# A Network Polypharmacology Approach to Drug Repurposing for Diffuse Intrinsic Pontine Glioma

Finlay MacLean[1*,4†] Javad Nazarian[2] Pan Pantziarka[3] Jabe Wilson[1]

[1]Elsevier[‡]

[2]University Children's Hospital Zurich[‡]

[3]Anticancer Fund[‡]

[4]BenevolentAI

[*]work performed at; [†]current affiliation; [‡]active partners in collaboration

## Abstract

Despite five decades of clinical investigations, there is currently no effective treatment for children diagnosed with Diffuse Intrinsic Pontine Glioma (DIPG). We now understand that DIPGs share the same histone 3 mutation and fatal prognosis as other diffuse midline gliomas (DMGs), which led to the introduction of a new entity referred to as DMG, H3 K27M mutant. Indeed, therapeutics indicated for other brain neoplasms have proven ineffective for DIPGs. We posit that by using a polypharmacological approach to determine drug combinations that target distinct mechanistic pathways of DIPG, it is more likely that an efficacious treatment will be developed. Using FDA-approved drugs as an indication for viability of a drug pair for combinatorial use, we developed an embedding-based link-prediction model trained on a drug-disease regulation network, to predict the complementary relationship between a drug pair. To further understand the regulatory mechanisms of these potential drug combinations, we explored the regulatory action of these drugs upon statistically significant biological processes and molecular functions that are in turn misregulated in the disease state of phylogenetically similar gliomas to DIPG. Next, we constructed a regulatory network of these regulators illustrating a negative correlation between network proximity and overlap of direct targets, regulators, pathways and biological processes. We replicate research that finds an optimal distance between drugs in this regulatory network for both FDA-approved and synergistic drugs, although we deem the embedding model more performant. Finally, we highlight the individual importance of all disease-host protein regulators in notable pathways, by constructing protein-protein interaction networks for each pathway and calculating the centrality of each regulatory gene. We believe this newly formed knowledge of drug pairs and their regulatory pathways, will help researchers in identifying an efficacious treatment for this disease.

# Introduction

To date, there is no approved and satisfactory means of treatment for DIPG. Repurposed drugs hold great promise as potential treatments; namely due to their established safety profiles and manufacturing routes, minimising time to use these drugs at the point-of-care. Whilst there have been over 200 clinical trials for DIPG,[1] none have shown great clinical promise. This is in part due to a lack of understanding of the pathophysiology of the disease, lack of effective drugs targeting tumorigenic pathways, and difficulties in drug delivery due to the blood brain barrier.

Systems pharmacology and network medicine (NM) approaches to drug discovery and drug repurposing have proved efficient methods to highlight potential drug candidates.[2,3,4] NM treats biological networks as heterogeneous information systems; correlating network topology and node properties with biological processes, functions, pathways and interactions. From a systems biology point of view, a disease can be seen as a selection of genes within a network, whose misregulation culminates in changes in biological processes and pathways. Similarly, drugs can be modelled by their drug targets, and the propagatory effect their perturbance has upon the network. The aim of NM is to identify drugs and diseases in which the network perturbance of the disease state is reversed by the perturbance of the drug.

It is highly largely improbable that a drug can be developed with the regulatory profile of a drug that exactly matches the genetic (or network) profile of a disease. It is even more improbable suggest a drug, designed for another indication, provides an exact match. Polypharmacology, the design or use of pharmaceutical agents that act on multiple targets or disease pathways, can be utilised to overcome this. Multiple drugs can be combined that each exhibit some overlap with differing subsets of the diseases regulatory genes. It has been shown that combinatorial drugs exist within the 'Goldilocks' zone in gene interaction networks.[5] Within the network, drugs must be sufficiently close to the disease as to illicit a change in gene function, but sufficiently distant from the other drugs to ensure there are no antagonistic effects between drugs. The above study showed that network proximity is correlated with many similarities between drugs, such as gene ontology terms, pathways, clinical, genetic and structural similarity. Ultimately drugs with larger network proximity will exhibit a larger disparity in terms of mechanism of action. Polypharmacology has seen success in many disease areas such as other cancers and viruses, most notably the human immunodeficiency viruses. We posit that a network polypharmacology approach could provide benefit in identifying treatments for DIPGs (see Figure 1).

The interactome (the totality of interactions within a cell), is largely unknown and fascinatingly complex. Drug target interactions constitute much less than one percent of small-molecules reported to bind to a protein. Drug regulatory interactions with disease follow a similar distribution. Our sparsity in known information holds promise that there are novel indications for current drugs; through the reapplication of drugs to new diseases (through known drug targets), or through novel drug targets.

In recent years, graph-based machine learning (GML) methods have been applied to the task of link prediction, to systemically 'fill in' these unreported interactions. Whilst many GML methods exists, the most intuitive is graph neural networks. In contrast to conventional neural networks which use arbitrary model architectures, GML models explicitly replicate the network relating to their prediction task. For example, using the aforementioned disease-gene network, and tasked to predict novel gene regulators of disease, genes which are functionally or physically similar are more

highly connected within the biological network and by extension more connected in the model architecture. Such genes will have greater influence on each other during the training process of the model. One of the applications of GML is the creation of node embeddings: transforming continuous high dimensional information of node neighbourhood to low-dimensional dense vectors, to be used in downstream machine learning tasks. Popular approaches include generating embeddings by sampling a network via random walks, treating the sample as corpus of words, and applying natural language processing (NLP) techniques. Similar NLP methods have been applied to amino acid sequences of proteins,[6] and molecular substructures of compounds[7] to generate embeddings representing primary structure and molecular substructure of proteins and compounds respectively. There is not an abundance of biological information pertaining to DIPG. For this work we use Diffuse Midline Gliomas (DMGs) as proxy diseases, with the belief that drugs that could regulate these would have a similar inhibitory effect on DIPG.

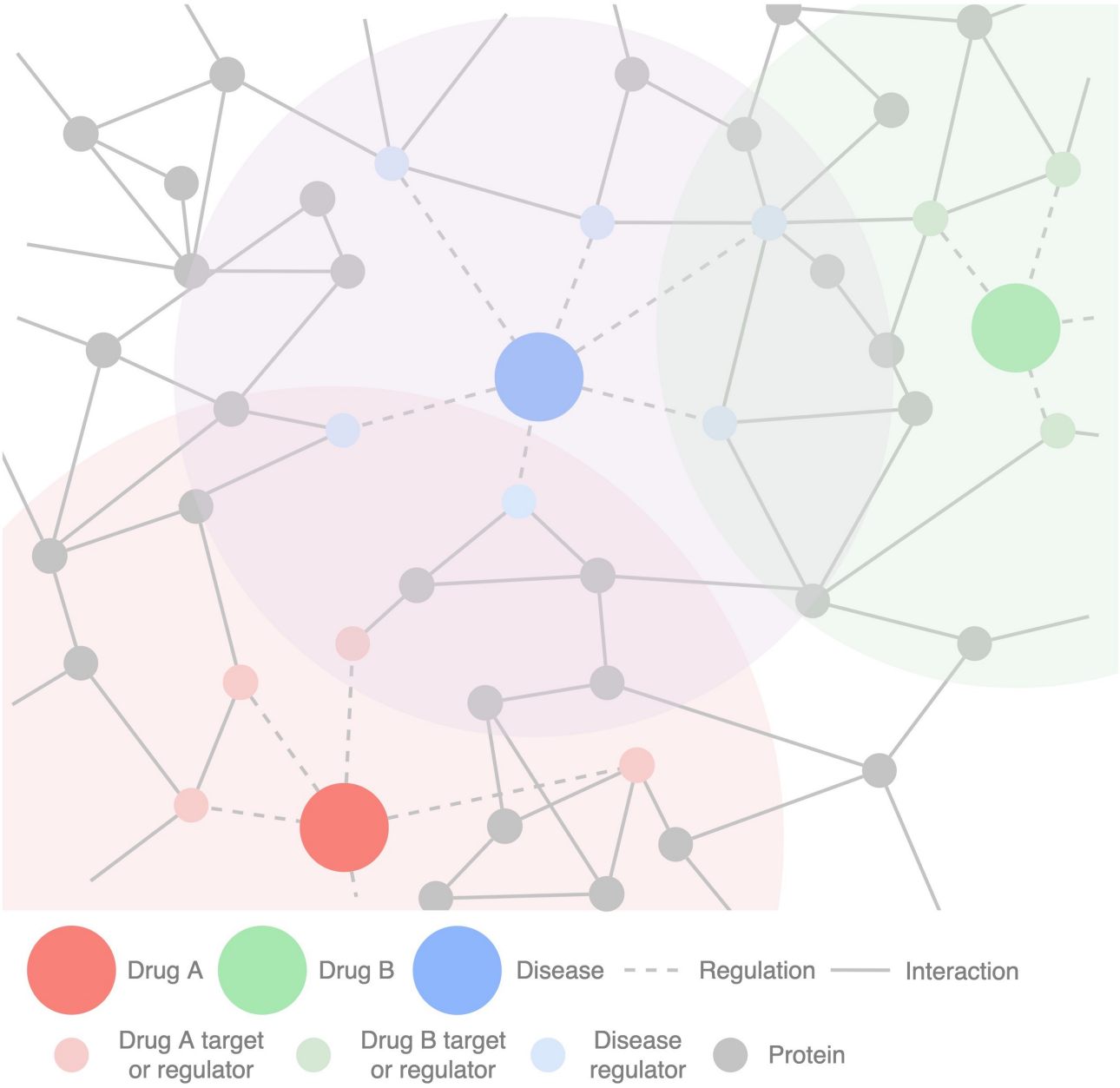


Figure 1: Polypharmacological drug repurposing strategy for DIPG. Illustration displays a tripartite regulatory network of diseases, gene-coded proteins and drugs. Edges between proteins represent direct protein modifications (eg. phosphorylation or cleavage). Edges between disease and proteins represent indirect regulation via any mechanism. Edges between drugs and proteins represent regulation via any direct or indirect mechanism. Drugs are shown to have circle of influence, centered around their drug targets and regulators. This diagram illustrates the ideal drug pair for treatment of DIPG. The two drugs do not share any drug targets or regulators, and their only overlap is genes not thought to regulate the disease.

In this study we present a network medicine approach to polypharmacological drug repurposing for DIPG. We create a biological knowledge graph predominantly based on literature, and develop a hybrid link prediction model using both network topology and physicochemical properties via multiple embedding methods. We use this model to predict down-regulators of various DMGs. Using the top predictions, we highlight known inhibitors and activators of autophagy, and rank these according to their predicted regulatory probability. Next, we use FDA-approved drug combinations to train a link-prediction model on a disease-drug regulatory network, to predict complementary combinations of these predicted disease regulators. To understand the regulatory actions of these drug pairs, we first establish biological processes and molecular functions misregulated in the disease state, and cross-reference these to processes and functions regulated by the candidate therapeutics. Next, we develop a local regulatory subgraph of these top molecules and DMGs, to determine the key gene regulators and their individual importance in the most enriched pathways of DIPG. To validate this subgraph and determine the most informative paths to predict therapeutic efficacy, we use a degree-weighted path count model. We show biological paths representative of a drugs mechanism of action also provide useful information. Finally, we posit that distance between drugs in the regulatory subgraph is correlated with similarities between nodes, such as shared biological process, drug targets and regulators, and pathways. We suggest distance between drugs in a regulatory graph, and information of regulation of DMGs for each drug, can be used to help identify combinatorial repurposing candidates for DIPG. We find a significant relationship between 1-dimensional network proximity and viability of drug pairs for combinatorial use in terms of both synergy and FDA approval. Whilst not as powerful as the embedding model to predict FDA approval, this method proves a valuable indicator of pharmacological and functional disparity in drug combinations.

# Results

## Drug Regulators

### Knowledge Graph

To suggest repurposable drugs candidates for DIPG, we must first understand the underlying mechanisms of oncogenesis. To suggest combinatorial drug candidates, we must further understand the biochemical interaction between the therapeutics and assess the synergism of the relationship. From a systems biology point of view, to model this we need to create a network capable of capturing these biochemical and regulatory interactions. To achieve this, we created a knowledge graph predominantly based on Pathway Studio, a biomedical database derived from relationships extracted from over 30 million academic manuscripts (see Methods). Each edge in the graph represents at least one occurrence in scientific literature that has stated a relationship between biological entities (for example that anisodamine down-regulates DIPG). In total, the graph possessed 1.36 million nodes and 8.52 million edges (weighted according to their frequency in literature).

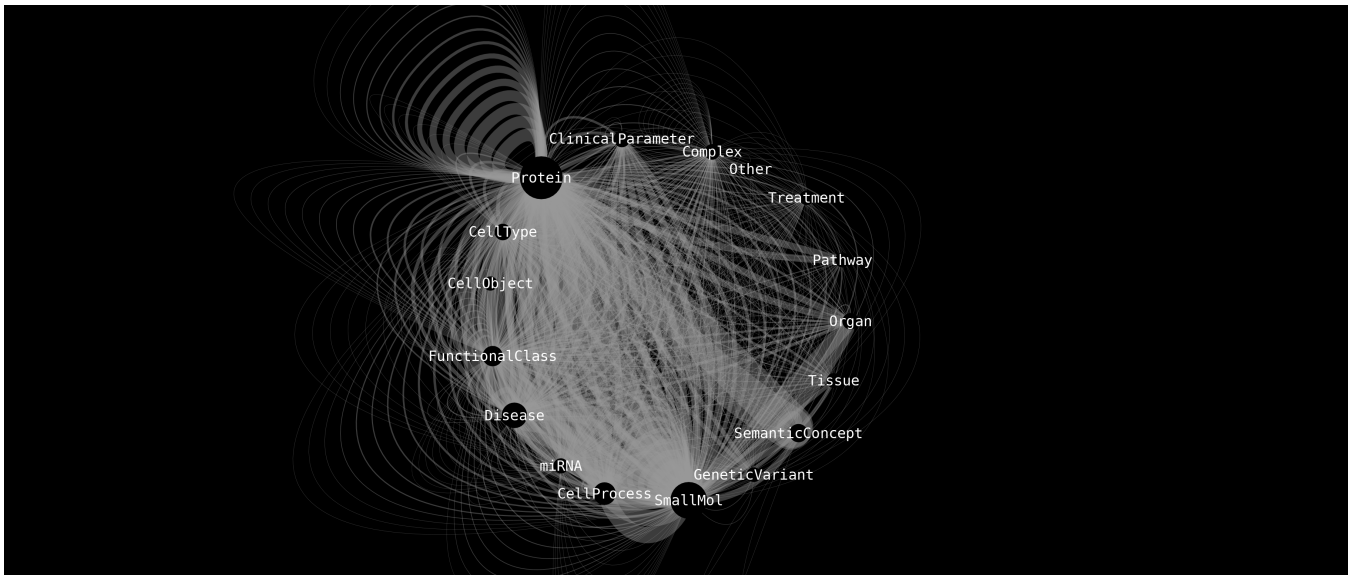


Figure 2: Knowledge graph visualisation using Cytoscape. Size of node represents number of nodes of this type in the graph. SmallMol (molecule), Protein and Disease are the largest, alongside SemanticConcept (gene ontology).

## Drug Regulation Prediction

To ascertain the therapeutic efficacy of drugs for treatment of DIPG, we developed an embedding-based link prediction model capable of predicting if a relationship between two nodes should exist. As our graph is based on occurrence of biological relationships in literature, a predicted link with high probability for the edge RegulationNegative between DIPG and Wee1 indicates that if a research group were to research the impact of inhibitory regulation of the tyrosine kinase on the cancer, there is high likelihood the regulation was sufficient to state this regulatory relationship in an academic paper.

Our model used node embeddings to create compressed discrete 100-dimensional representations of the neighbourhood of each node. After various evaluations, we chose to employ the state-of-the-art embedding method, GraRep,[8] due to the methods ability to capture both local and global neighbourhood information. In essence, GraRep calculates the singular value decomposition for direct neighbours of a node, and indirect neighbours up to a certain threshold distance. For a full explanation please see the original paper. For an illustrative explanation of the link prediction model please see the Methods section.

Alongside node information of neighbourhood, the model also utilised physicochemical structural information of proteins and molecules. We used Mol2Vec,[7] an unsupervised machine learning approach to learn vector representations of molecular substructures, to capture structural information of drugs. ProtVec[6] was used to represent the primary amino acid structure of proteins. For disease nodes, additional graph embeddings were generated based on two further networks - a disease-gene regulatory network and a disease-disease similarity network. The disease-disease similarity network was based on a proprietary disease ontology of Pathway Studio, to introduce the notion of disease semantic similarity to the model. A high-level overview of model performance can be seen below.

| Model | Source feature | Target feature | AUC | F1 score | Accuracy | Precision | Best Threshold | Best TPR | Best FPR |
|---|---|---|---|---|---|---|---|---|---|
| **Disease-RegulationNegative-Drug** | Disease-gene regulatory network | Mol2Vec | 0.954±0.002 | 0.888±0.003 | 0.888±0.003 | 0.844±0.004 | 0.628±0.018 | 0.891±0.007 | 0.115±0.007 |
| **Disease-RegulationNegative-Protein** | Disease-drug regulatory network | Prot2Vec | 0.936±0.002 | 0.868±0.004 | 0.87±0.004 | 0.826±0.004 | 0.543±0.004 | 0.857±0.007 | 0.116±0.003 |
| **Protein-BiomarkerDiagnostic-Disease** | Prot2Vec | Disease-drug regulatory network | 0.952±0.001 | 0.883±0.001 | 0.886±0.001 | 0.847±0.005 | 0.51±0.022 | 0.868±0.013 | 0.096±0.013 |
| **Protein-BiomarkerPrognostic-Disease** | Prot2Vec | Disease-drug regulatory network | 0.942±0.003 | 0.877±0.004 | 0.879±0.004 | 0.838±0.006 | 0.568±0.008 | 0.865±0.009 | 0.106±0.01 |
| **Protein-GeneticChangeGeneAmplification-Disease** | Mut2Vec | Disease-drug association network | 0.937±0.011 | 0.872±0.012 | 0.876±0.008 | 0.838±0.006 | 0.526±0.061 | 0.851±0.042 | 0.098±0.029 |
| **Protein-GeneticChangeGeneDeletion-Disease** | Mut2Vec | Disease-drug association network | 0.875±0.002 | 0.796±0.005 | 0.812±0.004 | 0.772±0.009 | 0.575±0.029 | 0.735±0.022 | 0.109±0.025 |
| **Protein-GeneticChangeMutation-Disease** | Mut2Vec | Disease-drug association network | 0.957±0.001 | 0.892±0.002 | 0.893±0.001 | 0.855±0.001 | 0.53±0.007 | 0.88±0.006 | 0.094±0.004 |

Figure 3: Model performance for all link prediction models. The $(Disease) - [RegulationNegative] \rightarrow (Drug)$ model was used in the above task to predict down-regulators of DMGs.

