbations to their biological effect on specific cellular models. Thus, we proposed the deepSIBA pipeline that can be used to infer a compound's signaling pathway signatures, without available expression data, using only its structure as input.

## DeepSNEM: Deep Signaling Network Embeddings for compound mechanism of action identification using deep learning

### DeepSNEM methods

Transcriptomic signatures following compound treatment were retrieved from the CMap dataset. After assigning a quality score to its experiment, the highest quality data were selected and transformed into compound-induced signaling networks using the CARNIVAL pipeline. CARNIVAL solves an ILP optimization problem to infer a family of highest scoring subgraphs, from a prior knowledge network of signed and directed protein-protein interactions, which best explain the experimental data, subject to specific constraints. In total, more than 700000 networks were created from 7781 transcriptomic signatures of 3005 compounds across 70 cell lines. Afterwards, an

unsupervised deep learning model that takes as input a compound-induced signaling network and outputs a fixed-length high dimensional representation was developed. The signaling networks were represented as input to the model using three matrices that contain information regarding the nodes of each network, their activity and the network's connectivity (sign and direction). The core of deepSNEM is a graph transformer network that pays specific attention to each node's neighborhood, when extracting the important features that describe the signaling mechanism. Finally, a whole-graph representation is created by summarizing the node embeddings, using a 2-layer Long Short Term Memory network. The model was trained by maximizing the mutual information between the nodes and subgraphs that belong to the same or duplicate experiments. Thus, the deepSNEM model creates similar representations for similar compound-induced signaling networks. The resulting 256-dimensional embeddings of the compounds' signaling networks were clustered using the k-means algorithm, into 200 unique clusters. These clusters were then analyzed based on their MoA composition, using labels provided by the Drug Repurposing Hub dataset. Furthermore, we developed a node importance pipeline, using the saliency approach, to identify the nodes that the model pays attention to, when creating the network representations. Finally, this node importance score was integrated with each node's frequency in a cluster to extract important subgraphs, in the original networks, that cause the representations to cluster together.

**DeepSNEM results**

The deepSNEM approach was evaluated based on the validity of the resulting embeddings on two separate tasks. The first task examines the models' ability to produce similar embeddings from signaling networks that are created from the same differential gene expression signature, while the second task evaluates the similarity of graph embeddings created from duplicate gene signatures, as compared to the similarity of embeddings from random gene signatures. On this front, the embeddings produced by the graph transformer architecture, termed deepSNEM-GT-MI, were compared to embeddings created from three additional models. These models include, a graph transformer trained to predict the edge presence between nodes (termed deepSNEM-GT-LP), a siamese GCN model to predict the graph edit distance between signaling networks (termed deepSNEM-GED) and the widely used graph2vec model (termed deepSNEM-G2V). Across both tasks, all deepSNEM model variations were able to identify embeddings produced from similar signaling networks, with the deepSNEM-GT-MI variation showing the best performance. The embeddings of the deepSNEM-GT-MI model were clustered, and the resulting clusters were analyzed based on their MoA composition. On this front, we were able to identify distinct clusters that are significantly enriched for mTOR, topoisomerase, HDAC and protein synthesis inhibitors respectively. Additionally, by applying the importance pipeline, on the clusters enriched for mTOR inhibitors, we were able to identify important nodes and subgraphs that are directly related to the mTOR/PI3K signaling mechanism. As a case study, deepSNEM was used to assign clusters to signaling networks created from compounds' gene expression profiles from various experimental platforms (MicroArrays and RNA sequencing). The results show that the majority of the compounds' signaling networks were correctly assigned to clusters that were enriched for their respective MoA. For the compounds in the use case, we also compared the cluster assignment of deepSNEM to a clustering of the compounds' differential expression gene measurements into the same number of clusters (k=200). Comparing the two approaches, 3/8 compounds were assigned to clusters composed of similar mechanisms. However,

the remaining compounds were assigned to clusters not enriched for any particular MoA, when the gene-clustering pipeline is used. Finally, for each compound of the use case, we calculated the Jaccard similarity index between the perturbations of the identified clusters using the two methods (deepSNEM and gene-based clustering). The similarity between the clusters was very low, with only two clusters showing a similarity higher than 0.1.

**DeepSNEM discussion**

DeepSNEM was not only able to identify clusters of network representations that were enriched for specific MoAs, but also identify important subgraphs that are related to them. However, the majority of the compounds' MoA labels are still unknown, which can result in a different MoA composition for the identified clusters, when they are taken into account. By comparing the deepSNEM pipeline to a simple gene-based clustering approach, we showed that the two approaches result in a different clustering of the perturbations. We argue that is due to the different biological hierarchy of information provided by the compound-induced signaling networks and differential gene expression signatures. Thus, the deepSNEM pipeline, by using the knowledge of molecular interactions, can identify similarities and differences in the compounds' signaling networks that are hidden in their transcriptomic signatures.

## DeepSIBA: Chemical Structure-based Inference of Biological Alterations using deep learning

**DeepSIBA methods**

Gene expression profiles following compound treatment were collected from the L1000 Connectivity Map resource (GSE92742). For this study, only level 5 expression data (z-scores) of the landmark genes were considered. For each compound perturbation, enrichment scores (ES) of GO terms related to biological processes were calculated using Gene Set Enrichment Analysis. Afterwards, a Kolmogorov-Smirnov based distance function was used to calculate the pairwise distance at the GO-term level. For each pair in the dataset, the distance was calculated using 5 different thresholds for up-regulated and down-regulated GO-terms. Finally, the distances were averaged. This pairwise GO-term distance score was then used as the target variable for the learning model. The learning model takes as input the chemical structures of compound pairs, represented as undirected graphs, with nodes being the atoms and edges the bonds between them. The architecture consists of two Siamese (identical) graph convolutional and convolutional encoders, one for each compound, that embed the chemical structure into a high dimensional latent space. The absolute difference of the embeddings is then fed through 2 convolutional layers followed by 2 fully connected layers. The final layer is a Gaussian regression layer that outputs a mean and standard deviation of the biological effect distance between the pair. By treating the distance as a sample from a Gaussian distribution with the predicted mean and variance, the model is trained by minimizing the Negative Log Likelihood. Model ensembles are created by taking the uniformly-weighted mixture of each model's Gaussian. Furthermore, we developed a novel inference approach method, similar to k-NN that can be used to infer a signaling pathway signature from a compound's chemical structure. To this end, the trained deepSIBA ensemble

is used to predict the biological effect distances between the target compound and all the compounds in the dataset. Then, training compounds with the lowest distance are selected as neighbors and a target signature is inferred by using a voting scheme between the neighbors' pathway signatures. Finally, in order to increase the interpretability of our approach, we developed an importance pipeline, based on graph gradients, to identify the important substructures that deepSIBA pays attention to, when inferring the signaling pathway signature of a compound.

**DeepSIBA results**

The performance of deepSIBA was evaluated with a realistic drug discovery scenario in mind, where gene expression data are available for only one compound per pair. Additionally, deepSIBA's performance was compared to machine learning models for pairwise distance learning tasks. Across all cell lines, deepSIBA was able to outperform the machine learning models and exhibited a very high precision and low MSE. By utilizing a transfer learning method, we were able to expand the coverage of our approach to 7 additional cell lines, with fewer data points, but with similar performance. Furthermore, deepSIBA was able to maintain high precision and low MSE, regardless of the chemical similarity of the input compounds. We also evaluated the performance of our approach in a scenario, where the test compounds are completely different, in terms of chemical structure, to the ones used in training. We showed that in this case the precision of the model decreases, while the MSE remains comparable. However, by utilizing deepSIBA's uncertainty estimate, we were able to focus on a specific subset of samples that the model was more certain, which led to a dramatic increase in performance. As a case study, deepSIBA was tasked to infer the signaling pathway signature of FDA approved anticancer drugs that were not present in our dataset, using only their structure as input. For the drugs of the use case, deepSIBA was able to infer a signaling pathway signature that is directly related to the compounds' MoA and subsequently identify the correct chemical substructures as important for the inference.

**DeepSIBA discussion**

DeepSIBA was able to encode molecular graph pairs and identify structurally dissimilar compounds that affect similar biological processes with high precision. Additionally, by utilizing deep ensembles to estimate uncertainty, we were able to provide reliable and accurate predictions for chemical structures that are very different from the ones used during training. However, there were many compound pairs with similar biological effect that were missed by the model. We argue that this happens due to the limited chemical coverage of the CMap dataset and we believe that as more data become available, the performance of our approach will increase as well.

## Conclusion

We believe that deepSNEM and deepSIBA have the potential to augment *in-silico* drug discovery, either by identifying a compound's MoA, using its signaling network effect, or by exploring on a massive scale the biological effect of compounds/libraries without available GEx data and suggesting new chemical structures with desired biological effect.

# Περίληψη

Οι μέθοδοι που εφαρμόζονται στη φαρμακολογία συστημάτων βασίζονται στην επιλογή φαρμάκων που επιδεικνύουν τη μέγιστη δύναμη πρόσδεσης στην πρωτεϊνη-στόχο και συνάμα οδηγούν στο επιθυμητό βιολογικό αποτέλεσμα. Μια συγκεκριμένη τεχνική, η οποία έχει συγκεντρώσει υψηλό ενδιαφέρον, είναι η μοντελοποίηση του κυττάρου ως ένα δίκτυο μοριακών αλληλεπιδράσεων, με στόχο την ανακάλυψη του σηματοδοτικού μηχανισμού, που περιγράφει με βέλτιστο τρόπο τα πειραματικά δεδομένα. Στην παρούσα διατριβή, πρώτα παρουσιάζουμε μια εκτενή συλλογή τόσο βάσεων δεδομένων τύπου omics όσο και μοριακών αλληεπιδράσεων, συνοδευόμενες από τις αντίστοιχες μεθόδους δικτύων για την ανάλυση αυτών. Περαιτέρω, παρουσιάζουμε δύο νέες μεθόδους deep learning, που ονομάζονται deepSNEM και deepSIBA, οι οποίες έχουν ως στόχο να ερευνήσουν τον τρόπο με τον οποίο τόσο το σηματοδοτικό δίκτυο όσο και η χημική δομή ενός φαρμάκου συσχετίζονται με το μηχανισμό δράσης και το βιολογικό αποτέλεσμα του φαρμάκου σε κυτταρικά μοντέλα.

Το μοντέλο deepSNEM είναι ένα unsupervised deep learning δίκτυο το οποίο εκπαιδεύεται προκειμένου να κωδικοποιήσει και να οριοθετήσει το σηματοδοτικό δίκτυο ενός φαρμάκου σε σταθερού μεγέθους πολυδιάστατες αναπαραστάσεις. Το μοντέλο εκπαιδεύεται ώστε να μεγιστοποιήσει την αμοιβαία πληροφορία μεταξύ αναπαραστάσεων δικτύων και υποδικτύων, που προκύπτουν από παρόμοια πειράματα. Ομαδοποιώντας τις αναπαραστάσεις, καταφέραμε να ανακαλύψουμε συγκεκριμένες ομάδες, οι οποίες είναι εμπλουτισμένες σημαντικά με φάρμακα, που μοιράζονται ένα συγκεκριμένο μηχανισμό δράσης. Με στόχο την καλύτερη επεξήγηση των αποτελεσμάτων αναπτύχθηκε μια μέθοδος ανάδειξης των σημαντικών υποδικτύων, τα οποία οδηγούν στην εκάστοτε ομαδοποίηση των αναπαραστάσεων. Η μέθοδος deepSNEM εφαρμόστηκε για την ομαδοποίηση των αναπαραστάσεων που προκύπτουν από σηματοδοτικά δίκτυα φαρμάκων, τα οποία βασίζονται σε δεδομένα γονιδιακής έκφρασης από διάφορες πειραματικές πλατφόρμες (Microarrays and RNA-sequencing). Θέλοντας να λάβουμε υπόψην και τη χημική δομή των ουσιών, παράλληλα με το deepSNEM αναπτύχθηκε και το μοντέλο deepSIBA, με στόχο τη διερεύνηση της συσχέτισης τους με το βιολογικό αποτέλεσμα των φαρμάκων.

Το μοντέλο deepSIBA είναι ένα supervised deep learning μοντέλο το οποίο εκπαιδεύεται για την πρόβλεψη της απόστασης μεταξύ των βιολογικών διεργασιών ενός ζεύγους φαρμάκων, χρησιμοποιώντας τα μοριακά τους δίκτυα ως είσοδο. Το μοντέλο κωδικοποίησε και ανακάλυψε ζεύγη φαρμάκων με διαφορετική χημική δομή, τα οποία επηρεάζουν παρόμοιες βιολογικές διεργασίες, με υψηλή ευστοχία και ακρίβεια. Εν συνεχεία, χρησιμοποιώντας πλήθος μοντέλων, προκειμένου να εκτιμήσουμε την αβεβαιότητα των προβλέψεων, καταφέραμε να έχουμε εύστοχες προβλέψεις για φάρμακα, τα οποία έχουν εντελώς διαφορετική δομή από εκείνα που χρησιμοποιήθηκαν κατά τη διάρκεια της εκπαίδευσης. Εν κατακλείδι, παρουσιάζουμε μια νέα μέθοδο για την εξαγωγή χαρακτηριστικών σηματοδοτικών μονοπατιών, χρησιμοποιώντας ως είσοδο μόνο τη χημική δομή των φαρμάκων. Σαν εφαρμογή, το μοντέλο χρησιμοποιήθηκε για την ανακάλυψη σημαντικών σηματοδοτικών μονοπατιών και χαρακτηριστικών χημικών δομών σε ένα σύνολο εγκεκριμένων αντικαρκινικών φαρμάκων.

# Εκτενής περίληψη

## Εισαγωγή

Η ανακάλυψη νέων φαρμάκων είναι μια πολύπλοκη και χρονοβόρα διαδικασία, η οποία αποσκοπεί στην εύρεση νέων χημικών δομών, με σκοπό να προσδέσουν στον κατάλληλο στόχο και να καταπολεμήσουν την εκάστοτε ασθένεια. Η ανάπτυξη της φαρμακολογίας συστημάτων και των τεχνολογιών τύπου omics, έχει οδηγήσει στην δημιουργία υπολογιστικών εργαλείων, που έχουν ως στόχο τη μοντελοποίηση της βιολογικής επίδρασης ενός φαρμάκου στο κυτταρικό σύστημα. Οι μέθοδοι βασίζονται σε σηματοδοτικά μονοπάτια και δίκτυα και προσφέρουν μια ολιστική αντιμετώπιση του προβλήματος της διερεύνησης του μηχανισμού δράσης των φαρμάκων. Στην παρούσα διατριβή, πρώτα παρουσιάζουμε μια εκτενή συλλογή τόσο βάσεων δεδομένων τύπου omics όσο και μοριακών αλληεπιδράσεων, συνοδευόμενες από τις αντίστοιχες μεθόδους δικτύων για την ανάλυση αυτών. Αρχικά, αναλύονται οι μέθοδοι και εν συνεχεία παρουσιάζονται παραδείγματα εφαρμογής αυτών, στην ανακάλυψη του μηχανισμού δράσης φαρμάκων. Παρά την πληθώρα των πλεονεκτημάτων τους, τα σηματοδοτικά δίκτυα εξακολουθούν να αποτελούν πολύπλοκες αναπαραστάσεις κι ως εκ τούτου, μεγάλες βάσεις δεδομένων καθώς και μέθοδοι για τη σύγκριση αυτών είναι περιορισμένες. Μια μέθοδος για την αντιμετώπιση αυτών των περιορισμών, είναι η χρήση μοντέλων deep learning, εφαρμοσμένα σε δίκτυα, που αποσκοπούν στη μετατροπή τους σε διαχειρίσιμες αναπαραστάσεις. Για αυτό το λόγο, δημιουργήθηκε το μοντέλο deepSNEM, ώστε να κωδικοποίησει σηματοδοτικά δίκτυα φαρμάκων και να τα συσχετίσει με το μηχανισμό δράσης τους. Στη συνέχεια, με γνώμονα τη σημασία της χημικής δομής των φαρμάκων, δημιουργήθηκε το μοντέλο deepSIBA, με στόχο τη διερεύνηση της σχέσης μεταξύ της χημικής δομής ενός φαρμάκου και της βιολογικής του επίδρασης στο κυτταρικό σύστημα.

## DeepSNEM: Deep Signaling Network Embeddings for compound mechanism of action identification using deep learning

### DeepSNEM μέθοδοι

Μέσω της βάσης δεδομένων CMap λήφθησαν δεδομένα γονιδιακής έκφρασης, τα οποία εκφράζουν την κατάσταση κυτταρικών μοντέλων μετά τη χρήση διαφόρων φαρμάκων. Στη συνέχεια, τα πειράματα γονιδιακής έκφρασης με την υψηλότερη ποιότητα μετατράπηκαν σε χαρακτηριστικά σηματοδοτικά δίκτυα, χρησιμοποιώντας τη μέθοδο CARNIVAL. Η μέθοδος CARNIVAL λύνει ένα πρόβλημα βελτιστοποίησης με τη μέθοδο του ακέραιου προγραμματισμού, ώστε να εξάγει ένα χαρακτηριστικό σηματοδοτικό δίκτυο που εκφράζει στο μεγαλύτερο βαθμό τα πειραματικά δεδομένα. Συνολικά δημιουργήθηκε μια βάση δεδομένων με περισσότερα από 70000 δίκτυα, τα οποία αντιστοιχούν σε 7781 πειράματα γονιδιακής έκφρασης μεταξύ 3005 φαρμάκων και 70 κυτταρικών μοντέλων. Για την ανάλυση τους δημιουργήθηκε ένα unsupervised deep learning μοντέλο, το οποίο δέχεται ως είσοδο τα σηματοδοτικά μονοπάτια και τα αναπαραστεί σε ένα πολυδιαστάτο χώρο. Για την είσοδο τους στο μοντέλο τα σηματοδοτικά δίκτυα κωδικοποιήθηκαν χρησιμοποιώντας τρεις πίνακες που περιγράφουν έκαστος, τους κόμβους, την κατάσταση του κάθε κόμβου και τη συνδεσμολογία του δικτύου. Το μοντέλο είναι ένας transformer δικτύων που δίνει ιδιαίτερη προσοχή σε κάθε γειτονιά του δικτύου, προκειμένου να δημιουργεί τις αναπαραστάσεις

των κόμβων. Η τελική αναπαράσταση του δικτύου δημιουργείται χρησιμοποιώντας ένα μοντέλο Long Short Term Memory προορισμένο για την κωδικοποίηση των αναπαραστάσεων των κόμβων. Το μοντέλο εκπαιδεύεται ώστε να μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ δικτύων και υπο-δικτύων που προκύπτουν από παρόμοια πειράματα γονιδιακής έκφρασης. Τοιουτοτρόπως, η ομοιότητα των αναπαραστάσεων μαρτυρά και συνάμα την ομοιότητα των σηματοδοτικών δικτύων. Οι τελικές αναπαραστάσεις στις 256 διαστάσεις ομαδοποιήθηκαν με τον αλγόριθμο k-means, και οι ομάδες που προέκυψαν αναλύθηκαν ως προς τη σύσταση τους σε μηχανισμούς δράσης φαρμάκων. Ομοίως, δημιούργηθηκε μια νέα μέθοδος για την ανακάλυψη σημαντικών υποδικτύων που οδηγούν στην εκάστοτε ομαδοποίηση, χρησιμοποιώντας τη μέθοδο saliency.

**DeepSNEM αποτελέσματα**

Οι αναπαραστάσεις των σηματοδοτικών δικτύων που δημιουργήθηκαν επαληθεύθηκαν ως προς τη δυνατότητα διαχωρισμού αυτών σε δίκτυα που προέρχονται από το ίδιο πείραμα και από παρόμοια πειράματα. Σε αυτό το στάδιο, η μέθοδος deepSNEM transformer συγκρίθηκε με άλλες 3 μεθόδους deep learning για δίκτυα. Όλες οι μέθοδοι κατάφεραν να διαχωρίσουν τα σηματοδοτικά δίκτυα που προέρχονται από παρόμοια πειράματα, με τη μέθοδο deepSNEM transformer να παρουσιάζει τον καλύτερο διαχωρισμό. Κατά την ομαδοποίηση των αναπαραστάσεων και την ανάλυση των ομάδων, ανακαλύφθηκαν συγκεκριμένες ομάδες, οι οποίες ήταν εμπλουτισμένες με αναστολείς mTOR, topoisomerase, HDAC και protein synthesis. Επιπλέον, χρησιμοποιώντας τη μέθοδο σημαντικών υπο-δικτύων, καταφέραμε να ανακαλύψουμε σημαντικά υποδίκτυα για τις ομάδες των αναστολέων mTOR, τα οποία είναι άμεσα συσχετισμένα με το σηματοδοτικό μονοπάτι mTOR/PI3K. Η μέθοδος deepSNEM εφαρμόστηκε για το χαρακτηρισμό του μηχανισμού δράσης νέων φαρμάκων με δεδομένα γονιδιακής έκφρασης από διάφορες πειραματικές διαδικασίες. Τα αποτελέσματα δείχνουν ότι τα νέα φάρμακα εντάχθηκαν σε ομάδες οι οποίες είναι εμπλουτισμένες με φάρμακα, τα οποία μοιράζονται το μηχανισμό δράσης των νέων φαρμάκων. Επίσης, η εφαρμογή της μεθόδου deepSNEM συγκρίθηκε με μια μέθοδο ανάθεσης ομάδων χρησιμοποιώντας τα δεδομένα γονιδιακής έκφρασης των νέων φαρμάκων. Κατά τη σύγκριση αυτών, παρατηρήθηκε ότι 3/8 φάρμακα, ανατέθηκαν σε ομάδες με παρόμοια σύσταση μηχανισμών δράσης, ενώ τα υπόλοιπα σε ομάδες με διαφορετική σύσταση. Εν τέλει, υπολογίστηκε η ομοιότητα Jaccard μεταξύ των ομάδων που δημιουργούνται χρησιμοποιώντας τις δύο διαφορετικές τεχνικές ομαδοποίησης και βρέθηκε ότι είναι πολύ χαμηλή, με μόνο δύο ομάδες να έχουν ομοιότητα μεγαλύτερη από 0.1.

**DeepSNEM συζήτηση**

Το μοντέλο deepSNEM κατάφερε να αναγνωρίσει ομάδες αναπαραστάσεων, οι οποίες είναι εμπλουτισμένες με φάρμακα συγκεκριμένου μηχανισμού δράσης καθώς και να χαρακτηρίσει σημαντικά υπο-δίκτυα, τα οποία οδηγούν στη συγκεκριμένη ομαδοποίηση. Παρόλα αυτά, το μεγάλο πλήθος φαρμάκων, τα οποία έχουν άγνωστο μηχανισμό δράσης μπορεί να οδηγήσει σε διαφορετική σύσταση των ομάδων που αναγνωρίστηκαν. Επιπλέον, συγκρίνοντας την ανάθεση ομάδων στα φάρμακα της εφαρμογής μεταξύ των μεθόδων deepSNEM και γονιδιακής έκφρασης, γίνεται εμφανές ότι αυτές καταλλήγουν σε διαφορεική ομαδοποίηση των πειραμάτων. Συνεπώς, η μέθοδος deepSNEM, χρησιμοποιώντας την πληροφορία των σηματοδοτικών δικτύων, δύναται να ανακαλύψει ομοιότητες και διαφορές, οι οποίες είναι κρυμμένες στα αρχικά δεδομένα έκφρασης των γονιδίων.

## DeepSIBA: Chemical Structure-based Inference of Biological Alterations using deep learning

**DeepSIBA μέθοδοι**

Δεδομένα γονιδιακής έκφρασης, προερχόμενα από πειράματα χρήσης φαρμάκων, λήφθησαν από την πλατφόρμα L1000 της βάσης δεδομένων CMap. Για κάθε πείραμα, υπολογίστηκαν τα enrichment scores (ES) σημαντικών βιολογικών διεργασιών με τη μέθοδο Gene Set Enrichment Analysis. Εν συνεχεία, η απόσταση μεταξύ των χαρακτηριστικών βιολογικών διεργασιών για ένα ζεύγος φαρμάκων υπολογίστηκε χρησιμοποιώντας μια συνάρτηση βασισμένη στα στατιστικα στοιχεία Kolmogorov-Smirnov. Για κάθε ζέυγος φαρμάκων, η απόσταση υπολογίστηκε για 5 διαφορετικά όρια σημαντικών βιολογικών διεργασιών και ο μέσος όρος τους χρησιμοποιήθηκε ως η μεταβλητή πρόβλεψης ενός μοντέλου deep learning. Το μοντέλο δέχεται σαν είσοδο τις χημικές δομές ενός ζεύγους φαρμάκων και προβλέπει την απόσταση μεταξύ των βιολογικων τους διεργασιών. Οι χημικές δομές ως είσοδοι στο μοντέλο κωδικοποιούνται από τρεις πίνακες που έκαστος περιγράφει τα άτομα, τον τύπο των δεσμών και τη συνδεσμολογία του μοριακού γράφου. Η αρχιτεκτονική του μοντέλου αποτελείται από δύο πανομοιότυπους encoders οι οποίοι χρησιμοποιούν δύο layers graph convolution και ένα layer convolution. Η απόσταση μεταξύ των διεργασιών προβλέπεται χρησιμοποιώντας convolution και fully connected deep learning layers. Η τελική έξοδος του μοντέλου αποτελείται από τη μέση τιμή και τυπική απόκλιση της πρόβλεψης, χρησιμοποιώντας ένα Gaussian regression layer. Θεορώντας ότι η τελική τιμή πρόβλεψης αποτελείται απο μία κατανομή, το μοντέλο εκπαιδεύεται ώστε να ελαχιστοποίησει τη Negative Log-Likelihood. Εν τέλει, εκπαιδεύοντας πολλά μοντέλα, δημιουργείται ένα μοντέλο τύπου ensemble λαμβάνοντας το μείγμα των κατανομών πρόβλεψης. Επίσης, αναπτύχθηκε μία νέα μέθοδος πρόβλεψης των σηματοδοτικών μονοπατιών ενός φαρμάκου, βασιζόμενη στη μέθοδο kNN, χρησιμοποιώντας μόνο τη χημική τους δομή ως είσοδο. Σε αυτή τη μέθοδο, το μοντέλο deepSIBA, χρησιμοποιείται για την ανακάλυψη γειτονικών φαρμάκων τα οποία επηρεάζουν παρόμοιες βιολογικές διαδικασίες με το φάρμακο προς ανάλυση. Στη συνέχεια, χρησιμοποιώντας μία μέθοδο ψηφοφορίας μεταξύ των γειτόνων, δημιουργείται μια πρόβλεψη σηματοδοτικών μονοπατιών για το αρχικό φάρμακο. Τέλος με σκοπό την καλύτερη ερμηνέια των αποτελεσμάτων, δημιουργήθηκε μία μέθοδος βασιζόμενη στις παραγώγους του μοντέλου ώς προς την είσοδο του, η οποία μπορεί να ανακαλύψει ποιες χημικές υποδομές συνεισφέρουν περισσότερο στη δεδομένη πρόβλεψη σηματοδοτικών μονοπατιών.

**DeepSIBA αποτελέσματα**

Οι επιδόσεις του μοντέλου εξετάστηκαν σε ζεύγη γνωστών και άγνωστων φαρμάκων. Τα άγνωστα φάρμακα αποτελούν χημικές ενώσεις οι οποίες δεν χρησιμοποιήθηκαν κατά τη διαδικασία εκπαίδευσης του μοντέλου. Επιπλέον, οι επιδόσεις του μοντέλου deepSIBA συγκρίθηκαν με τις επιδόσεις τριών μοντέλων machine learning, που είναι ειδικά σχεδιασμένα για προβλήματα πρόβλεψης αποστάσεων. Σε όλους τους ελέγχους, το μοντέλο deepSIBA παρουσίασε καλύτερα αποτελέσματα σε σύγκριση με τις υπόλοιπες τεχνικές. Με σκοπό την αύξηση του εύρους εφαρμογής της μεθόδου, χρησιμοποιήθηκε μια τεχνική transfer learning, και εφαρμόστηκε με επιτυχία σε κυτταρικές σειρές με λιγότερα δεδομένα. Εν συνεχεία, εξετάστηκαν οι επιδόσεις του μοντέλου σε ένα ειδικό σενάριο που τα φάρμακα-είσοδοι είναι εντελώς διαφορετικά σε σχέση με αυτά που

χρησιμοποιήθηκαν κατά τη διάρκεια της εκπαίδευσης. Σε αυτή την ειδική περίπτωση, δείξαμε ότι χρησιμοποιώντας την πρόβλεψη της αβεβαιότητας, μπορούμε να επικεντρωθούμε σε προβλέψεις που το μοντέλο δείχνει μεγαλύτερη σιγουρία και οι οποίες έχουν μεγαλύτερη πιθανότητα να είναι εύστοχες. Τέλος, κατά την εφαρμογή του deepSIBA στις χημικές δομές εγκεκριμένων αντικαρκινικών φαρμάκων, ανακαλύφθηκαν χαρακτηριστικά σηματοδοτικά μονοπάτια και χημικές υποδομές οι οποίες είναι άμεσα συνδεδεμένες με το μηχανισμό δράσης των φαρμάκων.

**DeepSIBA συζήτηση**

Το μοντέλο κατάφερε με επιτυχία και ακρίβεια να ανακαλύψει διαφορετικές χημικές δομές, οι οποίες παρουσιάζουν παρόμοια βιολογική επίδραση σε συγκεκριμένα κυτταρικά μοντέλα. Επιπλέον, χρησιμοποιώντας την πρόβλεψη της αβεβαιότητας, το μοντέλο κατάφερε να διατηρήσει υψηλές επιδόσεις σε ζεύγη φαρμάκων τα οποία διαφέρουν πολύ απο τα παραδείγματα κατά τη διαδικασία της εκπαίδευσης. Παρόλα αυτά, υπήρχαν αρκετά ζεύγη φαρμάκων με παρόμοια βιολογική επίδραση τα οποία δεν ανακαλύφθηκαν από το μοντέλο. Πιστεύουμε πως αυτό οφείλεται στην σχετικά χαμηλή κάλυψη του χώρου των δυνατών χημικών ουσιών που προσφέρουν τα διαθέσιμα δεδομένα της βάσης CMap. Τέλος, πιστεύουμε πως η χρήση περισσότερων δεδομένων γονιδιακής έκφρασης θα οδηγήσει στην άμεση βελτίωση των αποτελεσμάτων του μοντέλου.

## Συμπεράσματα

Τα μοντέλα που παρουσιάστηκαν έχουν τη δυνατότητα να βοηθήσουν τη διαδικασία υπολογιστικής ανακάλυψης φαρμάκων είτε συσχετίζοντας τα σηματοδοτικά δίκτυα με το μηχανισμό δράσης των φαρμάκων, είτε ανακαλύπτοντας νέες χημικές δομές που επηρεάζουν επιθυμητές βιολογικές διεργασίες.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

## 1.1 Background

Drug discovery is a complex and time-consuming process that aims to identify the right drug, for the right target and disease. The drug discovery pipeline stretches from target identification to hit discovery and lead optimization up to preclinical and clinical trials [1]. The first step of the pipeline is the target identification, where a disease is studied in order to understand its mechanism and identify key molecules that act as drivers of the disease. These key molecules have the potential to act as targets for small molecule perturbations in order to stop the disease progression or reverse its state [2]. After a potential set of targets has been identified, they have to be prioritized based on their ability to inhibit the disease mechanism(s) and to eliminate targets that are associated with adverse effects. This step of the pipeline, called target validation, is performed experimentally using small interfering RNA (siRNA) or CRISPR-cas9 editing to specifically silence or catalyze the specified targets [3]. The next step of the pipeline is the hit identification phase, where a small molecule that inhibits the identified target(s) has to be discovered. The goal of this process is to discover chemical compounds that exhibit strong binding affinity to the selected targets. Traditionally, the most widely employed method for *in-vitro* hit identification is High Throughput Screening (HTS). *In-vitro* HTS can produce hits with strong binding affinity that may later be developed into lead compounds through lead optimization [4]. However, due to the vast chemical space of possible chemical structures, even large scale in-vitro HTS offers limited chemical coverage and does not guarantee the biological efficacy and low toxicity of the identified hit compounds. During lead optimization the chemical compounds are optimized in order to improve their chemical properties and study their off-target effects, which could potentially cause adverse reactions, such as unwanted side effects and toxicity. In order to improve the success rate of the drug discovery pipeline, computational methods have been developed that aim to prioritize compounds in the hit identification and lead optimization phases.

The development of Computer Aided Drug Design (CADD) methods allows the virtual High Throughput Screening (vHTS) of large compound datasets, thus effectively increasing the search space of hit identification. CADD methods for vHTS prioritize compounds, which are likely to have activity against the target, for further experiments and are broadly categorized into structure-based and ligand-based [5]. Structure-based CADD approaches require the solved 3D structure of the target protein, either through X-ray crystallography or NMR spectroscopy and focus on docking simulations to assess protein-ligand complexes. On the

other hand, ligand-based virtual screening is used when the 3D structure of the target is unknown and involves the calculation of 2D or 3D structural similarities between a known active ligand and a virtual library. Structural similarity screening is based on the hypothesis that similar chemical structures will cause similar response. Even though CADD methods have revolutionized the drug development pipeline, the attrition rates of the process still remain high. The majority of the newly identified compounds fail at the preclinical trial stages due to poor efficacy or unfavorable side effects and toxicity profiles [6]. This happens in part, due to the nature of CADD methods, where efforts are focused on identifying compounds with maximal binding affinity to the target protein, often disregarding the effect that the compound will have on a biological system. A field that holds promise to improve the attrition rates of drug discovery, by understanding the compound's effect on the biological system, is systems pharmacology.

Systems pharmacology utilizes omics data coupled with computational methods in order to study the mechanism of action (MoA) of a chemical perturbation in a biological system. One specific omics approach that has gained considerable attraction is the analysis of gene expression data (transcriptomics) following treatment with a compound. Analyzing the post transcriptional state of a cellular system after a chemical perturbation has the potential to elucidate the compound's effect in term of which key genes are over- or under-expressed compared to the normal state. On this front, a large number of public gene expression data repositories have been developed, such as the Gene Expression Omnibus (GEO) from NCBI and ArrayExpress [7,8]. These repositories can be accessed to retrieve gene expression data from various cellular models following compound treatment that can then be analyzed using computational tools. For their analysis the Bioconductor library in R offers a large collection of packages for preprocessing and differential expression that can be applied on gene expression data from various platforms [9]. The output of the differential expression analysis is a set of genes that are over- or under-expressed in the condition under study. However, a compound's effect in a cellular system is rarely the effect of the change in expression of specific genes, rather the compound's effect is caused by changes in the expression of genes that interact with each other to form specific biological processes [10]. Additionally, since the analysis focuses on the most over- or under-expressed genes, the smaller but significant change in expression of genes that belong to the same biological pathways are often disregarded. For these reasons, a majority of computational methods that aim to identify enriched biological processes affected by a compound, using gene expression data, have been developed [11].

Pathway analysis methods utilize the results of the differential expression analysis, coupled with a form of prior knowledge of molecular interactions, in order to identify which pathways are affected by a compound. Today there exist several knowledge bases of pathway interactions, such as the KEGG database, Reactome, MsigDB, etc [12-14]. Computational methods that aim to identify enriched pathways are based on the assumption that the change in a gene's expression is translated to changes in the proteins' that are encoded by it. The majority of computational tools for pathway analysis aim to extract a score statistic for each pathway that signifies its enrichment, accompanied by a p-value that compares the

enrichment score to chance. One of the most widely used methods for pathway analysis is Gene Set Enrichment Analysis (GSEA). In GSEA the list of differentially expressed genes is ordered based on their expression and a statistic based on the Kolmogorov-Smirnoff test is calculated for pathway enrichment [15]. Additionally the VIPER algorithm utilizes the mean of ranks of genes' expression values with an analytic Rank-based Enrichment Analysis method to compute the enrichment of proteins from gene expression data. VIPER has been utilized to infer Transcription Factor (TF) activity scores using gene expression data and the appropriate Regulons, which are networks that show the relationship between the TFs and the genes' expression [16]. In order to overcome the limitations of pathway analysis, regarding the hypothesis of the connection between gene and protein expression, a new class of methods that model the compound's effect as a network of protein-protein interactions (PPI), has been developed [10]. These methods couple gene expression data with prior knowledge networks in a causal reasoning scheme to identify which sub-networks better explain the observed experimental data. Initially network creation methods have utilized phosphoproteomic data to describe the compound's effect [17]. Since large scale phosphoproteomic datasets following compound treatment are very rare, there has been a concentrated effort to develop methods for signaling network creation based on transcriptomics [18-20]. On this front, the CARNIVAL method is a causal reasoning framework to identify signaling networks that best explain a set of transcription factor (TF) activity scores, calculated from differential GEx data using VIPER [21]. The resulting networks are complex representations of the compounds' effect, since they incorporate the prior knowledge of molecular interactions in the form of a PPI network.

There have been many studies that utilize gene expression data along with pathway and network analysis methods to investigate the compounds' MoA in biological systems. One of the most influential approaches that have been widely used in this field is the Connectivity Map (CMap) approach [22]. CMap (CMap) and the LINCS project have been a cornerstone of transcriptomic-based approaches by providing a large scale database of transcriptomic signatures from compound perturbations along with essential signature matching algorithms. CMap can be accessed to query a large database of transcriptomic signatures in order to find compounds that have similar gene expression profiles. CMap's approach is based on the hypothesis that compounds with similar transcriptomic signatures will cause similar physiological effects on the cell and has been widely adopted by the field of drug repurposing. The original Microarray CMap dataset along with the more recent L1000 dataset have been used by many systems pharmacology studies to investigate the effect of compounds on the biological system. On this front, Iorio et al. analyzed similarities between drugs' transcriptional responses from CMap to create a drug network and identified the mechanism of action of new drugs based on their position in the network [23]. Furthermore, Verbist et al. showed how GEx data were able to influence decision making in eight drug discovery projects by uncovering potential adverse effects of the lead compounds [24]. Although, systems pharmacology approaches can be used to identify a compound's MoA at the later stages of the drug development pipeline, they don't take into account the structural elements of the chemical compound, which are crucial during the hit identification phase. Given the large

datasets generated by vHTS and chemical libraries, as well as the rise in computing power, machine learning and especially deep learning methods have been developed and applied across all stages of drug discovery. Deep learning models have the potential to incorporate both systems-based and structural-based approaches in order to identify drugs with optimal binding affinity and effect on the biological system.

Machine learning (ML) models are trained on data in order to improve their predictions for a specific task. The rise of big data and computing power has led to the development of a special class of ML models, called deep learning (DL). DL models offer the advantage of automatic feature extraction in order to learn the important features that are associated with the specific task. This advantage has important applications in the field of drug discovery, where the complexity of the problem is very high and pre-computed features are usually associated with a specific task. There have been many studies for the development and application of deep learning models in drug discovery. The majority of these approaches utilize the compound's chemical structure as input and have been developed for various tasks, including binding affinity, toxicity, side effect and chemical property prediction. In this regard, the DeepChem library along with MoleculeNet have been a cornerstone of DL approaches in drug discovery by providing a plethora of architectures along with benchmark datasets for their comparison [25]. For example, Ozturk et al. developed a deep learning model that encodes a compound's SMILE representation and a protein's amino acid sequence in order to predict the binding affinity, in terms of the dissociation constant Kd, of the drug-target pair [26]. For the binding affinity and property prediction task, deep learning models have achieved state of the art results and have outperformed traditional ML methods like random forests and support vector machines [27]. Despite their improved performance, DL models are still very sensitive to the training dataset and have shown generalization errors, when tested on chemical structures that are very different from the ones used to train them. This effect is mostly caused by the vast size of the chemical space and its small coverage by the training datasets. In order to address this issue, methods like one-shot learning and uncertainty estimation are crucial. One-shot learning techniques, such as Siamese and Matching networks, aim to learn a meaningful distance function between related inputs and have shown increased performance over traditional methods in tasks with few data points [28-31]. Furthermore, uncertainty estimation methods can be used in order to quantify the model's confidence in the predictions and avoid unnecessary experimental testing of new compounds. Methods that quantify uncertainty in deep learning models include test-time Dropout, deep ensembles and Bayesian NNs [32-35]. One particular deep learning approach that has achieved state of the art results in structure-based tasks is the representation of the chemical structure as a molecular graph and the use of graph deep learning models to encode them.

Graph deep learning models operate on graph structured data and aim to extract features that are representative of the graph's nodes and connectivity. One of the most widely used graph DL models is the graph convolutional neural network (GCNN). GCNNs apply filters on the neighborhoods of the graph and utilize a message passing algorithm to aggregate this information into representations of the graph's node attributes and connectivity [36]. GCNNs have been successfully applied to various drug discovery tasks, achieving state of the art

results. The input to molecular GCNNs is a compound undirected graph, with atoms being the nodes of the graph and bonds being the edges. As an example, Torng et al. applied GCNNs to encode both the compound's structure and the protein's binding pocket to predict the protein-ligand binding strength, outperforming traditional DL approaches [37]. Furthermore, Kearnes et al. developed the Weave graph convolution module, which encodes both atom and bond representations and combines them using fuzzy histograms to extract meaningful molecule-level representations [38]. GCNNs can be combined with one-shot learning methodologies in order to learn representations and distance functions between compound graphs, aiming to improve their generalization capabilities on new chemical structures that are very different from the ones used to train them. Altae-Tran et al. implemented one-shot learning for drug discovery by combining graph convolutions and Long Short Term Memory (LSTM) networks with attention and achieved better results than traditional GCNNs [39]. Recently, DL models that had been originally developed for Natural Language Processing (NLP) tasks have been modified and applied on graphs. These models encode the nodes of the graph as words in a sentence and their positioning based on the graph's connectivity. For example, the graph2vec model was inspired by the doc2vec approach for NLP tasks. Graph2vec treats the entire graph as a document and each node's neighborhood as a word and aims to learn a fixed-length representation of the entire graph in a fully unsupervised task [40]. Furthermore, the graph transformer model was developed that utilizes an attention mechanism for each node that is a function of the neighborhood's connectivity, rather than a message passing algorithm [41]. Although deep learning models have been applied in various structure-based learning tasks, their application in systems-based approaches is still very limited. DL and graph DL models can be applied for various systems pharmacology tasks in order to investigate the relationship between omics data following compound treatment and the compound's MoA.

Recently, there has been increased interest in the application of deep learning models for systems pharmacology approaches that utilize cellular response data. This is evident by the recent release of the $CTD^2$ Pancancer Drug Activity DREAM Challenge, which tasked the community to predict a compound's MoA based on post-transcriptional and cell viability data [42]. Additionally, deep learning models have been applied to predict the IC50 of compounds on specific cellular models by using transcriptomic data along with structural data [43]. Deep learning models present an interesting modeling opportunity for interdisciplinary drug discovery problems, by being able to incorporate information both from the structural and cellular domain. On this front, Jeon et al. developed the ReSimNet model to predict the transcriptional similarity score between compound perturbations using their molecular fingerprints as input [44].

The aim of this thesis is to investigate the ability of graph deep learning models to model a compound's MoA in terms of affected biological processes and molecular targets, by combining information from both the structural and systems domain. The thesis is organized into three distinct but complementary chapters. In the first chapter of the thesis we present a concise review of network pharmacology approaches for early drug discovery, while in the second and third chapters we present two novel graph deep learning approaches, called deepSIBA and deepSNEM. More specifically, the first chapter contains a thorough review of

available omics databases and knowledge bases of molecular interactions, along with network-based methods for their analysis. In the first chapter we also review a large number of network pharmacology studies ranging across all stages of the drug development pipeline.

In the second chapter we employ graph deep learning to develop a pipeline that can assess a compound's affected biological processes based on its chemical structure (deepSIBA). DeepSIBA is a Siamese GCNN that takes as input pairs of compound structures, represented as graphs and outputs their biological effect distance, in terms of enriched biological processes (BPs) along with an estimated uncertainty. The performance of DeepSIBA is evaluated in realistic drug development scenario, where GEx data are available for only one compound in a pair. Additionally, we present a novel inference pipeline to identify the affected signaling pathways of a chemical compound along with the important substructures that cause the effect. As a use case, DeepSIBA was used to successfully infer the signaling pathway signature of FDA approved anticancer drugs. DeepSIBA can be used in combination with existing *in-silico* drug discovery pipelines to identify structures that not only exhibit maximal binding affinity but also cause a desired biological effect.

In the third chapter of the thesis we employ graph deep learning to investigate the relationship between compound-induced signaling networks and the compounds' MoA. We present deepSNEM, a novel pipeline that encodes a compound's signaling network into a unique representation and assesses its relationship with the compound's MoA. The core of deepSNEM is an unsupervised graph transformer trained to maximize the mutual information between representations of graphs' substructures that belong to signaling networks created from similar perturbations. The network embeddings were clustered with the k-means algorithm and the resulting clusters were analyzed and characterized based on their MoA composition. Furthermore, a subgraph importance method was developed in order to identify which nodes and subgraphs in the original signaling networks cause the embeddings to cluster this way. As a use case, deepSNEM was used to assign clusters and assess the MoA of compounds with Gene expression data collected from various experimental platforms. DeepSNEM can be applied to generate hypotheses regarding the MoA of new lead compounds or suggest new potential mechanisms for already existing drugs.

# 1.2 References

1. Dickson, Michael, and Jean Paul Gagnon. "The cost of new drug discovery and development." Discovery medicine 4.22 (2009): 172-179.

2. Schenone, Monica, et al. "Target identification and mechanism of action in chemical biology and drug discovery." Nature chemical biology 9.4 (2013): 232-240.

3. Moore, Jonathan D. "The impact of CRISPR–Cas9 on target identification and validation." Drug discovery today 20.4 (2015): 450-457.

4. Rishton, Gilbert M. "Reactive compounds and in vitro false positives in HTS." Drug discovery today 2.9 (1997): 382-384.

5. G. Sliwoski, S. Kothiwale, J. Meiler and E. W. Lowe, Pharmacological reviews, 2014, 66, 334–395.

6. Kola, Ismail, and John Landis. "Can the pharmaceutical industry reduce attrition rates?." Nature reviews Drug discovery 3.8 (2004): 711-716.

7. Clough, Emily, and Tanya Barrett. "The gene expression omnibus database." Statistical genomics. Humana Press, New York, NY, 2016. 93-110.

8. Brazma, Alvis, et al. "ArrayExpress—a public repository for microarray gene expression data at the EBI." Nucleic acids research 31.1 (2003): 68-71.

9. Gentleman, Robert C., et al. "Bioconductor: open software development for computational biology and bioinformatics." Genome biology 5.10 (2004): 1-16.

10. Fotis, Chris, et al. "Network-based technologies for early drug discovery." Drug discovery today 23.3 (2018): 626-635.

11. García-Campos, Miguel A., Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. "Pathway analysis: state of the art." Frontiers in physiology 6 (2015): 383.

12. Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28.1 (2000): 27-30.

13. Fabregat, Antonio, et al. "The reactome pathway knowledgebase." Nucleic acids research 46.D1 (2018): D649-D655.

14. Liberzon, Arthur, et al. "Molecular signatures database (MSigDB) 3.0." Bioinformatics 27.12 (2011): 1739-1740.

15. Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences 102.43 (2005): 15545-15550.

16. Alvarez, Mariano J., et al. "Functional characterization of somatic mutations in cancer using network-based inference of protein activity." Nature genetics 48.8 (2016): 838-847.

17. Mitsos, Alexander, et al. "Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data." PLoS computational biology 5.12 (2009): e1000591.

18. Bradley, Glyn, and Steven J. Barrett. "CausalR: extracting mechanistic sense from genome scale data." Bioinformatics 33.22 (2017): 3670-3672.

19. Chindelevitch, Leonid, et al. "Causal reasoning on biological networks: interpreting transcriptional changes." Bioinformatics 28.8 (2012): 1114-1121.

20. Melas, Ioannis N., et al. "Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury." Integrative Biology 7.8 (2015): 904-920.

21. Liu, Anika, et al. "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL." NPJ systems biology and applications 5.1 (2019): 1-10.

22. Lamb, Justin, et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." science 313.5795 (2006): 1929-1935.

23. F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri and A. Isacchi, Proceedings of the National Academy of Sciences, 2010, 107, 14621–14626.

24. B. Verbist, G. Klambauer, L. Vervoort, W. Talloen, Z. Shkedy, O. Thas, A. Bender, H. W. Göhlmann, S. Hochreiter and QSTAR Consortium, Drug discovery today, 2015, 20, 505–513.

25. DeepChem: Deep-learning models for Drug Discovery and Quantum Chemistry.

26. Öztürk, Hakime, Arzucan Özgür, and Elif Ozkirimli. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

27. Nguyen, Thin, Hang Le, and Svetha Venkatesh. "GraphDTA: prediction of drug–target binding affinity using graph convolutional networks." BioRxiv (2019): 684662.

28. F. Schroff, D. Kalenichenko and J. Philbin, 2015, pp. 815–823.

29. L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. Torr, Springer, 2016, pp. 850–865.

30. O. Vinyals, C. Blundell, T. Lillicrap and D. Wierstra, 2016, pp. 3630–3638.

31. Y. Bai, H. Ding, Y. Sun and W. Wang, *arXiv preprint arXiv:1810.10866*.

32. Y. Gal and Z. Ghahramani, 2016, pp. 1050–1059.

33. B. Lakshminarayanan, A. Pritzel and C. Blundell, 2017, pp. 6402–6413.

34. S. Jain, G. Liu, J. Mueller and D. Gifford, arXiv preprint arXiv:1906.07380.

35. A. Kendall and Y. Gal, 2017, pp. 5574–5584.

36. Duvenaud, David, et al. "Convolutional networks on graphs for learning molecular fingerprints." arXiv preprint arXiv:1509.09292 (2015).

37. W. Torng and R. B. Altman, Journal of Chemical Information and Modeling, 2019, 59, 4131–4149.

38. S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, Journal of computer-aided molecular design, 2016, 30, 595–608.

39. H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, ACS central science, 2017, 3, 283–293.

40. Narayanan, Annamalai, et al. "graph2vec: Learning distributed representations of graphs." arXiv preprint arXiv:1707.05005 (2017).

41. Yun, Seongjun, et al. "Graph transformer networks." Advances in Neural Information Processing Systems 32 (2019): 11983-11993.

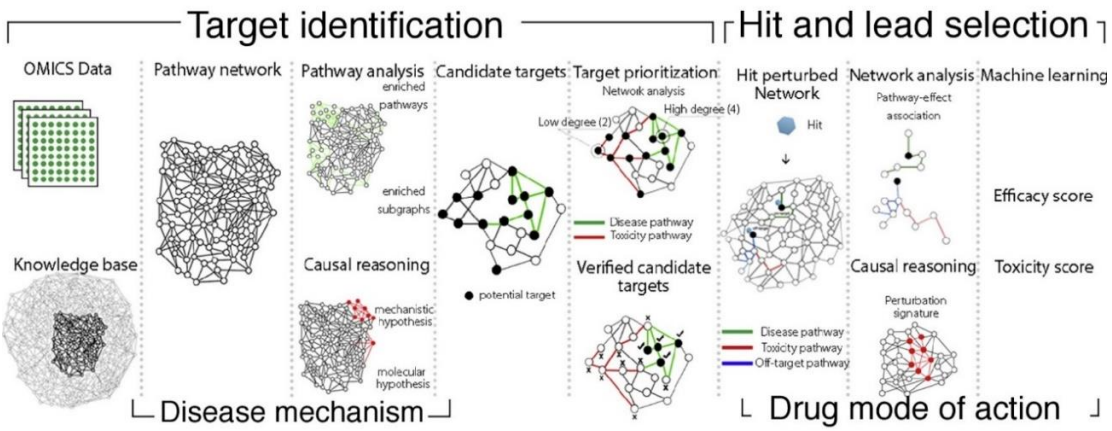
42. Douglass, Eugene F., et al. "A community challenge for pancancer drug mechanism of action inference from perturbational profile data." (2020).

43. Joo, Minjae, et al. "A deep learning model for cell growth inhibition ic50 prediction and its application for gastric cancer patients." International journal of molecular sciences 20.24 (2019): 6276.

44. Jeon, Minji, et al. "ReSimNet: drug response similarity prediction using Siamese neural networks." Bioinformatics 35.24 (2019): 5249-5256

# Chapter 2

# Network-based technologies for early drug discovery




## 2.1 Chapter abstract

Although the traditional drug discovery approach has led to the development of many successful drugs, the attrition rates remain high. Recent advances in systems-oriented approaches (systems-biology and/ or pharmacology) and 'omics technologies has led to a plethora of new computational tools that promise to enable a more-informed and successful implementation of the reductionist, one drug for one target for one disease, approach. These tools, based on biomolecular pathways and interaction networks, offer a systematic approach to unravel the mechanism(s) of a disease and link them to the chemical space and network footprint of a drug. Drug discovery can draw upon this holistic approach to identify the most-promising targets and compounds during the early phases of development.

# 2.2 Introduction

Drug discovery is a complex and time-consuming process that stretches from target selection and validation, through preclinical screening, to clinical trials and regulatory agency approval [1]. Traditionally, pharmaceutical companies have adopted a reductionist approach regarding the discovery of a new drug, which focuses on identifying one drug for one target for one disease. Although this approach has led to the development of many successful drugs, the attrition rates associated with this pipeline remain high [2]. In a recent study of the cause of attrition of drug candidates from four major pharmaceutical companies by Waring et al., the authors highlighted that efficacy issues accounted for 9% and 35% of overall terminated compounds during Phases I and II, respectively, whereas clinical safety problems accounted for 25% of terminated compounds across both Phases [3]. During the early phases of clinical trials, efficacy and clinical safety issues arise mostly because of the inherent complexity of the biological system and partly because of a lack of in-depth knowledge of the mechanism(s) of disease or the mode of action (MoA) of a drug [4]. Breakthroughs in systems-approaches (systems biology and systems pharmacology), driven by the latest technological advancements both in experimental technologies and computational methods, allow researchers to consider the system as a whole, with individual biomolecules interacting with each other and, thus, permitting their integration into classes of higher order. These classes, called pathways, are sets of molecules acting in concert, usually involved in a particular function or process [5]. Currently, there are several computational methods for the analysis of biological data at the pathway and network level that connect molecular data from a variety of 'omics databases (genomics, transcriptomics, proteomics, and metabolomics) to their biological functions using knowledge bases as templates to build associations between them. These methods can help decipher the mechanism(s) of disease or the MoA of a drug by offering a more- holistic view of the whole system and should enable one to pick the most-promising targets and model the effect of their modulation during the early phases of drug discovery. On this front, the focus on pathways as functional units rather than on individual biomolecules also increases explanatory power and eases result interpretation [6].

In this review, we focus on the combination of 'omics experiments and pathway- and network-based approaches for early drug discovery, to connect basic research on pathway models to the actual needs of the early drug discovery pipeline. We explore the plethora of tools that can be used to unravel the signaling mechanism(s) of a disease, tools that can decipher the signaling footprint of a target or a drug and how these can be used together for: (i) target identification; (ii) target verification and validation; and (iii) hit discovery and validation (Fig. 2.1). Three main elements are common to all these tools: molecular data related to the condition under study (in-house data and public/private repositories), knowledge bases (functional annotation of molecules and information about drugs, clinical trials, and biological pathways), and the appropriate computational method to analyze this complex combination.

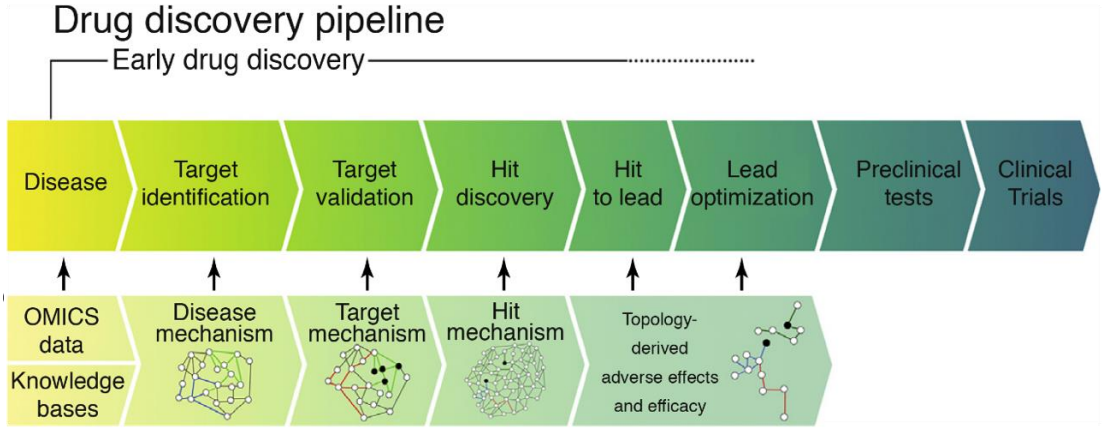



Figure 2.1. Schematic overview of the drug discovery pipeline (a) depicting the area of application of pathway-and network-based technologies (b). By using pathway technologies, the decision from target to hit to lead is based on network footprints. From target identification and validation, to hit discovery and lead selection, pathway- and network-based tools can help elucidate the disease mechanism, prioritize targets belonging to identified deregulated pathways, assess the importance of those targets, and unravel the MoA of a hit compound to predict adverse effects and efficacy.

# 2.3 Data gathering and integration

A drug discovery project usually starts with a focus on a disease that requires a new or a better therapeutic intervention. In the era of systems approaches, this can be achieved by identifying the diverse mechanisms that lead to disease as well as the optimal targets and/or drugs that can eradicate those mechanisms. The first step in building a quantitative representation of the biological processes of a cell and its alterations in the disease is data gathering. The different types of 'omics data (DNA, RNA, proteins, and metabolites) [7, 8] report on different levels of cell or organismal function. 'Omics data can be generated in-house or gathered from many publicly available repositories, such as the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/), the Proteomics IDEntifications (PRIDE; www.ebi.ac.uk/pride/archive/) or ArrayExpress (www.ebi.ac.uk/arrayexpress/) or mined from the literature by using text-mining methods [9,10] (Table 2.1). The different technologies used to generate these types of data result in a diverse set of quantitative data that makes their normalization and integration a daunting task [i.e., picks on liquid chromatography–mass spectrometry data (LC–MS), fluorescent intensity of affinity multiplexed assays and expression microarrays, copy number on next-generation sequencing, etc.].

The analysis of a single type of data (e.g. gene expression) is relatively easy and might provide useful information on its own. However, successful integration and modeling of different types of biomolecule together in healthy and disease states is a major endeavor of many computational efforts and can enhance the knowledge of the mechanism of a disease or the MoA of a drug [9]. In a recent review, Cisek et al. highlighted that, although genomics attempts to map phenotypic features to genetic background with genome-wide association studies (GWAS), it is only able to identify single nucleotide polymorphisms (SNPs), but not risk genes [11,12]; and, although transcriptomics can identify risk genes, it does not include information on protein expression, interaction, and post-translational modification [13]. By contrast,

proteomics can provide information about protein interactions, but cannot capture the function of a protein in its metabolic pathway [14]. Metabolomics is the final missing link that completes the circle of 'omics, providing functional information about proteins when influenced by their cellular environment [15]. System-based approaches require functional annotation of the molecules participating in biological processes, but such annotation can be challenging even for biomolecules belonging to the same 'omic domain. To facilitate data integration, a large number of public databases (also known as knowledge bases) exist, whose aim is to collect causal, correlational, functional, and contextual information about biomolecules and serve as templates for the association of individual biological entities into the aforementioned classes of higher order (pathways) [8].

Table 2.1. List of public omics repositories. [1]

| Name | Link | Brief description | Category [2] |
|---|---|---|---|
| GEO | https://www.ncbi.nlm.nih.gov/geo/ | Gene Expression Omnibus | gene expression |
| ArrayExpress | https://www.ebi.ac.uk/arrayexpress/ | Archive of Functional Genomics Data | gene expression |
| GeneSigDB | https://www.genesigdb.org/genesigdb/index.jsp | Gene Signature DataBase | gene expression |
| Oncomine | https://www.oncomine.org/resource/login.html | Cancer microarray data by gene or cancer type | gene expression |
| Expression Atlas | https://www.ebi.ac.uk/gxa/home | Gene expression across species and biological conditions | gene expression |
| miRGator | http://mirgator.kobic.re.kr/ | microRNA target prediction, functional analysis and gene expression data | gene expression |
| PRIDE | https://www.ebi.ac.uk/pride/archive/ | Proteomics peptide identification database | proteomics |
| PaxDB | https://pax-db.org/ | Protein Abundance Database | proteomics |
| The Human Protein Atlas | https://www.proteinatlas.org/ | MS-based proteomics, transcriptomics and antibody-based imaging | proteomics |

| | | | |
|---|---|---|---|
| Open Proteomics Database | https://www.hsls.pitt.edu/obrc/index.php?page=URL1152112355 | MS-based proteomics | proteomics |
| ProteomeXchange | http://www.proteomexchange.org/ | Proteomics resources portal | proteomics |
| HMDB | http://www.hmdb.ca/ | Curated human metabolism and metabolite data | metabolomics |
| BiGG Models | http://bigg.ucsd.edu/ | Biochemically, Genetically and Genomically structured metabolic networks | metabolomics |
| MetaboLights | https://www.ebi.ac.uk/metabolights/ | A database for metabolomics experiments and the associated metadata | metabolomics |
| Metabolomics Workbench | http://www.metabolomicsworkbench.org/ | Data repository for metabolomics data and metadata | metabolomics |
| GenBank | https://www.ncbi.nlm.nih.gov/genbank/ | An annotated collection of all publicly available nucleotide and protein sequences | DNA |
| dbSNP | https://www.ncbi.nlm.nih.gov/SNP/ | Database of single nucleotide polymorphisms | DNA |
| Ensembl | https://www.ensembl.org/index.html | Annotated information on eukaryotic genomes | DNA |
| ENA | https://www.ebi.ac.uk/ena | European Nucleotide Archive | DNA |
| UniProt | http://www.uniprot.org/ | All known protein sequences | Protein |

| | | | |
|---|---|---|---|
| PDB | https://www.rcsb.org/pdb/home/home.do | Protein structures | Protein |
| PDBe | https://www.ebi.ac.uk/pdbe/ | European resource for protein structures | Protein |

[1]Indicative list of open-access databases. For a full list of databases see Nucleic Acids Research Database Summary (http://www.oxfordjournals.org/nar/database/a/).

[2]A database may correspond to more than one categories but only the major one is shown here

# 2.4 Knowledge bases

Knowledge bases can associate the data gathered during the drug discovery and development pipeline and include clinical outcomes, drug information (i.e., chemical structure, adverse effects, etc.) or biological knowhow of drugs. Here, we focus on pathway and interactome databases that aim to capture the biological knowledge of molecular interactions and have been assembled from experimental data or through text mining followed by manual curation (Table 2.2). The result is an interaction model, either represented as a biological network or a bipartite graph [16]. There are different types of interaction models depending on the relationships that they represent. Signaling networks account for cellular processes, whereas metabolic networks represent the bio- chemical reactions of metabolism as well as the regulatory interactions that guide these reactions. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases (www.genome.jp/kegg/) providing systems, chemical, and health information that serves as a computer representation of the biological system. Most notably, the KEGG Pathway database contains manually drawn pathway maps in the form of directed graphs representing the interactions between genes and proteins. By contrast, Protein–Protein Interaction (PPI) maps represent physical interactions on a molecular level. As an example, InWeb_IM (www.intomics.com/inbio/map/#home) is a recently proposed interactome resource in the form of a human PPI network that can be used for accurate and costefficient functional interpretation of massive genomic datasets. IntAct is another manually curated open-access database focusing on PPIs (www. ebi.ac.uk/intact/), which was recently merged with the Molecular INTeraction (MINT) database (http://mint.bio.uniroma2.it/) and now includes more than 700000 binary protein interactions. In addition, the Reactome [17] Functional Interaction network (Reactome FI) is a manually curated protein functional interaction network organized in pathways covering over 60% of human proteins. The connection between two disjoint sets of nodes is considered In bipartite graphs, such as between genes and diseases or between drugs and their targets. As an example, the connectivity map (Cmap) is a widely used database (https://portals. broadinstitute.org/cmap/) connecting drugs and gene expression profiles that allows researchers to identify connections between drug candidates, and changes in gene expression profiles and diseases by using the similarity detection tool that the database offers. The Biological General Repository for Interaction Datasets (BioGRID) is a public database dedicated to manually curated functional interactions between genes and physical interactions between proteins, reported in biomedical publications (https://thebiogrid.org/). Recently, BioGRID expanded its coverage to

incorporate chemical–protein interactions and established themed curation projects based on particular biological processes and diseases to further facilitate early drug discovery. Besides open-source initiatives and free web-based tools, several commercial providers also focus on pathways relevant to human disease. Companies such as Ingenuity Systems (IPA), GeneGo, Ariadne Genomics, and Cambridge Cell Networks provide manually curated and usually high-quality pathway databases and analysis tools.

Those knowledge bases are the cornerstone of many computational tools that aim to extract the most-valuable information and deliver it to the end user.

Table 2.2. List of widely used knowledge bases.[1]

| **KEGG PATHWAY** | **http://www.genome.jp/kegg/pathway.html** | **Wiring diagrams of molecular interactions, reactions and relations** | **Metabolic and signaling pathways** |
|---|---|---|---|
| **Reactome** | https://reactome.org/ | Manually curated and peer reviewed pathway database | Metabolic and signaling pathways |
| **Pathway Commons** | http://www.pathwaycommons.org/ | A web resource for biological pathway data | Metabolic and signaling pathways |
| **NetPath** | http://www.netpath.org/ | A curated resource of signal transduction pathways | Signaling pathways |
| **OmniPath** | http://omnipathdb.org/ | A collection of curated signaling pathways | Signaling pathways |
| **BioCyc** | https://biocyc.org/ | Pathway/Genome database collection [2] | Genome and metabolic pathways |
| **InWeb_IM** | https://www.intomics.com/inbio/map/#home | A human protein-protein interaction network to catalyze genomic interpretation | PPI |
| **IntAct** | https://www.ebi.ac.uk/intact/ | Molecular interaction database | PPI |
| **BioGRID** | https://thebiogrid.org/ | Biological General Repository for Interaction Datasets | PPI, functional gene interactions |

| | | | |
|---|---|---|---|
| **Pathguide** | http://www.pathguide.org/ | Collection of biological pathway resources | PPI, pathways |
| **UniHI** | http://www.unihi.org/ | Collection of PPI and regulatory transcriptional interactions | PPI |
| **ConsensusPathDB** | http://cpdb.molgen.mpg.de/ | Protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions | Collection of interaction networks |
| **TTD** | https://db.idrblab.org/ttd/ | Therapeutic Target Database | Therapeutic target information |
| **DisGeNET** | http://www.disgenet.org/web/DisGeNET/menu/home | A collection of gene and variants associated to human disease | Gene-disease associations |
| **Open Targets** | https://www.opentargets.org/ | Platform for Target identification and prioritisation | Target-disease associations |
| **PHAROS** | https://pharos.nih.gov/idg/index | Knowledge base for the Druggable Genome (DG) | Target-disease associations |
| **SuperTarget** | http://insilico.charite.de/supertarget/index.php | Relations between drugs,proteins and side effects | Drug-target associations |
| **Drug2gene** | http://www.drug2gene.com/ | A resource to explore the drug-target relation network | Drug-target associations |
| **cmap** | https://portals.broadinstitute.org/cmap/ | Collection of gene expression profiles following drug perturbation | Drug-gene expression associations |
| **PharmGKB** | https://www.pharmgkb.org/ | Knowledge base about clinically actionable gene-drug and genotype-phenotype relationships | Drug-gene associations |

| | | | |
|---|---|---|---|
| **DGIdb** | http://www.dgidb.org/search_interactions | The Drug Gene Interaction Database | Drug-gene associations |
| **DrugBank** | https://www.drugbank.ca/ | A bioinformatics and chemoinformatics resource that combines drug and drug-target information | Drugs and targets information |
| **CTD** | http://ctdbase.org/ | A database to advance understanding about environmentally influenced diseases | Chemical-gene-disease interactions |
| **STITCH** | http://stitch.embl.de/ | Known and predicted interactions between chemicals and proteins | Chemical-protein interactions |
| **SIDER** | http://sideeffects.embl.de/ | Information on marketed medicine and their recorded adverse drug reactions | Drug-side effect interactions |
| **IntSide** | https://intside.irbbarcelona.org/ | A web server to elucidate the molecular processes involved in drug side effects | Drug-side effect interactions |

[1] All knowledge bases listed are open-access or part open-access. This collection, however, by no means pictures the whole range of knowledge bases regarding pathways and drug-target-disease Interactions that are currently available. The above knowledge bases were selected based on their relevance to the content of the review. For a full list please refer to Nucleic Acids Research Database Summary (http://www.oxfordjournals.org/nar/database/a/).

[2] BioCyc has now moved to a subscription based access plan.

# 2.5 Computational tools for target identification

The target identification is the first step of the drug discovery pipeline and aims to identify the magic molecular target that ideally cures or stops the progression of a disease. For a more informed implementation of the target identification process, a detailed disease mechanism becomes essential and network- and pathway-based approaches can be of great use in that regard [18]. Computational tools for the analysis of disease-specific 'omics data, at the pathway and network level rather than the molecular profile level, can identify ill-functioning cellular routes and altered biological functions and, thus, help draw potential targets for therapeutic intervention that will reverse those deregulated processes (Fig. 2.2) [19].