We applied our link prediction models to predict drugs that down-regulate multiple DMGs: diffuse intrinsic pontine glioma, glioblastoma, glioblastoma multiforme (a child disease of glioblastoma), anaplastic astrocytoma, gliomatosis cerebri and brainstem Neoplasm. For each prediction, we cross-referenced and highlighted known drug regulators reported in literature. The top 25 predictions for DIPG included 8 known down-regulators. We deemed the top 25 predicted regulators for DMGs as sufficiently probable, and took these forward for further analysis. We also appended this list with all known inhibitors. An important note is that the above diseases have vastly differing degrees of connectivity in the graph. Whilst glioblastoma has 716 connected down-regulators in the knowledge graph, brainstem neoplasm and DIPG have 9 and 28 respectively. Such differences in degree indicate different prior probabilities of treatment. In other words, nodes in a graph with higher connectivity are statistically more likely to be connected. This prior probability is reflected in the probability distributions. Glioblastoma, DIPG and brainstem glioma have a mean probability of 0.62, 0.18 and 0.07. As such, probabilities should not be directly compared between diseases, and instead used as an indication of relative position in ranked probability lists.

| Probability<br>Drug | Glioblastoma | Probability<br>Drug | GBM | Probability<br>Drug | Anaplastic Astrocytoma |
|---|---|---|---|---|---|
| fasudil | 0.99840* | 5-fluorouracil | 0.99480* | temozolomide | 0.92853* |
| nicotinamide | 0.99800* | carboplatin | 0.99360* | lomustine | 0.88520* |
| quercetin | 0.99800* | irinotecan | 0.99280* | procarbazine | 0.86240* |
| 5-fluorouracil | 0.99800* | gemcitabine | 0.99040* | carboplatin | 0.78530 |
| melatonin | 0.99800* | doxorubicin | 0.99000* | cisplatin | 0.77800 |
| captopril | 0.99760* | etoposide | 0.98880* | ifosfamide | 0.74327 |
| niacin | 0.99760 | lomustine | 0.98840* | doxorubicin | 0.73160 |
| glutathione | 0.99720* | camptothecin | 0.98840* | irinotecan | 0.72940 |
| dopamine | 0.99720* | sunitinib | 0.98800* | BCNU | 0.72880 |
| thalidomide | 0.99680* | temozolomide | 0.98720* | capecitabine | 0.72607 |
| dasatinib | 0.99640* | retinoic acid | 0.98680* | dacarbazine | 0.72120 |
| D-penicillamine | 0.99600* | lapatinib | 0.98640* | pazopanib | 0.71887 |
| suberoylanilide hydroxamic acid | 0.99600* | vincristine | 0.98600* | gemcitabine | 0.71333 |
| capsaicin | 0.99600 | sorafenib | 0.98480* | topotecan | 0.70530 |
| hydroxychloroquine | 0.99560* | mitoxantrone | 0.98480* | crizotinib | 0.70435 |
| ellagic acid | 0.99560* | capecitabine | 0.98120 | camptothecin | 0.70223 |
| pazopanib | 0.99560* | chloroquine | 0.98120* | etoposide | 0.69760 |
| valproic acid | 0.99560* | teniposide | 0.98000* | bleomycin | 0.68840 |
| troglitazone | 0.99520* | imatinib | 0.98000* | 5-fluorouracil | 0.68743 |
| caffeine | 0.99520* | suberoylanilide hydroxamic acid | 0.97920* | vincristine | 0.68547* |
| indomethacin | 0.99480 | topotecan | 0.97840 | oxaliplatin | 0.68460 |
| temozolomide | 0.99480* | vandetanib | 0.97840 | pemetrexed | 0.67863 |
| lipoic acid | 0.99480 | cabozantinib | 0.97840* | axitinib | 0.67270 |
| tamoxifen | 0.99480* | cyclophosphamide | 0.97760* | methicillin | 0.66760* |
| pirfenidone | 0.99440 | quercetin | 0.97760* | docetaxel | 0.66551 |

Figure 4: Predicted down-regulators of glioblastoma, glioblastoma multiforme and anaplastic astrocytoma. Known negative regulators are highlighted in red and with an asterisk. Known regulators are drugs in which the relationship *drug x negatively regulates disease or gene y has been reported in literature.*

| Probability Drug | DIPG | Probability Drug | Gliomatosis Cerebri | Probability Drug | Brainstem Neoplasm |
|---|---|---|---|---|---|
| carboplatin | 0.98440* | temozolomide | 0.92853* | vincristine | 0.97573* |
| 5-fluorouracil | 0.97787 | lomustine | 0.88520* | carboplatin | 0.95920* |
| etoposide | 0.97600* | procarbazine | 0.86240* | cisplatin | 0.95000* |
| cisplatin | 0.96520* | carboplatin | 0.78530 | irinotecan | 0.91790* |
| doxorubicin | 0.96400 | cisplatin | 0.77800 | doxorubicin | 0.91360 |
| ifosfamide | 0.95680 | ifosfamide | 0.74327 | gemcitabine | 0.90173 |
| BCNU | 0.95653* | doxorubicin | 0.73160 | etoposide | 0.89476 |
| irinotecan | 0.95307 | irinotecan | 0.72940 | bleomycin | 0.89240 |
| gemcitabine | 0.95040 | BCNU | 0.72880 | temozolomide | 0.88693 |
| oxaliplatin | 0.94840 | capecitabine | 0.72607 | 5-fluorouracil | 0.87340 |
| LBH589 | 0.94573* | dacarbazine | 0.72120 | ifosfamide | 0.87307 |
| imatinib | 0.94560* | pazopanib | 0.71887 | mitomycin C | 0.83640 |
| temozolomide | 0.94520* | gemcitabine | 0.71333 | docetaxel | 0.80036 |
| lomustine | 0.94280 | topotecan | 0.70530 | vinblastine | 0.79927 |
| mitomycin C | 0.94280 | crizotinib | 0.70435 | lomustine | 0.79040 |
| thiotepa | 0.94120 | camptothecin | 0.70223 | BCNU | 0.78627 |
| capecitabine | 0.93933 | etoposide | 0.69760 | nimustine | 0.78560* |
| docetaxel | 0.93627 | bleomycin | 0.68840 | topotecan | 0.77500 |
| vandetanib | 0.93480* | 5-fluorouracil | 0.68743 | oxaliplatin | 0.76837 |
| bleomycin | 0.93480 | vincristine | 0.68547* | actinomycin D | 0.75968 |
| topotecan | 0.93480 | oxaliplatin | 0.68460 | capecitabine | 0.74880 |
| vincristine | 0.93280 | pemetrexed | 0.67863 | melphalan | 0.74587 |
| sorafenib | 0.92810 | axitinib | 0.67270 | cytarabine | 0.73460 |
| pazopanib | 0.91980 | methicillin | 0.66760* | daunorubicin | 0.73207 |
| melphalan | 0.91920 | docetaxel | 0.66551 | dacarbazine | 0.73160 |

Figure 5: Predicted down-regulators of DIPG, gliomatosis cerebri and brainstem neoplasm. Known negative regulators are highlighted in red and with an asterisk. Known regulators are drugs in which the relationship *drug x negatively regulates disease or gene y has been reported in literature.*

## Drug Combinations

The basis of this proposed drug repurposing methodology is drug combinations will prove more efficacious treatments than compounds administered alone. Having established potential therapeutics, we next wished to predict if a drug pair is complementary. Researchers[5] pioneered a systems polypharmacology methodology, demonstrating how using protein-protein interaction networks, network proximity between drug targets of multiple drugs and disease gene regulators can be used to determine the suitability of those two drugs in treatment of the disease. In their work, two proximity measures are employed; distance between drugs (or more precisely distance between the closest drug targets of each drug) and distance between each drug and the disease (distance between the closest drug targets and disease regulators). We posit that by generating node embeddings of a disease-drug regulatory bipartite network, the embeddings will be able to capture both drug-drug proximity and drug-disease proximity. There are notable differences between the original work and ours. Whilst the original metrics represented drug-drug-disease relationships in three dimensions (three one-dimensional proximity measures), we believe an embedding-based approach will be able to capture more latent information of node neighbourhood in the higher dimensions of the embeddings. The previous work used protein-protein interactions alongside drug targets and disease regulators to connect drugs and diseases. We simply used drug-disease (indirect) regulation. In previous work, we observed extremely high correlation between drug-disease regulation, drug-drug target, and protein-protein interaction bipartite subgraphs (in terms of node embedding similarity). From this we believe that whilst this information is not directly supplied to the model, it is intrinsically captured on the drug-disease regulatory network alone.

To perform this workflow, we ingested all FDA approved drug combinations as edges between drugs in our knowledge graph, and generated embeddings based on a drug-disease regulatory network. We deemed that all FDA approved combinations would exhibit a synergistic or complementary relationship and thus proved a valuable indicator for the efficacy of drug pairs. We then trained a link-prediction model to predict if drug pairs matched this complementary relationship of FDA approved drugs. Model metrics are tabulated below. Link-prediction models based are often inherently biased. Models are most performant in areas of the network in which they have most information. In terms of this prediction task, this signifies that the model will perform best when i) the drug (or similar drugs) has been reported to regulate many diseases and ii) the drug pairs (or similar pairs) have been approved for combinatiorial use. In other words, predictions will be more accurate for predicting well researched diseases and extensively documented drugs (or similar diseases and drugs).

| Metric | Score |
|---|---|
| AUC | 0.927±0.002 |
| F1 score | 0.879±0.001 |
| Accuracy | 0.876±0.003 |
| Precision | 0.823±0.012 |
| Best Threshold | 0.511±0.039 |
| Best TPR | 0.909±0.031 |
| Best FPR | 0.148±0.037 |

Figure 6: Model performance for link prediction model to predict FDA-approved drug combinations.

Using the above model, we predicted the likelihood that all possible pairs of the drug candidates (drugs predicted or known to down-regulate DIPG) would exhibit the complementary relationship of FDA-approved drugs. Next, we calculated the transitive probability of combinatiorial use of a drug pair to treat a disease. To clarify, the transitive probability of carboplatin and 5-fluorouracil for treatment is: the summation of i) the individual probability of down-regulation of DIPG from administration of carboplatin, ii) the individual probability of down-regulation from use of 5-fluorouracil, and iii) the probability that both drugs are complementary and comply with the relationships of known FDA-approved drug pairs. The top 20 predictions, ranked according to transitive probability are shown below.

| Drug 1 (D1) | Drug 2 (D2) | Inhibit Disease (D1) | Inhibit Disease (D2) | Combination | Transitive |
|---|---|---|---|---|---|
| carboplatin | 5-fluorouracil | 0.98440* | 0.97787 | 0.96213* | 0.92616 |
| | oxaliplatin | 0.98440* | 0.94840 | 0.96013* | 0.89639 |
| | cisplatin | 0.98440* | 0.96520* | 0.94143* | 0.8945 |
| | docetaxel | 0.98440* | 0.93627 | 0.97001* | 0.89402 |
| 5-fluorouracil | doxorubicin | 0.97787 | 0.96400 | 0.9414* | 0.88742 |
| cisplatin | doxorubicin | 0.96520* | 0.96400 | 0.9385* | 0.87323 |
| etoposide | doxorubicin | 0.97600* | 0.96400 | 0.928* | 0.87312 |
| carboplatin | irinotecan | 0.98440* | 0.95307 | 0.92047* | 0.86358 |
| 5-fluorouracil | docetaxel | 0.97787 | 0.93627 | 0.94313* | 0.86348 |
| | oxaliplatin | 0.97787 | 0.94840 | 0.92927* | 0.86181 |
| gemcitabine | oxaliplatin | 0.95040 | 0.94840 | 0.95267* | 0.8587 |
| doxorubicin | gemcitabine | 0.96400 | 0.95040 | 0.93444* | 0.85612 |
| carboplatin | paclitaxel | 0.98440* | 0.90830 | 0.95621* | 0.85497 |
| 5-fluorouracil | capecitabine | 0.97787 | 0.93933 | 0.9264* | 0.85094 |
| cisplatin | oxaliplatin | 0.96520* | 0.94840 | 0.9293* | 0.85068 |
| carboplatin | doxorubicin | 0.98440* | 0.96400 | 0.89457* | 0.84891 |
| doxorubicin | docetaxel | 0.96400 | 0.93627 | 0.93927* | 0.84775 |
| | ifosfamide | 0.96400 | 0.95680 | 0.9156* | 0.84451 |
| irinotecan | gemcitabine | 0.95307 | 0.95040 | 0.93067* | 0.84299 |
| carboplatin | gemcitabine | 0.98440* | 0.95040 | 0.90081* | 0.84277 |

Figure 7: Probability of combinatorial treatment for DIPG. In order from left to right, the columns indicate: drug name, probability that each individual drug down-regulates the disease, probability that the drug pair comply with the drug-disease neighbourhood relationship of FDA-approved drugs, the transitive probability of the preceding three columns. Any known relationship is highlighted in red and with an asterisk. For example, carboplatin has been reported to inhibit DIPG, 5-fluorouracil has not, but they are both FDA-approved (for any disease) as part of a multidrug regimen.

The majority of multidrug therapies are indicated for cancers. This predilection towards oncology drug combinations introduces a heavy knowledge bias towards higher probabilities for cancer drugs and lower for non-oncology drugs. It has been shown the abundant antineoplastic potential of many non-oncology drugs.[9] We next filtered the drug combination predictions for combinations currently unapproved for indication of any disease, to highlight potential novel combinations.

| Drug 1 (D1) | Drug 2 (D2) | Inhibit Disease (D1) | Inhibit Disease (D2) | Combination | Transitive |
|---|---|---|---|---|---|
| carboplatin | capecitabine | 0.98440* | 0.93933 | 0.90417 | 0.83607 |
| etoposide | docetaxel | 0.97600* | 0.93627 | 0.88414 | 0.80792 |
| | oxaliplatin | 0.97600* | 0.94840 | 0.86957 | 0.8049 |
| 5-fluorouracil | etoposide | 0.97787 | 0.97600* | 0.84144 | 0.80307 |
| carboplatin | etoposide | 0.98440* | 0.97600* | 0.82473 | 0.79238 |
| etoposide | capecitabine | 0.97600* | 0.93933 | 0.84555 | 0.77519 |
| | gemcitabine | 0.97600* | 0.95040 | 0.83502 | 0.77456 |
| carboplatin | topotecan | 0.98440* | 0.93480 | 0.84118 | 0.77407 |
| 5-fluorouracil | gemcitabine | 0.97787 | 0.95040 | 0.82768 | 0.76921 |
| carboplatin | imatinib | 0.98440* | 0.94560* | 0.82567 | 0.76857 |
| | lapatinib | 0.98440* | 0.89363 | 0.87063 | 0.76589 |
| | melphalan | 0.98440* | 0.91920 | 0.8461 | 0.7656 |
| | mitomycin C | 0.98440* | 0.94280 | 0.82424 | 0.76497 |
| etoposide | cisplatin | 0.97600* | 0.96520* | 0.80905 | 0.76215 |
| doxorubicin | oxaliplatin | 0.96400 | 0.94840 | 0.82197 | 0.75149 |
| 5-fluorouracil | ifosfamide | 0.97787 | 0.95680 | 0.80247 | 0.75081 |
| carboplatin | ifosfamide | 0.98440* | 0.95680 | 0.79537 | 0.74914 |
| docetaxel | sorafenib | 0.93627 | 0.92810 | 0.85815 | 0.74569 |
| | topotecan | 0.93627 | 0.93480 | 0.85139 | 0.74516 |
| etoposide | paclitaxel | 0.97600* | 0.90830 | 0.84012 | 0.74476 |

Figure 8: Probability of combinatorial treatment for DIPG, where drug pair are not known to be part of a multidrug regimen (indicated for any disease). In order from left to right, the columns indicate: drug name, probability that each individual drug down-regulates the disease, probability that the drug pair comply with the drug-disease neighbourhood relationship of FDA-approved drugs, the transitive probability of the preceding three columns. Any known relationship is highlighted in red and with an asterisk. For example, carboplatin has been reported to inhibit DIPG, capecitabine has not, and they are both not FDA-approved (for any disease) as part of a multidrug regimen.