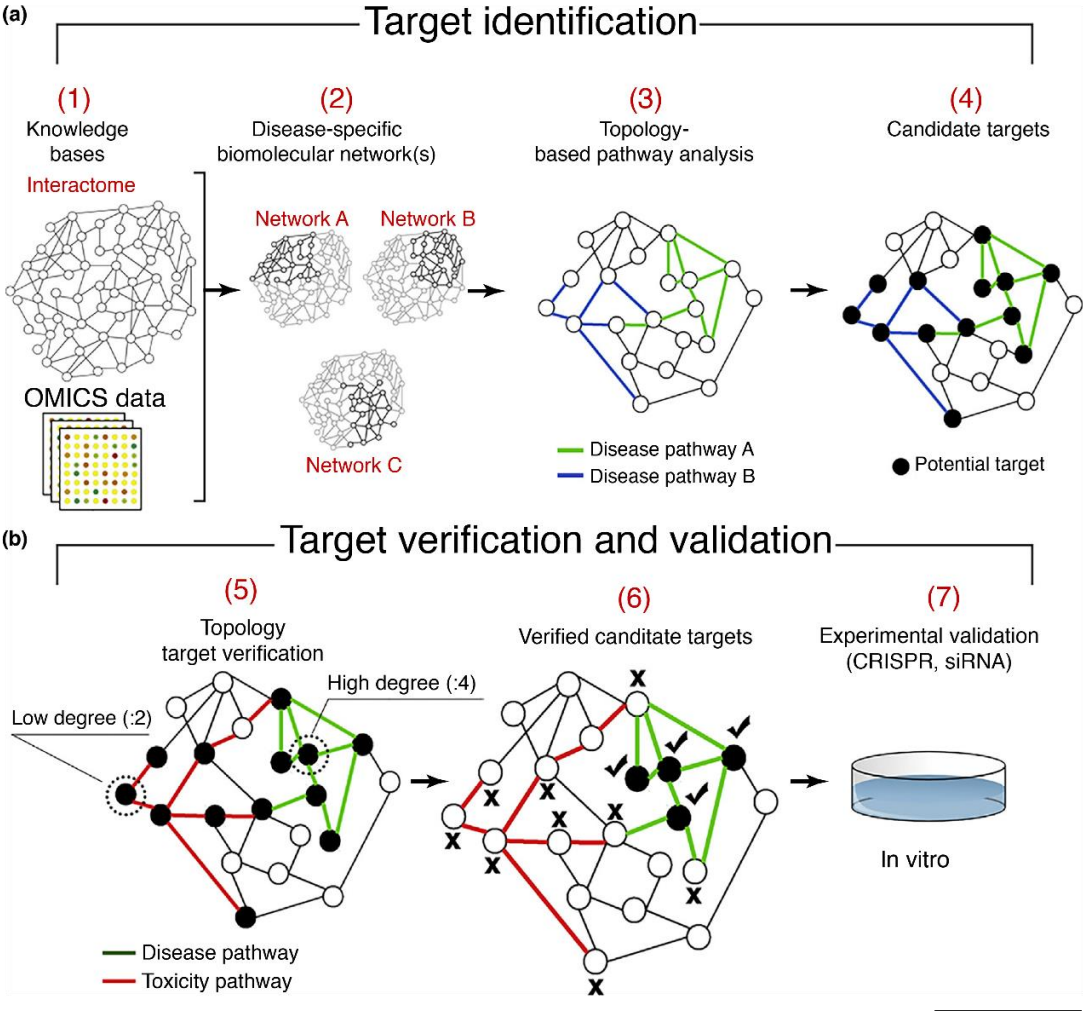



Figure 2.2. (a) The topology-based target identification process. Topology-based pathway analysis deciphers the disease mechanism(s) and identifies molecular pathways characteristic of the disease. Then, candidate targets are selected belonging to those pathways that, when perturbed, have the ability to reverse the disease state. (b) Verification and validation of proposed targets based on topological analysis. Targeting nodes with high degrees (representing hubs), although efficacious, can be linked to adverse effects. By contrast, targets with very few connections are not preferred because their perturbation usually will not reverse the disease state (low efficacy). Finally, candidate targets proximal to known toxicity pathways should not be selected. Based on those principles, machine-learning computational tools can prioritize candidate targets before their experimental validation.

There are several publicly available pathway and network analysis tools that can be applied for target identification (Table 2.3). NetworkAnalyst [20] is a web-based tool for the visualization, meta-analysis, and interpretation of gene expression data, and can be used to elucidate the disease mechanism. In NetworkAnalyst, genes of interest are identified from the user's submitted data, through differential expression analysis. Then, the identified genes are mapped to a PPI database to construct the whole network. Finally, hub or module analysis followed by topology-based pathway analysis can identify pathways characteristic of the disease. A similar web-based tool that allows users to integrate data from two commonly performed 'omics experiments (i.e., gene expression and metabolomics) is MetaboAnalyst. By combining the evidence based on changes in both gene expression and metabolite

concentrations, one is more likely to pinpoint the pathways involved in the underlying biological process [21]. MetaboAnalyst maps the user-submitted data to KEGG metabolic pathways for over-representation analysis and pathway topology analysis. The Database for Annotation, Visualization for Integrated Discovery (DAVID) is a tool that uses the KEGG pathway database as a knowledge base for functional interpretation of large set of genes derived from different genomic studies [22]. Another widely used open-source package to interpret genomic data and study the disease mechanism is PANTHER's analysis software [23]. Using PANTHER's gene list analysis software, users can analyze gene list expression data files and map them to multiple annotation data sources from the Gene Ontology Consortium, as well as biological pathways [24].

Table 2.3. Classification of computational tools to help decipher the disease mechanism and the drug MoA.

| Name | Method [1] | Reasoning | Data input | Knowledge base | Use topology | Approach | Key points | Main application | Availability | Ref. |
|---|---|---|---|---|---|---|---|---|---|---|
| **NetworkAnalyst** | ORA [2], Network analysis | Forward | Gene/protein list Gene expression | InnateDB(PPI) KEGG Reactome | Yes | Mechanistic | Interactive visualization, easy to use | Disease mechanism | Web-based | [20] |
| **MetaboAnalyst** | MSEA[3] Topology-based pathway analysis | Forward | Metabolite concentrations /lists ,gene list | HMDB, KEGG | Yes | Mechanistic | Integrative pathway analysis , biomarker analysis | Disease mechanism | Web-based | [21] |
| **PANTHER** | ORA, GSEA | Forward | gene list | PANTHER, Reactome | No | Mechanistic | Phylogenetic trees of protein coding genes | Disease mechanism | Web-based | [23] |
| **Whistle** | Causal reasoning | Backward | Gene expression | Causal network in BEL[4] | Yes | Mechanistic | Qualitative mechanistic hypotheses | Disease mechanism Drug MoA | Local Installation | [28] |
| **CRE** | Causal reasoning | Backward | Gene expression | Causal network in BEL[4] | Yes | Mechanistic | Qualitative molecular hypotheses, improved robustness | Disease mechanism | Local Installation | [27] |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **TopoNPA** | Causal reasoning | Backward | Gene expression | Causal network in BEL[4] | Yes | Mechanistic | towards noise Quantitative perturbation assessment, diagnostic signature extraction | Systems toxicology, disease mechanism, drug MoA | Local Installation* | [43] |
| **FS-MLKNN** | Multi-label K nearest neighbors | Predictive | Benchmark data sets[5] | KEGG, SIDER, DrugBank | No | Data driven | Multi-label learning | Side effect prediction | Local Installation | [46] |
| **DrugClust** | Clustering, GSEA | Predictive | Benchmark data sets[5] | KEGG, SIDER, Matador, DrugBank | No | Data driven[6] | Novel clustering algorithm, adoption of Bayesian score, easily modified, mechanistic insight | Side effect prediction | Local Installation | [49] |

[1]Although a computational tool may utilize more methods than the ones listed we chose to include those highly relevant to the contents of the article (pathway and network based).

[2]Extended Over Representation Analysis using topology.

[3]Metabolite Set Enrichment Analysis is similar to Gene Set Enrichment Analysis (2nd generation pathway analysis method).

[4]A cause-effect network model is used as a knowledge base in the form of Biological Expression Language (BEL) statements (see: causalbionet.com for a database of cause-effect biological networks).

[5]For training, validation and testing of the side effect prediction algorithms the data sets of Mizutani et al. [47], Liu et al. [48], Zhang et al. [46].

[6]Can also provide mechanistic insight utilizing GSEA on genes most prevalent in a cluster of drugs.

The merits of the pathway- and network-based approach are also prominent when analyzing ‘omics data belonging to different domains. As an example, in a recent study by Perco et al. [25], several of the aforementioned tools were combined for the integrated analysis of three transcriptomics and a proteomics data set for chronic kidney disease (CKD) (Fig. 2.3). Whereas separate gene and protein profile analysis identified only a limited number of features altered in both data sets, cojoint pathway and network analysis using KEGG PATHWAY, DAVID, and

omicsNET served in the functional interpretation of the data and identified a significant overlap of enriched biological functions (pathways) between both data sets, thus providing mechanistic insights into CKD.

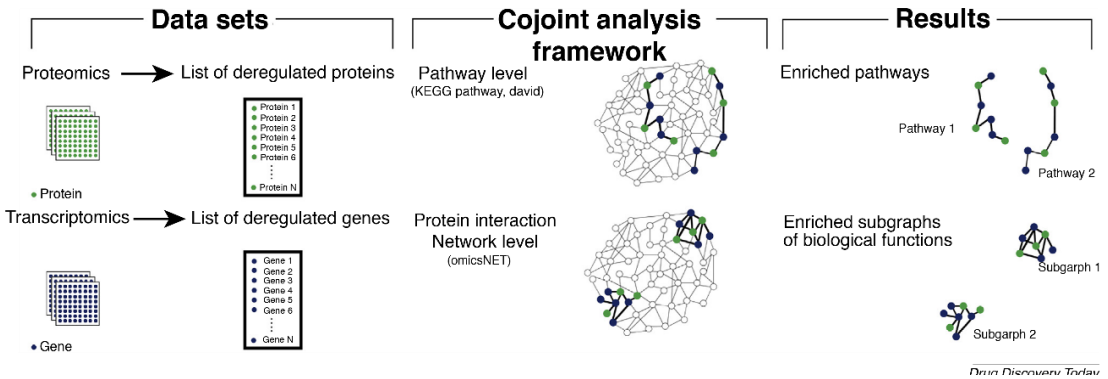



Figure 2.3. Schematic overview of the computational framework developed by Perco et al. [25] for the integrated analysis of four chronic kidney disease (CKD)-relevant data sets. Cojoint analysis at the pathway and network level revealed cell structure, cell adhesion, as well as immunity and defense mechanisms as jointly populated with deregulated features from both the proteomics and transcriptomics data sets

Another possible avenue for understanding the mechanism(s) of disease, based on molecular profiling data and interactome networks, is the Causal Reasoning (CR) method. Although CR is distinct from pathway analysis, in the sense that it does not rely on the assumption that differential RNA expression equates to differential protein activity (forward reasoning), the two methods can be applied in complementary fashion. CR is used to infer hypotheses mostly from gene expression data by detecting upstream regulators that could have led to the observed changes in gene expression between two states (backward reasoning). CR methods require as input a knowledge base of cause–effect relationships along with the gene expression data and can output a ranked list of causal drivers, called hypotheses [26]. Some notable examples of open-or part-open access computational tools utilizing CR are the Whistle algorithm (https://github.com/Selventa/whistle) by Selventa and the Causal Reasoning Engine (CRE) by Chindelevitch et al. (R source code available upon request to the original authors) [27,28]. The main characteristic of Whistle is that it infers mechanistic hypotheses in a qualitative manner as activated or inhibited and produces statistical metrics to evaluate their significance. By contrast, CRE infers molecular hypotheses from the data and offers a unique way to calculate the significance of the identified molecular drivers, with improved robustness towards noise. In addition, several commercial providers, such IPA and Thomson Reuters, also offer CR-based analysis tools [29].

In a different category, more recently, organizations have established open web-based platforms specifically tailored towards target identification, following a different pipeline than the above pathway topology-based and network analysis tools. This pipeline is more investigative than analytical, focusing on the integration and visualization of evidence gathered from available knowledge bases to score the association between biological targets and diseases. Based on the score provided, novel biological targets can be identified and prioritized for follow-up. For example, the Open Targets partnership established the Target

Validation Platform, which allows investigation and visualization of the evidence that associates targets and diseases [30]. The evidence that is integrated into the platform comes from public knowledge bases and includes rare and common disease genetics, transcriptomics, approved drugs and clinical candidates, animal models, Reactome biochemical pathways, and text mining from the medical literature. Similar to Open Targets, the NIH has launched the PHAROS platform to identify potential new drug targets within the four most-commonly drug-targeted protein families (G-protein- coupled receptors, nuclear receptors, ion channels, and protein kinases). PHAROS follows a similar approach to Open Targets by integrating multiple sources of biomedical data, albeit concentrating on these four protein families [31].

## 2.6 Target verification and validation

Following the identification of disease-specific mechanisms, a set of targets representative of those mechanisms can be selected. However, not all of those targets are equally potent and safe points of therapeutic intervention and their prioritization is the next logical step of the drug discovery process [1]. On that front, considering the position of the target node(s) inside the network along with the global properties of biological networks, arising from graph theory, can provide an early estimation of their safety and implication in potential adverse effects [32] (Fig. 2.2). Biological networks are usually scale free, which means that, although many nodes have a small degree, there are nodes, called hubs that have a great number of connections and have a key role in the information flow of the system [33]. Given that hubs drive the network traffic, targeting them has a significant impact on the cell behavior, including not only the increased chance of lethality, but also adverse effects [34]. That is why we need to be aware of the degree and betweenness of the target (i.e., the number of shortest paths traversing a node) and select between influential high-degree nodes (with potential adverse effects) versus nodes with middle to low degrees [35]. For example, NetworkAnalyst [20] estimates target importance using the network metrics of degree centrality and betweenness centrality. Furthermore, inside the network, pathway modules are connected to each other via bridging nodes indicating that, in the same way that the expression of each gene (node) is not independent from each other, the activity of each pathway is not independent from each other (a phenomenon known as cross-talk). Thus, even though a candidate target could belong to a non-vital and/or secondary pathway, because of cross-talk, alterations could be caused in another interlinked pathway, which could, in turn, cause adverse effects. In a recent study, Donato et al. developed a method to detect and quantify the crosstalk effect and identify novel functional sub-pathways involved in the condition under study [36].

Several studies have used network models to identify the optimal set of targets that affect specific disease mechanisms while minimizing adverse effects [37–39]. Ruths et al. used graph theory algorithms on signaling networks to identify sets of nodes that, when targeted, inhibited the function of a subnetwork while preserving signal flow to a different subnetwork [37]. Similarly, Dasika et al. proposed an optimization framework for signaling networks to

identify targets that block a specific mechanism while minimizing adverse effects [38]. Lu et al. developed an integer linear programming (ILP)-based method for multiple metabolic networks to identify the minimum set of reactions whose removal would block the production of a target in one network but not in the others [39].

Finally, the resulting set of targets that has been identified to inhibit the disease-specific mechanism(s), following the in-silico network-based target prioritization, needs to be verified and validated in the lab. Experimentally, this can be done with small interfering RNA (siRNA) or CRISPR-cas9 gene editing, specifically designed to silence and/or catalyze the specified target(s).

## 2.7 Hit and lead selection

Once the target(s) have been set, specific chemical or biological molecules need to be found that either catalyze or inhibit such target(s) [1]. This hit discovery phase is traditionally addressed by high- throughput screening (binding and phenotypic assays) where compounds (known as hits) with the desired effect over the targets are selected. For small molecules, many cheminformatics tools have been developed [18], with the aim to identify the best chemical structure. Experimentally, the Innovative Medicines Initiative (IMI) has established the European Lead Factory (ELF), offering a large collection of compounds and a state-of-the-art screening center to connect innovative drug targets to high-quality compounds (www. imi.europa.eu/content/european-lead-factory). Despite their importance, in-depth analysis of experimental and chemical optimization tools is out of the scope of this review. Instead, here we focus on the step after hit identification, where the most promising candidates should be selected from the identified hit compounds. In this step, biological networks and pathways can again serve as a link between the chemical space and the biological space to elucidate how hit compounds affect pathways and how those pathways can be associated with adverse effects and efficacy. On this front, multiomic experimental data of the hit compounds are essential for use with several pathway-based methods and to infer potential adverse effects and efficacy predictions [40–42].

Adverse effects of drugs are common and are either caused by off-target effects, (i.e., unforeseen direct physical drug–protein interactions because of drug and/or protein promiscuity), or indirect effects because of signal propagation after the direct interaction [4]. On this front, a study by Mitsos et al. identified via ILP how drugs or compounds alter the signaling pathways [40]. During the hit to lead stage, pathway effects (not just targets) can be used to predict efficacy and toxicity and enhance the drug discovery pipeline by providing better compounds to the optimization stage. Current efforts are focusing in relating drug effects to perturbed biological pathways. Biological annotation of adverse effects with associated pathways is a key step that allows one to predict the adverse effects of a compound. To provide biological insight into the generation process of adverse effects, several methods have been proposed utilizing molecular interaction data gathered from various knowledge bases, such as DrugBank (https://www. drugbank.ca/) and KEGG Drug

(www.genome.jp/kegg/drug/) for drug-affected pathway(s) interactions, the Therapeutic Target Database (TTD; http://bidd.nus.edu.sg/group/cjttd/) for target– pathway(s)–disease interactions, SIDER (http://sideeffects.embl. de/) and IntSide (source: https://intside.irbbarcelona.org/) for drug–adverse effect interactions, and Cmap for drug-transcriptional profile associations. Lee et al. used gene set enrichment analysis (GSEA), a second-generation pathway analysis tool, to reveal enriched pathways from the Cmap drug-induced transcriptional profiles and utilized Gene Ontology ontologies to connect them to biological processes [41]. The authors then built a tripartite net- work of biological processes–drugs–adverse effects by using SIDER to discover connections between biological processes and adverse effects. In another study by Bauer-Mehren et al., the authors gathered drug–target associations from DrugBank and protein– adverse effect associations from DisGeNET and, by intersecting them, were able to identify pathways from Reactome-containing proteins present in both sets as the links explaining the adverse effect generation mechanisms [42] (Fig. 2.4).

Another possible avenue for mechanistic toxicology, besides pathway analysis, is the usage of cause–effect network models to identify and quantify characteristic network signatures following perturbation by a drug. The quantification of the perturbation of a network is important in toxicology and pharmacology, where dose and time response are studied. In a recent study by Martin et al., the authors established a computational method, TopoNPA, for the analysis of gene expression data using cause-and-effect net- works as prior knowledge to identify, interpret, and quantify the perturbation of the network [43]. TopoNPA was successfully applied to infer a mechanistic hypothesis for the unequal efficacy of an anti-inflammatory drug and to generate a robust network signature for predicting individual patient responses.

Based on key insights generated into adverse effects, several machine-learning (ML) tools have been developed to predict the adverse effects of a hit compound based on the affected pathway(s) in conjunction with its chemical information (i.e., chemical structure) gathered from databases such as PubChem (https:// pubchem.ncbi.nlm.nih.gov/) (Fig. 2.4) [44,45]. As an example, Zhang et al. proposed a method named ‘feature selection-based multi-label k-nearest neighbor method’ (FS-MLKNN) for adverse effect predictions [46]. In this study, a feature selection approach identified predictors based on pathways, targets, enzyme, transporters, and chemical structure that were later combined in a weighted scoring system to predict adverse effects. In terms of area under the precision-recall curve (AUPR) and other ML performance metrics, FS-MLKNN performed better on benchmark data sets than did other similar ML methods [44, 47, 48]. Zhang et al. also tested FS-MLKNN on a new data set from SIDER and reported that the average of recall scores for the test-set drugs was 0.463, which means that, on average, 46.3% of the adverse effects of a candidate drug could be predicted in-silico. Recently, DrugClust [49], a new ML tool for adverse effect prediction with a similar hybrid approach, albeit only using drug–target and chemical substructure as features for the prediction, was developed and is freely available as an R package (https://cran.r-project.org/web/packages/DrugClust/ index.html). Comparing DrugClust and FS-MLKNN, Dimitri et al. reported that DrugClust produced a slightly higher AUPR for the Zhang data set,

when all respective features were considered, and a slightly lower AUPR for the Liu data set. Additionally, DrugClust can provide mechanistic insight into the drug adverse effect(s) generation process by performing pathway analysis on proteins that more frequently appear in a certain cluster of drugs.

Computational tools and algorithms based on pathway networks can similarly be used to evaluate the efficacy of hit compounds and prioritize them before experimental validation [50]. In a recent study by Gu et al., the authors proved that the degrees of the decrease of network efficiency and network flux, which are both measures of the connectivity of a pathway network, could evaluate the efficacy of a compound. The authors followed this approach to predict the drug–response curves of drugs on the pathway network of LPS-induced PGE2 production and their prediction agreed with the experimental results [50]. In another study by Guney et al., a disease–gene network was built and a drug– disease proximity measure was introduced using various distance metrics between the target of 238 drugs and 78 disease modules. The study concluded that proximity is a good measure to assess drug efficacy and that drug to pathway proximity, calculated from the distance of drug targets to proteins belonging in a pathway, can elucidate the drug MoA. [51].

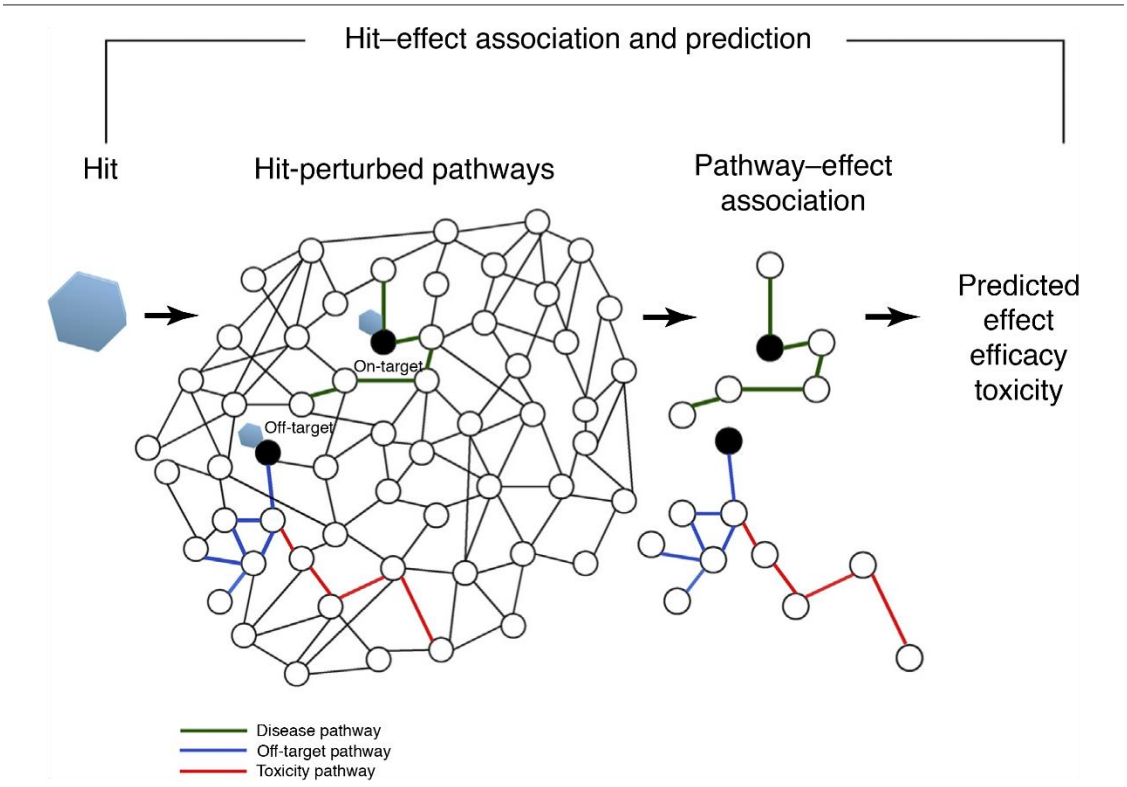



Figure 2.4. The hit–effect association and prediction process. Adverse effects of hit compounds are either caused by off-target effects or indirect effects because of signal propagation after the direct interaction. Thus, biological annotation of hit effects with associated pathways is a key step to predict the adverse effects and efficacy of a compound based on the affected pathway(s)

# 2.8 Conclusion

Drug discovery is a complex and time-consuming process that involves several steps, ranging from target discovery to clinical trials. In this review, we have shown that pathway- and network- based technologies help foster the conception of mechanism-based drug discovery and enhance the early stages of the drug discovery pipeline by shedding light on the underlying biology of drugs and diseases. On this front, topology-based pathway analysis tools can help decrease the number of false positives in the target identification step, prioritize target validation, select optimal hits, and help the hit to lead step.

Although the incorporation of pathway- and network-based drug discovery can improve the hit success rate, there are still limitations and challenges that need to be addressed. On the one hand, network topology databases have many conflicting reports and, in several instances, their quality is also debatable. On the other hand, computational tools that extract information from knowledge bases might have a bias towards pathways or molecules that are better studied and more present in the data and knowledge bases. On the data front, rapid changes on experimental technologies, lack of data formats, and lack of standardizations of experimental designs, influence the overall data quality. Finally, the simplistic static approach that most computational methods adopt, overlooks the dynamic behavior of biological systems, which limits the capability to model in detail the disease or drug state. We need to keep an eye on both technological and methodological advancements because they can help bridge the gap between in-silico verification and experimental validation. We anticipate that, as more disease- and drug-specific ‘omics  data are generated and shared, and as biological information is better annotated, knowledge bases will keep expanding their coverage and pathway- and network-based computational methods will capitalize on those advancements, paving a new path towards an evolved drug discovery pipeline with lower attrition rates.

# 2.9 References

1. Hughes, J.P. et al. (2011) Principles of early drug discovery. Br. J. Pharmacol. 162, 1239–1249

2. Apic, G. et al. (2005) Illuminating drug discovery with biological pathways. FEBS Lett. 579, 1872–1877

3. Waring, M.J. et al. (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat. Rev. Drug Discov. 14, 475–487

4. Iorio, F. et al. (2013) Network based elucidation of drug response: from modulators to targets. BMC Syst. Biol. 7, 139

5. Westerhoff, H.V. and Palsson, B.O. (2004) The evolution of molecular biology into systems biology. Nat. Biotechnol. 22, 1249

6. Khatri, P. et al. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 8, e1002375

7. Robinson, S.W. et al. (2014) Current advances in systems and integrative biology. Comput. Struct. Biotechnol. J. 11, 35–46

8. Antoranz, A. et al. (2017) Mechanism-based biomarker discovery. Drug Discov. Today 22, 1209–1215

9. Rebholz-Schuhmann, D. et al. (2012) Text-mining solutions for biomedical research: enabling integrative biology. Nat. Rev. Genet. 13, 829–839

10. Galperin, M.Y. et al. (2017) The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res. 45, D1–D11

11. Cisek, K. et al. (2016) The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. Nephrol. Dial. Transpl. 31, 2003–2011

12. Wray, N.R. et al. (2007) Prediction of individual genetic risk to disease from genome- wide association studies. Genome Res. 17, 1520–1528

13. Larance, M. and Lamond, A.I. (2015) Multidimensional proteomics for cell biology. Nat. Rev. Mol. Cell Biol. 16, 269

14. Nabieva, E. et al. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 (Suppl. 1), i302–i310

15. Zhang, A.D. et al. (2013) Reconstruction and analysis of human kidney-specific metabolic network based on omics data. BioMed Res. Int. 2013, 187509

16. Geppert, T. and Koeppen, H. (2014) Biological networks and drug discovery—where do we stand? Drug Dev. Res. 75, 271–282

17. Wu, G. et al. (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 11, R53

18. Katsila, T. et al. (2016) Computational approaches in target identification and drug discovery. Comput. Struct. Biotechnol. J. 14, 177–184

19. Garc´ıa-Campos, M.A. et al. (2015) Pathway analysis: state of the art. Front. Physiol. 6, 383

20. Xia, J. et al. (2015) NetworkAnalyst for statistical, visual and network-based meta- analysis of gene expression data. Nat. Protoc. 10, 823

21. Xia, J. et al. (2015) MetaboAnalyst 3.0 — making metabolomics more meaningful. Nucleic Acids Res. 43, W251–W257

22 Jiao, X. et al. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics 28, 1805–1806

23. Mi, H. et al. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41, D377–D386

24. Ashburner et al. (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29

25. Perco, P. et al. (2010) Linking transcriptomic and proteomic data on the level of protein interaction networks. Electrophoresis 31, 1780–1789

26. Barrett, S.J. et al. (2013) Generation of causal hypotheses via reasoning on interaction networks: methodological assessment and future possibilities. In Proceedings of the 14th International Conference on Systems Biology (ICSB), Aug 29–Sep 4; Copenhagen, Denmark

27. Chindelevitch, L. et al. (2012) Causal reasoning on biological networks: interpreting

28. Catlett, N.L. et al. (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics 14, 340

29. Kramer, A. et al. (2013) Causal analysis approaches in ingenuity pathway analysis. Bioinformatics 30, 523–530

30. Koscienly, G. et al. (2017) Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 45, D985–D994

31. Nguyen, D.-T. et al. (2017) Pharos: collating protein information to shed light on the druggable genome. Nucleic Acids Res. 45, D995–D1002

32. Ma'yayan, A. et al. (2007) Network analysis of FDA approved drugs and their targets. Mount Sinai J. Med. 74, 27–32

33. Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev . Genet. 5, 101

34. Jeong, H. et al. (2001) Lethality and centrality in protein networks. Nature 411, 41–42

35. Hase, T. et al. (2009) Structure of protein interaction networks and their implications on drug design. PLoS Comput. Biol. 5, e1000550

36. Donato, M. et al. (2013) Analysis and correction of crosstalk effects in pathway analysis. Genome Res. 23, 1885–1893

37. Ruths, D.A. et al. (2006) Hypothesis generation in signaling networks. J. Comput. Biol. 13, 1546–1557

38. Dasika, M.S. et al. (2006) A computational framework for the topological analysis and targeted disruption of signal transduction networks. Biophys. J. 91, 382–398

39 Lu, W. et al. (2015) Computing smallest intervention strategies for multiple metabolic networks in a boolean model. J. Comput. Biol. 22, 85–110

40. Mitsos, A. et al. (2009) Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. PLoS Comput. Biol. 5, e1000591

41. Lee, S. et al. (2011) Building the process–drug–side effect network to discover the relationship between biological processes and side effects. BMC Bioinformatics 12, S2

42. Bauer-Mehren, A. et al. (2012) Automatic filtering and substantiation of drug safety signals. PLoS Comput. Biol. 8, e1002457

43. Martin, F. et al. (2014) Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. BMC Bioinformatics 15, 238

44. Pauwels, E. et al. (2011) Predicting drug side-effect profiles: a chemical fragment- based approach. BMC Bioinformatics 12, 169

45. Huang, L.C. et al. (2011) Predicting adverse side effects of drugs. BMC Genomics 12, S11

46. Zhang, W. et al. (2015) Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinformatics 16, 365

47. Mizutani, S. et al. (2012) Relating drug–protein interaction network with drug side effects. Bioinformatics 28, i522–i528

48. Liu, M. et al. (2012) Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J. Am. Med. Inf. Assoc. 19, e28–e35

49. Dimitri, G.M. and Lio , P. (2017) DrugClust: a machine learning approach for drugs side effects prediction. Comput. Biol. Chem. 68, 204–210

50. Gu, J. et al. (2015) Quantitative modeling of dose–response and drug combination based on pathway network. J. Cheminformatics 7, 19

51. Guney, E. et al. (2016) Network-based in silico drug efficacy screening. Nat. Commun. 7, 10331

# Chapter 3

# DeepSNEM: Deep Signaling Network Embeddings for compound mechanism of action identification

## 3.1 Chapter abstract


### Motivation

The analysis and comparison of compounds' transcriptomic signatures can help elucidate a compound's Mechanism of Action (MoA) in a biological system. In order to take into account the complexity of the biological system, several computational methods have been developed that utilize prior knowledge of molecular interactions to create a signaling network representation that best explains the compound's effect. However, due to their complex structure, large scale datasets of compound-induced signaling networks and methods specifically tailored to their analysis and comparison are very limited. Our goal is to develop graph deep learning models that are optimized to transform compound-induced signaling networks into high-dimensional representations and investigate their relationship with their respective MoAs.

### Results

We created a new dataset of compound-induced signaling networks by applying the CARNIVAL network creation pipeline on the gene expression profiles of the CMap dataset. Furthermore, we developed a novel unsupervised graph deep learning pipeline, called deepSNEM, to encode the information in the compound-induced signaling networks in fixed-length high-dimensional representations. The core of deepSNEM is a graph transformer network, trained to maximize the mutual information between whole-graph and sub-graph representations that belong to similar perturbations. By clustering the deepSNEM embeddings, using the k-means algorithm, we were able to identify distinct clusters that are significantly enriched for mTOR, topoisomerase, HDAC and protein synthesis inhibitors

respectively. Additionally, we developed a subgraph importance pipeline and identified important nodes and subgraphs that were found to be directly related to the most prevalent MoA of the assigned cluster. As a use case, deepSNEM was applied on compounds' gene expression profiles from various experimental platforms (MicroArrays and RNA sequencing) and the results indicate that correct hypotheses can be generated regarding their MoA.

### Availability and Implementation

The source code and pre-trained deepSNEM models are available at https://github.com/BioSysLab/deepSNEM.

## 3.2 Introduction

Characterizing a compound's Mechanism of Action (MoA) in a cellular system is a very important step in the development of new drugs or the repurposing of existing ones. On this front, several systems-based computational methods that utilize omics data, following treatment with a compound, have been developed [1]. One approach that has gained considerable attraction for the MoA identification task is the analysis of post-transcriptional data from compound perturbations [2]. These approaches analyze compounds' transcriptomic signatures in order to identify key genes and signaling mechanisms that either cause the compound's therapeutic effect or are associated with specific adverse effects [3]. Furthermore, the comparison of transcriptomic signatures can be used to elucidate the MoA of new compounds, by associating them with compounds of known MoA, or propose new indications for already existing drugs.

There have been many studies that utilize differential gene expression (GEx) data to characterize a compound's MoA [4]. The Connectivity Map (CMap) and the LINCS project have played a pivotal role in this field, by providing large datasets of compounds' transcriptomic signatures and methods for their analysis, comparison and interpretation [5,6]. As an example, Iorio et al. utilized compounds' transcriptomic signatures from the CMap dataset to build a network, where perturbations are connected if they have similar transcriptional profiles [7]. This network was then analyzed to find communities and clusters that consisted of perturbations with similar MoA. Since a compound's phenotypic effect is usually caused by changes in the expression of interacting genes/proteins, combining transcriptomic data with a prior knowledge-base of molecular interactions, e.g. signaling pathways, can result in a more mechanistic explanation of a compound's MoA [1]. On this front, a promising modeling technique is the representation of a compound's effect as a network of signaling proteins (nodes), showing their activity and how these interact with each other to transfer the signal of the perturbation in the system [8].

Signaling network creation methods combine omics data with a prior knowledge network of protein-protein interactions (PPI) in order to extract a graph that best explains the experimental data. Mitsos et al. developed an Integer Linear Programing (ILP) optimization task to identify the signaling network that characterizes a compound's effect based on

phosphoproteomic data [9]. Since large scale phosphoproteomic datasets following compound treatment are very rare, there has been a concentrated effort to develop methods for signaling network creation based on transcriptomics [10-12]. Liu et al. developed CARNIVAL, a causal reasoning framework to identify signaling networks that best explain a set of transcription factor (TF) activity scores, calculated from differential GEx data [13]. Compound-induced signaling networks are information-rich and complex representations of the compounds' effect, since they incorporate the prior knowledge of molecular interactions in the form of a PPI network. However, this complexity poses limitations for their large scale analysis and comparison of networks from different compounds using traditional network similarity algorithms, i.e. graph kernels. More specifically, graph similarity algorithms, such as the Graph Edit Distance (GED), graphlet-based methods or graph kernels, utilize hand crafted features and are not optimized for signaling networks, which can result in reduced generalization performance and reduced scalability [14-16]. An interesting approach is to employ deep learning models for graphs in order to encode the complete information of the signaling network into high dimensional fixed-length representations [17]. These representations can then be compared using traditional algorithms in order to identify similarities between compound-induced signaling networks that could translate to similarities in the compounds' MoA.

There have been many studies for the development of deep learning models for graph data in a variety of fields. These models are usually neural networks that aim to learn new task-specific node and graph representations by using the graph's connectivity [18]. For example, the graph convolutional model utilizes a message passing algorithm to learn neighborhood-level representations of the input graph. Recently, the successful transformer architecture for natural language processing (NLP) problems has been modified and applied on graph data [19,20]. Graph transformers utilize an attention mechanism for each node that is a function of the neighborhood's connectivity, rather than a message passing algorithm. Similarly, the graph2vec model was inspired by the doc2vec approach for NLP tasks. Graph2vec treats the entire graph as a document and each node's neighborhood as a word and aims to learn a fixed-length representation of the entire graph in a fully unsupervised task [21]. Another important unsupervised approach for graph representation learning is the InfoGraph model [22,23]. InfoGraph aims to maximize the mutual information between graph-level representations and representations of the graph's substructures at different levels, e.g. nodes, edges and triangles. These unsupervised graph representation learning methods can be modified for compound-induced signaling networks in order to extract fixed-length feature vectors that can then be associated with the compound's MoA.

In this paper, we developed a novel deep learning framework, called deepSNEM, to learn new representations (embeddings) of signaling networks and investigate their relationship with the compound's MoA. Compounds' signaling networks were created using the CARNIVAL pipeline and the transcriptomic signatures of the CMap dataset, resulting in a large scale dataset of signaling networks that can aid future studies. The core of deepSNEM is an unsupervised graph transformer trained to maximize the mutual information between representations of graphs' substructures that belong to signaling networks created from

similar perturbations. The resulting embeddings were evaluated based on their ability to identify similar signaling networks and compared with representations created by different graph-based models. Subsequently, the embeddings were clustered with the k-means algorithm and the resulting clusters were analyzed based on their MoA composition. Furthermore, a subgraph importance method was developed to identify the most important nodes for each graph-level representation and the subgraphs that cause the signaling networks to cluster together. As a use case, deepSNEM was tasked to assign clusters to compounds' signaling networks generated using gene expression profiles from various experimental platforms. Analyzing the MoA composition of a compound's assigned cluster, deepSNEM can generate hypotheses regarding the MoA of new lead compounds or suggest new potential mechanisms for already existing drugs.

# 3.3 Results

## 3.3.1 The deepSNEM approach

The overview of our approach is presented in Figure 3.1. Differential gene expression signatures following compound treatment across cell lines were retrieved from the L1000 dataset (GSE92742) [6]. In total, 7722 signatures from 3005 compounds across 70 cell lines were utilized. The first step of the deepSNEM pipeline is the creation of signature specific signaling networks following the CARNIVAL framework [13]. In this framework, the gene expression signatures are first transformed into transcription factor activity scores and then an ILP model is tasked to extract the optimal subgraph from a global PPI network that best fits the calculated activity scores (see Methods 3.5.1). The created network is a labeled (protein activity), signed (edge activation or inhibition) and directed PPI graph that captures the signaling network effect of the drug-induced transcriptomic signature. The core of deepSNEM is a DL model, trained in an unsupervised setting, which takes as input the drug-induced signaling networks, created with CARNIVAL, and outputs a high dimensional embedding that best captures the information contained in the input graph. Regarding the DL models, we evaluated the use of a graph transformer trained to either maximize the mutual information of nodes belonging to the same signature (termed deepSNEM-GT-MI) or predict the edge presence between nodes (termed deepSNEM-GT-LP), a siamese GCN model to predict the graph edit distance between signaling networks (termed deepSNEM-GED) and the widely used graph2vec model (termed deepSNEM-G2V) (see Methods 3.5.2).

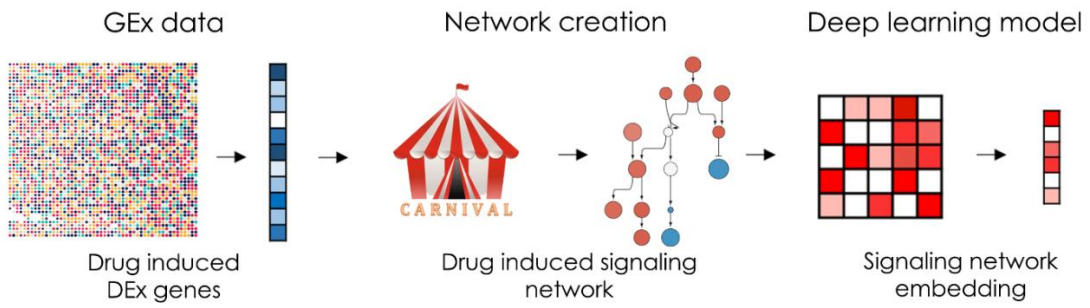



Figure 3.1. Schematic overview of deepSNEM. For each compound-induced differential expression signature, a signaling network is created using the CARNIVAL framework. Then an unsupervised DL model is tasked to encode the created signaling network in a high dimensional embedding that best captures the input graph information.

## 3.3.2 Model-embedding evaluation

The different deepSNEM model variations were evaluated based on the validity of the produced embeddings on two separate tasks. The first task examines the models' ability to produce similar embeddings from signaling networks that are created from the same differential gene expression signature. On this front, we utilized the slightly different but feasible network solutions of CARNIVAL's ILP model for the same signature and investigated the distributions of Euclidian distances between embeddings belonging to the same signature and between embeddings from different signatures (Figure 3.2A). As it can be seen in Figure 3.2A, there is a clear distinction between the distance distributions of embeddings from the same and different signatures. Thus, all models are able to produce embeddings that are significantly more similar for graphs created from the same measurements of differential expression. In the second task, we evaluated the similarity of graph embeddings created from duplicate gene signatures as compared to the similarity of embeddings from random gene signatures. Duplicate signatures indicate transcriptomic signatures from the same compound perturbation, cell line, dose and time point that were assayed on different L1000 plates [24]. Figure 3.2B shows the distributions of Euclidian distances between embeddings belonging to duplicate signatures and between embeddings of random signatures. For all models, the difference between the distributions is significant, as indicated by a two sample t-test (p-values < 0.001). Thus, all models are able to produce similar graph embeddings for gene signatures that share the same experimental conditions. Based on these results, we chose to perform a clustering analysis on the embeddings produced by the deepSNEM-GT-MI architecture, in order to examine the connection between a drug's induced signaling network and its reported MoA.

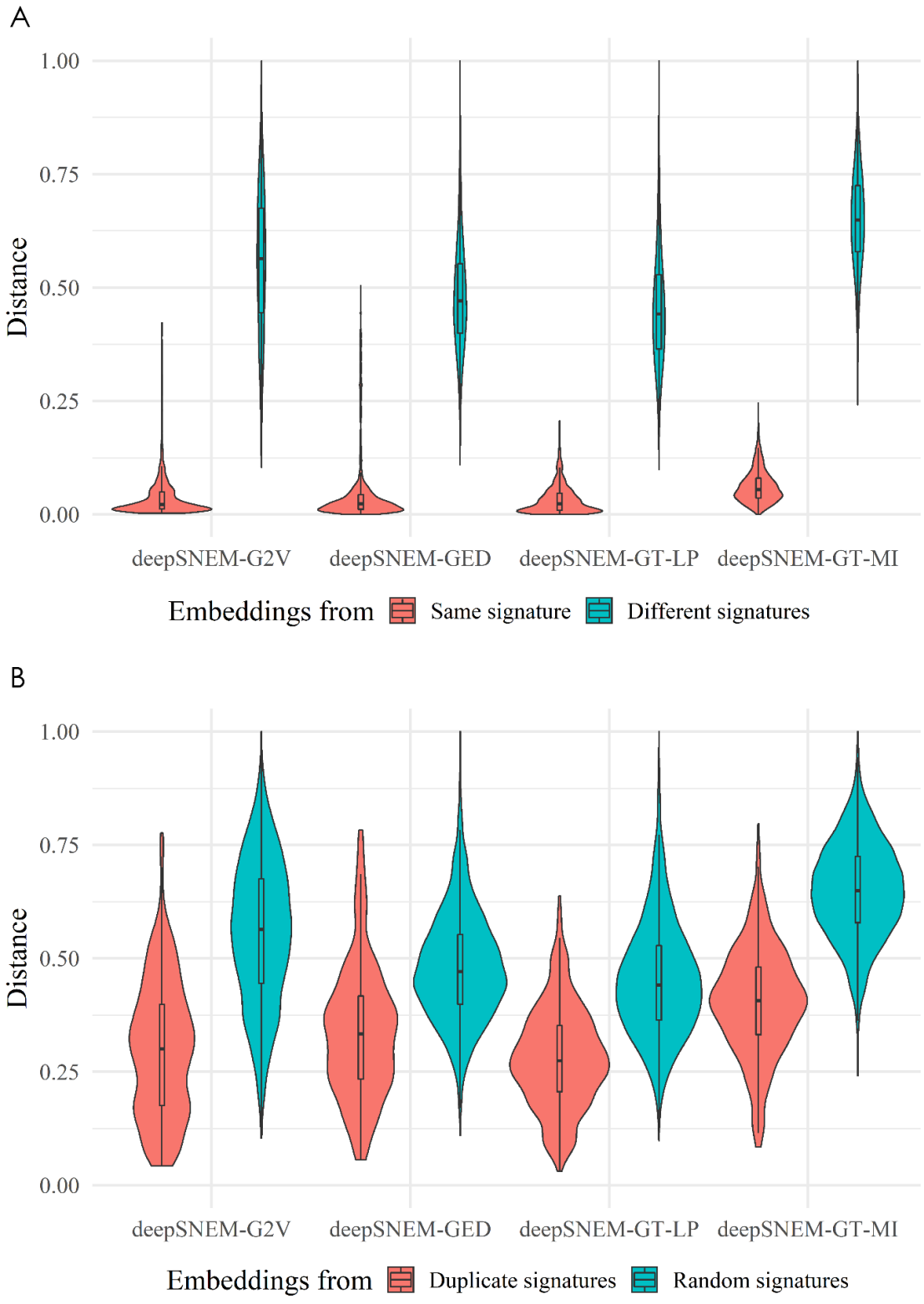



Figure 3.2.Model-embedding evaluation tasks. (A) Normalized Euclidian distances between embeddings from the same signature and different signatures for all deepSNEM model variations. (B) Normalized Euclidian distances between embeddings duplicate and random gene expression signatures for all model variations.

### 3.3.3 Clustering analysis for MoA identification

The signaling network effect of a compound perturbation in a cellular model presents a systematic view into the compound's MoA. In order to investigate this relationship, we first identified groups of perturbations with similar network effect, by clustering the deepSNEM network embeddings, and then analyzed the resulting clusters based on the reported MoA of the compounds. On this front, the 256-dimensional deepSNEM-GT-MI embeddings were clustered using the k-means algorithm. The optimal number of clusters was found to be 200, according to the k-means elbow plot (see Supplementary Material (SM) 6). Additionally, in order to analyze and characterize the resulting clusters, we utilized the MoA labels provided by the Broad's Institute Repurposing Hub [25]. Out of the 3005 unique compounds, 912 were mapped to 261 unique MoA labels using the Repurposing Hub dataset (see SM 1). Figure 3.3A shows the 2-dimensional t-SNE projections of all available signaling network embeddings. Additionally, the signaling network embeddings that belong to the top 9 most prevalent MoA labels in the dataset are presented with different colors (Figure 3.3A). In order to characterize the identified clusters, we focused on the subset of clusters that are significantly enriched for at least one mechanism (Figure 3.3B). The selected clusters have at least 25% of their compound perturbations belonging to the same MoA, with a p-value lower than $10^{-6}$ compared to a random selection. Figure 3.3B shows the breakdown of the available MoA in the selected clusters. As it can be seen, the identified clusters are enriched for the same mechanisms that are most prevalent in the labeled dataset. As a result, DeepSNEM was able to identify 11 clusters that are significantly enriched for specific mechanisms, i.e. mTOR, HDAC, topoisomerase, protein and ATP synthesis inhibitors. We have to note that clusters that are enriched for MTOR inhibitors are also enriched for PI3K inhibitors, which is expected due to the PI3K/mTOR signaling pathway. However, the majority of the compounds in each cluster still do not have available labels regarding their MoA (represented with grey color in Figure 3.3B). Thus, due to the unknown labels, the distribution of MoA between clusters that are enriched for the same MoA can still be quite different.

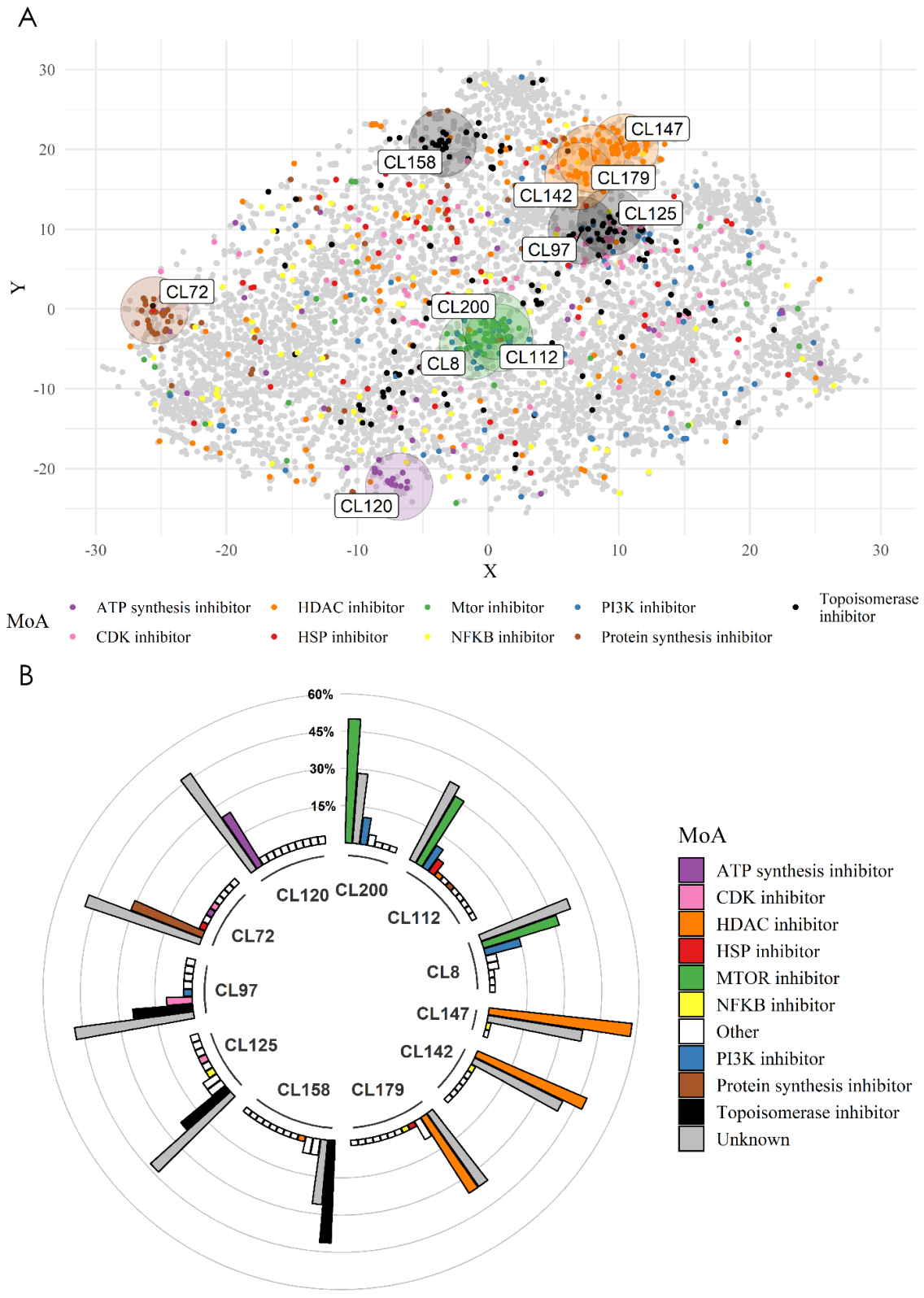



Figure 3.3. Clustering analysis. (A) T-SNE projection of the 256-dimensional signaling network embeddings of deepSNEM-GT-MI. Different colors represent the 9 most prevalent MoA in the dataset, while the grey color represents perturbations with either unknown or other MoA. Additionally, the centers of the identified clusters are represented with circles (CL: cluster). (B) MoA composition of the

analyzed clusters. The Y axis represents the frequency, as a percentage, of each MoA in the cluster (CL: cluster).

### 3.3.4 Subgraph importance

The analysis of compound-induced signaling networks for MoA identification offers the benefit of easier result interpretation. In order to utilize this benefit and increase the interpretability and explainability of deepSNEM, we created a framework to identify the important subgraphs for the subset of clusters analyzed in the previous section. For each cluster, important nodes were identified using an aggregate score based on their importance to the embedding model and the nodes' prevalence in the cluster's graphs (see Methods 3.5.3). Figure 3.4A shows the overlap, as a percentage, between the 20 most important nodes of the analyzed clusters. As it can be seen, clusters that are enriched for the same MoA, have a higher similarity between their most important nodes. Thus, the proposed importance framework can identify nodes of high importance in each cluster that show a connection to the cluster's most prevalent mechanism of action. For visualization purposes, the most important nodes in each cluster were connected by selecting the shortest paths between them, from the Omnipath PPI that also maximize the overall sum of importance scores in the path. Figure 3.4B shows an example of the important subgraphs for the clusters that are enriched for mTOR and PI3K inhibitors. The common most important nodes across the presented networks include the mTOR regulated transcription factors NRF1 and TFDP1 and the CSKNK2A1, RHOA, PRKACA and LCK proteins, which are involved in the PI3K-Akt-mTOR signaling pathway [26-30]. Finally, across all clusters, AKT1 and MAPK1 serve as central nodes that connect the most important nodes (Figure 3.4B). The important subgraphs for all analyzed clusters are presented in SM 7.

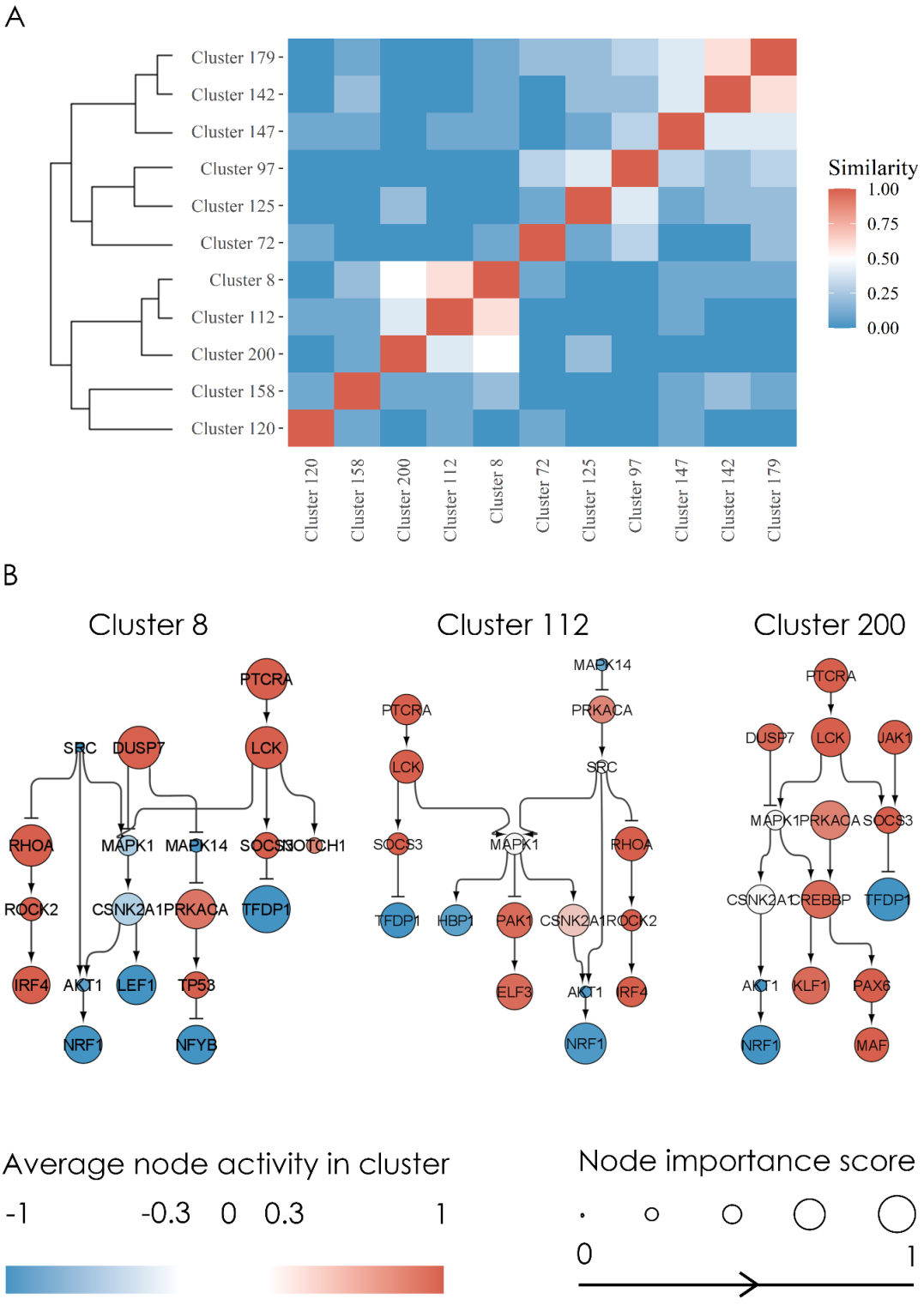



Figure 3.4. Cluster subgraph importance. (A) Heatmap showing the similarity, as percentage overlap, between the 20 most important nodes of each cluster. (B) Important subgraphs identified for the clusters enriched for mTOR and PI3K inhibitors (Clusters 8, 112 and 200). The average activity of each node in the cluster is color coded from blue to red. Blue nodes are inhibited, while red are activated. Each node's importance score, ranging from 0 to 1, is represented by the size of the node's circle.

### 3.3.5 Use case: cluster assignment

Gene expression data from 7 additional compounds with known mechanism of action were retrieved from the GEO database. The details regarding the experimental data used in the use case are presented in Table 3.1. Overall, the data were collected from 6 different studies, 4 cell lines and 3 different experimental platforms, i.e. Affymetrix/Agilent Microarrays and Illumina next generation sequencing. Following the deepSNEM pipeline, each differential gene expression signature was transformed into a compound induced signaling network with CARNIVAL and embedded using the deepSNEM-GT-MI model. Finally, each embedding was assigned to one of the already identified clusters (Table 3.1). Figure 3.5A shows the assigned clusters and the distribution of each cluster's available MoA. The topoisomerase inhibitor SN38 and the HDAC inhibitors Sodium-Butyrate, Panobinostat and Belinostat were assigned to clusters significantly enriched for topoisomerase and HDAC inhibitors respectively. Furthermore, the topoisomerase inhibitor Doxorubicin and the mTOR inhibitor Sirolimus were assigned to clusters enriched for their respective MoA, albeit having a large number of compounds with unknown MoA. Finally, the compound CDK-887 was assigned to a cluster that was not enriched for any particular MoA. Thus, the deepSNEM pipeline can be used to assign a cluster to a compound-induced gene expression signature, independent of the experimental platform, and provide insight into the compound's potential MoA. For the compounds in the use case, we also compared the cluster assignment of deepSNEM to a clustering of the compounds' differential expression gene measurements into the same number of clusters (k=200) (see SM 8) (Figure 3.5B). Comparing the two approaches, SN38, Belinostat and Panobinostat were assigned to clusters composed of similar mechanisms. However, this is not the case for Sirolimus, Doxorubicin and Sodium Butyrate, which are assigned to clusters not enriched for any particular MoA, when the gene-clustering pipeline is used. Finally, for each compound of the use case, we calculated the Jaccard similarity index between the perturbations of the identified clusters using the two methods (deepSNEM and gene-based clustering) (Table 3.2). As it can be seen in Table 3.2, across all compounds the similarity of the clusters is very low, with only the clusters assigned to the SN38 having a slightly higher Jaccard index. Thus, the deepSNEM and gene-based pipeline result in a different clustering of the perturbations, due to the different biological hierarchy of information provided by the compound-induced signaling networks and differential gene expression signatures.

Table 3.1. Information regarding the perturbations used in the use case and their assigned clusters.

| Compound | MoA | Cell line | GSE | Platform | Cluster (CL) |
|---|---|---|---|---|---|
| Sirolimus | mTOR inhibitor | MCF7 | GSE116447 | Affymetrix Microarray | 53 |
| CDK-887 | CDK inhibitor | MCF7 | GSE19638 | Affymetrix Microarray | 163 |

| | | | | | |
|---|---|---|---|---|---|
| Panobinostat | HDAC inhibitor | A375 | GSE145447 | Illumina NextSeq | 22 |
| Sodium-Butyrate | HDAC inhibitor | HT29 | GSE61429 | Agilent Microarray | 22 |
| Belinostat | HDAC inhibitor | A549 | GSE96649 | Illumina NextSeq | 188 |
| SN38 | Topoisomerase I inhibitor | MCF7 | GSE18552 | Affymetrix Microarray | 158 |
| Doxorubicin | Topoisomerase II inhibitor | MCF7 | GSE19638 | Affymetrix Microarray | 33 |

Table 3.2. Jaccard similarity index between the clusters that the use case compounds were assigned to, using the gene-based and deepSNEM pipelines.

| | |
|---|---|
| Sirolimus | 0.004 |
| CDK-887 | 0 |
| Panobinostat | 0 |
| Sodium-Butyrate | 0.006 |
| Belinostat | 0.029 |
| SN38 | 0.162 |
| Doxorubicin | 0.012 |

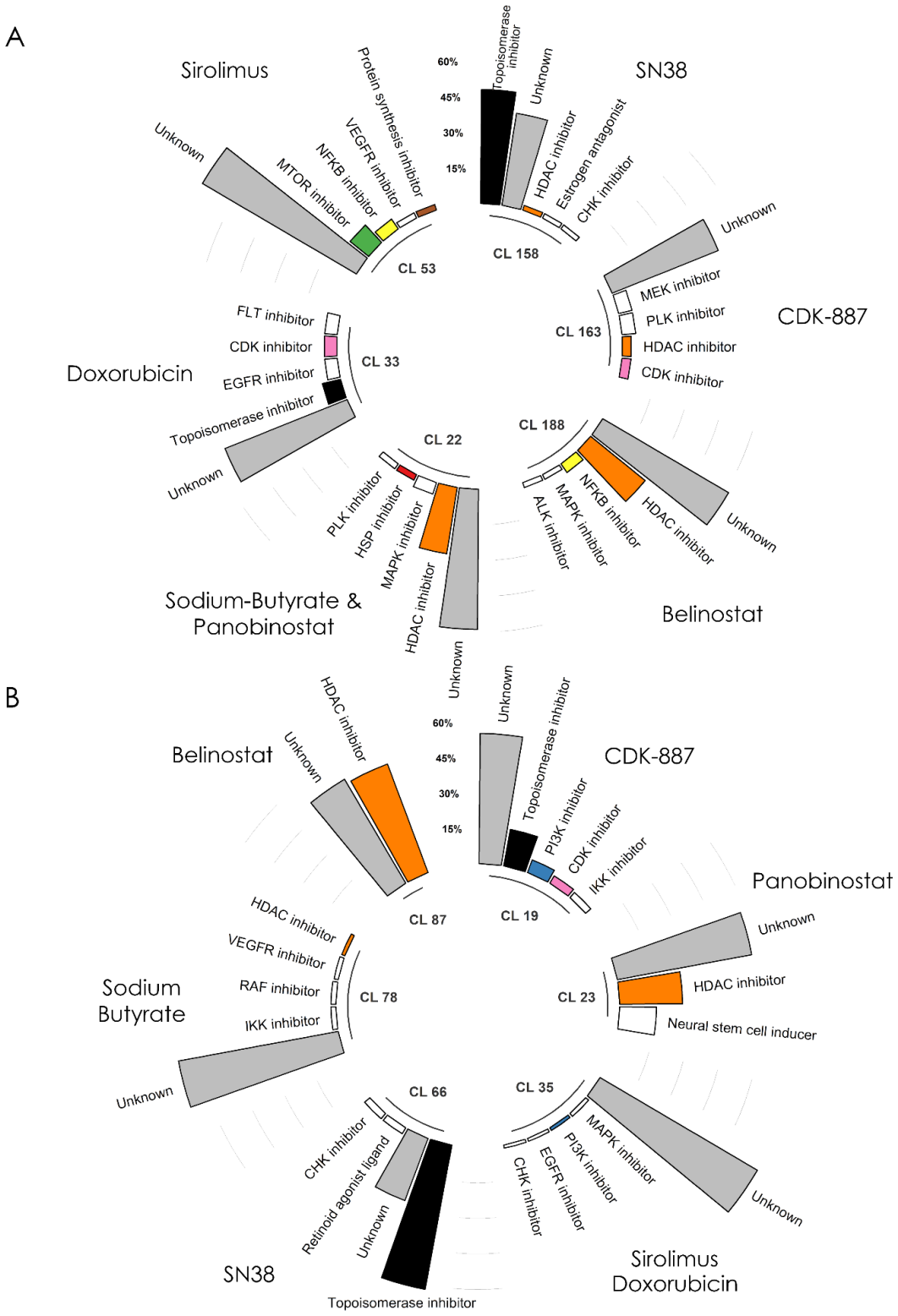



Figure 3.5. MoA composition of the compounds' clusters. (A) Bar plot of mechanism of action prevalence for the clusters that were assigned to the use case perturbations using the deepSNEM pipeline. (B) Similar bar plot for the assigned clusters using the gene-based clustering pipeline.

## 3.4 Discussion

The changes in the protein signaling network caused by a compound perturbation can aid in studying the compound's mechanism of action in the cellular system. However, analyzing compound-induced signaling networks on a massive scale is a very complex problem, not only due to the limited availability of large datasets containing such networks but also due to the complex structure of the data. This complex structure of signaling networks limits their representation abilities and poses a challenge in identifying similarities or differences between them. In this study, we created a large dataset of compound-induced signaling networks from the CMap dataset, using the CARNIVAL network creation pipeline and developed an unsupervised deep learning model to transform them into high-dimensional and information-rich representations. This novel approach, called deepSNEM was used to identify clusters of perturbations with similar network representations and offer insight into the compounds' MoA by analyzing the distribution of MoA in the clusters.

The prediction of a compound's MoA from biological response data has gained considerable attraction in the machine learning community [31,32]. This is evident by the recent release of the CTD$^2$ Pancancer Drug Activity DREAM Challenge, which tasked the community to predict a compound's MoA based on post-transcriptional and cell viability data [32]. Even though the learning task of MoA prediction is frequently modeled as supervised, in our approach we decided to develop deepSNEM in a fully unsupervised fashion. This decision was based on the nature of the learning task and the compounds' MoA, wherein if a compound has a reported MoA based on binding affinity data, we can't know with absolute certainty that it doesn't have additional MoA labels due to other binding targets or interactions between the proteins in a pathway. Thus, for some compounds the negative labels for all possible MoA indications might not be truly negative, rather they might be simply unknown. Additionally, another important benefit of using an unsupervised approach, is that we can greatly increase the amount of available data by including transcriptomic signatures following treatment with compounds that have no reported MoA. In deepSNEM the learning model is tasked to produce meaningful representations that capture the information included solely in the compound-induced signaling networks without taking into account the compounds' reported MoA. However, this unsupervised task makes the evaluation of the different models and the resulting embeddings quite challenging.

The evaluation of the validity of the resulting embeddings was based upon two tasks that test if the models can produce embeddings that capture the similarities of the input perturbation. Those tasks however, more closely resembling pass/fail tasks, rather than quantitative metrics (Figure 3.2). Thus, we cannot know with certainty which deepSNEM model variation, i.e. graph transformers, graph convolutions or graph2vec is better in terms of the resulting embeddings. For the downstream task of mechanism of action identification, we decided to use the embeddings of the graph transformed trained to maximize the mutual information between nodes that belong to networks created from the same or duplicate gene expression signatures. We argue that this deepSNEM variation is better suited to capture the information of the signaling networks, due to the graph transformer architecture and due to the mutual

information task that forces networks created from the same perturbation to have similar embeddings (see Methods 3.2). Finally, we have to note that the resulting 256-dimensional graph embeddings contain all the information of the input signaling networks, which makes it difficult for the t-SNE algorithm to project them in 2 dimensions, as it can be seen in Figure 3.3A.

The clustering analysis and MoA identification using the deepSNEM-GT-MI embeddings was performed by analyzing the MoA labels provided by the Broad Institute in the drug repurposing hub. Using this dataset, 912 out of the 3005 total compounds were mapped to 261 unique labels. We argue that this diversity of mechanisms and large number of compounds with unknown MoA in the dataset resulted in the large number k (k = 200) of clusters that were identified using the elbow plot of the k-means algorithm. Additionally, due to the large number of unlabeled compounds, in order to analyze the resulting clusters, we focused on a specific subset that is significantly enriched for at least one specific MoA (Figure 3.3B). Using this approach, we identified 11 clusters that each were enriched for the most prevalent mechanisms in the dataset. However, even for the clusters enriched for the same MoA, the large number of unknown compounds could result in different cluster compositions, which potentially further signifies the importance of analyzing biological response from different points of view, e.g. genes, pathways, signaling networks.

There have been many studies for the identification of a compound's MoA using biological response data. The majority of these approaches utilize post-transcriptional data and have been utilized successfully in the fields of systems pharmacology and drug repurposing [34,35]. Since the initial part of deepSNEM relies on transcriptomic data, similarities between the results and clustering of gene signatures and signaling networks are expected. This effect is evident in the presented use case, where some of the compounds were assigned to clusters with similar MoA composition between the gene-based and network-based pipeline. However, some compounds were assigned to clusters enriched for different MoA between the two approaches (Figure 3.5). Most importantly, between the two methods, each compound was assigned to clusters that had a very low Jaccard similarity index, meaning that the transcriptomic signatures and signaling network embeddings of deepSNEM cluster in a different way (Table 3.2). Thus, even though transcriptomic signatures do provide meaningful insight into a compound's MoA, there are cases, where analyzing the signaling networks can reveal complex relationships that are hidden in the original expression data. We argue that this is because a compound's effect on a biological system is usually caused by changes in the expression of genes that interact with each other to form specific biological processes. By supplying deepSNEM with this required prior knowledge of interactions in the form of the Omnipath PPI, the compound-specific signaling networks can provide a mechanistic view of the compound's effect and translate to the identification of its MoA [36]. Additionally, deepSNEM's signaling network creation via the CARNIVAL pipeline can provide a robust normalization factor to analyze and incorporate data from different experimental platforms (Table 3.1). Finally, the analysis of compound-induced signaling networks has the inherent benefit of increasing the interpretability of results.

The interpretability and explainability of machine learning models is a concept that has gained considerable attraction since the creation and application of powerful and complex deep learning models in various fields [37]. This is especially true in the fields of drug discovery and systems pharmacology, where understanding why the model made specific decisions and predictions can not only validate and help interpret the results, but also generate new knowledge and hypotheses regarding the complex systems under study [38]. Here, we developed a node and subgraph importance method to identify which nodes the model pays attention to when creating the embeddings and which nodes in the original networks cause the embeddings to cluster together. This resulted in the better understanding and interpretation of the novel representations that were extracted from the DL model. Using this approach, we showed that the models pay attention to similar nodes in order to cluster together compounds with similar MoA and were able to identify important signaling subgraphs that are characteristic of each cluster (Figure 3.4). For example, in the clusters enriched for mTOR inhibitors, even though mTOR as a node was not present in the input signaling networks of the cluster, deepSNEM was able to extract important subgraphs that are related to the mTOR signaling pathway.

The deepSNEM pipeline serves as proof of concept that compound-induced signaling networks can be analyzed on a massive scale, using deep learning and provide insight into the compound's effect. In a real-world application, deepSNEM would be used in combination with existing methods, utilizing transcriptomic data or pathway signatures, for a consensus-based assignment of compound perturbations into clusters that are enriched for specific MoA. Subsequently, deepSNEM could be used to identify which nodes and subgraphs mostly influenced the proposed cluster assignment, thus increasing its interpretability and help generate new hypotheses. We believe that our signaling network dataset and the proposed pipeline can help pave the way towards more studies that utilize the inherent knowledge of the changes in the signaling cascade of a system to better elucidate a compound's mechanism of action.