## Blood Brain Barrier Permeability Prediction

The blood brain barrier (BBB) is a microscopic structure that protects the brain from toxins in the blood. As such, this presents a major challenge in the development of clinically successful therapeutics for DMGs, due the prevention of pharmacological exposure neoplastic tissues have to the drugs. We developed an embedding based model, using pretrained Mol2Vec embeddings based on Morgan substructures, to predict if a drug was BBB permeable (BBB+) or impermeable (BBB-). Having predicted the permeability of all drugs, we then calculated the transitive probability that for a drug pair: i) each drug inhibits DIPG, ii) the drug pair match the FDA-approved relationship, and iii) each drug can cross the BBB. This can be seen as an extended version of the transitive probability in Drug Combinations. We evaluated the model on 10-fold stratified cross-validation, and the model achieved an AUC score of 0.97.

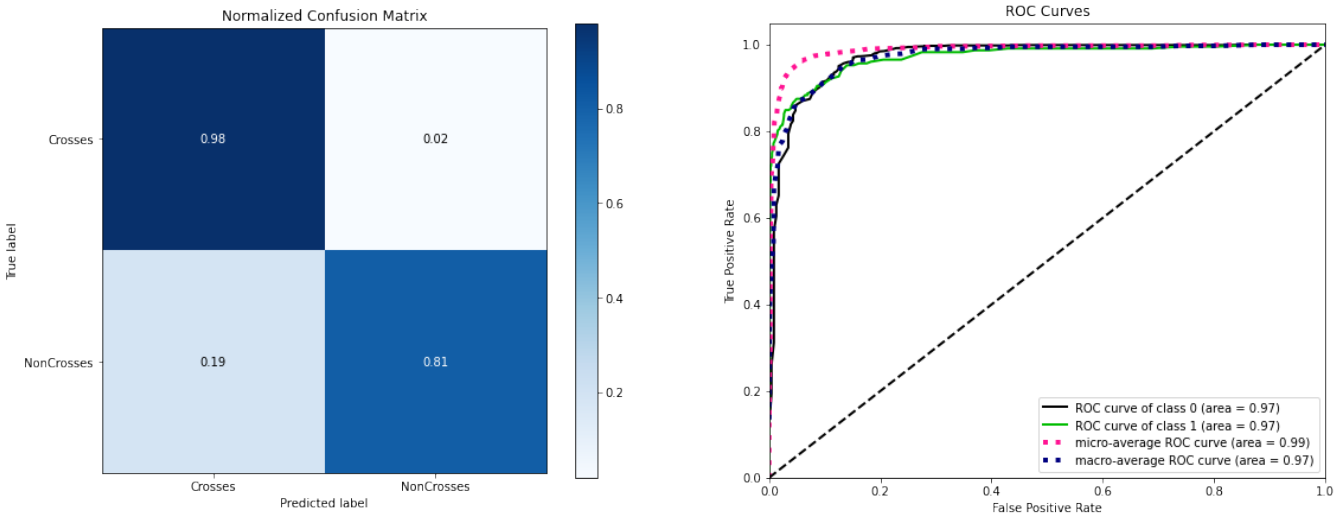


Figure 9: Confusion matrix and area under the receiver operating characteristic curve for BBB permeability prediction model. Calculated over 10-fold stratified cross-validation.

Previous studies[10] have shown how drug clinical phenotype can be used to determine BBB permeability, as brain-associated phenotypes are indicative that a drug crosses the BBB. We experimented replacing the mol2vec structural embedding with embeddings based on drug-disease association and regulatory networks. Both these embeddings were evaluated on 10-fold stratified cross validation and obtained AUC scores of 0.92 and 0.91 for association and regulatory networks respectively.

| Drug 1 (D1) | Drug 2 (D2) | Inhibit Disease (D1) | Inhibit Disease (D2) | Combination | BBB+ (D1) | BBB+ (D2) | Transitive |
|---|---|---|---|---|---|---|---|
| carboplatin | cisplatin | 0.98440* | 0.96520* | 0.94143* | 0.8332 | 0.9576 | 0.7137 |
| cisplatin | temozolomide | 0.96520* | 0.94520* | 0.88247* | 0.9576 | 0.9204 | 0.70958 |
| | ifosfamide | 0.96520* | 0.95680 | 0.91235* | 0.9576 | 0.8724 | 0.70388 |
| | oxaliplatin | 0.96520* | 0.94840 | 0.9293* | 0.9576 | 0.7942 | 0.64696 |
| | cyclophosphamide | 0.96520* | 0.84680 | 0.94173* | 0.9576 | 0.8288 | 0.61089 |
| carboplatin | oxaliplatin | 0.98440* | 0.94840 | 0.96013* | 0.8332 | 0.7942 | 0.59317 |
| cisplatin | imatinib | 0.96520* | 0.94560* | 0.90373* | 0.9576 | 0.7464 | 0.58955 |
| carboplatin | 5-fluorouracil | 0.98440* | 0.97787 | 0.96213* | 0.8332 | 0.75696 | 0.58413 |
| 5-fluorouracil | cisplatin | 0.97787 | 0.96520* | 0.81524* | 0.75696 | 0.9576 | 0.55775 |
| cisplatin | thalidomide | 0.96520* | 0.91700* | 0.68523 | 0.9576 | 0.956* | 0.55522 |
| carboplatin | temozolomide | 0.98440* | 0.94520* | 0.76551 | 0.8332 | 0.9204 | 0.54623 |
| | ifosfamide | 0.98440* | 0.95680 | 0.79537 | 0.8332 | 0.8724 | 0.54454 |
| cisplatin | gemcitabine | 0.96520* | 0.95040 | 0.87153* | 0.9576 | 0.70493 | 0.53968 |
| 5-fluorouracil | oxaliplatin | 0.97787 | 0.94840 | 0.92927* | 0.75696 | 0.7942 | 0.5181 |
| cisplatin | lenalidomide | 0.96520* | 0.82500 | 0.73693 | 0.9576 | 0.9208 | 0.51742 |
| | tegafur | 0.96520* | 0.77307 | 0.92553* | 0.9576 | 0.7795 | 0.5155 |
| | gefitinib | 0.96520* | 0.87013 | 0.90131* | 0.9576 | 0.7062 | 0.51191 |
| carboplatin | thalidomide | 0.98440* | 0.91700* | 0.71193 | 0.8332 | 0.956* | 0.5119 |
| ifosfamide | oxaliplatin | 0.95680 | 0.94840 | 0.8075 | 0.8724 | 0.7942 | 0.50769 |
| temozolomide | suberoylanilide hydroxamic acid | 0.94520* | 0.80500 | 0.84667* | 0.9204 | 0.84853 | 0.50313 |
| cyclophosphamide | lenalidomide | 0.84680 | 0.82500 | 0.93533* | 0.8288 | 0.9208 | 0.49867 |
| oxaliplatin | vandetanib | 0.94840 | 0.93480* | 0.87865* | 0.7942 | 0.804 | 0.49741 |
| thalidomide | cyclophosphamide | 0.91700* | 0.84680 | 0.8072* | 0.956* | 0.8288 | 0.49663 |
| 5-fluorouracil | ifosfamide | 0.97787 | 0.95680 | 0.80247 | 0.75696 | 0.8724 | 0.49581 |
| carboplatin | gemcitabine | 0.98440* | 0.95040 | 0.90081* | 0.8332 | 0.70493 | 0.495 |

Figure 10: Probability of combinatorial treatment for DIPG, including BBB+ probability. In order from left to right, the columns indicate: drug name, probability that each individual drug down-regulates the disease, probability that the drug pair comply with the drug-disease neighbourhood relationship of FDA-approved drugs, probability that each drug is BBB+, and the transitive probability of the preceding five columns. Any known relationship is highlighted in red and with an asterisk. For example, cisplatin has been reported to inhibit DIPG, thalidomide has also, they are both not FDA-approved (for any disease), cisplatin is not known to be BBB+, but thalidomide is.

## Notable Therapeutics

### Monodrug

Having predicted inhibitors for multiple DMGs (Figures 4 and 5), we collated the best candidates by taking the top 50 predictions for each DMG, alongside all known inhibitors. This yielded 135 unique compounds, referred to hereafter as the monodrug drug set.

### Multidrug

Following the various analyses predicting drug pair viability for treatment of DIPG, we collated the compounds in the top 20 predictions from each analysis into a list of most noteworthy drug candidates. The analyses used to generate this list were i) the drug combinations analysis (ranked only by transitive probability of multidrug treatment: Figure 7) , ii) the drug combinations analysis (in which only novel combinations were permitted: Figure 8) and iii) drug combinations including BBB+ probability for each drug (Figure 10). From these 60 drugs, 21 unique compounds were highlighted, referred to hereafter as the multidrug drug set, shown below.

| | | |
|---|---|---|
| 5-Fluorouracil | Etoposide | Lenalidomide |
| Capecitabine | Gefitinib | Melphalan |
| Carboplatin | Gemcitabine | Mitomycin_C |
| Cisplatin | Ifosfamide | Oxaliplatin |
| Cyclophosphamide | Imatinib | Paclitaxel |
| Docetaxel | Irinotecan | Sorafenib |
| Doxorubicin | Lapatinib | Temozolomide |

Figure 11: Drugs taken from the top 20 drug pairs from i) the drug combinations analysis, ii) the drug combinations analysis with novel drug pairs, iii) drug combinations including BBB permeability.

We also performed a larger search for notable therapeutics, this time including the top 50 predictions for each DMG. In total, this yielded 135 unique compounds.

## Graph Enrichment

Having established possible drug combinations for DMGs, we next wished to understand the mechanisms through which the drugs regulated these cancers. To achieve this, we first calculated statistically significant biological processes misregulated by DIPG (see Methods). To determine the significance of the processes, we summed the processes that followed the path in our knowledge graph $(Disease) - [Regulation(Positive/Negative)] - (Protein) - [Associates] - (BiologicalProcess)$, and referenced these to diseases with a similar degree distribution. Over three hundred biological processes were found to be over the p-value threshold of 0.05. The top 20 processes are displayed below. Enriched processes included histone modifications and spindle organisation and regulation of cell cycle checkpoints.

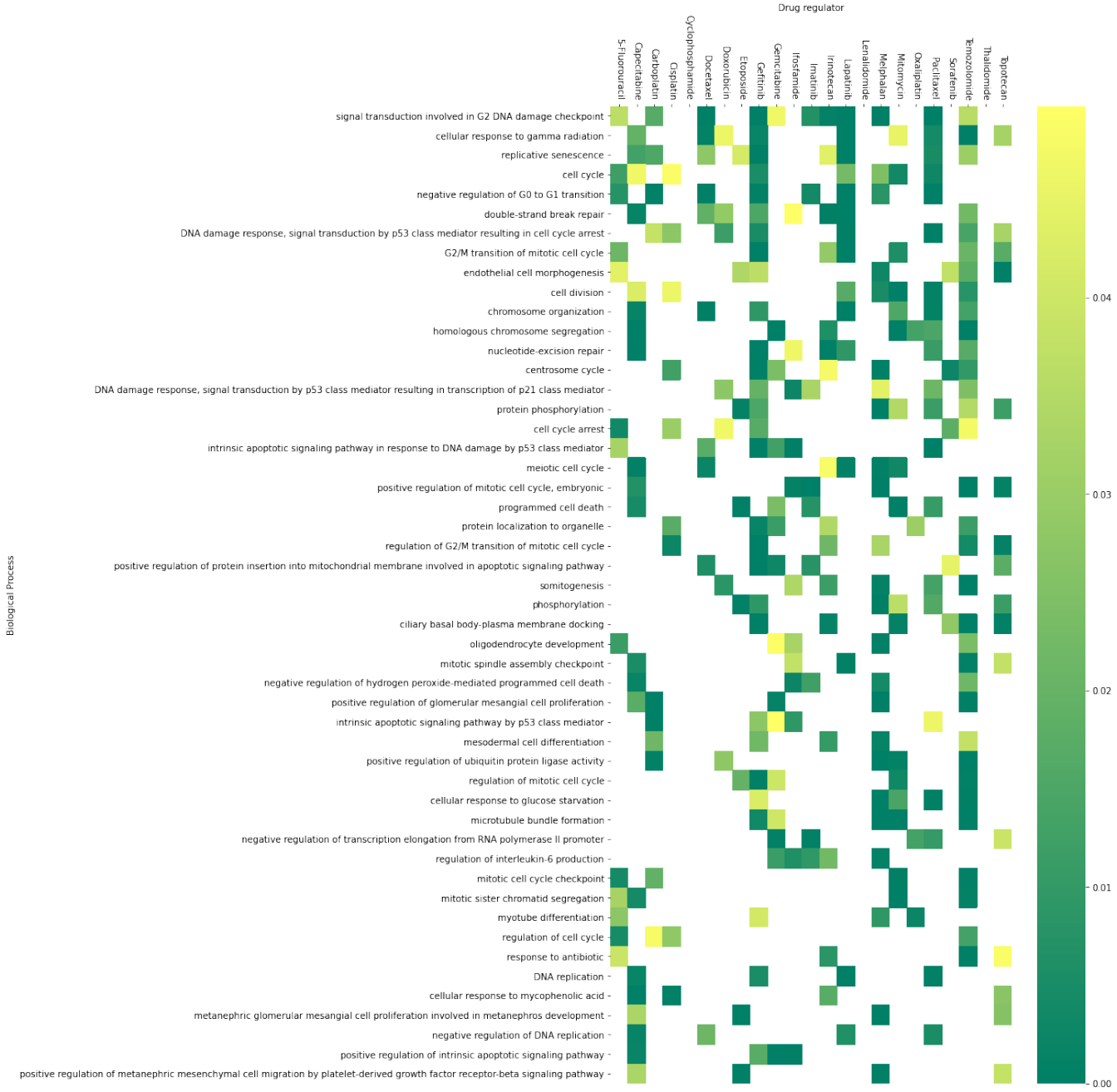


Figure 12: Enriched biological processes misregulated by DIPG. Statistically significant biological processes misregulated by DIPG. Top 20 processes shown, ranked by p-value. Under the p-value of 0.05, there were 305 enriched biological processes.

Next, we performed a similar analysis on the candidate therapeutics. We determined the enriched biological processes following the path $(Drug)-[Regulation*] \rightarrow (Protein)-[Associates] \rightarrow (BiologicalProcess)$. We then cross-referenced the PLACEHOLDER processes enriched in the disease state, against the processes regulated by each drug, to highlight processes through which the drugs could act.

| | Z-score | p-value |
|---:|---:|---:|
| **mitotic spindle midzone assembly** | 27.694788 | 8.067917e-169 |
| **attachment of spindle microtubules to kinetochore** | 27.694788 | 8.067917e-169 |
| **mitotic cytokinesis checkpoint** | 27.694788 | 8.067917e-169 |
| **spindle checkpoint** | 27.694788 | 8.067917e-169 |
| **negative regulation of cytokinesis** | 27.694788 | 8.067917e-169 |
| **abscission** | 19.570419 | 2.763998e-85 |
| **regulation of mitotic metaphase/anaphase transition** | 19.570419 | 2.763998e-85 |
| **regulation of follicle-stimulating hormone secretion** | 19.570419 | 2.763998e-85 |
| **protein localization to nuclear envelope** | 19.570419 | 2.763998e-85 |
| **synaptonemal complex disassembly** | 19.570419 | 2.763998e-85 |
| **spinal cord oligodendrocyte cell fate specification** | 19.570419 | 2.763998e-85 |
| **polar body extrusion after meiotic divisions** | 19.570419 | 2.763998e-85 |
| **histone H3-S28 phosphorylation** | 19.570419 | 2.763998e-85 |
| **spinal cord oligodendrocyte cell differentiation** | 19.570419 | 2.763998e-85 |
| **progesterone secretion** | 19.570419 | 2.763998e-85 |
| **histone modification** | 19.570419 | 2.763998e-85 |
| **striatal medium spiny neuron differentiation** | 19.570419 | 2.763998e-85 |
| **anaphase-promoting complex-dependent catabolic process** | 17.508965 | 1.223997e-68 |
| **spindle organization** | 15.968760 | 2.109373e-57 |
| **GABAergic neuron differentiation** | 15.968760 | 2.109373e-57 |

Figure 13: Top 50 most frequent enriched biological processes misregulated by DIPG and regulated by candidate therapeutics, ranked by number of drugs that regulate the process. Candidate therapeutics were determined as top drugs predicted for combinatorial use against DIPG. All processes are under a p-value of 0.05.

We performed the same analysis substituting molecular function for biological processes. PLACEHOLDER were found below the p-value of 0.05. The top 20 function are shown below.

Figure 14: Enriched molecular functions misregulated by DIPG. Statistically significant molecular functions misregulated by DIPG. Top 20 processes shown, ranked by p-value. P-value of 0.05 and under.

We then cross-referenced misregulated molecular functions with those regulated by the candidate drugs.

| index | Z-score | p-value |
|---|---|---|
| type II activin receptor binding | 19.570419 | 2.763998e-85 |
| histone serine kinase activity | 12.353190 | 4.681757e-35 |
| inhibin binding | 10.426677 | 1.873259e-25 |
| superoxide-generating NADPH oxidase activator activity | 10.426677 | 1.873259e-25 |
| histone methyltransferase activity | 9.746861 | 1.902602e-22 |
| hepatocyte growth factor-activated receptor activity | 9.183389 | 4.177330e-20 |
| histone methyltransferase activity (H3-K27 specific) | 8.706394 | 3.136944e-18 |
| protein-lysine N-methyltransferase activity | 8.706394 | 3.136944e-18 |
| RNA polymerase II CTD heptapeptide repeat kinase activity | 6.456046 | 1.074740e-10 |
| ribonucleoprotein complex binding | 5.826152 | 5.671985e-09 |
| primary miRNA binding | 5.694158 | 1.239819e-08 |
| anaphase-promoting complex binding | 5.140999 | 2.732814e-07 |
| magnesium ion binding | 4.704482 | 2.545109e-06 |
| promoter-specific chromatin binding | 4.451873 | 8.512465e-06 |
| phospholipase binding | 4.445347 | 8.775027e-06 |
| histone-lysine N-methyltransferase activity | 4.334803 | 1.458909e-05 |
| TFIID-class transcription factor complex binding | 4.323611 | 1.534962e-05 |
| Wnt-protein binding | 4.157766 | 3.213755e-05 |
| histone deacetylase regulator activity | 4.106308 | 4.020333e-05 |
| platelet-derived growth factor binding | 4.060810 | 4.890282e-05 |

Figure 15: Enriched molecular functions misregulated by DIPG and regulated by candidate therapeutics, ranked by number of drugs that regulate the process. Candidate therapeutics were determined as top drugs predicted for combinatorial use against DIPG. All processes are under a p-value of 0.05.

Next, we performed the analysis for pathways.

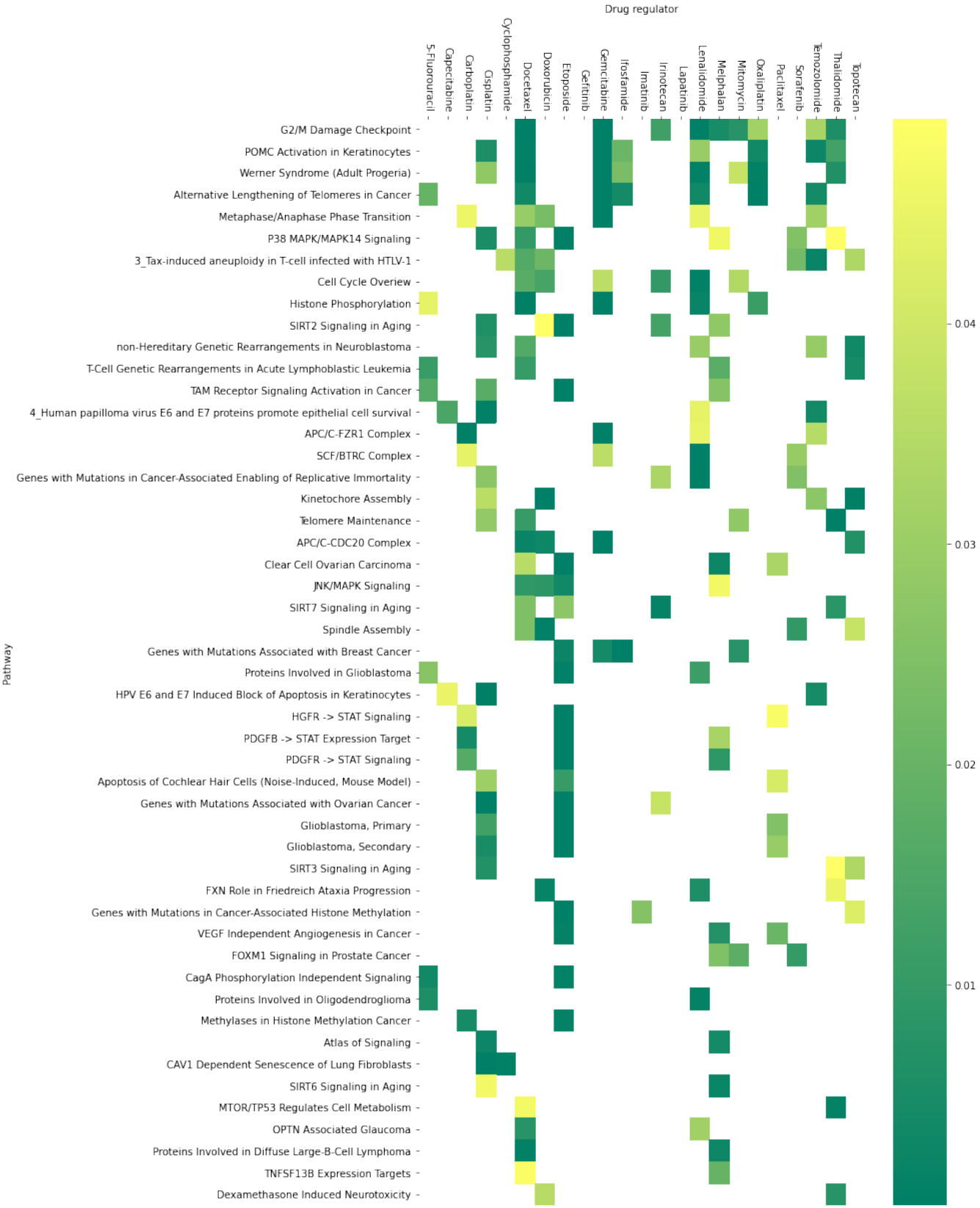


Figure 16: Statistically significant pathways. Top 20 processes shown, ranked by p-value.

| | Z-score | p-value |
|---|---|---|
| **mRNA Transcription and Processing** | 27.604371 | 9.860803e-168 |
| **Kinetochore Assembly** | 22.533959 | 1.929172e-112 |
| **Spindle Assembly** | 18.396273 | 1.407161e-75 |
| **Metaphase/Anaphase Phase Transition** | 18.396273 | 1.407161e-75 |
| **APC/C-FZR1 Complex** | 15.916490 | 4.869411e-57 |
| **Sister Chromatid Cohesion** | 14.748485 | 3.146507e-49 |
| **Chromosome Condensation** | 13.372147 | 8.796437e-41 |
| **APC/C-CDC20 Complex** | 12.312648 | 7.743971e-35 |
| **GPR3 in Oocyte Meiotic Arrest** | 9.714747 | 2.608962e-22 |
| **Proteins with Altered Expression in Cancer-Associated Histone Methylation** | 9.153092 | 5.532701e-20 |
| **3_Tax-induced aneuploidy in T-cell infected with HTLV-1** | 7.061726 | 1.644469e-12 |
| **Genes with Mutations in Cancer-Associated Histone Methylation** | 6.834081 | 8.253219e-12 |
| **Microtubule Cytoskeleton** | 6.832921 | 8.320280e-12 |
| **SCF/BTRC Complex** | 6.331722 | 2.424403e-10 |
| **Glioblastoma, Proneural Subtype** | 6.258634 | 3.883652e-10 |
| **Histone Phosphorylation** | 6.257732 | 3.906175e-10 |
| **G2/M Phase Transition** | 5.831141 | 5.504970e-09 |
| **T-Cell Genetic Rearrangements in Acute Lymphoblastic Leukemia** | 5.829411 | 5.562324e-09 |
| **Methylases in Histone Methylation Cancer** | 5.806616 | 6.374807e-09 |
| **MicroRNAs in Epithelial to Mesenchymal Transition in Cancer** | 5.803717 | 6.486085e-09 |

Figure 17: Enriched pathways misregulated by DIPG and regulated by candidate therapeutics, ranked by number of drugs that regulate the process. Candidate therapeutics were determined as top drugs predicted for combinatorial use against DIPG. All processes are under a p-value of 0.05, and all the top 20 are shown.

## Regulatory Graph

Next, we wished to investigate regulatory paths through which the above drugs perturbed the cancerous state of DIPG. Due to the size and high connectivity of the graph, we decided to focus upon a subgraph that we believe capture sufficient information pertaining to the fundamental regulatory mechanisms of disease. We filtered on the monodrug drug set and DMGs. We modelled a tripartite network, focusing on three key nodes: disease, protein, and drug. We filtered on the following edges: i) regulatory between drugs and genes, and diseases and genes, protein-protein direct interactions, and binding between drug and drug target. In the below visualisation, gene regulators of the diseases are labelled alongside all drug candidates. One can observe how drugs are localised near the genes with which they interact most. For example, kinase inhibitors such as imatinib, dasatinib and sorafenib are all clustered near Signal Transducer And Activator Of Transcription 1 (STAT1) and ABL Proto-Oncogene 1 (ABL1), due to the large number of paths

between drugs in this class and these genes. From this we can infer, that the further apart two drugs are in the graph, the lower likelihood that they will regulate the diseases via the same mechanisms. It should be noted this figure is to illustrate the core methodology of this paper; network layout was determined via charge and gravity and should not be used to directly infer network proximity.

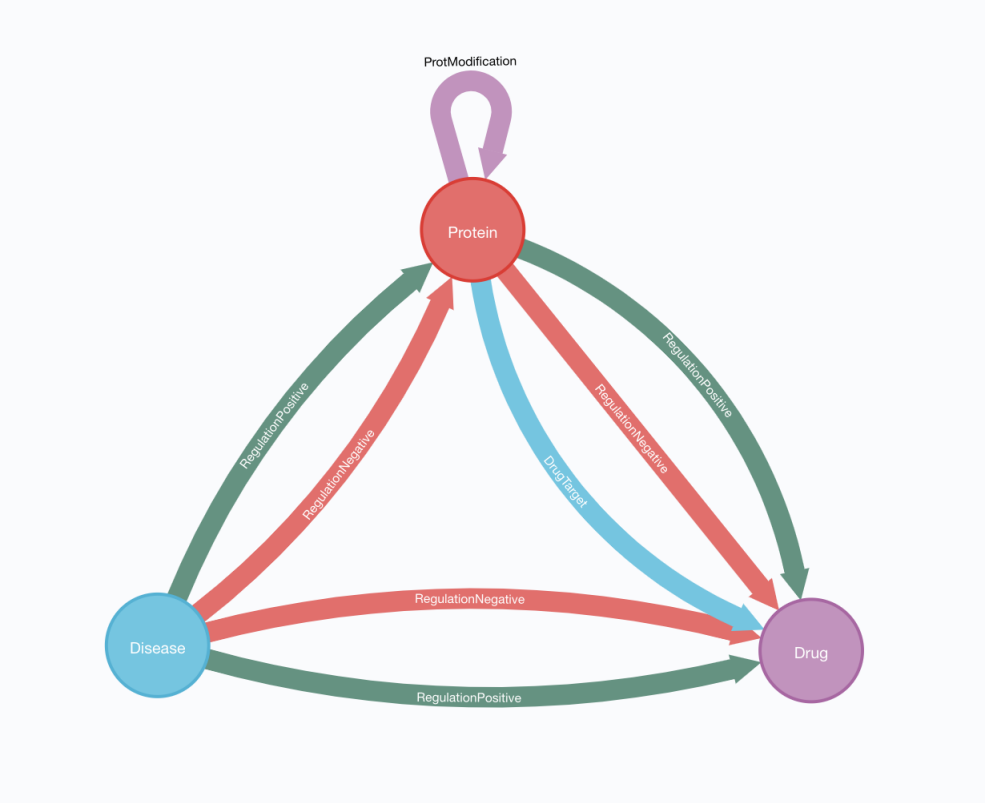


Figure 18: Regulatory subgraph schema. Schema of the regulatory subgraph from the Neo4j graph database. In our knowledge graph, there are distinct edges for all types of protein modifications ((de)phosphorylations, (de)sumoylation, cleavages etc). In this subgraph, we conflated all of these edges into one protein modification edge. Similarly, drug target edges were characterised by their mechanism of action (modulatory, opener, agonist, allosteric agonist etc). Similarly, these were conflated into one drug target edge. We enforced a direction in the graph: arrow direction indicates that the source node applies the edge type on the target node. For example, diseases are regulated by both proteins and drugs. Proteins are drug targets and regulated by drugs.

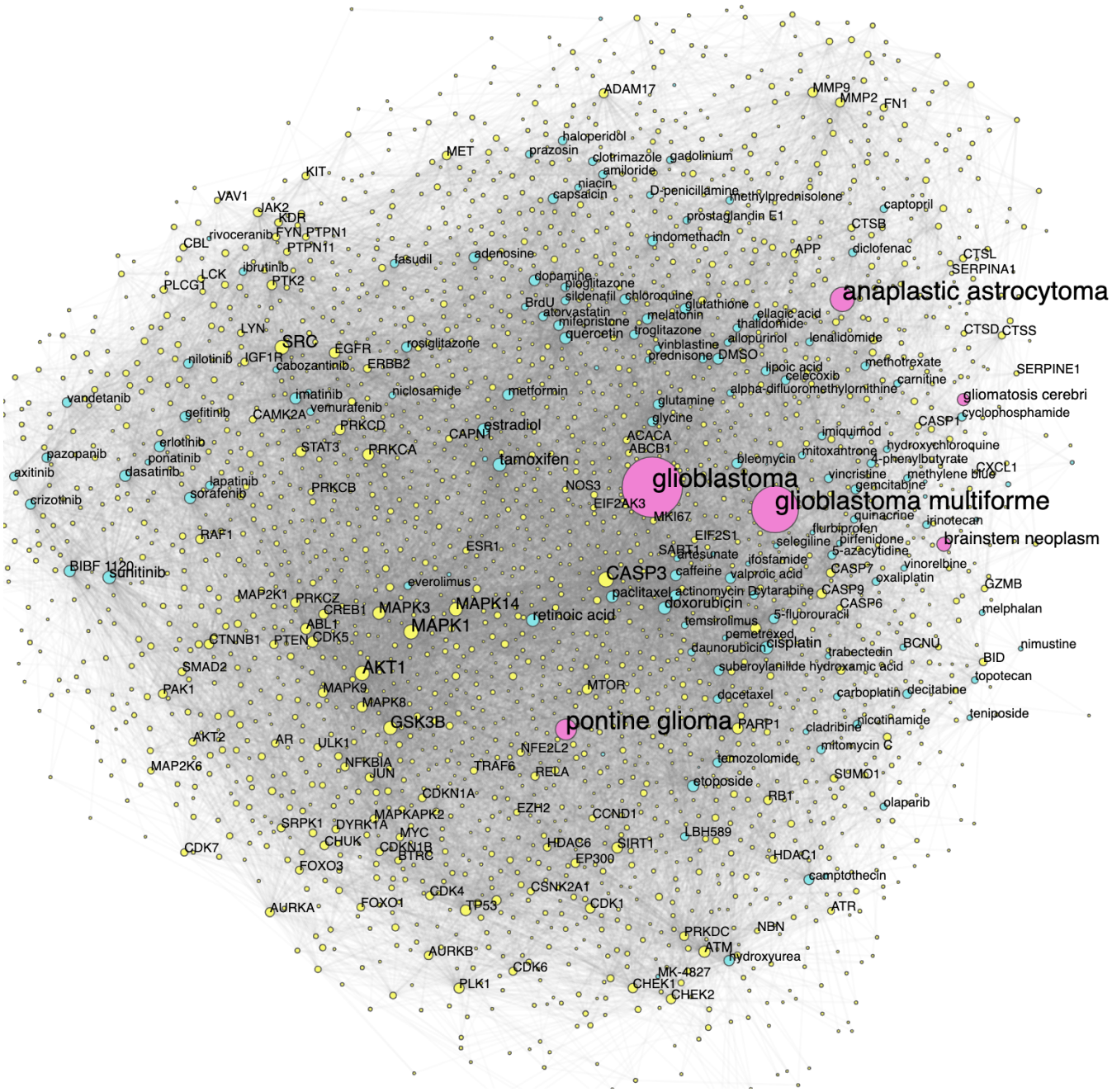


Figure 19: Local regulatory subgraph of DMGs. Subgraph filtered from the main knowledge graph. Protein-protein interactions, indirect regulations, and drug-target interactions were permitted. The subgraph was filtered on the top drug candidates and the brain cancers. Key regulator genes are labelled. Node size is determined by its connectivity within the network. Network layout was determined using charge force and gravity; it should not be used to directly correlate network proximity and overlap in regulatory mechanisms.

## Degree Weighted Path Count

Having constructed the subgraph we wished to i) validate that the graph sufficiently represented the regulatory paths between therapeutics and ii) investigate which paths were most informative in terms of predicting drug efficacy. To assess this we needed a graph sufficiently large to facilitate statistical analysis of path importance. We constructed a subgraph following the exact same node and edge heuristics as above but including all diseasees and infections in the knowledge graph. We then used a degree weighted path count (DWPC) model[11] to predict if drugs were reported to regulate the diseases. DWPC models exploit all paths that connect two nodes, to assess if a link exists between the two. To clarify, each feature used by the model refers to the summation of the degree-normalised paths of a specific type, that connect a source and target node. The main benefit of this is that one can investigate the relative importance of each path type, simply by assessing the importance of each feature.

A optimized random forest model achieved an area under the receiver operating characteristic (AUROC) of 0.89. Importantly, this score was achieved on an undirected graph: non-biological edges were permitted.[11] showed how for a drug indication prediction task, the most important features/paths inferred either drug or disease similarity. Figure 20 shows how the most important path infers compound similarity (DrC—CrP—PrC; disease is regulated by compound, which regulates a gene, which is regulated by a compound/drug). This first path infers drug similarity via association of gene regulation. On a similar vein, the third most important edge infers compound similarity via association disease regulation (DrC—CrD—DrC; disease is regulated by compound, which regulates a disease, which is regulated by a compound/drug). Biological (directed) paths were shown to be informative to the prediction task. The second most informative edge was DrP—PrP—PrC (disease is regulated by a gene, which interacts with another gene, which is regulated by a drug), and the fourth was DrP—PrC (disease is regulated by a gene, which is regulated by a drug). Both of these edges describe regulatory routes through which a drug can regulate a disease state. It should be noted that correlation between regulatory routes does not indicate this route is causative of the drugs efficacy.