# 3.5 Methods

## 3.5.1 Signaling network creation

Gene expression profiles (level-5 z-score transformed) of compound perturbations were downloaded from the L1000 CMap dataset [6]. In the current study, only measurements of the relative gene expression of the 978 landmark genes in the L1000 assay were used (GSE92742). For each gene expression signature, a quality score was derived, based on its transcriptional activity score (TAS), the number of biological replicates and whether the signature is considered an exemplar, similar to the deepSIBA approach [24]**.** Based on this quality score, only the signatures with the highest quality score were selected. An overview of the transcriptomic signatures used in this study can be found in SM 1. For each signature, transcription factor (TF) activity scores were inferred using the DoRothEA R package [39]. This method utilizes a knowledge base of signed TF-target interactions called Regulons and the

VIPER enrichment algorithm to calculate TF activity scores [40]. For each compound perturbation, the discretized TF activities of DoRothEA were transformed into signaling networks using the CARNIVAL pipeline [13]. CARNIVAL solves an ILP optimization problem to infer a family of highest scoring subgraphs, from a prior knowledge network of signed and directed protein-protein interactions, which best explain the TF activities, subject to specific constraints. In our approach the OmniPath network was used as the global prior knowledge network [36]. Furthermore, the CARNIVAL pipeline without using the perturbation targets as input was utilized (InvCARNIVAL method). Finally, the ILP formulation of the problem was solved using the IBM ILOG CPLEX solver, which is freely available through the Academic Initiative (https://www.ibm.com/products/ilog-cplex-optimization-studio). Details regarding the parameters of CARNIVAL can found in SM 2.

## 3.5.2 DeepSNEM model

### 3.5.2.1 DeepSNEM-GT-MI

Each compound-induced signaling network is represented as a labeled, signed and directed graph $G = (V, E)$, with nodes (V) being the proteins and edges (E) denoting the directed physical interaction between them. Additionally, the activity of each protein is represented as a node attribute, while the inhibition or activation of each edge is represented as an edge attribute. Each input graph to the deepSNEM-GT-MI consists of a node feature matrix ($X_{prot}$), a node activity embedding ($X_{act}$) and a node proximity embedding ($X_{dist}$). The node feature matrix contains the initial protein features of each graph, which were created using the SeqVeq protein sequence model [41]. For each protein, the node activity embedding is a projection of the node's activity to the dimensions of the SeqVeq features, using a single embedding layer. The node feature and node activity matrices are added before being processed by the graph transformer. Finally, the node proximity embedding is a relative positional embedding, where each shortest path distance between nodes is calculated using the Floyd Warshall Algorithm [42]. Thus, the proximity embedding contains information about the relative distance of each node to all other nodes in the graph. The input matrices are then passed through the self-attention mechanism of the graph transformer, resulting in a final feature matrix X [19,20]. Finally, this feature matrix is summarized using the Set2Set global pooling method into a trainable whole-graph representation [43]. The model is trained fully unsupervised by maximizing the mutual information between node and whole-graph embeddings that are created from the same or duplicate transcriptomic signatures, using the CARNIVAL pipeline, thus resulting in similar graph representations for the same perturbation. Similar to the InfoGraph approach, the Jensen-Shannon Mutual Information estimator was used, while an additional term was added to the total loss function in order to force the embeddings to be uniformly distributed [22]. More details regarding the deepSNEM-GT-MI model can be found in SM 5.

### 3.5.2.2 DeepSNEM model variations

The DeepSNEM-GED variation is a Siamese graph convolutional model that is trained to minimize the error between the predicted and calculated graph edit distance for a pair of

compound-induced signaling networks. Furthermore, the deepSNEM-GT-LP variation is a transformer model similar to deepSNEM-GT-MI, albeit trained to predict the presence of an edge between two proteins (nodes). Finally, the deepSNEM-G2V model is an application of the widely used graph2vec model for whole-graph representations [21]. Details regarding these model variations can be found in SM 3 and 4.

### 3.5.3 Node and subgraph importance

The average attribution of each node (protein) to the resulting signaling network embedding was calculated using the saliency map approach of the Captum library [44]. With the saliency approach the attributions are calculated based on the gradient with respect to the input [45]. This approach results in an attribution score for each node that shows the importance of the node to the model, when calculating the network embedding. Subsequently, a scoring function was designed in order to identify the important nodes in a specific cluster of signaling network embeddings. For each node, this scoring function calculates the product of the median rank of the node's attribution score in the cluster and the frequency that the node appears in the signaling networks of the cluster. Finally, this score is normalized between 0 and 1. For visualization purposes, the 20 most important nodes of each cluster were connected using the shortest paths from the OmniPath PPI network that maximize the overall sum of importance scores in the connected graph.

## 3.6 References


1. Fotis, Chris, et al. "Network-based technologies for early drug discovery." Drug discovery today 23.3 (2018): 626-635.
2. Verbist, Bie, et al. "Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project." Drug discovery today 20.5 (2015): 505-513.
3. Yang, Xiaonan, et al. "High-throughput transcriptome profiling in drug and biomarker discovery." Frontiers in genetics 11 (2020): 19.
4. Schenone, Monica, et al. "Target identification and mechanism of action in chemical biology and drug discovery." *Nature chemical biology* 9.4 (2013): 232-240.
5. Lamb, Justin, et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." *science* 313.5795 (2006): 1929-1935.
6. Subramanian, Aravind, et al. "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles." *Cell* 171.6 (2017): 1437-1452.
7. Iorio, Francesco, et al. "Discovery of drug mode of action and drug repositioning from transcriptional responses." *Proceedings of the National Academy of Sciences* 107.33 (2010): 14621-14626.

8. Butcher, Eugene C., Ellen L. Berg, and Eric J. Kunkel. "Systems biology in drug discovery." *Nature biotechnology* 22.10 (2004): 1253-1259.
9. Mitsos, Alexander, et al. "Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data." *PLoS computational biology* 5.12 (2009): e1000591.
10. Bradley, Glyn, and Steven J. Barrett. "CausalR: extracting mechanistic sense from genome scale data." Bioinformatics 33.22 (2017): 3670-3672.
11. Melas, Ioannis N., et al. "Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury." Integrative Biology 7.8 (2015): 904-920.
12. Chindelevitch, Leonid, et al. "Causal reasoning on biological networks: interpreting transcriptional changes." Bioinformatics 28.8 (2012): 1114-1121.
13. Liu, Anika, et al. "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL." NPJ systems biology and applications 5.1 (2019): 1-10.
14. Gao, Xinbo, et al. "A survey of graph edit distance." Pattern Analysis and applications 13.1 (2010): 113-129.
15. Pržulj, Nataša. "Biological network comparison using graphlet degree distribution." Bioinformatics 23.2 (2007): e177-e183.
16. Shervashidze, Nino, et al. "Weisfeiler-Lehman graph kernels." Journal of Machine Learning Research 12.9 (2011).
17. Georgousis, Stavros, Michael P. Kenning, and Xianghua Xie. "Graph deep learning: State of the art and challenges." IEEE Access (2021).
18. Duvenaud, David, et al. "Convolutional networks on graphs for learning molecular fingerprints." arXiv preprint arXiv:1509.09292 (2015).
19. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
20. Yun, Seongjun, et al. "Graph transformer networks." Advances in Neural Information Processing Systems 32 (2019): 11983-11993.
21. Narayanan, Annamalai, et al. "graph2vec: Learning distributed representations of graphs." arXiv preprint arXiv:1707.05005 (2017).
22. Sun, Fan-Yun, et al. "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization." arXiv preprint arXiv:1908.01000 (2019).
23. Veličković, Petar, et al. "Deep graph infomax." arXiv preprint arXiv:1809.10341 (2018).
24. Fotis, C., et al. "DeepSIBA: chemical structure-based inference of biological alterations using deep learning." Molecular Omics 17.1 (2021): 108-120.
25. Corsello, Steven M., et al. "The Drug Repurposing Hub: a next-generation drug library and information resource." Nature medicine 23.4 (2017): 405-408.
26. Wang, Chun-I., et al. "mTOR regulates proteasomal degradation and Dp1/E2F1-mediated transcription of KPNA2 in lung cancer cells." Oncotarget 7.18 (2016): 25432.

27. Zhang, Yinan, and Brendan D. Manning. "mTORC1 signaling activates NRF1 to increase cellular proteasome levels." Cell cycle 14.13 (2015): 2011-2017.
28. Jiang, Chao, et al. "CSNK2A1 promotes gastric cancer invasion through the PI3K-Akt-mTOR signaling pathway." Cancer Management and Research 11 (2019): 10135.
29. Gordon, Bradley S., et al. "RhoA modulates signaling through the mechanistic target of rapamycin complex 1 (mTORC1) in mammalian cells." Cellular signalling 26.3 (2014): 461-467.
30. Jewell, Jenna L., et al. "GPCR signaling inhibits mTORC1 via PKA phosphorylation of Raptor." Elife 8 (2019): e43038.
31. Gao, Shengqiao, et al. "Modeling drug mechanism of action with large scale gene-expression profiles using GPAR, an artificial intelligence platform." BMC bioinformatics 22.1 (2021): 1-13.
32. Menden, Michael P., et al. "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen." Nature communications 10.1 (2019): 1-17.
33. Douglass, Eugene F., et al. "A community challenge for pancancer drug mechanism of action inference from perturbational profile data." (2020).
34. Iorio, Francesco, et al. "A landscape of pharmacogenomic interactions in cancer." Cell 166.3 (2016): 740-754.
35. Napolitano, Francesco, et al. "Drug-set enrichment analysis: a novel tool to investigate drug mode of action." Bioinformatics 32.2 (2016): 235-241.
36. Türei, Dénes, Tamás Korcsmáros, and Julio Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources." Nature methods 13.12 (2016): 966-967.
37. Chakraborty, Supriyo, et al. "Interpretability of deep learning models: A survey of results." 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI). IEEE, 2017.
38. Jiménez-Luna, José, Francesca Grisoni, and Gisbert Schneider. "Drug discovery with explainable artificial intelligence." Nature Machine Intelligence 2.10 (2020): 573-584.
39. Garcia-Alonso, Luz, et al. "Benchmark and integration of resources for the estimation of human transcription factor activities." Genome research 29.8 (2019): 1363-1375.
40. Alvarez, Mariano J., et al. "Functional characterization of somatic mutations in cancer using network-based inference of protein activity." Nature genetics 48.8 (2016): 838-847.
41. Heinzinger, Michael, et al. "Modeling aspects of the language of life through transfer-learning protein sequences." BMC bioinformatics 20.1 (2019): 1-17.
42. Magzhan, Kairanbay, and Hajar Mat Jani. "A review and evaluations of shortest path algorithms." International journal of scientific & technology research 2.6 (2013): 99-104.

43. Vinyals, Oriol, Samy Bengio, and Manjunath Kudlur. "Order matters: Sequence to sequence for sets." arXiv preprint arXiv:1511.06391 (2015).
44. Kokhlikyan, Narine, et al. "Captum: A unified and generic model interpretability library for pytorch." arXiv preprint arXiv:2009.07896 (2020).
45. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).

# 3.7 Supplementary Material

## 1 Data preprocessing and quality control

The filtered CMap dataset contains 7722 transcriptomic signatures from 3005 compounds tested across 70 cell lines. During the filtering process, for each compound per cell line, its signature with the highest quality across different dosages and time points was selected. The assigned quality score based on TAS, number of replicates and whether the signature is considered an exemplar is presented in Table S3.1. Only signatures with Quality score of 1 were used.

**Table S3.1** Signature quality score

| Quality score | TAS | Number of replicates | Exemplar |
|---|---|---|---|
| Q1 | > 0.4 | > 2 | True |
| Q2 | 0.2 – 0.4 | > 2 | True |
| Q3 | 0.2 – 0.4 | ≤ 2 | True |
| Q4 | 0.2 – 0.4 | > 2 | True |
| Q5 | 0.2 – 0.4 | ≤ 2 | True |
| Q6 | < 0.1 | > 2 | True |
| Q7 | < 0.1 | ≤ 2 | True |
| Q8 | < 0.1 | < 2 | False |

## 2 CARNIVAL parameters

The CARNIVAL pipeline was ran in parallel and without using the perturbation's known targets as input (InvCARNIVAL). The signaling network dataset was created with an older version of CARNIVAL in R version 3.6, but the same parameters can be used in the latest version of CARNIVAL.

The main parameters, which can be found in Table S3.2, are the time limit until the optimization terminates (timelimit), the allowed number of solutions to be generated (limitPop), the allowed number of solution to be kept in the pool of solution (poolCap) and the external ILP Solver used. The rest parameters can be set to the default of each CARNIVAL version [1].

**Table S3.2** CARNIVAL pipeline parameters

| **Execution mode** | parallel |
|---|---|
| **inverseCR** | TRUE |
| **ILP Solver** | Cplex |
| **timelimit (in minutes)** | 1800 |
| **limitPop** | 500 |
| **poolCap** | 100 |

## 3 Graph2vec

Our approach was compared with a well-known and well-established model for the generation of graph embeddings, called graph2vec [2]. Graph2vec works like doc2vec by assuming that a graph is a document and the rooted subgraphs around every node in the graph are words that compose the document. Like two documents in doc2vec have similar embeddings if they consist of similar words, two graphs in graph2vec have similar embeddings if they consist of similar subgraphs, meaning that embeddings are generated in a way, both unsupervised and domain-agnostic, in which similar graphs would have similar embeddings. In the current study, signaling networks were considered undirected, so that they can be fed to the graph2vec model, and node labels are assigned as concatenated strings of the node name and the sign of the activity of each node so that the important feature of activity in a signaling network can be considered. We reason that the transformation of the graph from directed to undirected would not undermine the quality of the resulting embeddings completely, as in the case of signaling networks every connection encountered is unique for all graphs, meaning that every unique pair of nodes that exists in the dataset can have only one unique direction and sign. The graph2vec model was trained for 1 epoch and the embedding size was set to 128.

## 4 GED model

One approach to embed graphs into a high dimensional space, while maintaining the original graph-graph similarity in the high dimensional space too, is the utilization of a distance learning approach that employs Siamese encoders (shared weights) to construct graph embeddings. As proposed in the UGraphEmb framework by Bai et al., similarity or dissimilarity between graphs can be defined by domain-agnostic and unbiased distance metrics, such as

Graph Edit Distance (GED), which can be used to train the model in an supervised manner [3]. One definition of GED is that of the number of operations, such as node or edge insertions and deletions, needed in order to transform one graph G1 into another graph G2 [4]. To this end, a distance learning model consisting of siamese graph convolutional encoders is trained to minimize the Mean Squared Error (MSE) between the predicted cosine distance of paired graph embeddings and the GED of the pair of input graphs. The input representation and the architecture of the encoder is similar to the one used in the deepSIBA framework [5]. The encoder consists of three graph convolutional layers, as proposed by Duvenaud et al., followed by one convolutional layer, one pooling layer and one fully connected layer, while the final graph embeddings are L2-normalized. The graphs are represented by a node matrix, containing information about the nodes' features, and edge attribute matrix, containing information about the edges' features and a connectivity matrix.

## 5 DeepSNEM-GT-MI

The deepSNEM-GT-MI model encodes the input matrices of each signaling network using two multi-head attention layers. Each multi-head attention layer computes the attention score using the key, query and value matrices, which are later combined using a simple feed forward network. The output of this network is used to produce the whole-graph representations using the Set2Set LSTM model. The mutual information is approximated using simple discriminators in order to train the model. The final node embedding size is set to 128, while the whole-graph representation embedding size is set to 256.

## 6 Clustering with k-means

The deepSNEM-GT-MI embeddings were clustered using the k-means algorithm. The optimal number of clusters were selected using the elbow method. The elbow plot of the clustering is presented in Figure S3.1. Figure S3.1 shows the total within sum of squared distances between the centroids and the points of each cluster, for different values of k. We can see that the elbow starts to form around k=200. This comes in agreement with the internal diversity of the dataset, where we have 261 unique MoA labels assigned to 912 compounds. Based on the results of Figure S3.1, the number of clusters was set to 200.

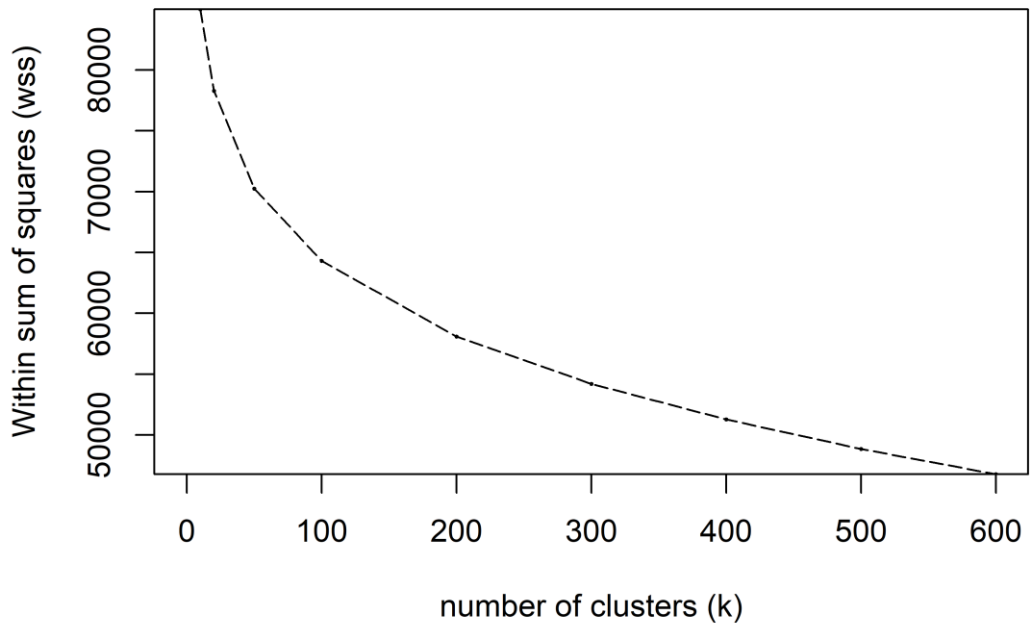



Figure S3.1. Elbow plot of the k-means clustering of the deepSNEM-GT-MI embeddings.

## 7 Subgraph importance

The important subgraphs for all analyzed clusters that were significantly enriched for a specific MoA are presented in Figure S3.2.

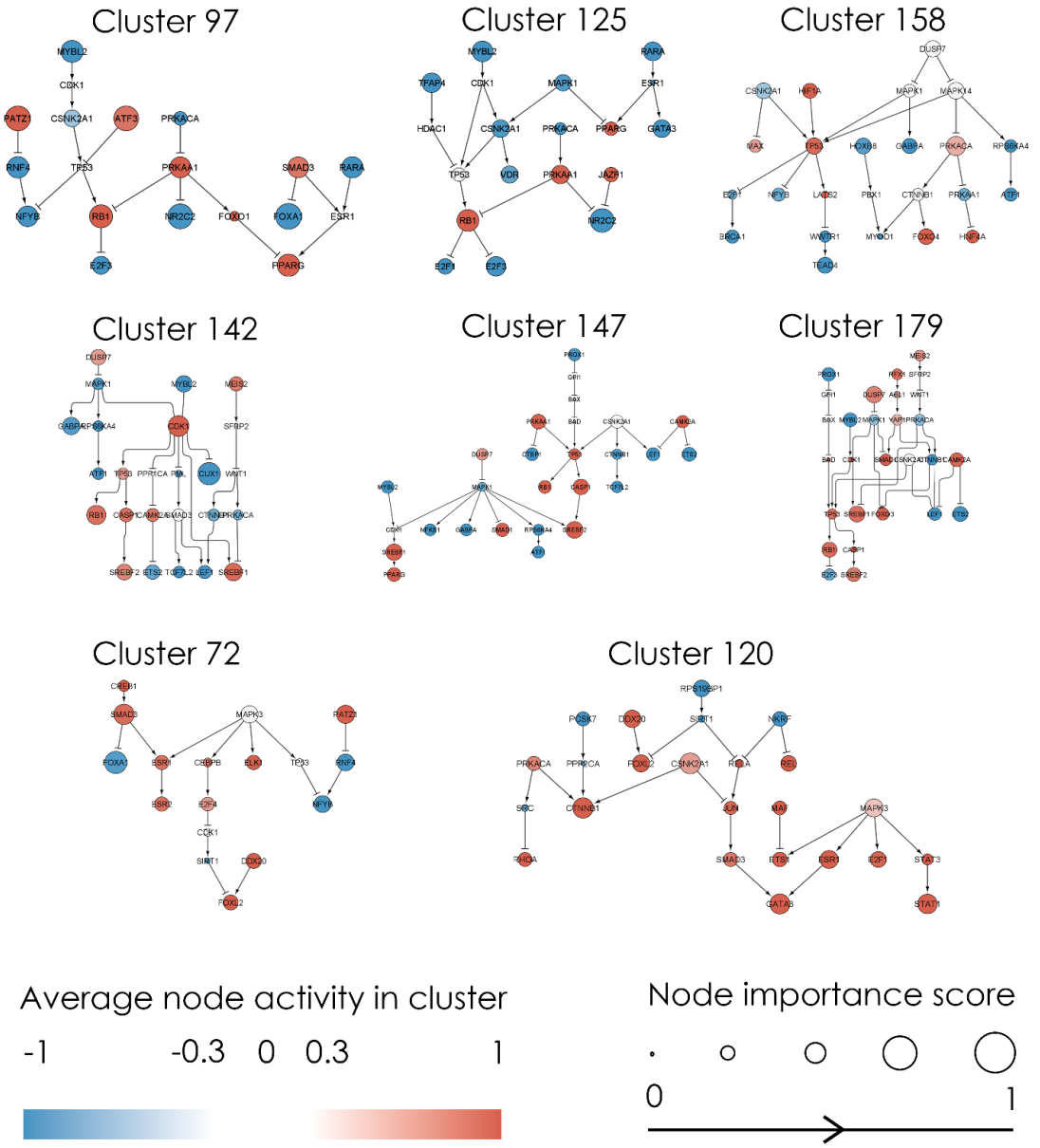



Figure S3.2. Important subgraphs identified for the clusters enriched for topoisomerase (Clusters 97, 125 and 158), HDAC (Clusters 142, 147 and 179), protein synthesis (Cluster 72) and ATP synthesis (Cluster 120) inhibitors. The average activity of each node in the cluster is color coded from blue to red. Blue nodes are inhibited, while red are activated. Each node's importance score, ranging from 0 to 1, is represented by the size of the node's circle.

## 8 Use case and gene-level clustering

The MicroArray gene expression profiles following compound treatment were preprocessed with the RMA algorithm, while the RNAseq data with the edgeR algorithm. The transcriptomic signatures of the CMap dataset were clustered with the k-means algorithm, similar to the signaling network embeddings. The elbow plot of the gene expression clustering is shown in Figure S3.3. Similar to the clustering of the deepSNEM embeddings, the number of clusters k

was set to 200. Furthermore, Figure S3.4 shows the t-SNE projections of the gene expression profiles, where the most prevalent MoA labels in the datasets are coded with different colors.

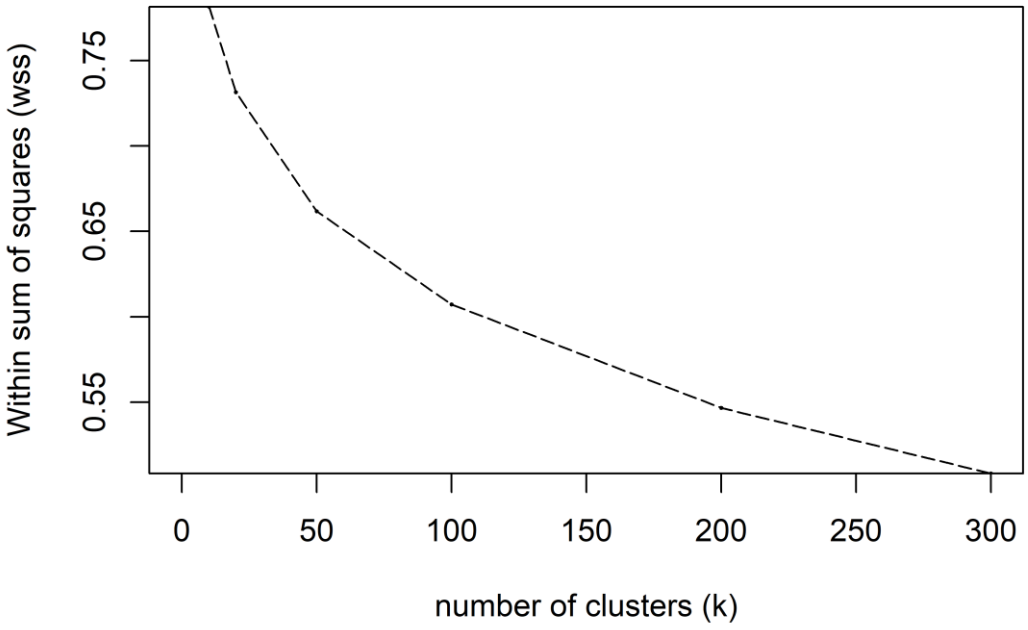



Figure S3.3. Elbow plot of the k-means clustering of the differential gene expression profiles.

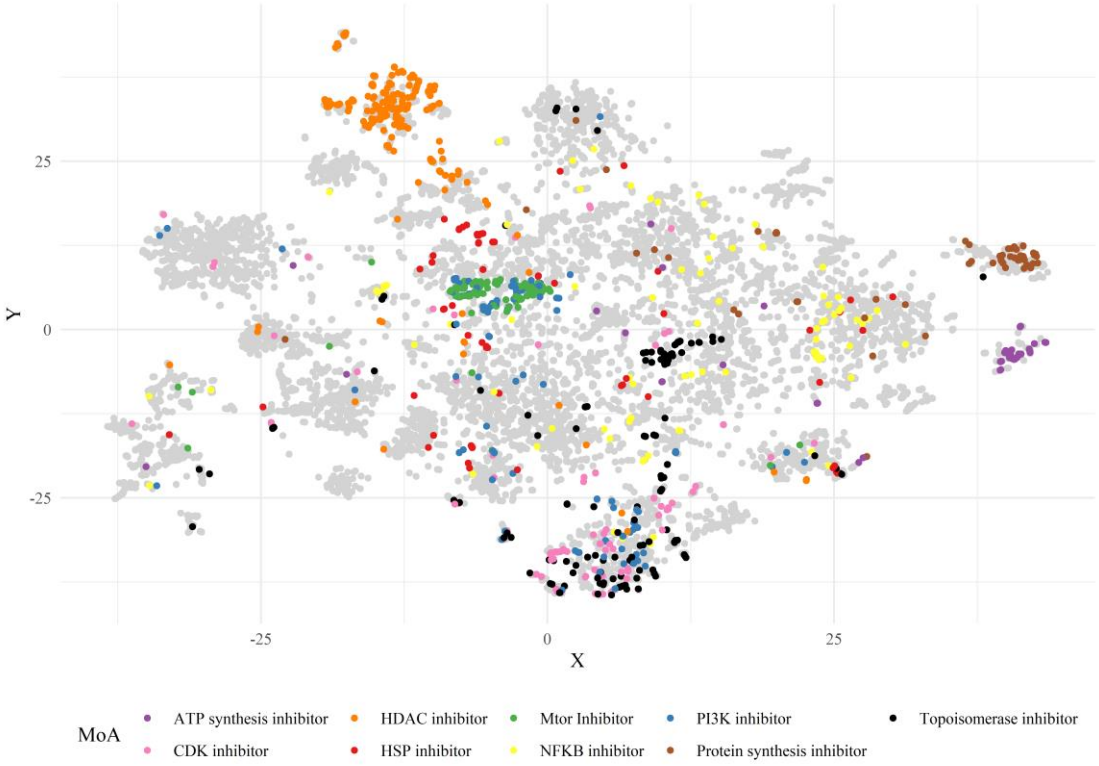


Figure S3.4. T-SNE projection of the gene expression profiles. Different colors represent the 9 most prevalent MoA in the dataset, while the grey color represents perturbations with either unknown or other MoA

## 9 References

1. Liu, Anika, et al. "From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL." NPJ systems biology and applications 5.1 (2019): 1-10.

2. Narayanan, Annamalai, et al. "graph2vec: Learning distributed representations of graphs." arXiv preprint arXiv:1707.05005 (2017).

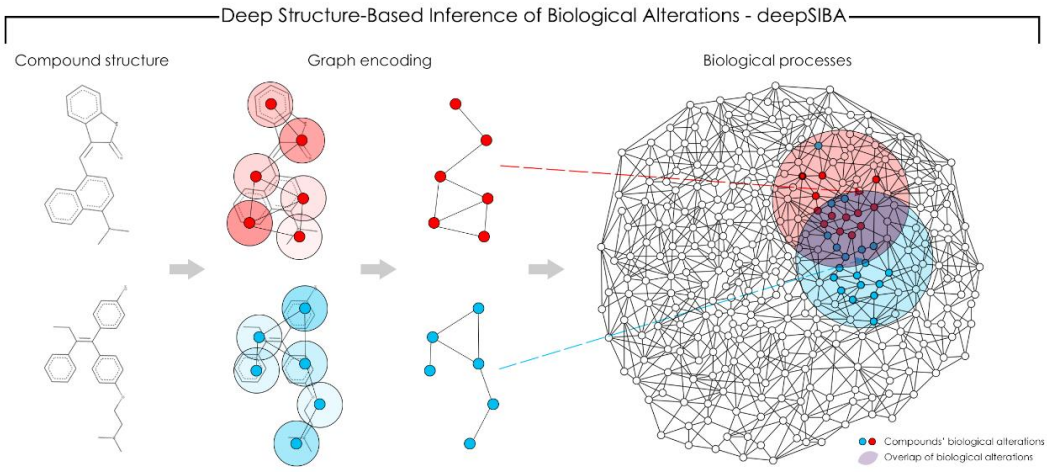
3. Bai, Yunsheng, et al. "Unsupervised inductive graph-level representation learning via graph-graph proximity." arXiv preprint arXiv:1904.01098 (2019).

4. Riesen, Kaspar, Sandro Emmenegger, and Horst Bunke. "A novel software toolkit for graph edit distance computation." International Workshop on Graph-Based Representations in Pattern Recognition. Springer, Berlin, Heidelberg, 2013.

5. Fotis, C., et al. "DeepSIBA: chemical structure-based inference of biological alterations using deep learning." Molecular Omics 17.1 (2021): 108-120.

# Chapter 4

# DeepSIBA: Chemical Structure-based Inference of Biological Alterations using deep learning




## 4.1 Chapter abstract

Predicting whether a chemical structure leads to a desired or adverse biological effect can have a significant impact for in-silico drug discovery. In this study, we developed a deep learning model where compound structures are represented as graphs and then linked to their biological footprint. To make this complex problem computationally tractable, compound differences were mapped to biological effect alterations using Siamese Graph Convolutional Neural Networks. The proposed model was able to encode molecular graph pairs and identify structurally dissimilar compounds that affect similar biological processes with high precision. Additionally, by utilizing deep ensembles to estimate uncertainty, we were able to provide reliable and accurate predictions for chemical structures that are very different from the ones used during training. Finally, we present a novel inference approach, where the trained models are used to estimate the signaling pathway signature of a

compound perturbation, using only its chemical structure as input, and subsequently identify which substructures influenced the predicted pathways. As a use case, this approach was used to infer important substructures and affected signaling pathways of FDA-approved anticancer drugs.

# 4.2 Introduction

Early stage drug discovery aims to identify the right compound for the right target, for the right disease. A very important step in this process is hit identification, in which compounds that exhibit strong binding affinity to the target protein are prioritized. Traditionally, the most widely employed method for *in-vitro* hit identification is High Throughput Screening (HTS). *In-vitro* HTS can produce hits with strong binding affinity that may later be developed into lead compounds through lead optimization. However, due to the vast chemical space, even large scale *in-vitro* HTS offers limited chemical coverage. On this front, the development of Computer Aided Drug Design (CADD) methods has enabled the virtual High Throughput Screening (vHTS) of vast compound libraries, thus effectively increasing the search space of hit identification. CADD methods for vHTS focus on compounds' chemical structures and prioritize those that are likely to have activity against the target, for further experiments.[1] More specifically, ligand-based approaches are based on the hypothesis that similar chemical structures will cause similar biological response, by binding to the same protein.[2] However, there are many cases of compounds and drugs, which although structurally dissimilar, cause similar biological effect, either because of off-target effects or by targeting proteins in the same pathway.[3] As a whole, CADD approaches focus on optimal binding affinity, by assessing a compound's structural attributes, often disregarding the effect of the perturbation on the biological system, which is closely related to clinical efficacy and toxicity.[4]

Advances in systems-based approaches and 'omics technologies have led to the development of systems pharmacology methods that aim to lower the attrition rates of early stage drug discovery. Systems pharmacology approaches couple 'omics data with knowledge bases of molecular interactions and network analysis methods in order to assess compounds based on their biological effect.[5] One approach that has gained considerable attraction is the use of gene expression (GEx) profiling to characterize the systematic effects of compounds. On this front, Verbist et al. showed how GEx data were able to influence decision making in eight drug discovery projects by uncovering potential adverse effects of the lead compounds.[6] Additionally, Iorio et al. utilized similarities between drugs' transcriptional responses to create a drug network and identified the mechanism of action of new drugs based on their position in the network.[7] Since its release, the Connectivity Map (CMap) and the LINCS project have been a cornerstone of transcriptomic-based approaches by providing a large scale database of transcriptomic signatures from compound perturbations along with essential signature matching algorithms.[8,9] CMap's approach is based on the hypothesis that compounds with similar transcriptomic signatures will cause similar physiological effects on the cell and has been widely adopted by the field of drug repurposing.[10] However, signature-based

approaches are not only limited in the search space of compounds with available GEx data but are also missing key structural information that is pivotal for drug design. Thus, an interdisciplinary framework that translates a compound's structural attributes to its biological effect holds promise in augmenting the application of both CADD and systems-based approaches for drug discovery. A computational approach that meets the requirements of such an interdisciplinary framework is Machine Learning (ML) and especially Deep Learning (DL)

The recent increase in available data and computing power has given rise to Deep Learning (DL) methods for various drug discovery tasks, including bioactivity and toxicity prediction as well as de-novo molecular design.[11–15] DL methods offer the advantage of flexible end-to-end architectures that learn task specific representations of chemical structures, without the need for precomputed features.[16] One particular DL architecture that has achieved state of the art results in several drug discovery benchmark datasets is the Graph Convolutional Neural Network (GCNN).[17,18] Molecular GCNNs operate on chemical structures represented as undirected graphs, with nodes being the atoms and edges the bonds between them. Kearnes et al. developed the Weave graph convolution module, which encodes both atom and bond representations and combines them using fuzzy histograms to extract meaningful molecule-level representations.[19] Despite their improved performance over traditional ML methods, end-to-end models including GCNNs are still prone to generalization errors on new chemical scaffolds. This is mainly because of the limited coverage of the chemical space by the training data.[20] In order to tackle this limited chemical coverage, methods like one-shot learning are promising candidates for drug discovery applications. One-shot learning techniques, such as Siamese networks, aim to learn a meaningful distance function between related inputs and have shown increased performance over traditional methods in tasks with few data points.[21–24] Altae-Tran et al. implemented one-shot learning for drug discovery by combining graph convolutions and Long Short Term Memory (LSTM) networks with attention and achieved better results than traditional GCNNs.[25] Furthermore, for drug discovery applications, uncertainty estimation is crucial, since incorrect predictions e.g. regarding toxicity can lead to incorrect prioritization of compounds for further experimental testing. [26–29] On this front, Ryu et al. developed Bayesian GCNNs for molecular property, bioactivity and toxicity predictions and showed that quantifying predictive uncertainty can lead to more accurate virtual screening results.[30] The flexibility provided by GCNN architectures along with one-shot learning and uncertainty estimation approaches can combine aspects from both systems and ligand-based methods into an interdisciplinary framework for early stage drug discovery.