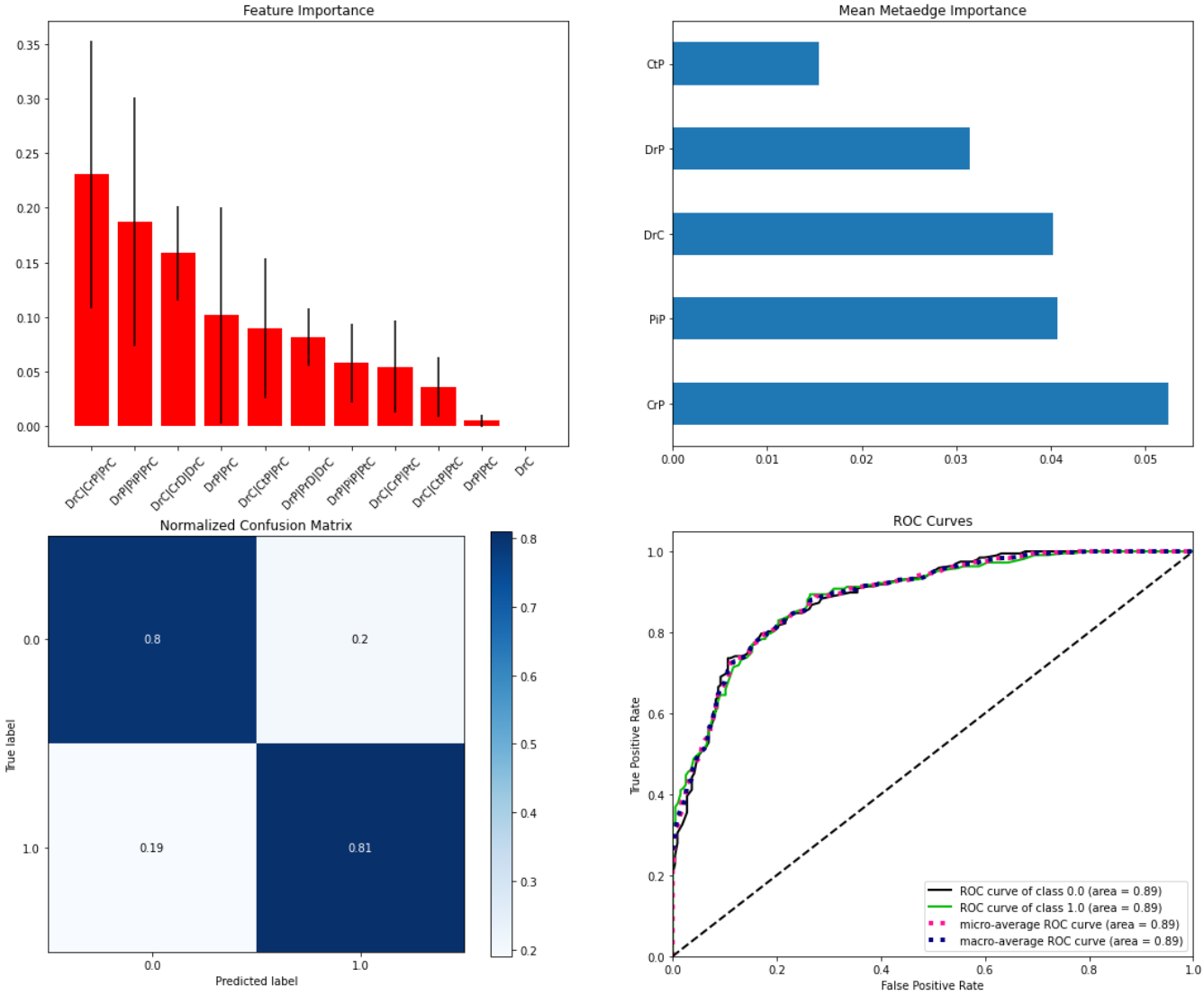


Figure 20: Results for degree weighted path count model. From left to right, and top to bottom: i) feature importance for paths in regulatory graph that connect diseases to drug via 0-2 protein intermediates, ii) mean importance of edges, extracted from feature importance of paths, iii) confusion matrix for holdout test set, iv) AUROC curve for holdout test set.

## Enrichment Analysis

To further understand these regulatory routes, we performed enrichment analysis on the disease regulators. Importantly, we only included the protein regulators that could be connected to a drug, and not simply any disease regulators (with no direct path to a drug). This was performed to give us a clearer idea of the pathways and biological processes through which known drugs are thought to regulate the diseases.

We performed enrichment analysis on both pathways and biological processes. Promisingly, the pathway analysis yielded multiple pathways involving cancer and apoptosis. Enriched biological processes included apoptotic processes and cell proliferation.

Figure 21: Enrichment Analysis. Top: Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis for all disease regulators that are connected to drugs via 1-3 hops in a directed acyclic regulatory graph. Bottom: Same analysis for Gene Ontology biological process. Performed using Enrichr.

## Gene Importance in Pathways

To assess importance of each gene in these pathways, we constructed protein-protein interaction (PPI) networks of all of these enriched pathways. According to the message passing principle, it has

been shown network topological importance of a nodes is correlated with how essential a protein is. Centrality measures such as eigenvector centrality have shown to be effective predictors of gene importance in PPI networks.[12] We constructed homogeneous gene networks, permitting edges of any reported direct modification between two genes. We then calculated the eigenvector centrality for each gene in each pathway. Finally, we tabulated the centrality of each disease regulator in each enriched pathway.

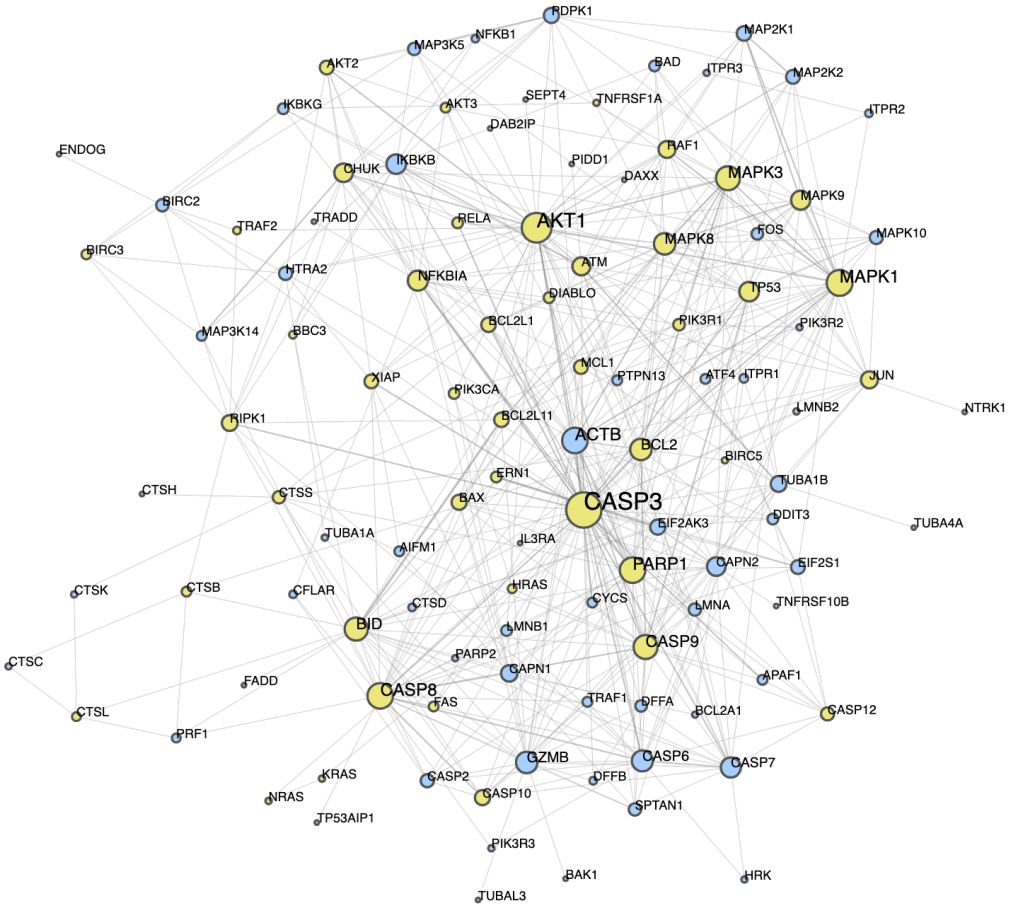


Figure 22: Protein-protein interaction network for all proteins in the apoptosis pathway (KEGG). Edges represent any direct modification such as (de)acetylation, (de)phosphorylation between two proteins that have been reported in literature.

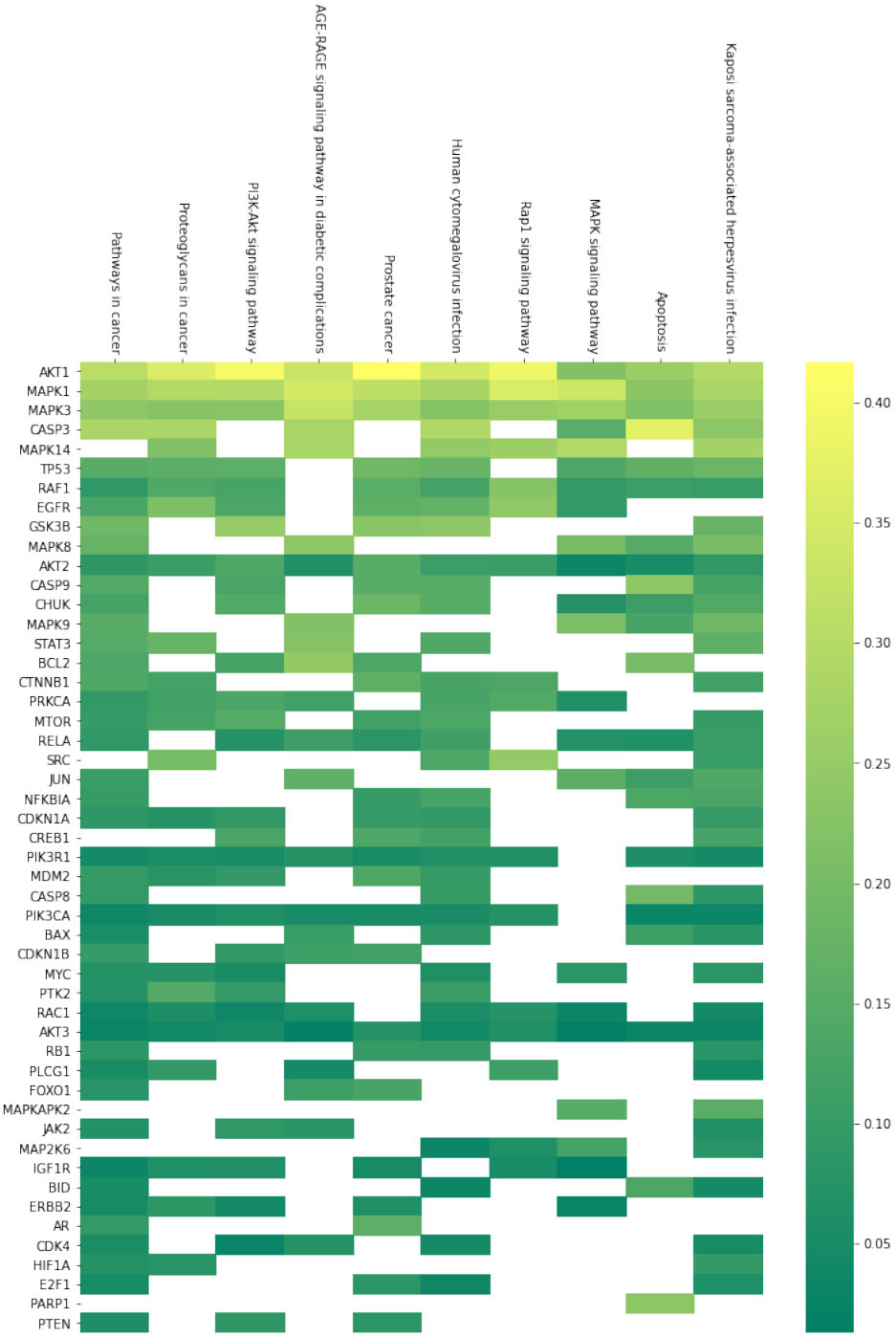


Figure 23: Network centrality for disease regulators in their enriched pathways. Pathway enrichment analysis was performed on the disease regulators that could be connected via 1-3 hops in a directed acyclic regulatory graph. Eigenvector centrality of these genes within the respective protein-protein interaction network for each pathway was calculated and tabulated above. Top 50 genes are displayed out of 200, ranked by total summed centrality over all pathways.

## Drug Proximity

As previously mentioned, the knowledge bias of FDA-approved drug combinations signified that the transitive probability of treatment in  were biased towards oncology (heavily immunomodulatory). We next investigated whether non-oncology therapeutics could be highlighted mitigating this bias. Again, we drew heavily from the research that inspired our embedding-based link prediction model of FDA drug combinations.[5] The group demonstrated how using PPI networks, network proximity between drug targets of multiple drugs and disease regulators can be used to determine the suitability of those two drugs in treatment of the disease. Specifically, the group showed how 1-dimensional measures of drug-drug and drug-disease proximity could be used to highlight drug pair therapeutic viability. Replication of this viability window would be considerably less biased than our machine learning model. We extend their methodology by allowing drugs to be connected to proteins via indirect regulation, as well as direct drug targets. As the degree-weighted path count model showed, regulatory edges between drugs and genes (PrC) proved significantly more informative than drug target edges (PtC). We reason that this increase in information proved to the network will facilitate more accurate and credible predictions of drug combinations.

There are notable differences in methodology between the original researchers who pioneered this model and our implementation. The researchers projected drug and disease information onto a PPI network, whilst we use a heterogeneous regulatory network. The original methodology used a network proximity measure based on 'closest' measure - the average shortest path length between targets of drug A and the nearest target of the drug A. We employed an embedding based approach: generating Node2vec embeddings for each drug in the graph, and computing the cosine distance between drug embeddings.[5] used any disease-associated genes to connect diseases to the PPI network, whilst we used genes specifically reported to regulate the disease. Finally, the paper demonstrated that two distance metrics were important in determining combinatorial drug candidates; distance between a drug and a disease, and distance between the two drugs. We posit that all drugs known or predicted to be down-regulators of DMGs have demonstrated they must be within the acceptable network distance for the drugs to illicit a therapeutic response on the disease state. Thus, we focus solely on distance between two drugs.

We used node2vec[13] to generate embeddings for every drug in the graph. Node2vec samples a graph via random walks, with the mantra that nodes closer in the graph will have more similar embeddings.[5] demonstrated how network proximity was correlated with many functional and physicochemical similarities of drugs and their targets, and disease-associated genes. The closer drugs are positioned within the graph to gene regulators, the more regulatory routes exist between the disease and drug, through this gene. Thus, network distance can also be used to indicate differences in the regulatory mechanisms of our candidate therapeutics. As the network in this study differs from,[5] and different network proximity measures were employed, we replicated their analysis to prove their findings were applicable to our regulatory graph.

To perform this analysis, we created embeddings for all drugs in the smaller subgraph, and computed the relationships between cosine distance of two drugs, and various biological and pharmacological relationships (see Figure 24). We show how a significant correlation exists between network proximity of two drugs and overlap of their drug targets. Similarly, we show the correlation extends to indirect drug regulators (literature-derived relationships stating the drug has a regulatory effect on the gene by any mechanism).Next, we show the biological processes and path-

ways of these drug regulators and targets, also exhibit this relationship with network distance. The drug indication overlap between drug pairs (diseases known to be down-regulated by the two drugs) showed little correlation with network proximity, as does overlap of drug-drug interactions.

To ascertain if network proximity was correlated with FDA-approved drugs, we compared cosine distance between drug pairs (Figure 26). We compared synergistic and antagonistic drugs to random pairs that followed similar degree distributions. Interestingly, we found synergistic drugs pairs tended to have a cosine distance of 0.7. We plotted FDA approved drug pairs against non-approved pair and found a similar distribution peak at 0.7, suggesting many FDA- approved drug pairs are synergistic. This supports the findings of the original paper, that a 'Goldilocks' zone exists between drugs. The original paper described two metrics that highlighted multidrug viability: i) drug-disease distance and ii) drug-drug distance. We replicated this drug-drug ideal distance using cosine distance between a drug pair. We did not replicate the drug-disease distance. Whilst the researchers computed these metrics on a global graph, and computed it on a specific subgraph. In this subgraph, every drug node was known or highly predicted to inhibit the DMGs. Thus, it was unnecessary to compute a drug-disease distance as all drugs in our subgraph had already been proven to be probable inhibitors. Interestingly, there was a slight increase in FDA approved drugs with a cosine distance near 0.2. All of these belonged to drug pairs of kinase inhibitors (lapatinib and pazopanib, sorafenib and sunitinib, and gefitinib and sunitinib.

To compare the machine learning model to predict FDA approved drugs with this network proximity measure, we plotted probability versus cosine distance (Figure 27). We filtered to show only drug pairs from the 135 candidate drugs list (the top 50 predicted and known DMG inhibitors). Whilst all FDA approved drugs had a clear threshold probability of 0.7, there was not as clear distinction for cosine distance (as we previously saw in above distribution figure). Whilst more in-depth quantitative analysis is needed on a larger disease set, it seems the model is a more performant indicator of drug pair viability. We do believe network proximity can be used as an additional indicator. We saw how network proximity is strongly negatively correlated with overlap of drug targets, regulators, biological processes and pathways. We also saw how drug pairs with a cosine distance around 0.2 are from the same drug class. Ensuring combination predictions have a cosine distance greater than say 0.6 would ensure drug pairs are synergistic, functionally disparate, and regulate the disease through different mechanisms.

Finally, we plotted the embeddings of each drug on a 2D plot preceding principal component analysis (Figure 25). It is often tempting to draw conclusions and generate hypothesis regarding clustering using such a plot, but it is important to remember that significant information was lost in this dimension reduction. Moreover, in such a densely connected graph, significant drug disparity (far apart in the graph) does not infer an absence of shared drug targets and regulators (and by extension mechanisms of disease regulation), merely lower probability of such an overlap. Nevertheless, one can observe notable groupings of pharmacological drug classes such as kinase inhibitors. Conversely, platinum drugs carboplatin, oxaliplatin and cisplatin are extremely disparate in the embedding space.

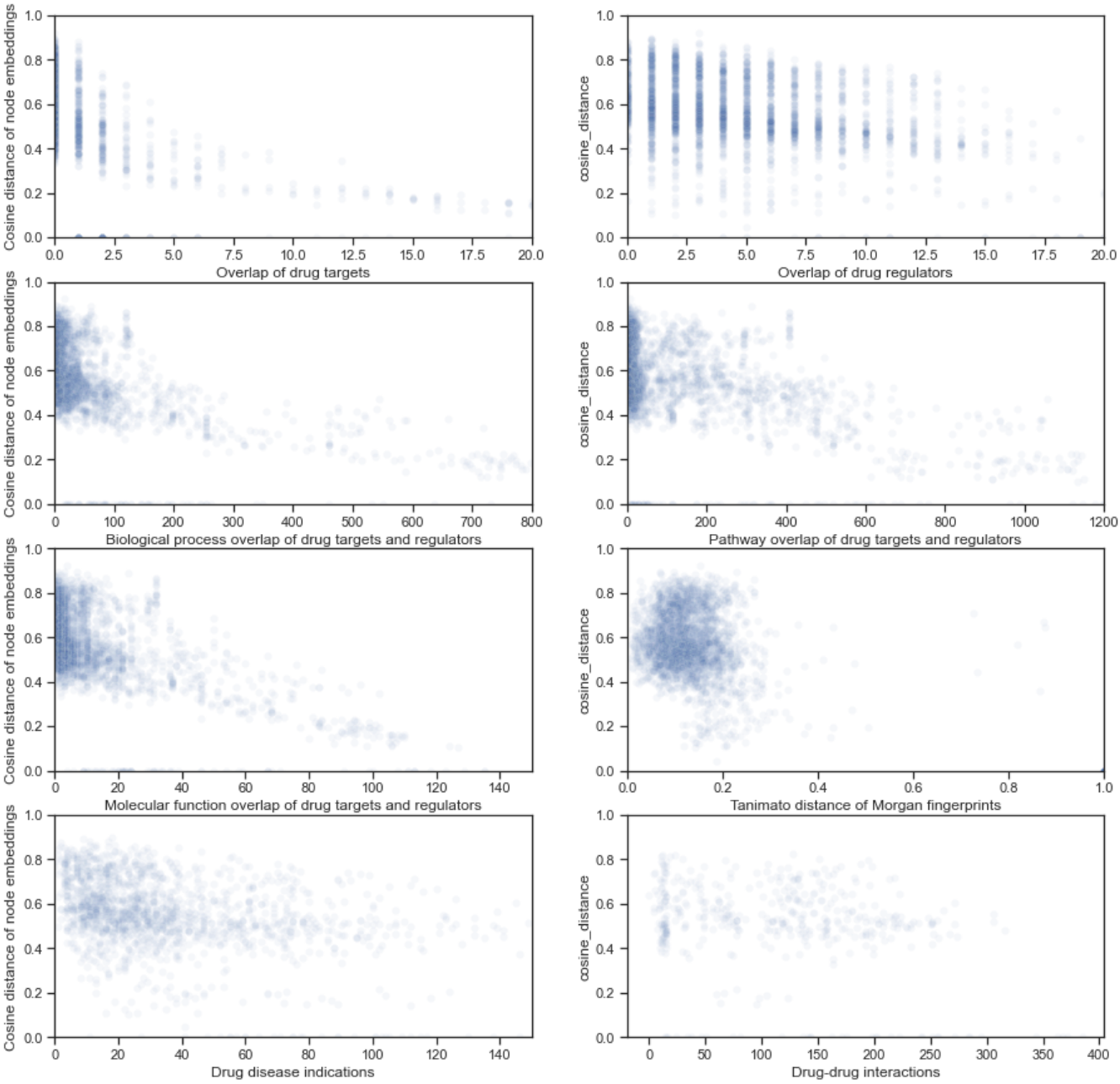


Figure 24: Network proximity analyses. Analyses investigating the relationships between cosine distance (a measure of network distance) and multiple biochemical and biological interactions. Embeddings were generated from the regulatory subgraph of all diseases and infections in the knowledge graph.

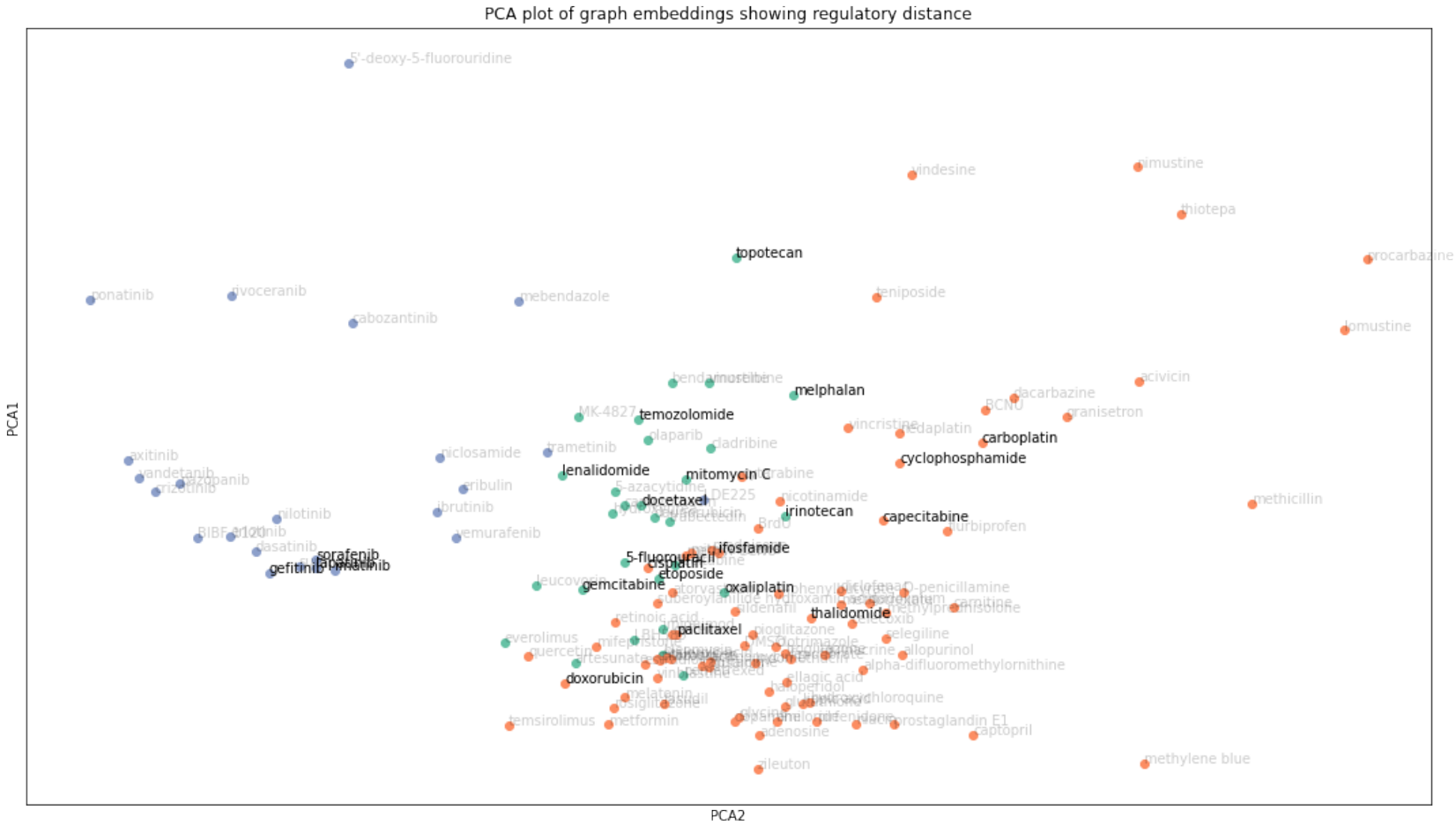


Figure 25: Embeddings of drugs based on the directed regulatory graph. The 21 notable drugs candidates are highlighted.

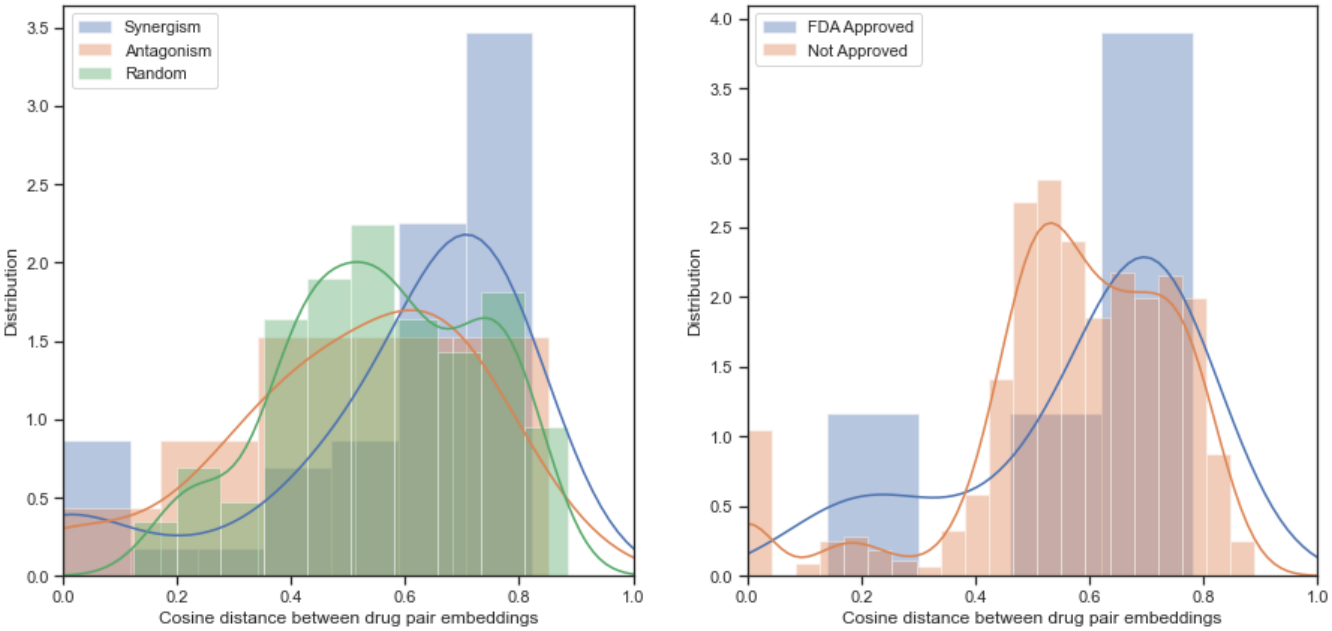


Figure 26: Left: distribution of cosine distance between synergistic and antagonistic drugs. Right: distribution of cosine distance between FDA approved drugs vs non-approved drugs.

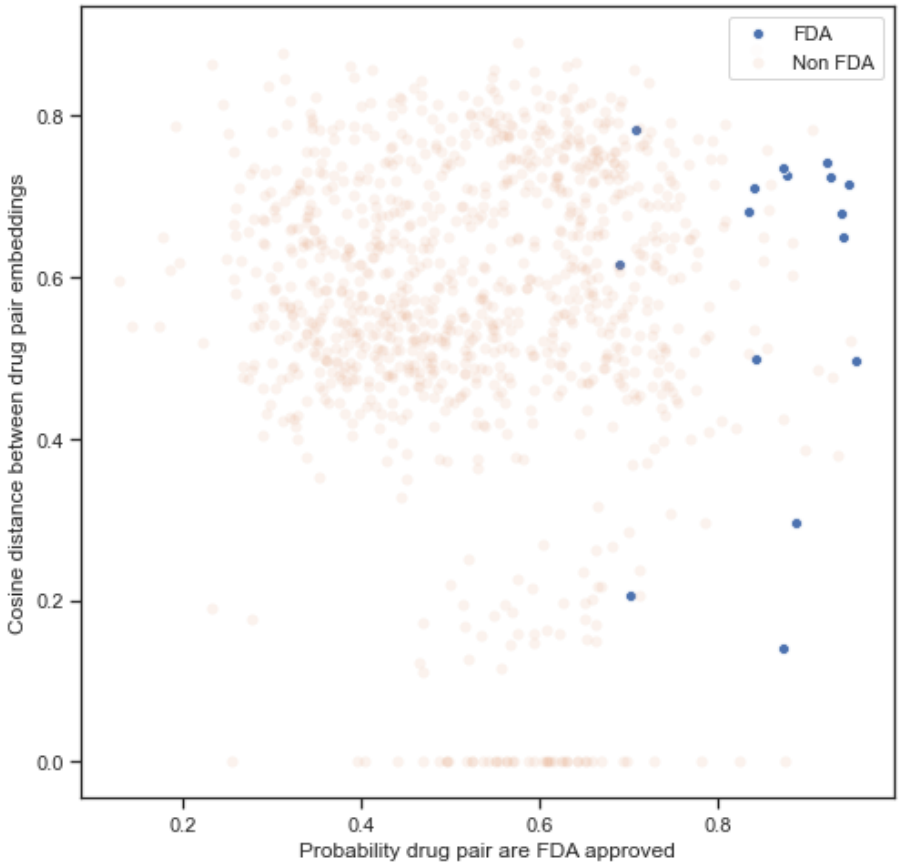


Figure 27: Cosine distance between drug pair embeddings as a function of probability two drugs are FDA approved. Only drug pairs for the top 50 predicted or known inhibitors of DMGs are shown.

# Discussion

## Methodology

There are many limitations to this workflow. An important note is that the above diseases have vastly differing degrees of connectivity in the knowledge graph. Whilst a well-researched cancer such as glioblastoma has 716 connected down-regulators in the knowledge graph, Diffuse Intrinsic Pontine Glioma, a rare pediatric brain cancer has 28. Such differences in degree indicate different prior probabilities of treatment. In other words, nodes in a graph with higher connectivity are statistically more likely to be connected. This prior probability is reflected in the probability distributions. Glioblastoma and DIPG have a mean probability of down-regulation of 0.62 and 0.07. As such, probabilities should not be directly compared between diseases, and instead used as an indication of relative position in ranked probability lists.

Link prediction on knowledge graphs also signifies that the area under the receiver operating characteristic is not 0.5 for random guesses of connecting source and target nodes: it is much higher. Another approach that overcomes this degree-induced threshold issue was used by the researchers who developed the degree weighted path count model. They determined prior probability of treat-

ment by brute-force - randomly permuting bipartite graphs multiple times (swapping edges but preserving node degree) and noting the average probability of an edge as a function of source and target degree. Whilst a significant improvement over our method (and essentially normalises class prediction) this still does not solve the problem that probabilities are unbalanced (probabilities are not normalised). We hypothesise a method which would normalise both probability and prediction. If one was to calculate mean prediction as a function of source and target node degree, one could calculate the function necessary to normalise all mean predictions, for all source and target node degrees, to 0.5. To the best of our knowledge this has not yet been implemented in academia.

Machine learning methods are often cursed by their lack of explainability. Specifically, the optimized parameters of supervised models are hard to interpret. Whilst embeddings methods have proved powerful strategies to capture continuous data into discrete and dense vector representations, they obfuscate explainability even further. Embeddings are used as the features of a model. As each dimension is the output of another vectorization model, they are completely uninterpretable. In contrast, the DWPC model used features based on specific paths. Using common feature importance techniques, it is much easier to explain how informative certain features are (and thus paths). It should be noted DWPC models are massively more computationally expensive, limiting their use to relatively small graphs.

To determine probability of multidrug treatment of DMGs, we used the transitive probability of i) the probability of inhibition for each individual drug and ii) the probability the drug pair were FDA-approved. Due to lack of data, there was no quantitative evaluation performed to ascertain if the transitive probability represented drug pair viability for a certain disease. Moreover, no coefficients were applied to optimize this equation. We performed qualitative analysis, by running this workflow for HIV, and found known drug combinations indicated for the HIV such as lopinavir and ritonavir were ranked highly. Quantitative analysis is needed to truly validate this approach.

FDA approved drugs were used as a proxy to measure the complementary action of two drugs. This may not be the best approach to ascertain this relationship. Whilst FDA approval was shown to be correlated with drug synergy (see Figure 26), the majority of drug pairs are different drug classes such as antivirals and kinase inhibitors are often indicated together. If one wished to predict complementary drugs acting upon different disease mechanisms, it may be beneficial to remove all drug pairs from the same drug class from the training set of the model. Future work, informed by planned *in vitro* analyses, may explore this strategy. Alternatively, this could easily be achieved by imposing a threshold of network proximity (cosine distance), to ensure functional and mechanistic dissimilarity between drugs. Drug synergy dramatically varies between cell lines. If one wished to predict synergy for certain drugs pairs in certain cell lines, there are more established and accurate methods.[14]

This workflow trained and predicted on only FDA-approved drugs. Whilst there are a few thousand drugs in the graph, there are over 1 million compounds. Rerunning the analysis on all compounds may highlight i) drug-like compounds with therapeutic potential and ii) potential nutraceutical candidates. We did not initially include compounds, as the degree of connectivity massively varies between drugs and compounds, to a much greater extent than diseases. We believed this increased prior probability of connection would have dominated the predictions. The prior probability issue should be resolved before rerunning the analysis.