In this paper, we employ deep learning to decipher the complex relationship between a compound's chemical structure and its biological effect. To make this complex problem computationally tractable, we focus on learning a combined representation and distance function that maps structural differences to biological effect alterations. For this task, we propose a deep Siamese GCNN model called deepSIBA. DeepSIBA takes as input pairs of compound structures, represented as graphs and outputs their biological effect distance, in terms of enriched biological processes (BPs) along with an estimated uncertainty. DeepSIBA is trained to minimize the loss between predicted and calculated distances of enriched BPs for

compound pairs with available GEx data. In order to account for the biological factors that influence the learning task, we train cell line-specific deep ensembles only on carefully selected chemical structures, for which high quality GEx data are available. The performance of our approach was evaluated with a realistic drug discovery scenario in mind, where gene expression data are available for only one compound per pair and compared with ML methods for pairwise (dyadic) data.[31,32] Finally, we present a novel inference approach, in which the trained models can be used to infer the signaling pathway signature of a target compound, without available GEx data. This inference approach is coupled with a novel method, based on graph saliency maps[33], which can identify substructures that are responsible for a compound's inferred biological footprint. As a use case, this approach was tasked to infer the signaling pathway signature and important substructures of approved anticancer drugs for which no transcriptomic signatures are available in our data sets, using only their chemical structure as input. DeepSIBA can be used in combination with existing *in-silico* drug discovery pipelines to identify structures that not only exhibit maximal binding affinity but also cause a desired biological effect. Thus, by incorporating deepSIBA's interdisciplinary approach, the drug discovery process can produce candidates with improved clinical efficacy and toxicity.

# 4.3 Material and methods

## 4.3.1 The deepSIBA approach

The overview of our approach is presented in Figure 4.1. Transcriptomic signatures from compound perturbations along with their respective chemical structures were retrieved from the CMap dataset.[9] For each compound perturbation, normalized enrichment scores (NES) of GO terms related to BPs were calculated using Gene Set Enrichment Analysis (GSEA). Afterwards, the lists of enriched BPs were ranked based on NES and a Kolmogorov-Smirnov based distance function, similar to GSEA, was used to calculate their pairwise distance (Figure 4.1A). During the learning phase, the proposed model is trained to predict the pairwise distance between compounds' affected BPs using only their chemical structure as input. The input chemical structures are represented as undirected graphs, with nodes being the atoms and edges the bonds between them and encoded using a Siamese GCNN architecture (Figure 4.1B). In our approach, compounds with available GEx data, representing a small portion of the chemical space, serve as reference for the inference phase. During inference, the model is tasked to predict the biological effect distance between reference and unknown compounds (without available GEx data).

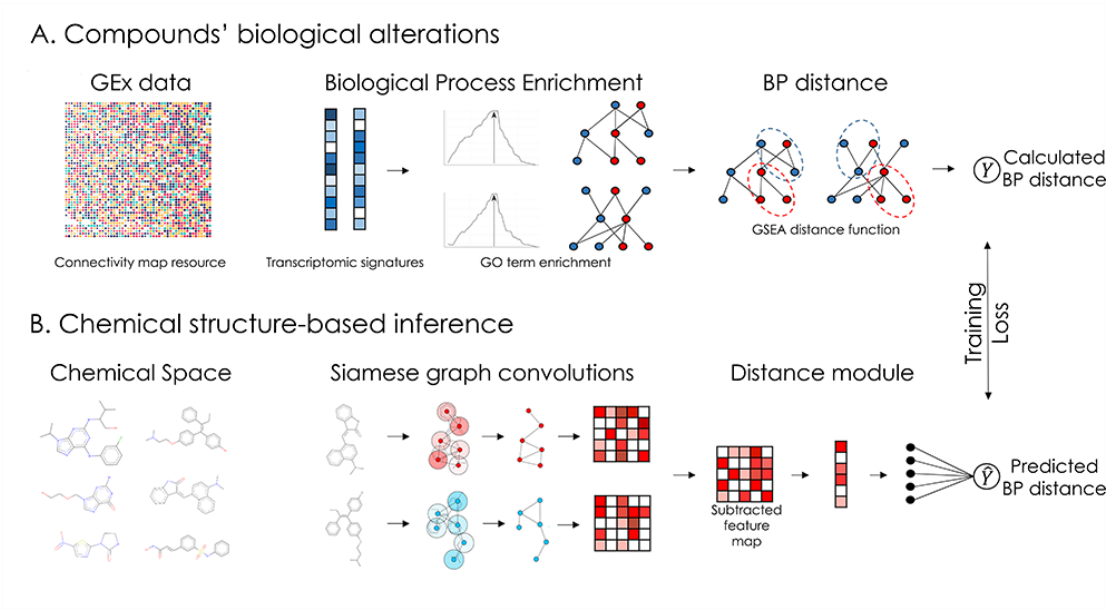



Figure 4.1 Schematic overview of deepSIBA. (A) Pairs of transcriptomic signatures following compound treatment are retrieved and enriched GO terms for BPs are calculated. The pairwise distance between enriched BPs is calculated using a Kolmogorov-Smirnov based function (Y). (B) Pairs of chemical structures are represented as molecular graphs and encoded by a deep learning model using Siamese graph convolutions. Compounds' feature maps are then subtracted and a score, which represents their distance between enriched BPs, is predicted ($\hat{Y}$). The deep learning model is trained by minimizing the loss between predicted ($\hat{Y}$) and calculated distance (Y).

## 4.3.2 Data preprocessing and quality control

Transcriptomic signatures (level 5 z-score transformed) following compound treatment were downloaded from the L1000 CMap resource.[34] In this project, only the differential expression of the 978 landmark genes in the L1000 assay was considered. For each signature, a quality score was derived, based on its transcriptional activity score (TAS), the number of biological replicates and whether the signature is considered an exemplar. This quality score ranges from Q1 to Q8, with Q1 representing the highest quality. TAS is a metric that measures a signature's strength and reproducibility and is calculated as the geometric mean of the number of differentially expressed (DEx) transcripts and the 75$^{th}$ quantile of pairwise replicate correlations. Furthermore, exemplar signatures are specifically designated for further analysis in the CLUE platform.[35] For each compound per cell line, among signatures from different dosages and time points, the signature with the highest quality was selected. An overview of the processed dataset is presented in Supplementary Information (SI) 1.1.

### 4.3.3 Biological process enrichment and pairwise distance calculation

Gene Ontology (GO) terms for biological processes (BP) involving the landmark genes of the L1000 assay were retrieved using the topGO R package in Bioconductor.[36] Only GO terms with at least 10 genes were considered. For each signature, GO term enrichment was calculated using the R package FGSEA in Bioconductor.[37] Thus, the gene-level feature vector of each perturbation was transformed to a BP-level feature vector of Normalized Enrichment Scores (NES). Pairwise distances between BP-level feature vectors were calculated similar to Iorio *et al.*[7], using the R package Gene Expression Signature in Bioconductor.[38] Given two feature vectors ranked by NES, A and B, GSEA is used to calculate the ES of the top and bottom GO terms of A in B and vice versa. The distance between the vectors is computed as $\mathbf{1} - \frac{\boldsymbol{ES_{A\ in\ B}+ES_{B\ in\ A}}}{\mathbf{2}}$ and ranges from 0 to 2. An important parameter that can introduce bias in the distance calculation is the number of top and bottom GO terms to consider during GSEA. On this front, an ensemble approach was developed, by calculating pairwise distances between BP-level feature vectors for 5 different numbers of top and bottom GO terms. The numbers we considered were selected based on the average number of significantly enriched GO terms across all perturbations in the dataset (see SI 1.3 for details). The distance scores were finally averaged and normalized between 0 and 1.

### 4.3.4 Siamese GCNN architecture

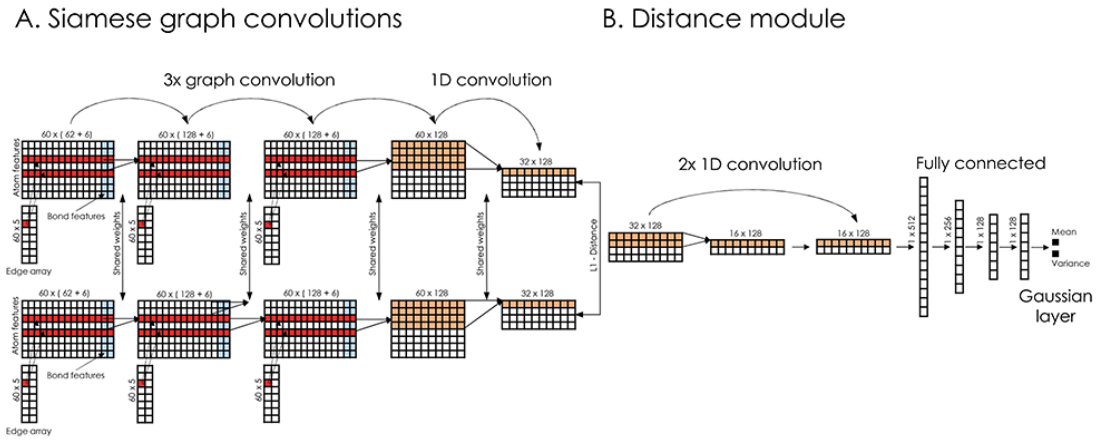



Figure 4.2 Schematic representation of the model's architecture. (A) Siamese graph convolutional encoders; compounds' molecular graphs are encoded using 2 encoders with shared weights (Siamese). Each encoder consists of 3 graph convolution and 1 convolution layers. (B) Architecture of the distance module; the distance module consists of 2 convolution, 3 fully connected and 1 Gaussian regression layers.

A schematic representation of our model's architecture is presented in Figure 4.2. The learning model takes as input the chemical structures of compound pairs and predicts their biological

distance, at the level of affected biological processes (GO terms). Regarding the input, chemical structures are represented as undirected graphs, with nodes being the atoms and edges the bonds between them. Each input is encoded using 3 matrices: the atom array, which contains atom-level features, the bond array, which contains bond-level features and the edge array, which describes the connectivity of the compound (see SI 2.1 for details). The learning model consists of two Siamese encoders (shared weights) that embed the input graphs into a high dimensional latent space and a trainable distance module that outputs the final distance prediction. The Siamese encoders consist of 3 graph convolutional layers that learn neighborhood-level representations, followed by a convolutional layer that extracts compound-level features (Figure 4.2A). Graph convolutions were implemented similar to Duvenaud *et al.*[17] (see SI 2.2 for details). The overall goal of the Siamese encoder is to learn task-specific compound representations. The feature maps of the last Siamese layer are then subtracted and their absolute difference is passed to the distance module. The distance module consists of 2 convolutional layers, which extract important features from the difference of the feature maps and 3 fully connected layers that aim to combine those features, while progressively reducing the dimensions (Figure 4.2B). Finally, a Gaussian regression layer outputs a mean and variance of the biological effect distance between the compound pair. By treating the distance as a sample from a Gaussian distribution with the predicted mean and variance, the model is trained end-to-end by minimizing the negative log-likelihood criterion[27] given by

$$-logp_\theta(y_n|X_n) = -\frac{1}{2}log\sigma_\theta^2(x) - \frac{1}{2\sigma_\theta^2(x)}\big(y - \mu_\theta(x)\big)^2 + constant.$$

For each cell line, an ensemble model combining 50 models was created. The ensemble's output is also a Gaussian, with mean and variance calculated from the uniformly weighted mixture of each model. The coefficient of variation (CV) of the Gaussian mixture is used as the model's estimate of predictive uncertainty. The model's hyperparameters, along with the equations for the Gaussian mixture's mean and variance are presented in SI 2.3 and 2.4.

## 4.3.5 Dataset splitting and evaluation metrics

For each cell line, available compounds were split into training and test. Each cell line specific training set consists of the pairwise distances between training compounds' affected BPs, while each test set contains distances between test and training compounds. Additionally, the Tanimoto similarity between the ECFP4 fingerprints of all training and test compounds was calculated and test compounds that exhibited a similarity higher than 0.85 to any training compound were excluded. An overview of the training and test sets is presented in SI 4.1. Across all test scenarios, model performance was evaluated in terms of Mean Squared Error (MSE), Pearson's r and precision. MSE and Pearson's r were calculated between the predicted and computed distance values. In order to calculate precision, the continuous distance values were transformed to binary form by comparing them with an appropriate distance threshold. Even though the learning task is a regression problem, given its nature and potential

applications, high precision ($\frac{true\ positives}{positives}$) is important in order to avoid false positive hits for validation experiments. The appropriate distance threshold for precision was set at 0.2, based on the distance distribution of duplicate compound signatures, the threshold equivalent to a 90% Connectivity Score and the relationship between the threshold and the actual average number of common enriched BPs. Duplicate signatures indicate transcriptomic signatures from the same compound perturbation, cell line, dose and time point that were assayed on different L1000 plates. Thus, the distribution of distances between duplicate signatures most closely approximates the reference distribution of truly similar biological effect. A thorough investigation of the distance threshold to distinguish compounds with similar biological effect is presented in SI 5.1.

### 4.3.6 Signaling pathway inference for target structure

The predictions of a trained deepSIBA model can be used to infer a pathway signature for a target structure without the need for GEx data, in terms of the most upregulated and downregulated signaling pathways. The inference approach is similar to the k-Nearest Neighbor algorithm (KNN). Given a target structure, a trained ensemble model for the cell line of choice is used to predict all pairwise distances between target and training compounds. The predicted distance represents the difference between compounds' enriched BPs (GO terms). Training set compounds with predicted distance less than a specified threshold $d_{th}$ are selected as the target's neighbors. If a target structure has more than $k$ neighbors, a signaling pathway signature can be inferred in the following way. For each neighbor $N_i$, the lists of the top 10 most upregulated and most downregulated pathways, based on NES, are constructed. Pathway enrichment is calculated using FGSEA with KEGG as a knowledge base.[39] KEGG signaling pathways were chosen for inference due to their interpretability. Signaling pathways that appear in the neighbors' lists with a frequency score higher than a threshold $f_{th}$ are selected. Additionally, to account for signaling pathways that are frequently upregulated or downregulated in the set of training compounds, a p-value for each inferred pathway is also calculated. On this front, sets of $k$ neighbors are randomly sampled 5000 times from the training set and a Null distribution of frequency scores for each pathway is derived. A p-value is computed as the sum of the probabilities of observing equally high or higher frequency scores. Finally, only pathways with p-value lower than a threshold $p_{th}$ are inferred. Thus, for each chemical structure, our approach infers two signatures of variable length (up to 10 each) of potentially downregulated and upregulated pathways respectively. For the MCF7 cell line, the aforementioned thresholds and parameters of the inference approach were selected by evaluating the results, in terms of precision and number of inferred pathways, on its respective test set (see SI 6.1 for details).

### 4.3.7 Substructure importance using graph-based gradients

A graph-based gradient approach, similar to saliency maps, was developed to identify important substructures that influence the biological effect similarity of chemical structure pairs. First, the derivative of deepSIBA's output w.r.t the input matrices that contain the atom features of each compound, in the input pair, is calculated using Tensorflow ($\left[\frac{\partial F}{\partial X_{atoms}}\right]$). Subsequently, for each compound atom importance is scored, using a directional derivative approach. Thus, similar to vector calculus, the directional derivative of a scalar f(X), with X being a matrix, in the direction of a matrix Y is

$$\nabla_Y f(X) = tr\left(\frac{\partial f}{\partial X} * Y\right),$$

where, $\frac{\partial f}{\partial X}$ is the gradient matrix, or in our case $\frac{\partial F}{\partial X_{atoms}}$, while $Y$ can be considered a matrix with zeros everywhere, except the row containing the specific atom's feature. Thus, an importance score for each atom of a compound can be calculated as

$$S_a = tr\left(\frac{\partial F}{\partial X_{atoms}} * Y_a\right),$$

where the only non-zero part of $Y_a$ is the one-hot encoded feature vector of atom $a$. For each atom the importance score $S_a$ was transformed to a count score $C_a$, based on how many times each atom was in the top 20% most important atoms for each model in a deepSIBA ensemble. When scoring atom importance during the pathway inference approach, a similar score was calculated based on the times an atom was present in the top 20% for each target-reference pair. Finally, due to the GCNN core module of deepSIBA, important substructures are formed by important atoms that are neighbors in the compound's molecular graph. Atom importance is visualized using the RDKit library.[40]

## 4.4 Results and discussion

### 4.4.1 Biological factors influence the model's learning task

The presented model is tasked to predict the biological effect distance between compounds, using their molecular graphs as input. Considering that this distance is calculated from experimental GEx data following compound treatment, there are specific biological factors that can influence the learning task. The CMap dataset contains over 110K transcriptomic signatures from over 20K compounds assayed across 70 cell lines. By carefully analyzing these signatures and their pairwise distances, we were able to pinpoint the most influential factors and identify their effect on the model's target value.

**4.4.1.1 The variation in quality of GEx data is reflected on the calculated distance value.** The quality of gene expression data, from which transcriptomic signatures in the Connectivity map were derived, varies across compound perturbations. In our case, this variation in data quality is especially important. On this front, a categorical quality score, ranging from Q1 to Q8, was

assigned to each signature, with a score of Q1 representing the highest quality (see SI 1.1). In order to assess the effect of signature quality, distributions of distances between duplicate transcriptomic signatures (same compound, cell line, dose, time) for different quality scores were examined and are presented in Figure 4.3A. As expected, Q1 duplicate signatures are very similar and their distances are centered near a small value. However, this is not the case for Q2 duplicate signatures, where differences in differentially expressed genes are prominent even when all the perturbation parameters are kept constant. It is clear that signature quality significantly affects the distribution of the model's target variable.

**4.4.1.2 Distances between transcriptomic signatures vary across cell lines.** Compound response, in terms of DEx genes, is highly dependent on the cellular model. Due to different genetic backgrounds and gene expression patterns the same compound perturbation will have different transcriptomic signatures across cell lines.[41] This dependence, directly affects the distance between compounds' transcriptomic signatures for different cell lines. The relationship between gene-level distances of compound pairs present in both the MCF7 and VCAP cell lines, with Q1 signatures, is shown in Figure 4.3B. In general, Q1 transcriptomic distances of the same compound pair in the 2 examined cell lines are moderately correlated (Pearson's r = 0.469). However, there is a significant number of compound pairs which have similar transcriptomic signatures in one cell line but not in the other (lower right and upper right quadrants of Figure 4.3B). Such cases are even more prominent for compound pairs with Q2 signatures (see SI 1.2). Thus, the cell line effect poses a problem for the proposed learning task by providing a one-to-many mapping between input (pair of chemical structures) and output (distance between signatures).

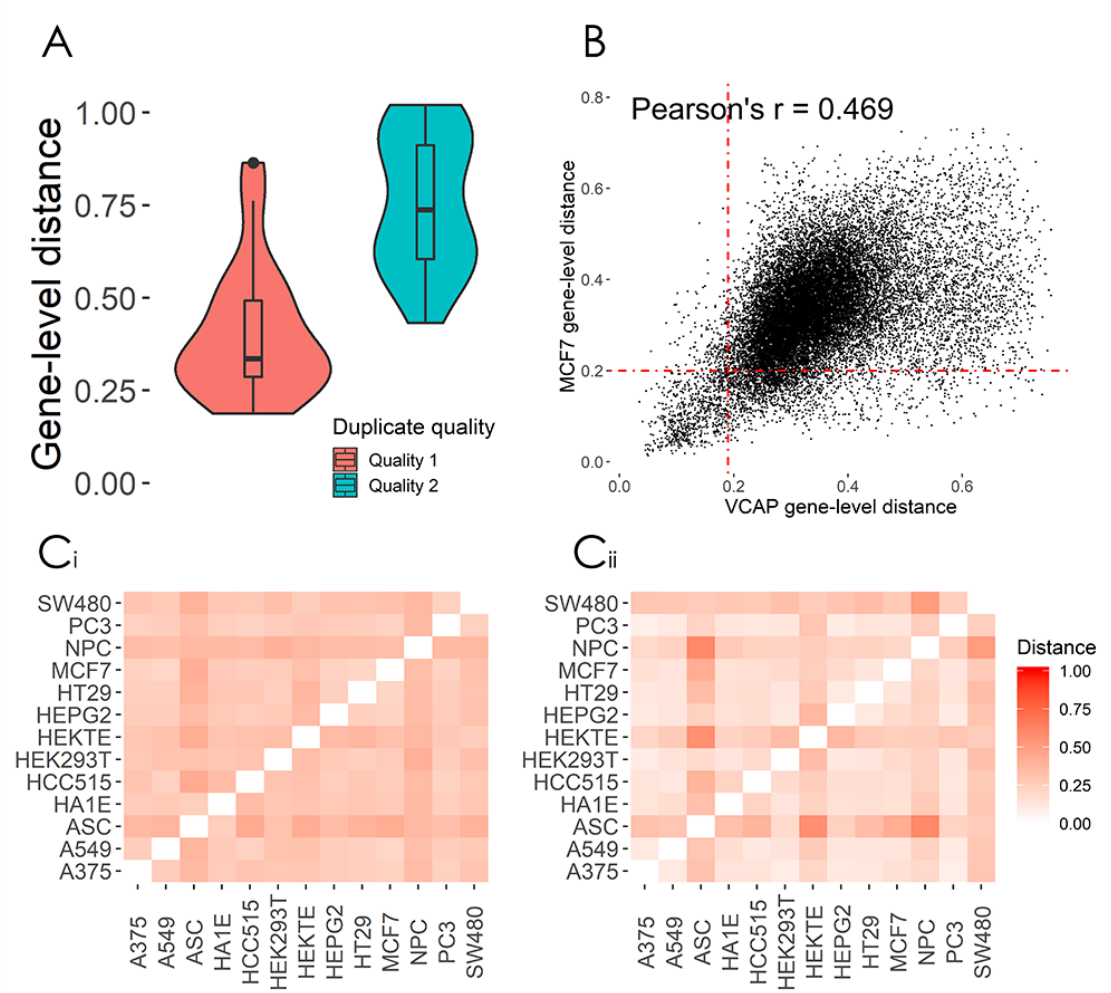



Figure 4.3 Influence of biological factors on the learning task. (A) Evaluation of data quality based on the gene-level distance between duplicate compound perturbations (same compounds) for the MCF7 cell line. (B) Scatterplot of distances between transcriptomic signatures (Quality 1) of compound pairs present in both the MCF7 and VCAP cell lines. The red lines, at 0.2 for MCF7 and 0.19 for VCAP indicate the mean + standard deviation of the distribution of distances between Q1 duplicate signatures for each respective cell line; ($C_i$ & $C_{ii}$) Heatmaps of gene and BP-level distances between cell lines for the knockdown of the MYC gene.

**4.4.1.3 Compounds' biological effects are better represented on a functional level.** A distance function that operates directly on transcriptomic signatures does not account for smaller differences in the DEx of genes that belong to the same biological pathway. Thus, the similar effect between perturbations, in terms of enriched BPs, might not be clearly reflected on their gene-level distance. On this front, a comparison of BP and gene-level distances between cell lines for the knockdown of the MYC gene (Q1 signatures) with shRNA is presented in Figure 4.3C. MYC is an oncogene that plays a key role in cell cycle, transformation and proliferation and was selected because its knockdown is expected to cause similar response across cancer cell lines. The smaller overall distance between cell lines in Figure 4.3$C_{ii}$

indicates that the expected similar effect of MYC knockdown is better highlighted on a functional level between enriched biological processes rather than between transcriptomic signatures (Figure 4.3C$_i$). Furthermore, we evaluated which distance metric, either between BPs or DEx genes, can better highlight the expected similar biological effect of structurally similar compounds.[42] In the CMap dataset, we identified pairs of similar chemical structure using the traditional Tanimoto coefficient between ECFP4 fingerprints and then calculated what percentage of those cause similar biological response at the BP and gene-level (Table 4.1). As it can be seen in Table 4.1, across all structural distance thresholds the percentage of structurally similar compounds with similar biological effect is significantly higher when distance is calculated between signatures of enriched BPs. A detailed comparison between structural and biological effect distances for all examined cell lines is presented in SI 1.4.

Table 4.1 Percentage of structurally similar compounds that cause similar biological effect, either at the gene or BP-level, in the MCF7 cell line

| Structural distance threshold | Pairs with similar chemical structure | Pairs affecting similar BPs (%)* | Pairs affecting similar genes (%)** |
|---|---|---|---|
| 0.10 | 91 | 76.9 | 68.1 |
| 0.15 | 114 | 75.4 | 65.7 |
| 0.20 | 200 | 74.0 | 61.0 |
| 0.25 | 316 | 69.9 | 57.6 |
| 0.30 | 494 | 65.3 | 51.0 |

* BP distance threshold to consider compounds similar = 0.2

** Gene distance threshold to consider compounds similar = 0.19

Through the careful analysis of the processed data sets, we showed that raw data quality greatly affects the distribution of distance values and that lower quality transcriptomic signatures of the same compound, with the same perturbation parameters (duplicates), often exhibit large differences in terms of DEx genes (Figure 4.3A). Based on these findings, we chose to develop deepSIBA using only compounds with available Q1 transcriptomic signatures. Furthermore, we showed that the transcriptomic distance of a compound pair can vary depending on the choice of cellular model (Figure 4.3B). One common approach to address this issue is to aggregate either transcriptomic signatures or distance values across cell lines. While aggregating enables the training of a general model on all available compound pairs, it can often produce misleading results and cause information loss. Thus, we decided to make our approach cell line specific and develop our models for cell lines that have the highest number of Q1 transcriptomic signatures following compound treatment. Finally, we highlighted that a distance function operating on enriched BPs, rather than genes, can better capture the expected biological effect similarities of perturbations with similar structure or biological nature (Table 4.1, Figure 4.3C). We reason that this is the case due to the BP enrichment analysis that precedes the distance calculation, which can capture smaller changes in the expression of genes that interact with each other to form a biological process. By analyzing the relationship between the aforementioned experimental factors and our target variable, we were able to make data-driven decisions to propose a learning task that minimizes their effect. In the following sections we evaluate the ability of deepSIBA to learn

the proposed task and test whether our approach can identify dissimilar structures that affect similar BPs in a meaningful way.

## 4.4.2 Performance evaluation

Model performance was evaluated on pairs of reference and test compounds. Test compounds were removed from the training sets and thus represent new chemical structures without available experimental GEx data. Additionally, the effect of the structural similarity between input compounds on performance, along with the utility of the model's estimate for uncertainty, were investigated. Finally, we evaluated the performance of our approach on test chemical structures that are very different from the ones used in training.

**4.4.2.1 Test set performance.** In each cell line specific test set, the performance of deepSIBA was compared to the performance of ReSimNet and TwoStepRLS. ReSimNet is a recently proposed deep Siamese MLP model, while TwoStepRLS is a regularized kernel-based regression method. Both methods are suitable for distance/similarity learning for pairwise (dyadic) data and were implemented using compounds' ECFP4 fingerprints as input (see SI 3.1 and 3.2 for details). As shown in Table 4.2, across all cell lines, deepSIBA achieved the lowest overall MSE and in the 1% of test samples with the lowest predicted values. The ReSimNet models for the A375 and MCF7 cell lines achieved the highest Pearson's r, while deepSIBA and TwoStepRLS had the highest Pearson's r, for the PC3 and VCAP cell lines respectively. In terms of precision, the deepSIBA models heavily outperformed the other methods across all cell lines. In order to calculate precision, an appropriate distance threshold of 0.2 was used for all approaches (see section 4.3.5 for details) While ReSimNet and TwoStepRLS exhibited low precision, they predicted that many more compound pairs will have similar biological effect. When examining the lowest 1% of predicted distances, their precision improves and in the MCF7 cell line TwoStepRLS' precision surpasses deepSIBA's. Additional 5-fold cross validation results for each cell line are presented in SI 5.2.

Table 4.2 Cell line specific test set performance

| Cell line | Model | MSE | MSE @1%* | Pearson's r | Precision (%) | Precision @1% (%)* | Predicted similar pairs |
|---|---|---|---|---|---|---|---|
| A375 | DeepSIBA | **0.008** | **0.006** | 0.59 | **98.22** | **98.22** | 169 |
| | ReSimNet | 0.012 | 0.022 | **0.60** | 32.23 | 56.80 | 18243 |
| | TwoStepRLS | 0.010 | 0.008 | 0.51 | 44.61 | 78.68 | 4024 |
| PC3 | DeepSIBA | **0.011** | **0.007** | **0.53** | **89.29** | **89.29** | 28 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ReSimNet | 0.017 | 0.032 | 0.49 | 25.02 | 46.89 | 14195 |
| | TwoStepRLS | 0.013 | 0.041 | 0.44 | 29.98 | 38.96 | 1758 |
| VCAP | DeepSIBA | **0.033** | **0.026** | 0.41 | **71.63** | **71.63** | 141 |
| | ReSimNet | 0.039 | 0.105 | 0.38 | 32.69 | 52.97 | 9245 |
| | TwoStepRLS | 0.034 | 0.049 | **0.43** | 32.34 | 31.12 | 3120 |
| MCF7 | DeepSIBA | **0.012** | **0.007** | 0.56 | **61.03** | 61.03 | 195 |
| | ReSimNet | 0.015 | 0.029 | **0.59** | 26.93 | 51.20 | 13420 |
| | TwoStepRLS | 0.015 | 0.010 | 0.47 | 33.55 | **70.14** | 4322 |

**4.4.2.2 Transferring knowledge to other cellular models**. Initially deepSIBA was trained and evaluated in the four cell lines that have the highest number of Q1 transcriptomic signatures following compound treatment. In order to expand the biological coverage of deepSIBA we utilized transfer learning to train our models on six additional cell lines which have the next highest number of Q1 signatures. On this front, we pre-trained a deepSIBA model on the entirety of the A375 cell line dataset and then applied it on additional cell lines by resuming training for 6 epochs. The performance of the transfer learning approach on each cell line specific test set is presented in Table 4.3. Across all additional cell lines deepSIBA was able to achieve similar performance to that of the A375, PC3, MCF7 and VCAP cell lines.

Table 4.3. Test set performance of the transfer learning approach

| Cell-line | MSE | MSE @1% | Pearson's r | Precision (%) |
|---|---|---|---|---|
| HT29 | 0.010 | 0.013 | 0.60 | 84.88 |
| A549 | 0.013 | 0.012 | 0.62 | 83.00 |
| HA1E | 0.015 | 0.009 | 0.58 | 100 |
| HEPG2 | 0.013 | 0.014 | 0.61 | 99.10 |
| HCC515 | 0.014 | 0.010 | 0.52 | 97.92 |

| | | | | |
|---|---|---|---|---|
| NPC | 0.006 | 0.005 | 0.67 | 73.64 |

**4.4.2.3 Performance as a function of the structural distance between input compounds**. As shown previously, similar chemical structures have similar signatures of enriched BPs. However, there are many cases of structurally dissimilar compounds that cause similar biological response. It is therefore important to evaluate the ability of deepSIBA to identify such cases, by calculating its performance for test pairs of varying structural distance. On this front, each cell line specific test set was split into parts based on the structural distance between compounds and in each part MSE and precision were calculated (Figures 4A and 4B). As a measure of structural distance/similarity, the traditional Tanimoto coefficient between ECFP4 fingerprints was utilized. The PC3, A375 and VCAP deepSIBA models maintain a high precision across all different structural distance ranges (Figure 4.4B). The exception is the MCF7 model, for which precision slightly decreases for structural distance higher than 0.7. Regarding MSE, only the VCAP model exhibits a slightly higher MSE as structural distance increases (Figure 4.4A). As a whole, the models' performance seems unaffected by the distance between the ECFP4 fingerprints of the input pairs.

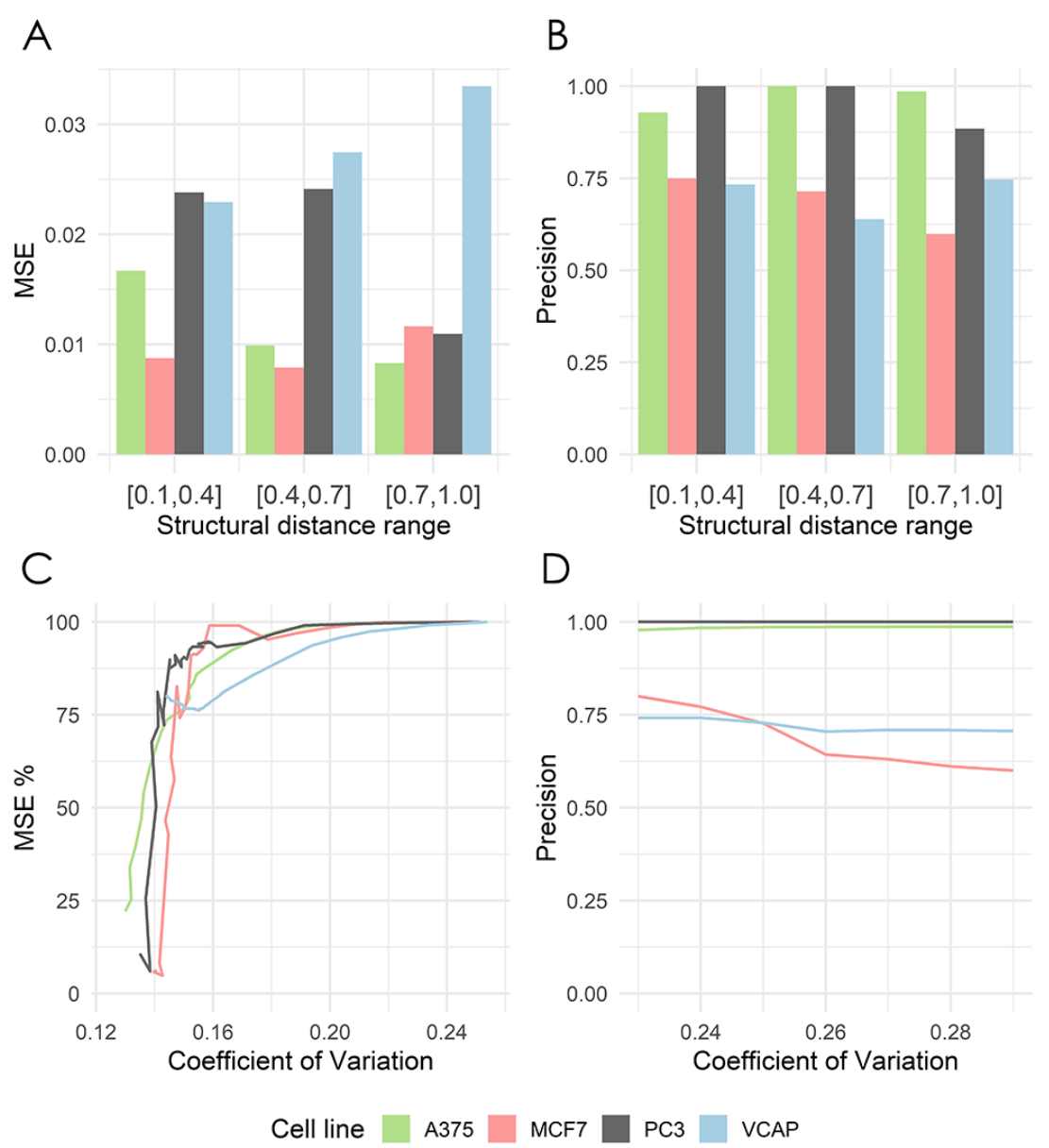



Figure 4.4 Performance as a function of structural distance and predictive uncertainty; (A) MSE for different ranges of structural distance between compound pairs; (B) Precision for different ranges of structural distance between compound pairs. (C) Percentage of total test MSE, calculated in samples with increasing CV; (D) Precision calculated in test samples with increasing CV.

**4.4.2.4 Performance as a function of predictive uncertainty**. It has been shown that quantifying predictive uncertainty can lead to more accurate results in virtual screening applications.[30] In this context, the relationship between the predictive uncertainty estimate and performance was investigated. In DeepSIBA we estimate predictive uncertainty as the coefficient of variation (CV) of the mixture of each model's Gaussian in the ensemble. MSE and precision were calculated for specific samples in the test set, which have CV lower than an increasing threshold and are presented in Figures 4C and 4D. As the CV threshold increases and more samples with higher CV are included in the evaluation, the MSE of the models

increases as well and eventually becomes the MSE of the entire test set (Figure 4.4C). On the other hand, due to the low number of false positives, for all the models, precision seems unaffected by the CV threshold. Only the MCF7 model, which has the lowest overall precision, exhibits a higher precision for samples with lower CV (Figure 4.4D). Overall, the results indicate that point predictions with lower uncertainty are closer to the true value, or that when the model is certain, it's usually not wrong.