Interestingly, 1-dimensional network proximity of drugs in the regulatory graph was well correlated with chemical similarity and functional and mechanistic similarity of their target. This analysis

successfully replicated the findings from the original paper.[5] There was no discernible correlation between 1-dimensional network proximity and FDA-approved drugs, suggesting the relationship of efficacious drug pairs is more complex than distance within a regulatory graph. In short, we were unable to replicate this 'goldilocks zone'- drug-drug 1-d dimensional network distance was not correlated with synergy of the drug pair, nor likelihood of the pair being FDA-approved. This may have been caused by notable differences in our methodologies: more research is needed in this area to explore FDA-approved drugs from a systems polypharmacology point of view.

The dataset[15] we used for the BBB permeability prediction only included compounds able to pass through the BBB via passive diffusion. Whilst our model was very performant, it has been shown a select few molecular descriptors suffice to accurately predict the permeability of a drug across the membrane. We posit that the reduced Mol2vec embeddings were able to capture the hydrophilicity or lipophilicity of a compound and thus able to predict if a drug was BBB- or BBB+ respectively. The other method[10] mentioned used brain associated clincal phenotypes are features. The dataset they used was notably smaller, but also included BBB+ drugs via multiple mechanisms such as selective transporters and pinocytosis. We deemed the dataset too small, and thus focused only on passive diffusion. To date, there is no satisfactory informative and well-sized dataset documenting BBB permeability via multiple mechanisms. Such a dataset would greatly improve efforts to develop drugs to circumvent the BBB and treat myriad diseases alongside DMGs.

Biomedical knowledge graphs are largely incomplete. On average, well under one percent of possible relationships are reported for edges such as regulation, binding and expression. This presents a problem for link prediction models - the true distribution of positive and negative classes is unknown, and it is impossible to differentiate a true (but unreported) positive from a false positive. It is preferable to train your model on a class balance that matches your real world distribution. In this application, this is not possible. Moreover, when your real world distribution is known, optimizing your classification threshold becomes much simpler task. We used maximal f-1 score to determine optimal threshold. This often meant an unrealistic increase in the number of positive predicted samples. Out of 3302 drugs and 19387 diseases, there are 163590 known inhibitors (0.25 percent). The average positive class for a random subset of predicted regulators above the optimal threshold of 0.634, was 28 percent - a 112-fold increase in regulators. This is obviously a gross overestimation in the number of unreported regulators and a major limitation in our approach. We encourage predictions to be analysed only on the context of the relative list of regulators for that disease. An alternative approach taken to decide classification threshold based largely on the known class distribution.[3]

The knowledge graph provides an incredible amount of biological information, and we only used an incredibly small amount of this information. Each edge in the graph was weighted according the absolute occurrence of this relationship appears in scientific literature. It is logical to believe the number of times a relationship is reported is correlated with its strength or confidence. This does however introduce a new level of knowledge bias. Genes such as TP53 have received tremendous attention. Does this signify its importance over less researched genes? In previous analyses, we attempted to utilise this weighting through numerous normalisation strategies such as frequency of node, edge and path and log-scaled reference counts. We found that such strategies only generalised solutions; well-researched areas performed better, whilst less-researched performed worse. An interesting approach would be to normalise nodes according to the weighted edges that connect to it (weighted by reference count). In this case, for a well-researched node with 10 edges with

an individual edge weight of 50, each edge would be equivalent to a node with 10 edges, with an edge weight of 5. This may also solve the prior probability issue stated earlier, as for nodes with fewer edges, each edge will be given higher importance, partially mitigating the dominating effect of nodes with high degrees.

Ultimately, this workflow was predominantly based on literature derived data. One of the principal challenges of DMGs is their intratumoral molecular heterogeneity. This work did not take either genetic or epigenetic disparity into account. The DIPG node and associated edges in the graph represent a generic view of the cancer. If researchers investigated and reported say regulation of the H3.3.K27M subgroup by a gene, this would be added the same DIPG node in the graph, as if another research group reported the regulation of another gene on the IDH1 subgroup.

The disease model used for our regulatory graph was incredibly simplistic. We chose a directed graph, with only 0-2 protein intermediates to connect disease and drugs. As a directed graph, this was incapable of capturing any feedback loops. Whilst some direct interactions were recorded through protein modifications and drug targets, the predominant regulatory edge in the graph encompassed indirect mechanisms of regulation. This graph does not therefore capture the type of information pertinent in conventional pathway analysis. Moreover, 3 hops is not sufficient to capture signal transduction, which can be described as long cascades that fan out. There was no attempt to capture these cascades, or the additive effects or up-stream genes. At best, the regulatory graph visualises known targets and their effect on possible hub proteins (down-stream proteins with the highest degree).

The degree weighted path count model proved useful in identifying the following directed path was the most informative to predict disease inhibition: $(Disease) - [Regulation*] \rightarrow (Protein) - [ProtModification*] \rightarrow (Protein) - [Regulation*] \rightarrow (Drug)$. Correlation does not infer causation. By far, this directed path was the most frequent in the graph. We speculate the sheer volume of this path may be responsible for its high feature importance; the higher volume of network topological information allowing the model to differentiate between diseases and drugs. Drug target information has always been the gold standard for drug discovery and design, and therefore must carry more causative information regarding mechanism of action than the summative regulatory path features in our model.

We performed to network analysis to try and shed light on the regulatory mechanisms through which the drug candidates regulated the disease state of each disease. Again it is pertinent to remember correlation is not causation, and whilst these paths have been shown to be informative in differentiating regulators from non-regulators, is does not indicate the drugs necessarily act through these genes. Ultimately we find it difficult to gain insight through these network analyses. All graphs are largely dominated by known hub proteins such as STAT1. Perhaps it would be more effective to normalise proteins by their relative degree in the human interactome, reducing the overpowering nature of the hub proteins ubiquitous in the majority of human pathways. These views are generic and do not represent the protein levels in the pulmonary and immune cells in which DIPG is most associated. An improvement on this approach may be to filter proteins according to RNA-seq gene expression data for pulmonary epithelial cells infected with the disease.

For the graph enrichment, we assessed how statistically significant drugs that regulate proteins were, according to the respective proteins Gene Ontology terms. Essentially, we were comparing the following paths for each drug: $(Protein) - [Regulation*] \rightarrow (Drug)$. Similarly, for the enriched diseases, we were comparing the disease-specific paths: $(Disease) - [Regulation*] \rightarrow (Protein)$. As

we then cross-referenced each enriched Gene Ontology term, this can be loosely seen as highlighting enriched proteins in the path $(Disease) - [Regulation*] \rightarrow (Protein) - [Regulation*] \rightarrow (Drug)$. We know from the degree weighted path count model, that whilst this path was informative, the path that contains one more protein intermediate was significantly more informative. Future analysis could be performed using this path instead. It should be investigated if this more abundant path would further explain the processes and functions through which the drugs act upon the diseases, or if the longer path in question merely obfuscates processes regulated by the disease or drug. The gene set enrichment analysis of disease regulators yielded high-level pathways such as Pathways in cancer. In contrast, the graph enrichment yielded pathways much more specific to DIPG.

# Methods and Materials

## Drug Regulators

### Knowledge Graph

We created a knowledge graph predominantly based on Pathway Studio, a literature-derived database that uses natural language processing techniques to leverage biological relationships from over 30 million literary sources. Pathway Studio also contains the relevant subset of Reaxys Medicinal Chemistry, a database of small-molecule protein bioactivities, pertaining the species Homo sapiens, Mus muccus and Rattus rattus and Rattus norvegicus. We appended this core graph with gene ontologies,[16] drug side effects from SIDER[17] and drug-target information from Drug Central.[18] We also created similarity links such as protein-protein similarity (local Smith-Waterman of over 0.5), and molecule substructure similarity (Tanimoto similarity of Morgan Fingerprint of over 0.5). After refactoring and harmonization, the graph possessed 1.36 million nodes and 8.52 million edges, weighted according to their occurrences in literature.

For each molecule with an InChI code or key within the graph, we generated a Mol2Vec embedding using the pretrained embeddings[7] which were trained on 20M compounds form the ZINC database.[19] As all other embeddings had a dimension of 100, and as the model requires embeddings of the same size, we used the scikit-learn version of principal component analysis to reduce the embedding down to the required size. Summation of the explained variance showed only 1.3 percent of information was lost. Similarly, we generated embeddings for proteins based on trimers of their amino acid sequence using the pretrained model of ProtVec,[6] trained on 551,754 proteins from Swiss-Prot. As trimers can start at the first, second or third amino acid in a proteins sequence, three embeddings were generated per protein. As per the methodology of the original paper, we took the element-wise average of these.

### Drug Regulation Prediction

To determine if a drug down-regulated a disease or gene, we developed an embedding-based link prediction model based on multiple disease regulatory bipartite networks and additional physicochemical and structural information of the source and target nodes. The embeddings were used by a random forest classifier (scikit-learn implementation), optimized via hyper-parameter bayesian optimization. We assessed multiple node embedding strategies in this project. For embedding

choice in the link prediction model, we assessed GraRep, nodevec, LINE and SVD model. All models used the BioNev[20] implementation, except node2vec, for which we used the C++ version from SNAP.[13] Random search was employed to determine models with the highest AUC score. We also investigated different mathematical functions to create an edge embedding from two node embeddings (concatenation, element-wise average, hadamard, L1 and and L2 loss). As our model uses two different types of embeddings for each edge (for example i) graph embedding for protein and molecule and ii) ProtVec and Mol2vec respectively), employing the hadamard edge function signifies, the hadamard was calculated for both graph embeddings, and the Mol2vec and ProtVec separately before concatenating. We also investigated stacking models to create soft-voting bagging classifiers. As a subset of edges must be removed and used to train the model, we postulated that training multiple models on different subsets, and combining their predictions via a weighted average according to the performance, would increase predictive power of the stacked model. Results showed however, that increase was extremely marginal (under 0.5 percent increase in AUC). Due to the doubling of training time, we deemed this increase unnecessary. The following metrics were used to determine model performance: AUC, F1 score, accuracy and precision. The best classification threshold was determined by the finding the value at which F1 score was highest (the point at which accuracy and precision intersect). Scores were averaged over 3 random folds.

For the protein-drug regulation prediction, node neighbourhood embeddings for source and target $n_{s,t}$, are complemented with structural information: $e_s$ protein ProtVec amino acid trimer embeddings and $e_t$ Mol2Vec Morgan fingerprint substructure embeddings, where:

$$n_s, n_t, e_s, e_t = R^d$$

Each pair of embeddings is combined via one of various functions to create an edge embedding, before the edge embeddings of source and target are combined:

$$n_{st} = f(n_s, n_t)$$
$$e_{st} = f(e_s, e_t)$$
$$x_{st} = f(n_s t, e_s t)$$

$$f_{l1}(x, y) = \|x - y\|$$
$$f_{l2}(x, y) = \|x - y\|^2$$
$$f_{hadamard}(x, y) = x \cdot y$$
$$f_{concatenate}(x, y) = x \oplus y$$
$$f_{average}(x, y) = \frac{x + y}{2}$$

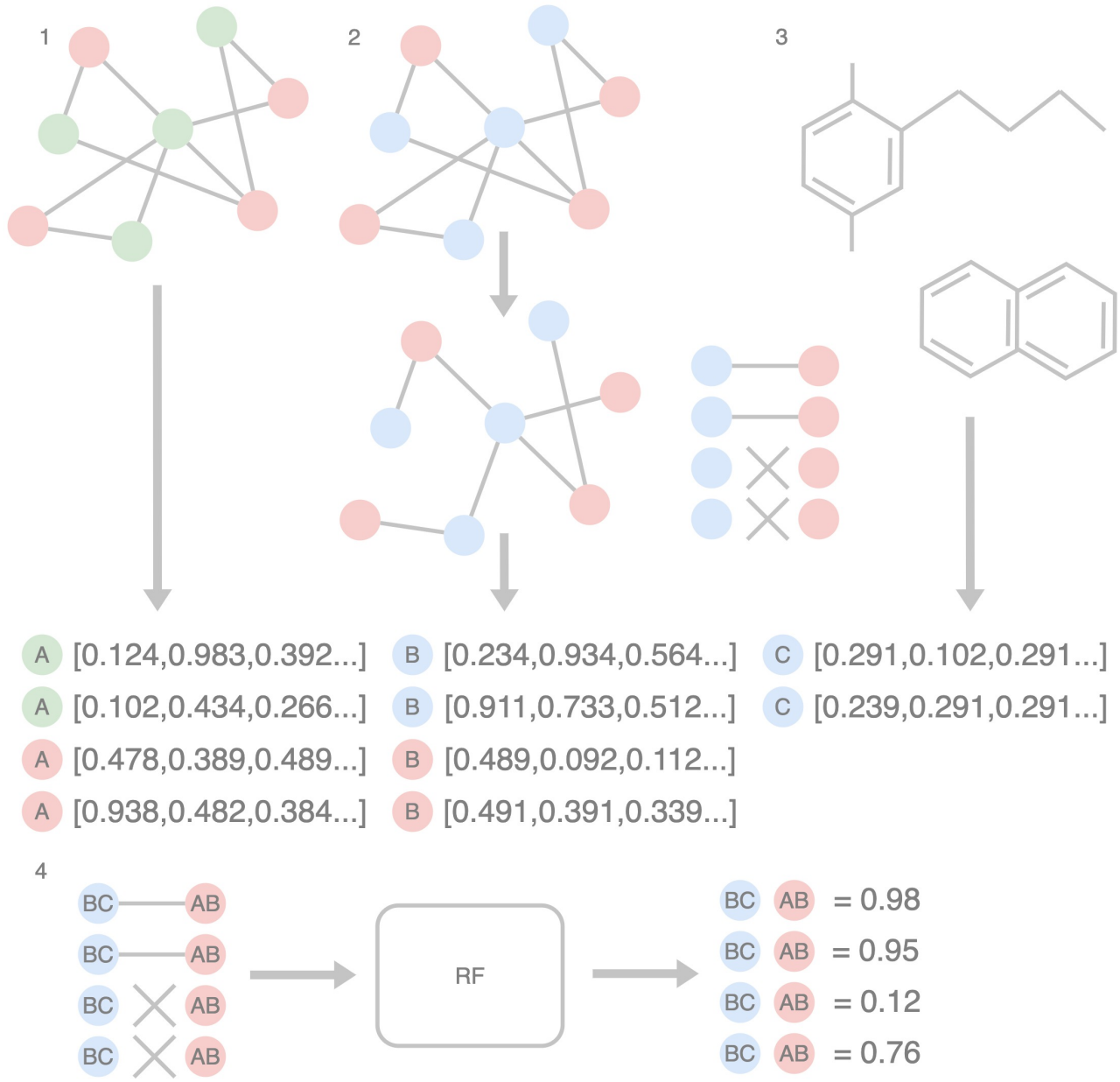


Figure 28: Workflow for model generation for $(Disease) - [RegulationNegative] \rightarrow (SmallMol)$ edge. First, we split the graph into two subgraphs, a disease-gene subgraph (Red and green nodes respectively), and a disease-mol subgraph (red and blue respectively). 1) We generate embeddings for disease and gene nodes (Embeddings A). 2) We remove a subset of the edges of the disease-mol subgraph and generate node embeddings for disease and molecules (Embeddings B). 3) We generate embeddings based on the molecular substructure (Embeddings C). We now have two embeddings for each disease, and two for each molecule. To summarise, Embeddings A describes the neighbourhood or proximity of disease and genes in a disease-gene regulatory network (how close is one disease to another). Embeddings B describes the neighbourhood or proximity of a disease and molecule in a disease-mol regulatory network. Embeddings C describe the physicochemical similarity of compounds. 4) We combine the two embeddings for each disease and molecule (AB and BC respectively), and use the edges removed from the disease-mol network to train a model capable of predicting if a link between a disease and molecule should exist. In the example below, we can see there are two known disease-gene pairs (the regulation has been stated in PathwayStudio), and two random disease-gene pairs. We can see the model has predicted the final pair to actually be an unreported regulation link.

## Drug Combinations

We ingested FDA approved drugs from the DrugCombDB dataset,[21] mapped to the drugs in our knowledge graph by PubChem CID. As multidrug therapies exist within the dataset (more than two drugs), we split these into all possible two pair combinations of these multidrug combinations, and individually added these to the graph. Next, we trained a link-prediction model to predict this FDA Combination relationship between two drugs. This model differed from the previous models (used to predict down-regulation of diseasees, genes and cellular processes). The previous models generated embeddings of the bipartite subgraph of the edge it was predicting, alongside an additional embedding based on an additional network. To clarify, the model to predict $(Disease) - [RegulationNegative] - (Drug)$ edges used embeddings based on a $(Disease) - [RegulationNegative] - (Drug)$ subgraph alongside embeddings based on a $Disease - RegulationNegative - Protein$ subgraph. In contrast, this model trained to predict the $(Drug) - [FDACombination] - (Drug)$ edge did not generate embeddings from a $(Drug) - [FDACombination] - (Drug)$ bipartite graph, and only a $(Drug) - [FDACombination] - (Drug)$ subgraph. As the GraRep embedding is transductive, one can only generate embeddings if the node exists within the graph. Thus, using embeddings based on $(Drug) - [FDACombination] - (Drug)$ would limit predictions to the 572 drugs currently approved for combinatiorial use. In contrast, using the $(Drug) - FDACombination - Drug$ subgraph permitted predictions for 2861 drugs. We investigated the effect of using the FDA combination edges alongside the drug-disease regulatory edges, which yielded a 2 percent increase in AUC. We deemed the decrease in utility and drug scope more important than this increase in model performance.

We calculated transitive probability of treatment $p_{T_{AB}}$ using the equation below, where $p_{A_{down}}$ is the probability drug A down-regulates the disease, $p_{B_{down}}$ is the probability drug B down-regulates the disease and $p_{C_{AB}}$ is the probability drug A and drug B are FDA-approved (complementary):

$$p_{T_{AB}} = p_{A_{down}} \times p_{B_{down}} \times p_{C_{AB}}$$

## Blood Brain Barrier Permeability Prediction

To predict if a drug was BBB+ or BBB-, we created an simple predictive model. For training data, we used Adenot and Lahana's dataset.[15] Once mapped onto molecules in our graph, there were 1220 BBB+ drugs and 291 BBB-. This class imbalance threatened to dominate the predictions towards the majority class (BBB+). To prevent this, we used Synthetic Minority Oversampling Technique (SMOTE) to artificially synthesise negative samples. We used pretrained Mol2Vec embeddings, based on Morgan substructures, trained on 20 million molecules from ZINC[19] as features for our model. The pretrained Mol2Vec embeddings had a dimension of 300. Since the minority class had only 291, we applied principal component analysis (PCA) to reduce the embeddings until 0.95 of the variance was still conserved. This created reduced embeddings of 35 dimensions; a significant reduction in vector size, whilst only losing 0.05 percent of information within the embeddings. We used an non-optimized random forest classifier for this model. To validate this model we used 10-fold stratified cross validation. Importantly, to prevent data leakage, SMOTE and PCA were applied within each fold of the training data. Artificially generated samples were only used for training and not testing.

A previous study[10] showed drug indications and side effects can be used to predict if a drug is BBB+ or BBB-. Their logic dictates that any drug that exhibits clinical phenotypes associated with the brain must be able to pass or circumvent the BBB via any mechanism. The study used grouped brain-associated phenotypes as features. We replicated this analysis by replacing Mol2Vec with embeddings based on i) drug-disease regulatory network and ii) drug-disease association network. We reasoned that the embeddings would be able to capture the brain-associated diseases and therefore would be informative to the model. The replacement with both of these embeddings represented an 5 percent decrease in AUC score.

We calculated the transitive probability of treatment $p_{C_{AB}}$ using a similar equation as in Drug Combinations, however extended to include the probability of BBB permeability for each drug ($p_{A_{BBB+}}$ and $p_{B_{BBB+}}$ for drug A and drug B respectively).

$$p_{T_{AB}} = p_{A_{down}} \times p_{B_{down}} \times p_{C_{AB}} \times p_{A_{BBB+}} \times p_{B_{BBB+}}$$

## Graph Enrichment

To determine the mechanistic pathways misregulated in the disease state of DIPG, we added Gene Ontology concepts as nodes in our graph, connected to appropriate proteins via an association edge. We calculated enriched concepts for both molecular functions and biological processes. We calculated the absolute count of each concept and for each of the 19387 diseases in the knowledge graph. Specifically, we determined the absolute count for each concept that followed the path $(Disease) - [Regulation*] \rightarrow (Protein) - [Associates] \rightarrow (BiologicalProcess/MolecularFunction)$. For a specific disease, we converted the absolute concept count to Z-score based on permutation tests. Tests were based on all diseases which had a degree distribution of regulatory edges within 20 percent of the disease in question (DIPG). We used a p-value of 0.05. We then cross-referenced the processes enriched in the disease state, against the processes regulated by each drug. We performed the same enrichment for molecular functions and pathways.

To calculate the Z-score $Z_{d_A}$ for Gene Ontology term A for a certain disease, the following equation was used, where $d_A$ is the absolute count of term A, and $\bar{d_r}$ and $\sigma_r$ are the mean and standard deviation respectively of term A based on random permutation tests.

$$Z_{d_A} = \frac{d_A - \bar{d_r}}{\sigma_r}$$

## Regulatory Graph

We filtered the main knowledge graph on DMGs and the monodrug drug set that could be connected via 1,2 or 3 hops. The monodrug set consisted of all known and the top 50 predicted inhibitors for all DMGs (see ). Importantly, all paths had to follow the directed route between disease and drug via 0-2 protein intermediates. We filtered on only regulation, protein modification and drug target edges, allowing any edge in the knowledge graph.

## Degree Weighted Path Count

To validate the predictive power of the i) overall regulatory network and ii) individual paths, we created a degree weighted path count model originally developed by Himmelstein et al..[11] We made a few changes to Himmelstein's methodology. Whilst they utilized a novel self-testing procedure, we used a simple 80/20 train/test split. We used a random forest classifier in place of a logistic classifier. Himmelstein ranked predicted edges according the prior probability of connection for a node pair. We chose not to use this, as the principle aim of this analysis was to rank the predictive power of each path in the graph (that connects drugs and diseases), not to use the probability of predictions to infer new repurposable candidates.

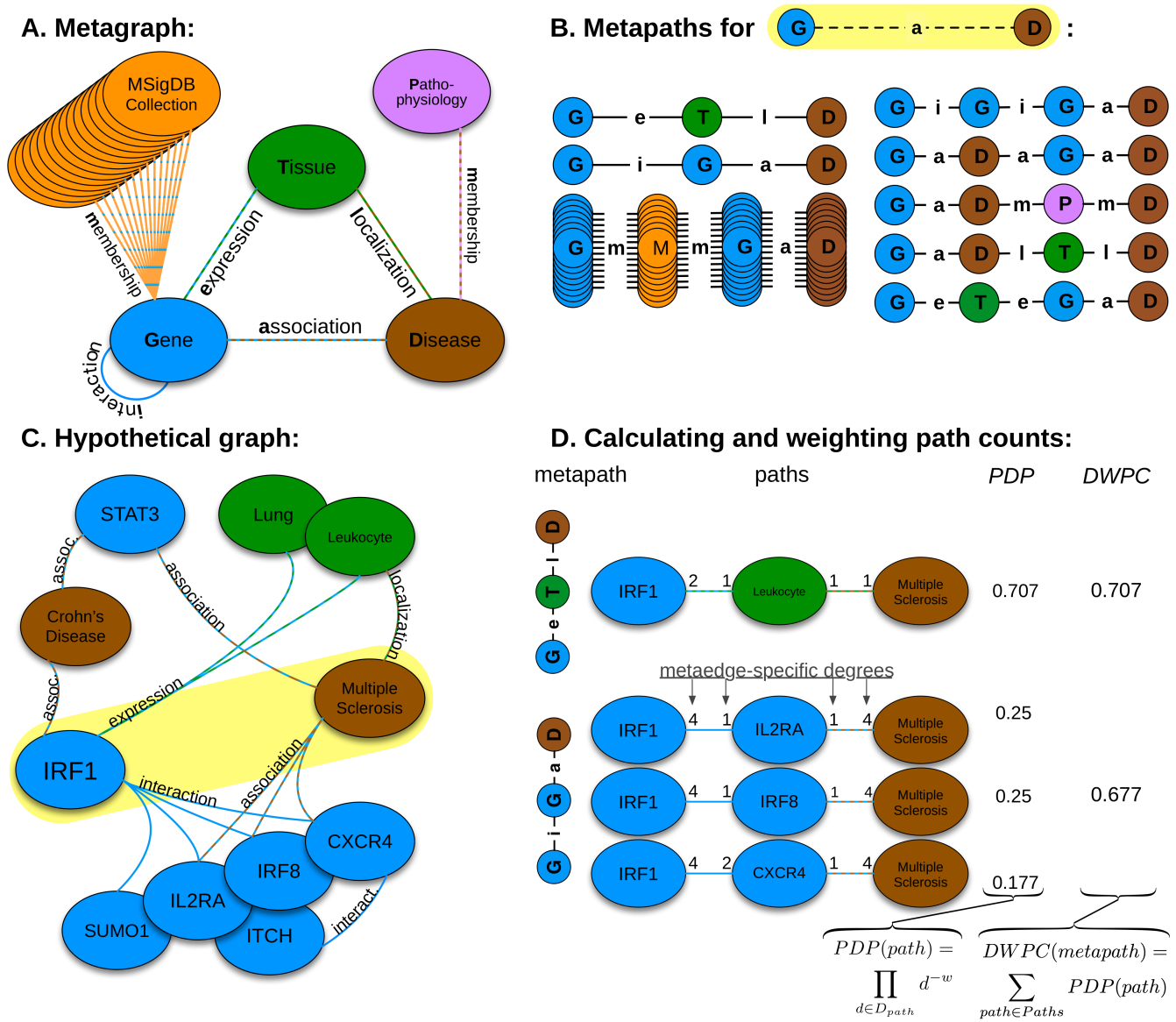


Figure 29: Degree weighted path count methodology. Weighted path count methodology pioneered by Himmelstein et al. Each path between a source and a target node (in this case IRF1 and Multiple Sclerosis respectively). One can see how the more ubiquitous gene-interacts-gene-associates-disease (G-i-G-a-D) path, the count is normalised and the individual contribution of each path is less. The less frequent gene-expressed-tissue-localises-disease (G-e-T-l-D) has a greater contribution to the degree weighted path count. This illustration was taken from the original paper.[11]

## Enrichment Analysis

To evaluate the validity of our regulatory subgraph, we evaluated the functional pathways and biological relevance of the known regulators (both inhibitors and activators) of DMGs with the subgraph. We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analyses using Enrichr,[22] with a p-value of 0.05.

## Gene Importance in Pathways

To evaluate the importance of each of the above genes (that regulate DMGs), within all of the above enriched pathways, we used eigenvector centrality as a measure of importance. Eigenvector centrality computes the centrality for a node based on the centrality of its neighbours.[12] We first took the top 25 pathways above the p-value threshold. For each KEGG pathway, we found all genes within the pathway. Next, we constructed a protein-protein interaction (PPI) networks for each pathway. We found protein-protein interactions from our biomedical knowledge graph. All protein modification interactions types that have been reported in literature were permitted in this network. Cleavage, (de)phosphorylation, (de)acetylation, (de)methylation, (de)ubiquitination, (de)sumoylation, and modifcations of unknown mechanisms were allowed. The PPI network constructed was homogeneous: all protein modification types were conflated into one interaction edge type. For each pathway, the centrality of every gene was calculated.

## Drug Proximity

First, we generated node embeddings for drugs in the regulatory graph using Standford's C++ implementation of node2vec.[13] We then computed the cosine distance between all drug pairs. We clustered drugs using k-means clustering, deciding the optimum number of clusters using the elbow method. Next, for each drug pair, we computed absolute overlap of drug targets, and overlap of drug regulators (proteins that followed the path $((Drug) - [Regulation*] \rightarrow (Protein)$ in the knowledge graph). For each drug pair, using these drug targets and regulators, we computed the overlap of biological processes that these genes shared. We performed the same for pathway overlap of these shared genes. For each drug pair, we computed overlap of drug indications (overlap of diseases in the path $((Drug) - [RegulationNegative] \rightarrow (Disease))$ in the knowledge graph). For each drug pair, we calculated the Tanimoto distance using the respective drug's Morgan Fingerprint using rdkit.[23] Drug-drug interactions were ingested from DrugBank, and represented the overlap of other overlapping drugs interactions for each drug pair. Finally we plotted the distribution of the cosine distance for all synergistic and antagonistic drug pairs, alongside random pairs.

The aim of this analysis was to assess if node embeddings can be used to ascertain if a 1-dimension distance between embeddings could be used to indicate drug combination viability. GraRep, the embedding method used in the embedding model, performs significantly better than node2vec on most down-stream machine learning tasks. The method uses neighbourhood information 1-3 hops away from the source node. We initially tried this method, however found the plots of cosine distance no longer had a clear negative correlation. The distances were staggered - we believe artifacts of the three structural layers of the embeddings. Node2vec, based on random sampling, did not exhibit this behaviour.

Cosine similarity $\cos(\mathbf{s}, \mathbf{t})$ of source and target embeddings $\mathbf{s}$ and $\mathbf{t}$ was calculated as shown below, where $\mathbf{s}_i$ is the $i^{th}$ element in the source embedding:

$$similarity = \cos(\mathbf{s}, \mathbf{t}) = \frac{\mathbf{st}}{\|\mathbf{s}\|\|\mathbf{t}\|} = \frac{\sum_{i=1}^{n} \mathbf{s}_i \mathbf{t}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{s}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2}}$$

# Conclusion

There are currently no satisfactory means of treatment of DIPG. Using a network polypharmacology approach, we highlighted their are many immunosupressive therapeutics with additional virological regulatory mechanisms. Many of these drugs are appropriate for combinatorial use as part of a multidrug regimen against DIPG. This workflow was a high-level review of therapeutic options. Amongst many admissions, it did not take in account the transient nature of the viral infection process, dose-response, pathways, and drug-drug interactions. Further research should be conducted to assess the pharmacological viability of these drug combinations. Alongside this analysis, we published a regulatory graph for researchers to use to understand the pathophysiology of the disease and regulatory targets for possible repurposable drug candidates.

# Data Licensing

The Neo4j knowledge graph uses multiple proprietary and public data sources. This graph was never intended to be used for commercial gain. The graph includes relationships from DrugCentral and SIDER (CC SA licenses). Pathway Studio is a separate commercial database and analytical solution.

# Availability

The data used in this project was primarily proprietary data from Elsevier. As such, the data can not fully be provided to replicate this workflow.

# References


[1] *Existing Drug May Treat the Deadliest Childhood Brain Tumor, Stanford-led Study Finds.* URL: `https://www.stanfordchildrens.org/en/about/news/releases/2015/existing-drug-may-treat-the-deadliest-childhood-brain-tumor`.

[2] Daniel Scott Himmelstein et al. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing". In: *eLife* 6 (2017). DOI: `10.7554/elife.26726`.

[3] Finn Womack, Jason Mcclelland, and David Koslicki. "Leveraging Distributed Biomedical Knowledge Sources to Discover Novel Uses for Known Drugs". In: (Nov. 2019). DOI: `10.1101/765305`.

[4] Yadi Zhou et al. "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2". In: *Cell Discovery* 6.1 (2020). DOI: 10.1038/s41421-020-0153-3.

[5] Feixiong Cheng, Kovács István A., and Barabási Albert-László. "Network-based prediction of drug combinations". In: *Nature Communications* 10.1 (2019). DOI: 10.1038/s41467-019-09186-x.

[6] Ehsaneddin Asgari and Mohammad R. K. Mofrad. "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics". In: *Plos One* 10.11 (Oct. 2015). DOI: 10.1371/journal.pone.0141287.

[7] Sabrina Jaeger, Simone Fulle, and Samo Turk. "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition". In: *Journal of Chemical Information and Modeling* 58.1 (Oct. 2018), pp. 27–35. DOI: 10.1021/acs.jcim.7b00616.

[8] Shaosheng Cao, Wei Lu, and Qiongkai Xu. "GraRep". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM 15* (2015). DOI: 10.1145/2806416.2806512.

[9] A.h. Calvert. "Discovering selective anticancer drugs: Succeeding slowly is better than failing fast". In: *Annals of Oncology* 26 (2015), p. ii1. DOI: 10.1093/annonc/mdv079.1.

[10] Zhen Gao et al. "Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and drug indications". In: *Bioinformatics* (2016). DOI: 10.1093/bioinformatics/btw713.

[11] Daniel S. Himmelstein and Sergio E. Baranzini. "Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes". In: *PLOS Computational Biology* 11.7 (Sept. 2015). DOI: 10.1371/journal.pcbi.1004259.

[12] Marco Schmitt. "Bonacich (1987): Power and Centrality: A Family of Measures". In: *Schlüsselwerke der Netzwerkforschung Netzwerkforschung* (Apr. 2018), pp. 59–61. DOI: 10.1007/978-3-658-21742-6_14.

[13] Aditya Grover and Jure Leskovec. "node2vec". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). DOI: 10.1145/2939672.2939754.

[14] Pavel Sidorov et al. "Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data". In: *Frontiers in Chemistry* 7 (2019). DOI: 10.3389/fchem.2019.00509.

[15] Marc Adenot and Roger Lahana. "Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates". In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004). PMID: 14741033, pp. 239–248. DOI: 10.1021/ci034205d. eprint: https://doi.org/10.1021/ci034205d. URL: https://doi.org/10.1021/ci034205d.

[16] *Gene Ontology Resource*. URL: http://geneontology.org/.

[17] Michael Kuhn et al. "The SIDER database of drugs and side effects". In: *Nucleic Acids Research* 44.D1 (2015). DOI: 10.1093/nar/gkv1075.

[18] Oleg Ursu et al. "DrugCentral 2018: an update". In: *Nucleic Acids Research* 47.D1 (2018). DOI: 10.1093/nar/gky963.

[19] Teague Sterling and John J. Irwin. "ZINC 15 – Ligand Discovery for Everyone". In: *Journal of Chemical Information and Modeling* 55.11 (Sept. 2015), pp. 2324–2337. DOI: 10.1021/acs.jcim.5b00559.

[20] Xiang Yue et al. "Graph embedding on biomedical networks: methods, applications and evaluations". In: *Bioinformatics* (Apr. 2019). DOI: 10.1093/bioinformatics/btz718.

[21] Lei Deng et al. "DrugCombDB: a comprehensive database of drug combinations toward network medicine and combination therapy". In: (2018). DOI: 10.1101/477547.

[22] Edward Y Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". In: *BMC Bioinformatics* 14.1 (2013), p. 128. DOI: 10.1186/1471-2105-14-128.

[23] greg Landrum. *Open-Source Cheminformatics Software*. URL: http://rdkit.org/.