**4.4.2.5 Generalization on different chemical structures.** End-to-end deep learning models for drug discovery have trouble generalizing on new compounds that are structurally very different from the ones used to train them. In order to evaluate the ability of our approach to generalize on different chemical structures, the performance of the A375 model was evaluated on 2 extra test sets and is presented in Table 4. These test sets were created by restricting the maximum allowed structural similarity between selected test compounds and all remaining training compounds and thus represent test scenarios of increasing difficulty (Figure 4.5A). As the minimum distance between test and training compounds increases, the performance of the model becomes worse. However, the performance decrease in terms of MSE and Pearson's r is smaller than the decrease in precision. In this case, the distance threshold to calculate precision was set to 0.22, because in the hardest test set (#3) there were no samples with predicted value lower than 0.2. Thus, even though the model's performance is comparable across test sets in terms of regression metrics, its ability to identify compounds with similar biological effect is hindered. In this case, it is important to estimate predictive uncertainty and evaluate its utility, by focusing on predictions with smaller CV (Figure 4.5B). In the third test set, which only contains compounds with maximum similarity to the training compounds less than 0.3, the model's precision is significantly higher for test predictions with low CV. More specifically, in test samples with CV lower than 0.16, the model's precision is upwards of 80%.

Table 4.4 Generalization performance on different chemical structures for A375

| Test set | Max similarity to training set | MSE | Pearson's r | Precision (%) | Predicted Similar Pairs |
|---|---|---|---|---|---|
| #1 | [0-0.85] | 0.0083 | 0.59 | 97.26 | 876 |
| #2 | [0.35-0.65] | 0.0092 | 0.52 | 76.48 | 330 |
| #3 | [0-0.3] | 0.0107 | 0.44 | 50.37 | 135 |

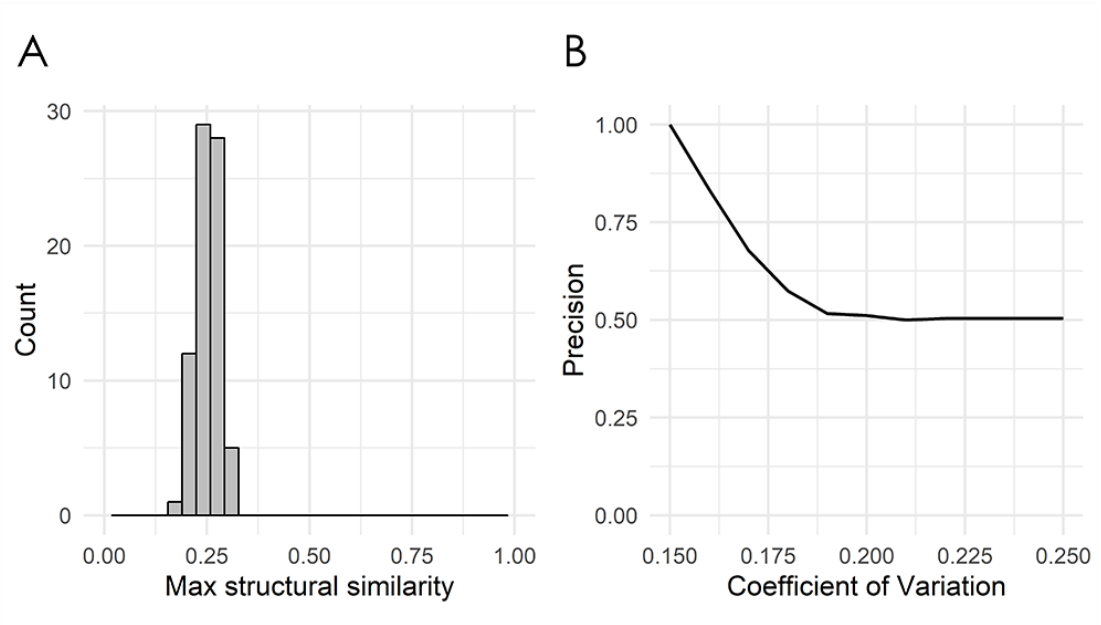



Figure 4.5 Precision and uncertainty estimation for test set number 3. (A) Histogram of maximum structural similarity between test and training compounds for test set number 3; structural similarity is calculated between compounds' ECFP4 fingerprints. (B) Precision calculated in test samples with CV lower than an increasing threshold.

Across all examined cell lines, deepSIBA was able to identify chemical structures that affect similar BPs, outperforming, especially in terms of precision, the distance learning methods that utilize compounds' ECFP4 fingerprints as input (Table 4.2). Even though the learning task is regression, we reason that precision is a crucial metric, considering the potential screening applications of deepSIBA in order to identify compounds that exhibit similar biological effect to a query. In this scenario, high precision, rather than a large number of identified hits, is required to correctly prioritize compounds for downstream experimental validation. We chose not to compare our approach with traditional machine learning methods, e.g. Random Forests and SVMs, because we argue that these are not optimal for a distance/similarity learning task. Furthermore, deepSIBA was able to maintain its high performance regardless of the structural similarity between input compounds and identify cases of structurally dissimilar compounds that affect similar BPs (Figure 4.4A and 4.4B). Thus, the employed GCNN architecture shows promise towards this highly interdisciplinary task. However, there were some cases of compounds affecting similar BPs that were missed by the model. These cases, in combination with the decrease in performance as the minimum structural distance between test and training compounds increases highlight key limitations in our approach (Table 4.4). On this front, limited coverage of the chemical space by compounds with available GEx data is a major issue that limits our ability to model in its entirety the complex function that translates changes in chemical structure to BP alterations. Even though each training set for each cell line contains on average around 320K samples, these are comprised from the pairing of around 800 compounds. The limitations that arise from this low coverage of the chemical space can't be solved by changes in deep learning architecture and require more

training compounds and/or extra input information. On this front, we applied a data augmentation technique, where each training set was augmented with randomly sampled pairs between Q1 and Q2 compound signatures (see SI 4.2). However, due to conflicting evidence between Q1 and Q2 transcriptomic signatures the performance of the models varied significantly across cell lines (see SI 5.3). A rather efficient workaround that we utilized in our approach is to quantify predictive uncertainty using deep ensembles. We showed that the model's performance, even when tested on compounds that are structurally different from the ones used in training, is higher for samples with lower uncertainty (Figure 4.5). Thus, the model's estimate of predictive uncertainty can be used to provide more reliable and accurate results. For instance, if an application imposes a constraint on the maximum allowed error, the appropriate uncertainty threshold can be identified and only point predictions with uncertainty lower than this threshold can be considered. Finally, we showed that transfer learning is a suitable approach to expand the biological coverage of deepSIBA to additional cellular models with fewer available data points (Table 4.3). For example, in the NPC cell line, which has approximately 50% fewer compound signatures than A375, deepSIBA was still able to achieve reasonable performance.

### 4.4.3 Signaling pathway inference for target structure

The predictions of deepSIBA can be used to infer a signaling pathway signature, in terms of the most upregulated and downregulated pathways, for a target chemical structure without available GEx data. The inference is performed following a KNN-like approach, in which reference compounds with the smallest distance to the target, as predicted by the model, are selected as its neighbors and their pathway signatures are retrieved. Then, pathways that frequently belong in the 10 most upregulated or downregulated pathways of the neighbors are inferred as the target's signature. The performance of the approach was evaluated on the test compounds of the MCF7 model and then, as a use case, it was tasked to infer the signaling pathways affected by FDA approved anticancer drugs, for which no GEx data are available in our dataset. Additionally, the chemical substructures that mostly influence the inferred pathways were identified and visualized using a graph gradient-based approach.

**4.4.3.1 Performance evaluation in the test set of MCF7**. For the test set of the MCF7 cell line, the average performance of the inference approach is presented in Table 4.5. On average 5 pathways per test compound were inferred to belong in its 10 most downregulated pathways with a precision of 73.3%. Regarding upregulation, an average of 2.5 pathways per compound with a precision of 69.7% were inferred. We have to note that the statistical significance of the inferred pathways is ensured by comparing the neighbor selection process using the trained model to a random selection.

Table 4.5 Pathway inference results for the test compounds of MCF7

| | Number of inferred pathways | Precision (%) |
|---|---|---|
| Downregulated | 5 | 73.3 |
| Upregulated | 2.5 | 69.7 |

**4.4.3.2 Use case: signaling pathway inference of FDA approved anticancer drugs.** Out of the 59 FDA-approved cytotoxic drugs presented by Sun *et al.*, 18 were present or had a structural analogue in the MCF7 training set (Tanimoto ECFP4 similarity > 0.85).[43] In order to simulate a realistic application for the signaling pathway inference, these 18 drugs were excluded from the use-case. From the remaining 39 drugs, only 3 had more than 5 neighbors each in the training set, as predicted by the model and the inferred pathways are presented in Table 4.6. Fludarabine and Clofarabine are direct nucleic acid synthesis inhibitors, while Pralatrexate is an indirect inhibitor of nucleotide synthesis through inhibition of the folate cycle.[44] In our use case, the inferred downregulated signaling pathways include cell cycle, purine and pyrimidine metabolism, RNA transport and spliceosome, which are closely related to the drugs' mechanism of action. Furthermore, because of the MCF7 cell line, pathways such as oocyte meiosis and progesterone-mediated oocyte maturation, that have been associated with the pathogenesis of breast cancer, were inferred as downregulated.[45] Regarding upregulation, pathways such as NF-kappa B signaling, natural killer cell mediated cytotoxicity, leukocyte transendothelial migration and TNF signaling, that are closely related to inflammation and apoptosis, were inferred.

Table 4.6 Pathway inference results for FDA approved anticancer drugs

| Drug | Mechanism of Action | Inferred Downregulated KEGG Signaling Pathways | Inferred Upregulated KEGG Signaling Pathways |
|---|---|---|---|
| Fludarabine | Nucleic Acid Synthesis Inhibitor | ***Purine metabolism, Pyrimidine metabolism***, **RNA transport**, **Spliceosome**, **Cell cycle**, **Oocyte meiosis**, **Progesterone-mediated oocyte maturation,** MicroRNAs in cancer | Leukocyte transendothelial migration, Oxytocin signaling pathway, Alzheimer's disease, Pertussis, Rheumatoid arthritis |
| Clofarabine | Nucleic Acid Synthesis Inhibitor | **RNA transport**, **Spliceosome**, **Cell cycle**, Ubiquitin mediated proteolysis, **Progesterone-mediated oocyte maturation**, MicroRNAs in cancer | **Natural killer cell mediated cytotoxicity**, Leukocyte transendothelial migration, Oxytocin signaling pathway, Pertussis, Rheumatoid arthritis |

| | | | |
|---|---|---|---|
| Pralatrexate | Inhibits dihydrofolate reductase (DHFR) and thymidylate synthase | **Purine metabolism, Pyrimidine metabolism**, Metabolic pathways, **RNA transport, Spliceosome** | **NF-kappa B signaling pathway, Natural killer cell mediated cytotoxicity, TNF signaling pathway**, Leukocyte transendothelial migration |

**4.4.3.3 Substructure importance.** The method described in section 2.7 was used to highlight important substructures that deepSIBA pays attention to when inferring the pathway signature of each anticancer compound presented in the use case (Table 4.6) (Figure 4.6). In Figure 4.6, red colored atoms represent atoms for which the model exhibits large directional derivatives across all pairs of target and neighbor compounds. Such atoms that are closely connected in the target compound's molecular graph are identified as influential to the inferred pathway signature. As shown in Figure 4.6, for Fludarabine and Clofarabine, deepSIBA highlights the 2-Fluoroadenine and 2-Chloroadenine substructures as important respectively, while the model mostly focuses on the Pteridine structure when inferring the pathways affected by Pralatrexate.

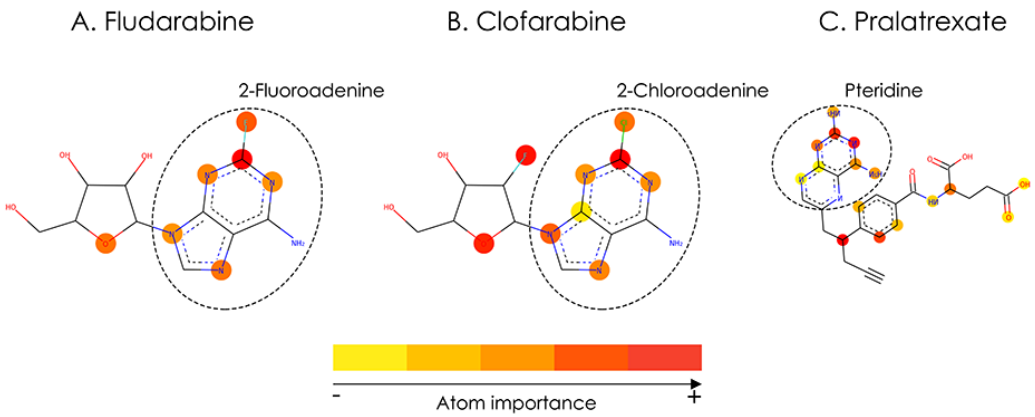



Figure 4.6 Important atoms related to the inferred biological footprint of the compounds of the use case, as identified by the deep learning model (the red color signifies the most important atoms)

In the presented use case, we demonstrated that by utilizing the training compounds as reference, the inferred signaling pathway signatures for each of the anticancer drugs were found to be closely connected to their respective MoA (Table 4.6). Thus, our inference method has the potential to provide an early estimate regarding the pathways affected by a compound, using only its chemical structure as input. Additionally, we showed that for each compound the highlighted substructures are also directly related to their respective MoA (Figure 4.6). This fact not only increases the interpretability of the model's predictions, which is a crucial topic of DL methods for drug discovery, but also shows that a GCNN model trained end-to-end on molecular graphs is able to learn meaningful structural representations that

are related to compounds' biological effects.[46–48] To the best of our knowledge, this is the first time a DL model was used to identify important substructures and infer the signaling pathway signature of a target compound without available experimental GEx data. A possible limitation of our approach might be its resolution capabilities in specific use-cases of compounds with similar chemical structure but different MoA. Although the comparison of the Fludarabine and Clofarabine use-cases suggests that our approach might be able to identify small structural differences between drugs with similar MoA (Figure 4.6), we haven't systematically compared use-cases of structurally similar compounds that affect different BPs. From the analysis of the CMap dataset we have showed that compounds with high structural similarity tend to have similar biological effect (see SI Figure S4.6). This lack of data regarding compounds that are derivatives but affect different BPs limits our ability to systematically perform the aforementioned comparison and pinpoint the maximum resolution of our approach. Furthermore, due to the nature of the inference method, limiting factors may also arise from the lack of diversity in affected BPs by the training compounds. This lack of diversity can influence the signaling pathway inference for an unknown target structure, when its true biological footprint is not represented in the reference compounds. In such cases, the inference of incorrect signatures can be avoided by focusing on target compounds with at least k reference neighbors (here k = 5) and only infer statistically significant pathways, using our method's calculated p-value.

# 4.5 Conclusion and availability

In this paper, we developed a deep learning framework to match the chemical structure of compound perturbations to their biological effect on specific cellular models. We showed, that the careful formulation of the learning problem and the flexibility of the Siamese GCNN architecture enabled our models to achieve high performance across all test scenarios. Additionally, we highlighted the utility of the uncertainty estimate, provided by deep ensembles, in test cases where the unknown chemical structures are very different from the structures used to train the models. Finally, we presented a novel inference pipeline, which can infer a signaling pathway signature for a target compound and subsequently identify which substructures mostly influenced the prediction. The novelty, performance and interpretability of our methods paves the way for further investigation in order to expand their coverage and utility.

Possible efforts for further investigation can be concentrated on the input representation, the biological response distance and the model's uncertainty estimate. Regarding the input, one interesting idea is to include binding information in order to capture the potential protein target of the input molecules. This extra information can be passed to the model either in the form of latent space embeddings from a trained binding affinity prediction model or in the form of predictions against a panel of protein kinases.[49] Regarding the biological distance between compound perturbations, this can be augmented by calculating the compound's effect on different levels of biological hierarchy, i.e. GEx, signaling pathways, transcription

factors and signaling networks.[50,51] Afterwards, these distances could be combined or separate models could be trained in order to better capture the similar effect of compounds. Additionally, instead of using a distance metric between all affected BPs, specific biological processes could be selected and application specific models could be developed to identify compounds that affect these biological processes. Regarding the model's uncertainty estimate, an interesting avenue for investigation is to take into account the transcriptomic signatures of replicates from the CMap dataset and calculate distributions of pairwise distances between compounds. Then, models could be trained on these distributions to better capture the variation of the experimental ground truth. Finally, collecting more data regarding derivative compounds with different MoA is an interesting avenue for further investigation in order to identify the resolution capabilities of the substructure importance approach.

The highly interdisciplinary framework of deepSIBA combines aspects from both the CADD and 'omics domains in order to incorporate the structural and systematic effects of small molecule perturbations, which are closely related to their efficacy and toxicity profiles. We believe that our methods have the potential to augment *in-silico* drug discovery, either by exploring on a massive scale the biological effect of compounds/libraries without available GEx data, or by suggesting new chemical structures with desired biological effect.

All analyzed data that were used to train our models and produce all tables and figures are available at https://github.com/BioSysLab/deepSIBA. Furthermore, the R source code to analyze the CMap dataset and create the training, validation and test sets is available at https://github.com/BioSysLab/deepSIBA/preprocessing. Finally, the Keras/TensorFlow implementation of our deep learning models, alongside trained ensemble models for each cell line are available at https://github.com/BioSysLab/deepSIBA/learning.

# 4.6 References

1 G. Sliwoski, S. Kothiwale, J. Meiler and E. W. Lowe, *Pharmacological reviews*, 2014, **66**, 334–395.

2 J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *Journal of Chemical Information and Modeling*, 2006, **46**, 462–470.

3 F. Sirci, F. Napolitano, S. Pisonero-Vaquero, D. Carrella, D. L. Medina and D. di Bernardo, *NPJ systems biology and applications*, 2017, **3**, 1–12.

4 J. P. Bai and D. R. Abernethy, *Annual review of pharmacology and toxicology*, 2013, **53**, 451–473.

5 C. Fotis, A. Antoranz, D. Hatziavramidis, T. Sakellaropoulos and L. G. Alexopoulos, *Drug discovery today*, 2018, **23**, 626–635.

6 B. Verbist, G. Klambauer, L. Vervoort, W. Talloen, Z. Shkedy, O. Thas, A. Bender, H. W. Göhlmann, S. Hochreiter and QSTAR Consortium, *Drug discovery today*, 2015, **20**, 505–513.

7 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri and A. Isacchi, *Proceedings of the National Academy of Sciences*, 2010, **107**, 14621–14626.

8 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian and K. N. Ross, *science*, 2006, **313**, 1929–1935.

9 A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli and J. K. Asiedu, *Cell*, 2017, **171**, 1437-1452. e17.

10 S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer and C. McNamee, *Nature reviews Drug discovery*, 2019, **18**, 41–58.

11 H. Chen, O. Engkvist, Y. Wang, M. Olivecrona and T. Blaschke, *Drug discovery today*, 2018, **23**, 1241–1250.

12 H. Öztürk, A. Özgür and E. Ozkirimli, *Bioinformatics*, 2018, **34**, i821–i829.

13 W. Jin, R. Barzilay and T. Jaakkola, *arXiv preprint arXiv:1802.04364*.

14 DeepChem: Deep-learning models for Drug Discovery and Quantum Chemistry.

15 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chemical science*, 2018, **9**, 513–530.

16 Y. LeCun, Y. Bengio and G. Hinton, *nature*, 2015, **521**, 436–444.

17 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, 2015.

18 W. Torng and R. B. Altman, *Journal of Chemical Information and Modeling*, 2019, **59**, 4131–4149.

19 S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *Journal of computer-aided molecular design*, 2016, **30**, 595–608.

20 I. Wallach and A. Heifets, *Journal of chemical information and modeling*, 2018, **58**, 916–932.

21 F. Schroff, D. Kalenichenko and J. Philbin, 2015, pp. 815–823.

22 L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. Torr, Springer, 2016, pp. 850–865.

23 O. Vinyals, C. Blundell, T. Lillicrap and D. Wierstra, 2016, pp. 3630–3638.

24 Y. Bai, H. Ding, Y. Sun and W. Wang, *arXiv preprint arXiv:1810.10866*.

25 H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, *ACS central science*, 2017, **3**, 283–293.

26 Y. Gal and Z. Ghahramani, 2016, pp. 1050–1059.

27 B. Lakshminarayanan, A. Pritzel and C. Blundell, 2017, pp. 6402–6413.

28 S. Jain, G. Liu, J. Mueller and D. Gifford, *arXiv preprint arXiv:1906.07380*.

29 A. Kendall and Y. Gal, 2017, pp. 5574–5584.

30 S. Ryu, Y. Kwon and W. Y. Kim, *Chemical Science*, 2019, **10**, 8438–8446.

31 M. Jeon, D. Park, J. Lee, H. Jeon, M. Ko, S. Kim, Y. Choi, A.-C. Tan and J. Kang, *Bioinformatics*, 2019, **35**, 5249–5256.

32 T. Pahikkala and A. Airola, *J. Mach. Learn. Res.*, 2016, **17**, 7803–7807.

33 K. Simonyan, A. Vedaldi and A. Zisserman, *arXiv:1312.6034 [cs]*.

34 GEO GSE92742, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742, (accessed 2020-03-20).

35 CLUE platform, https://clue.io/, (accessed 2020-03-20).

36 A. Alexa and J. Rahnenführer, *Bioconductor Improv*.

37 A. Sergushichev, *BioRxiv*, 2016, 060012.

38 F. Li, Y. Cao, L. Han, X. Cui, D. Xie, S. Wang and X. Bo, *Omics: a journal of integrative biology*, 2013, **17**, 116–118.

39 M. Kanehisa, Wiley Online Library, 2002, pp. 91–100.

40 RDKit: Open-source cheminformatics.

41 M. Iwata, R. Sawada, H. Iwata, M. Kotera and Y. Yamanishi, *Scientific reports*, 2017, **7**, 40164.

42 D. Rogers and M. Hahn, *Journal of chemical information and modeling*, 2010, **50**, 742–754.

43 J. Sun, Q. Wei, Y. Zhou, J. Wang, Q. Liu and H. Xu, *BMC systems biology*, 2017, **11**, 87.

44 O. Shuvalov, A. Petukhov, A. Daks, O. Fedorova, E. Vasileva and N. A. Barlev, *Oncotarget*, 2017, **8**, 23955.

45 D. Wu, B. Han, L. Guo and Z. Fan, *Journal of Obstetrics and Gynaecology*, 2016, **36**, 615–621.

46 P. Pope, S. Kolouri, M. Rostrami, C. Martin and H. Hoffmann, *arXiv:1812.00265 [cs, stat]*.

47 K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter and T. Unterthiner, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, eds. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen and K.-R. Müller, Springer International Publishing, Cham, 2019, pp. 331–345.

48 J. Jiménez-Luna, F. Grisoni and G. Schneider, *arXiv:2007.00523 [cs, stat]*.

49 A. Cichonska, B. Ravikumar, R. J. Allaway, S. Park, F. Wan, O. Isayev, S. Li, M. Mason, A. Lamb and Z. Tanoli, .

50 L. Garcia-Alonso, F. Iorio, A. Matchan, N. Fonseca, P. Jaaks, G. Peat, M. Pignatelli, F. Falcone, C. H. Benes and I. Dunham, *Cancer research*, 2018, **78**, 769–780.

51 A. Liu, P. Trairatphisan, E. Gjerga, A. Didangelos, J. Barratt and J. Saez-Rodriguez, *NPJ systems biology and applications*, 2019, **5**, 1–10.

# 4.7 Supplementary Information (SI)

## 1 Data preprocessing and quality control

**1.1 Dataset and quality overview.** The filtered CMap dataset contains 112994 transcriptomic signatures from 20254 compounds tested across 70 cell lines. During the filtering process, for each compound per cell line, its signature with the highest quality across different dosages and time points was selected. The assigned quality score based on TAS, number of replicates and whether the signature is considered an exemplar is presented in Table S4.1. The distribution of signature quality across cell lines is presented in Figure S4.1. Deep learning models were developed for the MCF7, PC3, VCAP and A375 cell lines, which have the highest number of compounds with Q1 signatures.

**Table S4.1** Signature quality score

| Quality score | TAS | Number of replicates | Exemplar |
|---|---|---|---|
| Q1 | > 0.4 | > 2 | True |
| Q2 | 0.2 – 0.4 | > 2 | True |
| Q3 | 0.2 – 0.4 | ≤ 2 | True |
| Q4 | 0.2 – 0.4 | > 2 | True |
| Q5 | 0.2 – 0.4 | ≤ 2 | True |
| Q6 | < 0.1 | > 2 | True |
| Q7 | < 0.1 | ≤ 2 | True |
| Q8 | < 0.1 | < 2 | False |

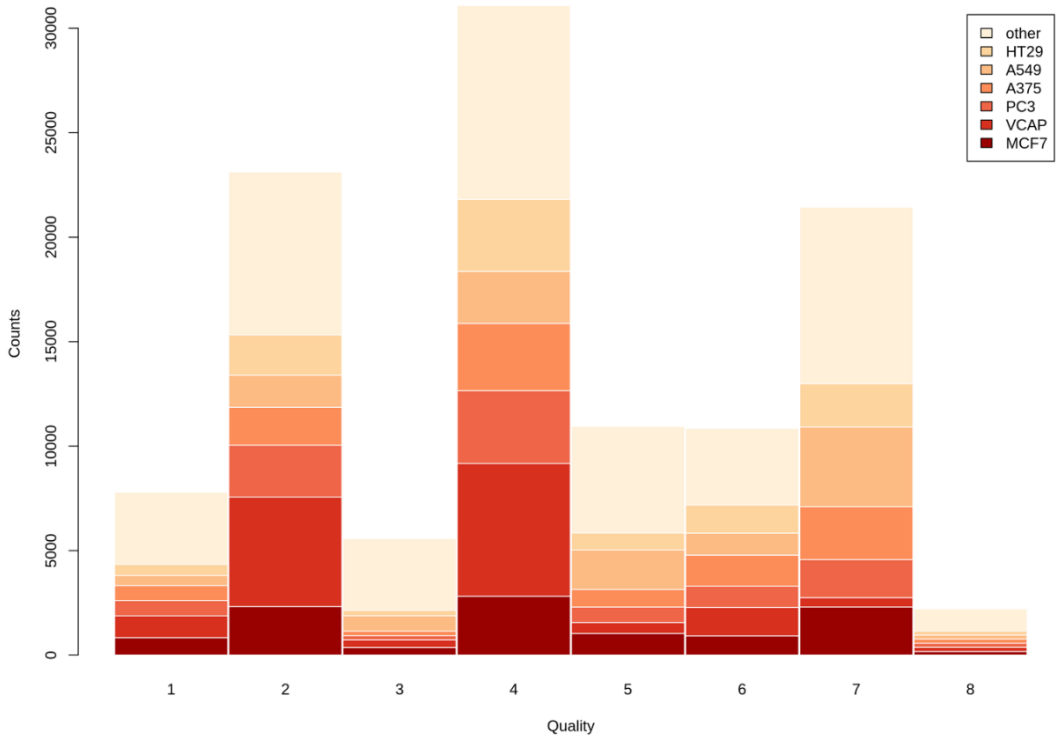



**Figure S4.1** Distribution of signature quality scores across cell lines. The "other" cell line category is formed by grouping together 63 cell lines with smaller number of available transcriptomic signatures.

---

**1.2 Distances between Q2 transcriptomic signatures across cell lines.**

As already discussed on the main paper, there is a significant number of compound pairs which have similar transcriptomic signatures in one cell line but not in the other (Figure S4.2). As it can be seen in Figure S4.2, this effect is much more prominent in the case of quality 2 (Q2) signatures.

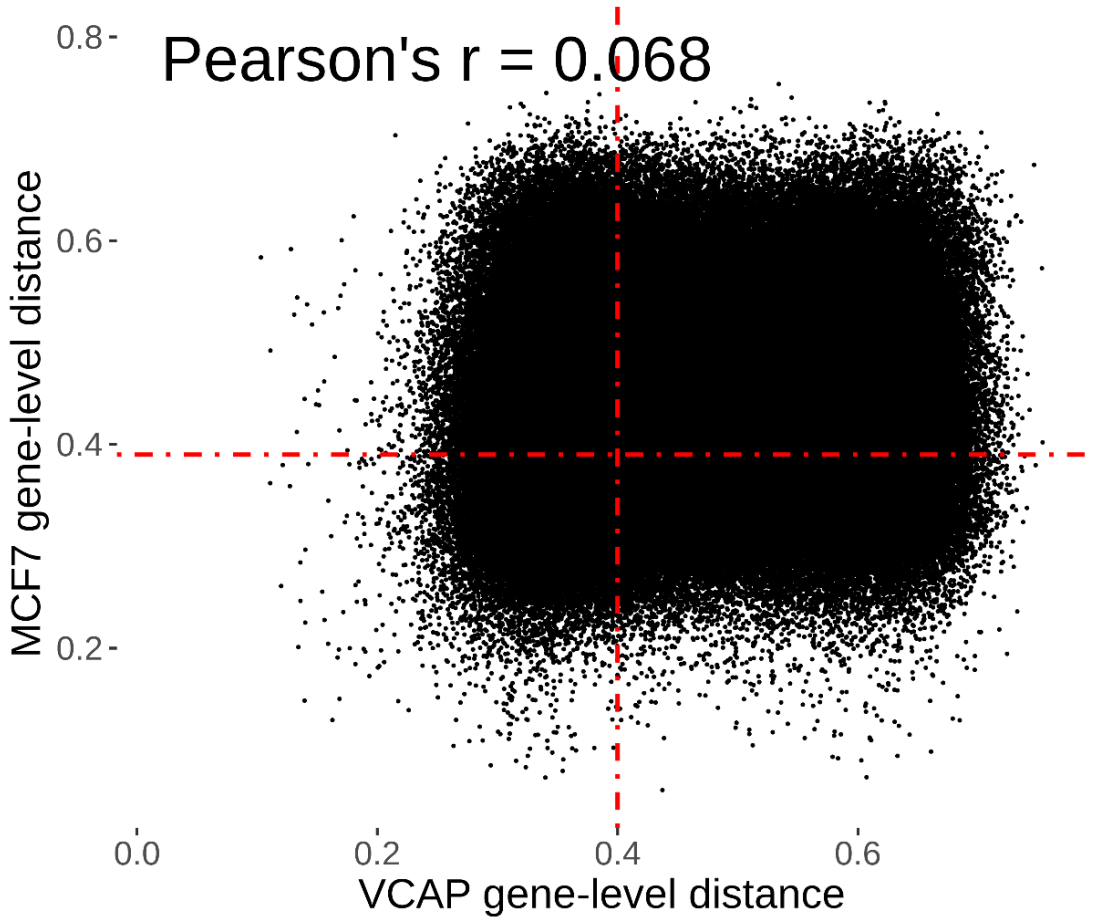



**Figure S4.2** Scatterplot of distances between Q2 transcriptomic signatures for the same compound pairs in the MCF7 and VCAP cell lines. Each point in the plot represents a pair of compounds with available transcriptomic signatures in both cell lines. The red lines, at 0.39 for MCF7 and 0.4 for VCAP indicate the mean + standard deviation of the distribution of distances between Q2 duplicate signatures for each respective cell line

---

**1.3 GO term enrichment and distance calculation.** For the MCF7, A375, VCAP and PC3 cell lines, the average number of significantly enriched GO terms in quality 1 signatures is presented in Table S4.2. Enrichment p-values were calculated with GSEA and adjusted using the Benjamini-Hochberg procedure. GO terms with an adjusted p-value less than 0.05 were considered significantly enriched. Based on Table S4.2 the number of top and bottom GO terms to consider during the ensemble distance calculation was selected (10, 20, 30, 40 and 50 GO terms). The ensemble distance approach outputs 5 distance scores for each signature pair, one for each of the numbers of top and bottom GO terms considered. The histogram of standard deviations of the calculated distances for each cell line is presented in Figure S4.3. The effect of the number of GO terms to consider during distance calculation is small, but not negligible. Furthermore, the relationship between pairwise distances between compounds at the GO term-level and at the gene-level was examined (Figure S4.4). Although distances are significantly correlated, the similar biological effect of chemical structures is better represented on a functional level between enriched GO terms. Finally, the ensemble distance approach of the GO term feature vectors was validated computationally. For the MCF7 cell

line, where enough quality 1 duplicate signatures are available (n = 20), their distance distribution was compared to a randomly selected subset of pairwise distances between different compound perturbations (Figure S4.5). It is clear that the proposed ensemble distance metric can easily separate duplicate compound perturbation pairs from random pairs.

**Table S4.2** Average number of significantly enriched GO terms following compound treatment

| Cell line | Average number of significant GO terms (p.adj < 0.05) |
|---|---|
| MCF7 | 20.1 |
| A375 | 29.8 |
| VCAP | 11.2 |
| PC3 | 30.0 |

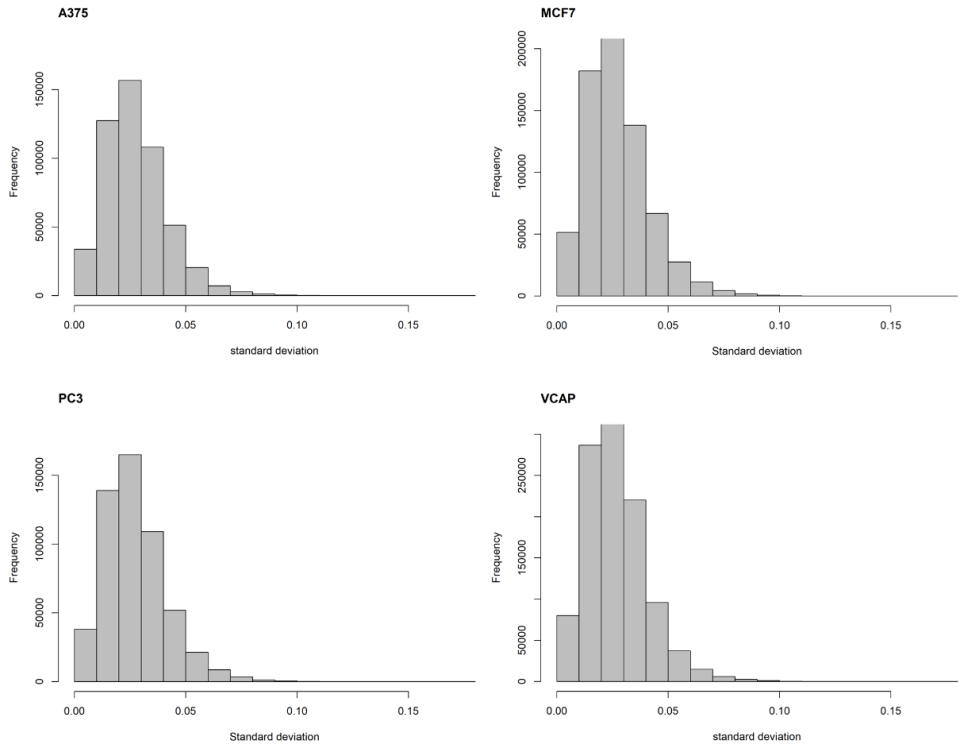



**Figure S4.3** Histograms of standard deviations of distances calculated between enriched GO terms for 5 different numbers of top and bottom GO terms (10, 20, 30, 40 and 50) for each cell line. Distances were calculated between compounds with Q1 signatures only.

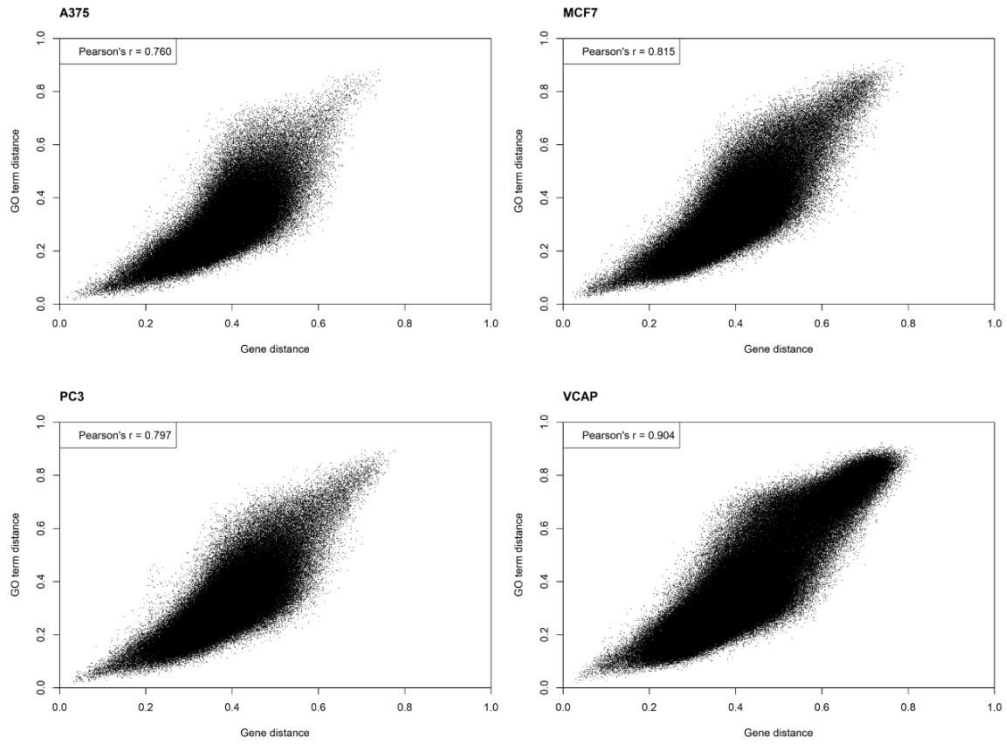



**Figure S4.4** Scatter plot of pairwise distances between compounds calculated at the gene and GO term-level for each cell line. Distances were calculated between compounds with Q1 signatures only.

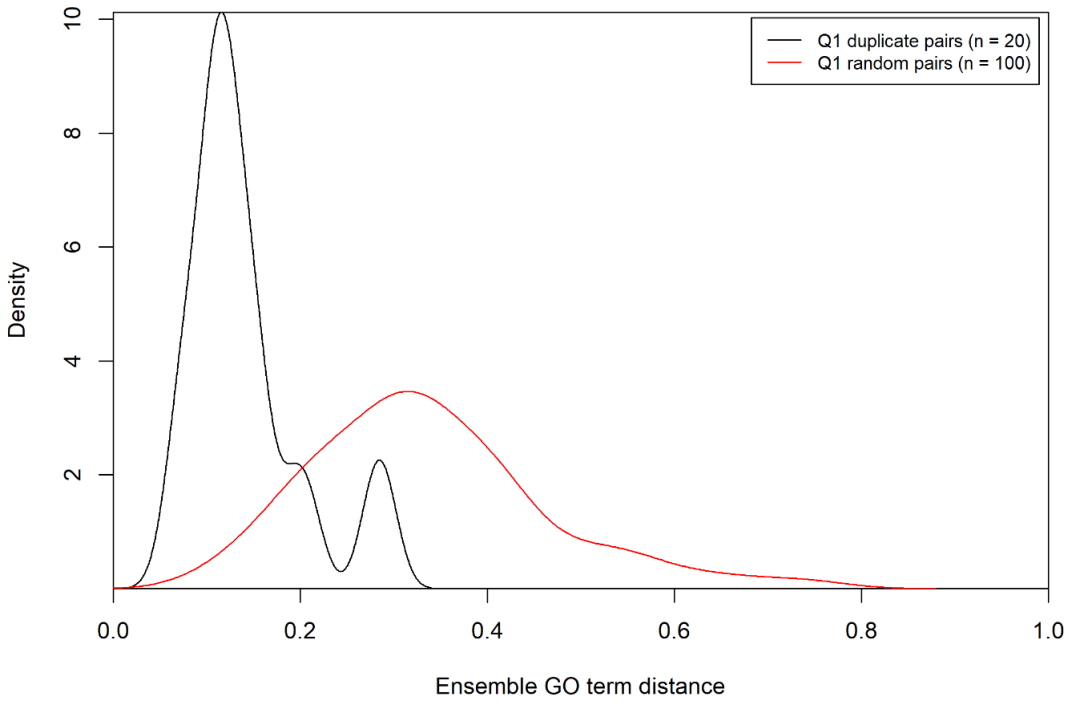



**Figure S4.5** Distribution of distances calculated with the ensemble GSEA score approach between compounds' affected BPs for the MCF7 cell line. The black line represents the distribution of pairwise distances between duplicate signatures, while the red line represents

the distances between signatures of random compound pairs. The separation between the two distributions indicates that the ensemble distance function can distinguish compounds that affect similar BPs (duplicates) from random compound pairs.

**1.4 Comparing structural and biological effect distance.** For each compound pair, Morgan circular fingerprints with radius 2 were generated using RDkit and their pairwise Tanimoto coefficient (Tc) was computed.[1] During fingerprint generation, the default atom invariants were used, making them similar to the widely used ECFP4.[2] Finally, pairwise compound distances were calculated, as $1 - Tc$. The relationship between pairwise compound structural distances and their distance in terms of affected biological processes (GO terms) in each cell line, was examined (Figure S4.6). We report similar results to Sirci *et al*.[3] and their analysis of transcriptomic and structural distances in the original CMAP dataset.[4] Indeed, compounds with similar ECFP4 fingerprints, tend to affect similar biological processes (lower left quadrant of Figure S4.6). However, there are many structurally dissimilar compounds that have similar biological footprint (upper left quadrant of Figure S4.6). Finally, the majority of compounds are structurally dissimilar and affect different biological processes (upper right quadrant of Figure S4.6).

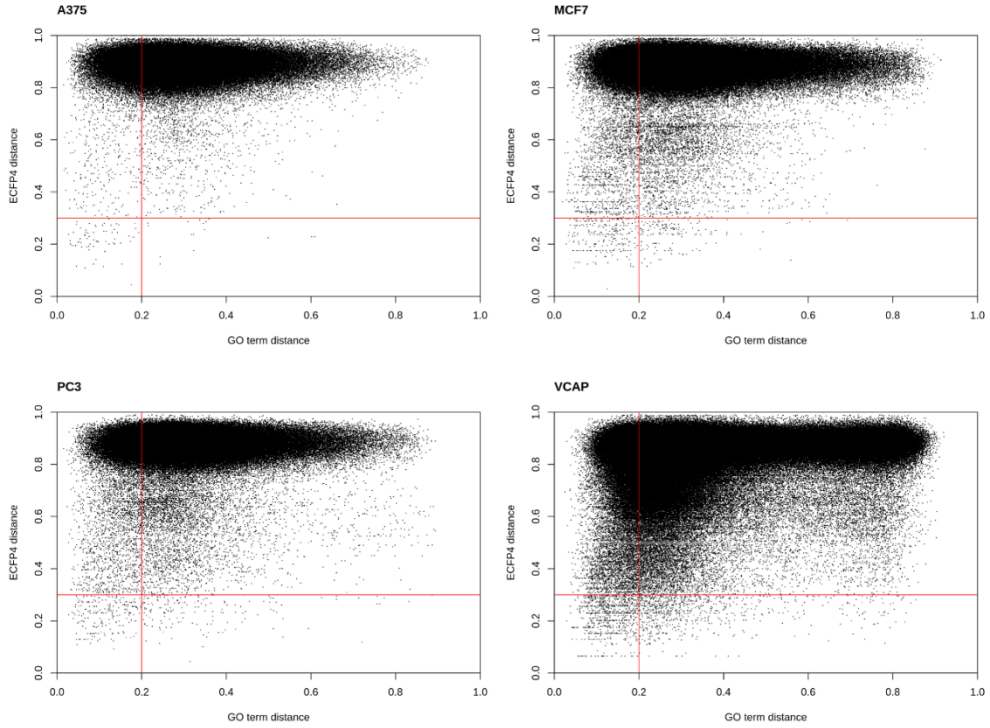



**Figure S4.6** Scatterplot of pairwise distances between compounds' ECFP4 fingerprints and between compounds' enriched BPs. The red lines represent reasonable thresholds to consider compounds similar in structure and similar in effect (0.3 for ECFP4 and 0.2 for BPs). Even though there is no correlation between compounds' structural and biological effect distances, the majority of structurally similar compounds tend to affect similar BPs (lower left quadrant of the plot).

## 2 Deep learning model

**2.1 Input representation.** Molecular graphs are presented to the model using the Atom array, the Edge array and the Bond array. The Atom array has as many rows as the max number of atoms across all compounds and each column represents an atom feature. In total, 62 atom features are utilized. The atom features consist of the concatenated vectors of 4 one hot encoded features and 1 binary feature, which describe:

- The symbol of the atom (one-hot).
- The degree of the atom (one-hot).
- The number of attached hydrogen atoms (one-hot).
- The valence of the atom (one-hot).
- If the atom is aromatic (binary).

The Edge array describes the connectivity of the graph representing the molecule. The Edge array consists of as many rows as the max number of atoms. Each row contains the atom's neighbors. The Bond array is 3-dimensional and contains the features of each bond. Each row represents an atom, while each column represents a neighbor, up to 5 for each atom. A bond is described by 6 binary features contained in the Bond array, which describe whether the bond is:

- Single
- Double
- Triple
- Aromatic
- Conjugated
- In a Ring

The Atom, Bond and Edge arrays were created using RDKit in python.

**2.2 Graph convolutions**. Graph convolutions were implemented in Keras as described by Duvenaud *et al*.[5] A graph convolutional layer aggregates information from the neighboring nodes of a node/atom in the molecular graph. For every atom, its bond features are summed and concatenated with its atom feature vector. The resulting feature vector of each atom is summed with the feature vectors of its neighbors, using the connectivity information of the Edge array, creating in this way a new feature vector for every atom with aggregated information from the atom's neighborhood. Then, every feature vector passes through a fully connected layer, based on the atom's degree, and a non-linear activation function. Typically, following a graph convolution layer, a function, such as sum, is used to aggregate node embeddings into whole graph embeddings. In our implementation we omitted the use of an aggregation function and instead utilized 1D convolutions to gather information across neighborhoods and produce a graph feature map.

**Graph Convolutional Layer Pseudocode:**

1: **Input:** Atom array $X_A$, Bond array $X_B$, Edge array *D*

2: **for** each atom $a_i$ in a molecule

3: $SX_{B_i}$=$\sum X_{B_i}$

4: $X'_{A_i} = concatenate(X_{A_i}, SX_{B_i})$

5: **for** each neighbor j from N neighbors

6: $SX_{B_j}$=$\sum X_{B_j}$

7: $X'_{A_j} = concatenate(X_{A_j}, SX_{B_j})$

8: $X''_{A_i} = X'_{A_i} + \sum_{j=1}^{N} X'_{A_j}$

9: $X_{A_i}^{new} = relu(W_{degree} * X''_{A_i} + b_{degree})$ #is the new concatenated atom and bond matrix

**2.3 Model hyperparameters.** The hyperparameters used to train the models are presented in Table S4.3. In our approach we utilized widely accepted hyperparameter values without performing hyperparameter optimization.

**Table S4.3** Model hyperparameters

| Optimizer | Adam |
|---|---|
| Learning Rate | 0.001 |
| Epochs | 20 |
| Batch size | 128 |
| Regularization | Dropout (rate = 0.3) |
| Batch Normalization Momentum | 0.6 |
| Weight Initializer | Glorot Normal |
| Activation Function | ReLu |

**2.4 Gaussian mixture.** By using a Gaussian regression layer, each model outputs a mean and variance of the biological effect distance between pairs of molecular graphs. The ensemble's output is also a Gaussian, with mean and variance calculated from the uniformly weighted mixture of each model. The mean and variance of the mixture are defined as

$$m_u = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i$$

$$sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(sd_i + \bar{y}_i^2\right) - m_u^2}$$

where,

- $\bar{y}_i$ is the output mean value of the biological effect distance of each model.

- $N$ is the number of models.
- $m_u$ is the final mean value of the uniformly weighted mixture.
- $sigma$ is the standard deviation of the uniformly weighted mixture.
- $sd_i$ is the output variance of the biological effect distance of each model.

Finally, the coefficient of variation of the Gaussian mixture is used as the model's estimate of predictive uncertainty and is defined as

$$CV = \frac{sigma}{m_u}.$$

## 3 Other similarity/distance learning methods

**3.1 ReSimNet.** Our approach was compared with a recently proposed architecture called ReSimNet.[6] ReSimNet takes as input the 2048-bit ECFP4 fingerprints of two chemical compounds and predicts their CMap score, which corresponds to the transcriptional response similarity of their GEx signatures. ReSimNet encodes the ECFP4 input to embedding vectors in the latent space using Siamese MLPs and predicts their CMap score as the cosine similarity of their embeddings. In our implementation, ReSimNet was trained to predict the similarity between compounds' affected BPs and afterwards the output similarity was transformed to a distance value for evaluation. The performance of randomly initialized ReSimNet ensemble models was evaluated for each cell line, on its respective test set (Table 4.2 of the main paper).

**3.2 TwoStepRLS.** In this study, TwoStepRLS, a Kronecker product kernel that utilizes a regularized least-squares (RLS) method, was used to predict the GO-term similarity of pairs of compounds, using as input the Tanimoto similarity between compounds' ECFP4 fingerprints. The regularization parameter was set to the proposed value of $2^{-15}$. This method was implemented using RLscore[7], an open-source python package for kernel-based machine learning, which includes implementations of RLS machine learning methods.

## 4 Dataset splitting and augmentation

**4.1 Dataset splitting.** An overview of the training and test sets for each cell line is presented in Table S4.4 and S4.5, while the distribution of the target variable is presented in Figure S4.7. For the proposed learning task, random splitting of compound pairs between training and test has no benefit, since if a compound is present on the training set its affected BPs are known and distances to other compounds can be calculated instead of predicted.

**Table S4.4** Cell line specific training sets

| Cell line | Number of Compounds | Number of Pairs |
|---|---|---|
| MCF7 | 713 | 253828 |
| PC3 | 608 | 184528 |
| A375 | 592 | 174936 |
| VCAP | 934 | 435711 |

**Table S4.5** Cell line specific test sets

| Cell line | Number of Compounds | Number of Pairs |
|---|---|---|
| MCF7 | 70 | 49910 |
| PC3 | 74 | 44992 |
| A375 | 77 | 45584 |
| VCAP | 63 | 58842 |

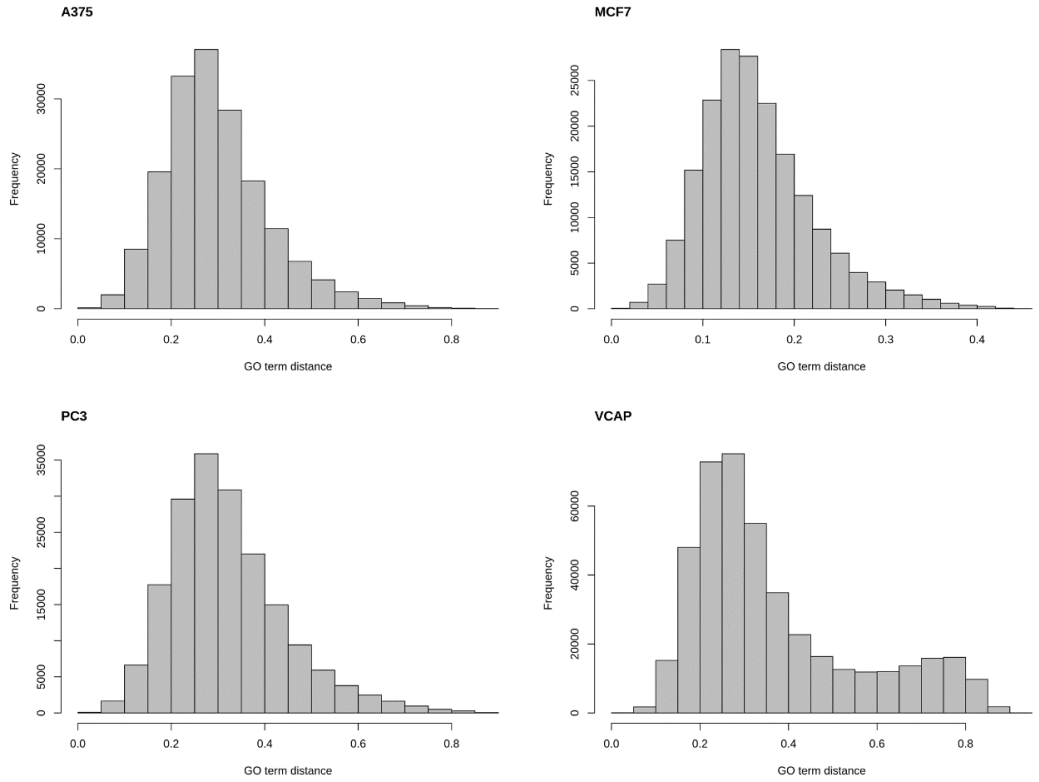



**Figure S4.7** Histogram of the model's target variable for each cell line.

**4.2 Data Augmentation.** When training augmented ensemble models, the original training set of each model, consisting of Q1 signature pairs, was augmented with randomly sampled pairs between Q1 and Q2 signatures. This technique was utilized in order to increase the number of compounds and the diversity of chemical structures available during training. Although this approach resulted in better performance for the MCF7 cell line compared to random initialization ensembles in terms of precision, it wasn't pursued further due to reliability issues of Q2 transcriptomic signatures. We observed many cases where the distance between Q1 signatures of compounds A and B was very small, e.g. 0.1, while the distance between signatures of compounds A and C, where C is a structural analogue of B (Tanimoto similarity > 0.85) and has a Q2 signature was high, e.g. 0.8. This kind of discrepancy between Q1 and Q2 signatures poses a problem for the learning model that only uses chemical structures as input.

## 5 Performance evaluation

**5.1 Distance threshold for precision.** The model outputs a continuous value between 0 and 1 for the distance between compounds' affected biological processes (GO terms). In order to evaluate the model's precision, a reasonable distance threshold has to be specified. Compounds with predicted distances below this threshold are considered similar in terms of affected biological processes. First, the connection between the distance threshold and the average number of common GO terms in the most upregulated and downregulated GO terms, respectively, for all compound pairs in the dataset was examined and is presented in Figure S4.8. The average number of common GO terms decreases linearly as the distance threshold increases (Figure S4.8). Additionally, for each compound per cell line, the distance threshold equivalent of a 90% Connectivity score was calculated (Figure S4.9). For a specific compound X, a threshold equivalent of a 90% score indicates that only 10% of other Touchstone compounds have a distance from X smaller that this threshold. The 90% CMAP score is a widely accepted threshold to identify compounds with similar transcriptomic signatures. Finally, for MCF7, for which enough quality 1 duplicate compound signatures are available, the distribution of pairwise distances between duplicate signatures is presented in Figure S4.5 (black line). Based on the information provided in Supplementary figures 8, 9 and 5, a threshold of 0.2 was selected when evaluating the models' precision across all cell lines. When calculating the model's precision on test compounds that exhibit maximum structural similarity to all training compounds less than 0.3, this threshold was adjusted to 0.22, because in this case no samples had a predicted distance less than 0.2.

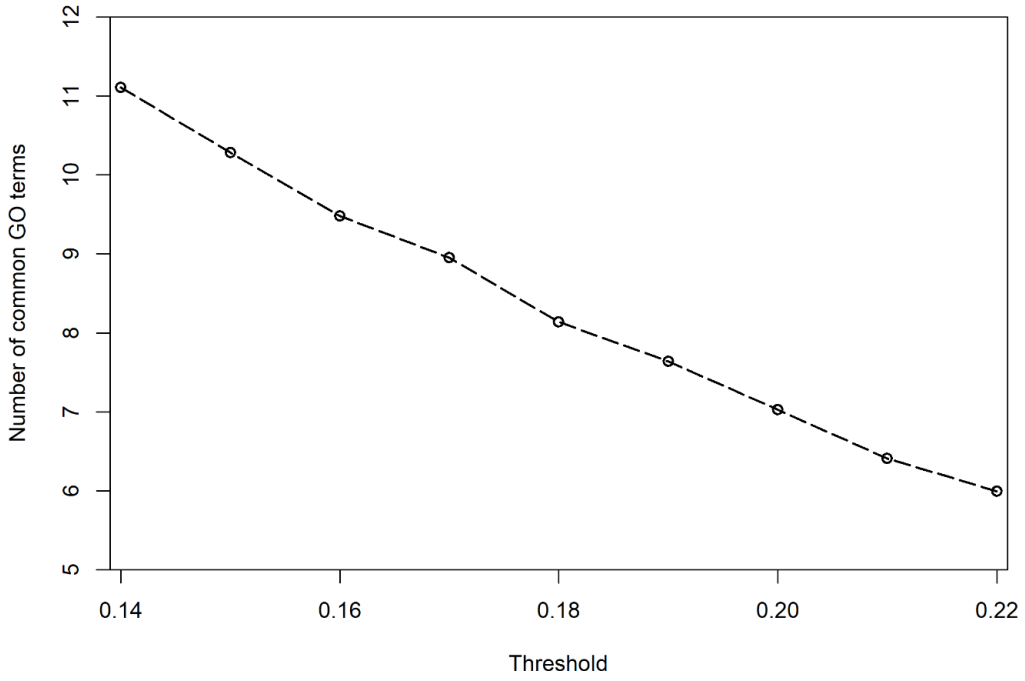



**Figure S4.8** The relationship between the biological effect distance threshold and the average number of common enriched BPs. In order to produce the above plot, signature pairs with GSEA distance below each threshold (x axis) are selected and the average number of common GO terms (BPs) in the 20 most upregulated and downregulated terms of all pairs is calculated (y axis).

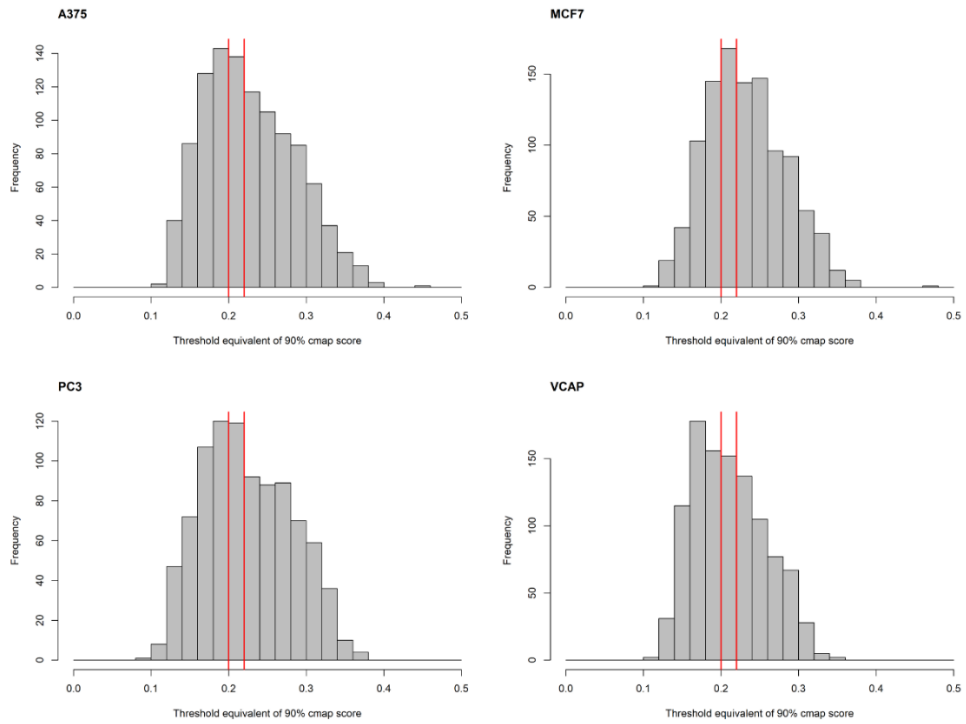



**Figure S4.9** Histogram of the threshold, which is equivalent to a 90% CMAP score, for all Touchstone compounds per cell line. The red vertical lines at 0.2 and 0.22 indicate the thresholds that were used to evaluate the models' precision. Across all Touchstone compounds, in all cell lines, the utilized thresholds are close to the mean of the threshold equivalent to a 90% CMAP score.

---

**5.2 Cross validation performance.** For each cell line, we evaluated the performance of a 10 model ensemble in a 5-fold cross validation split. Each validation set contains pairwise distances between BPs of non-overlapping sets of 80 validation compounds and all remaining training compounds. When extracting validation compounds, the maximum allowed Tanimoto similarity between ECFP4 fingerprints of validation and training compounds was set to 0.85. The results of the 5-fold cross validation are presented in Table S4.6. In all tested cell lines, our approach was able to produce consistently good results.

**Table S4.6** Cross validation performance of deepSIBA

| Cell-line | MSE | MSE @1% | Pearson's r | Precision (%) |
|---|---|---|---|---|
| A375 | **0.008** | **0.005** | **0.56** | 92.11 |
| VCAP | 0.025 | **0.005** | 0.54 | 64.28 |
| PC3 | 0.011 | 0.009 | 0.54 | **96.47** |
| MCF7 | 0.013 | 0.009 | 0.52 | 58.94 |

**5.3 Augmented deepSIBA performance evaluation.**

**Table S4.7** Cell line specific test set performance of augmented deepSIBA

| Cell-line | MSE | Pearson's r | Precision (%) | Predicted similar pairs |
|---|---|---|---|---|
| A375 | 0.009 | 0.61 | 91.54 | 272 |
| VCAP | 0.030 | 0.42 | 84.78 | 46 |
| PC3 | 0.011 | 0.54 | 25.97 | 77 |
| MCF7 | 0.015 | 0.45 | 88.00 | 25 |

## 6 Signaling pathway inference for target structure

**6.1 Parameter selection.** The most important parameters of the signaling pathway inference are the distance threshold $d_{th}$ and the frequency threshold $f_{th}$. Training set compounds with predicted distances from the target less than $d_{th}$ are selected as its neighbors, while pathways that appear in the neighbors' signatures with frequency higher than$f_{th}$ are inferred as the target's signature. The performance of the inference method was evaluated for different values of $d_{th}$ and $f_{th}$ in the test set of MCF7 (Figure S4.10). The average precision of the inferred pathway signatures as well as their length were chosen as evaluation metrics. The performance of the method decreases as $d_{th}$ increases and $f_{th}$ is kept constant at 0.65 for both the upregulated and downregulated signatures. In terms of precision, as $d_{th}$ increases the precision of the approach decreases, and for $d_{th} > 0.4$ it becomes 0 (Figure S4.10A and S4.10B). In terms of the length (average number) of the inferred signatures, for $d_{th}$ higher than 0.4, the length of inferred signatures becomes 0, as more distant compounds are considered neighbors (Figure S4.10C and S4.10D). After selecting 0.65 as a reasonable

threshold for $d_{th}$, we evaluated the performance of the approach for different frequency thresholds $f_{th}$, the results are presented in Figure S11. As $f_{th}$ increases, while $d_{th}$ is kept constant at 0.2 the inference becomes more strict. This results in increased precision (Figure S4.11A and S4.11B), but shorter in length inferred pathway signatures (Figure S4.11C and S4.11D). Based on these results, the selected parameters of the pathway inference for the MCF7 cell line and the respective use case, are presented in Table S4.8.

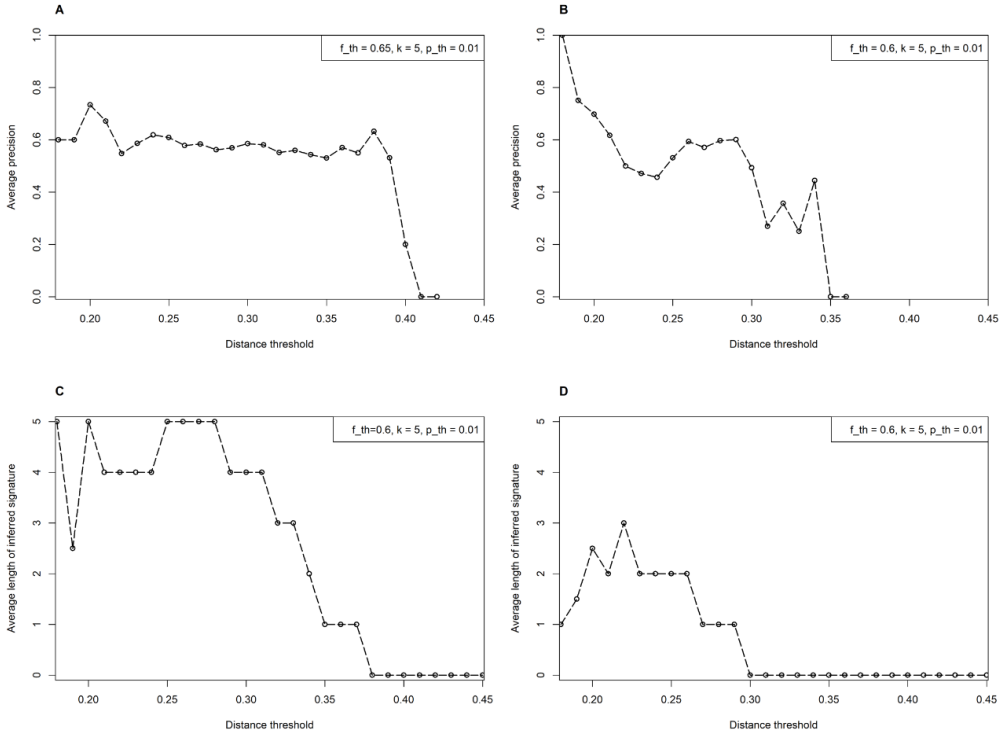



**Figure S4.10** Performance evaluation of the signaling pathway inference for different distance thresholds for the test compounds of the MCF7 cell line. Reference compounds (training) with predicted distance lower than the threshold are selected as the target's neighbors. The rest of the inference parameters are kept constant and their values are presented in the legend. (A) The average precision of the downregulated pathway signature as a function of the distance threshold; (B) The average precision of the upregulated pathway signature as a function of the distance threshold; (C) The average length of the inferred downregulated pathway signature; (D) The average length of the inferred upregulated pathway signature.

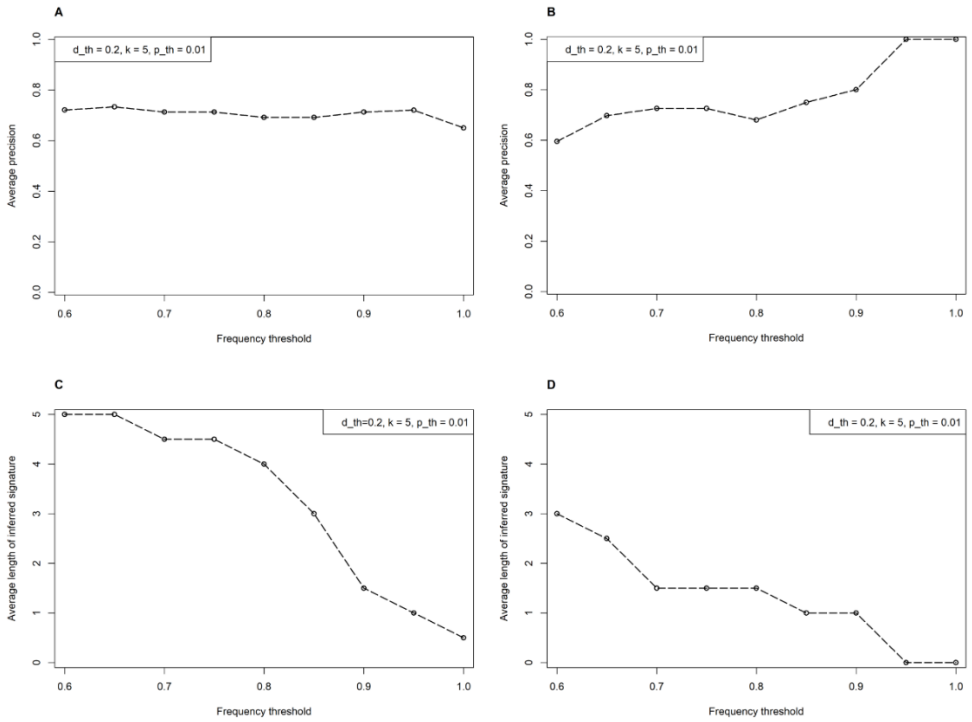



**Figure S4.11** Performance evaluation of the signaling pathway inference for different frequency thresholds for the test compounds of the MCF7 cell line. Signaling pathways that appear in the neighbors' signatures with frequency higher than $f_{th}$ are inferred as the target's signature. The rest of the inference parameters are kept constant and their values are presented in the legend. (A) The average precision of the downregulated pathway signature as a function of the distance threshold; (B) The average precision of the upregulated pathway signature as a function of the distance threshold; (C) The average length of the inferred downregulated pathway signature; (D) The average length of the inferred upregulated pathway signature.

**Table S4.8** Parameter values for the signaling pathway inference approach

| Parameter | Value |
|---|---|
| Distance threshold $d_{th}$ | 0.2 |
| Number of neighbors $k$ | 5 |
| Frequency threshold $f_{th}$ | 0.65 |
| P-value threshold $p_{th}$ | 0.01 |

# 7 References

1 RDKit: Open-source cheminformatics, http://www.rdkit.org/, (accessed 2020-03-20).

2 D. Rogers and M. Hahn, Journal of chemical information and modeling, 2010, 50, 742–754.

3 F. Sirci, F. Napolitano, S. Pisonero-Vaquero, D. Carrella, D. L. Medina and D. di Bernardo, NPJ systems biology and applications, 2017, 3, 1–12.

4 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian and K. N. Ross, science, 2006, 313, 1929–1935.

5 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, 2015.

6 M. Jeon, D. Park, J. Lee, H. Jeon, M. Ko, S. Kim, Y. Choi, A.-C. Tan and J. Kang, Bioinformatics, 2019, 35, 5249–5256.

7 T. Pahikkala and A. Airola, J. Mach. Learn. Res., 2016, 17, 7803–7807.

# Chapter 5

# Concluding remarks

## 5.1 Conclusion

In this thesis I developed and tested deep learning models for the field of systems pharmacology. The deepSIBA and deepSNEM pipelines serve as proof of concept that deep learning can provide a framework to successfully incorporate elements from both the structural and systems domain of drug discovery in order to reduce its attrition rates. I hope that the created datasets and methods can pave the way for more research in the field of deep learning for systems pharmacology. I also believe that as more data become available, the deep learning applications for systems pharmacology will become increasingly useful and find real world applications in the field of drug discovery. However, we have to be mindful regarding where further research in the field should be focused to. From my experience, the most important aspect of any developed pipeline is the problem statement, along with the data/features and preprocessing steps, rather than the deep learning method. During my research, most of my time was spent on developing the learning problem that can test the hypothesis, along with finding the right data to train the system. Building a deep learning pipeline based on experimental data, for tasks that cannot be solved by humans is a very complex problem. Given the complexity of the problem, I believe that the collaborative effort of machine learning scientists, systems scientists and biologists is paramount for the success of the field.

## 5.2 Data and code availability

All the datasets and code that were created as part of my research is available at the github page of the lab https://github.com/biosyslab.