école normale supérieure paris–saclay

ARIA Graph

Drug combination and their effect on patient health

Nathalie Heinzelmeier Chloé Scholent

November 20, 2025

Contents

Introduction

Context

In order to treat certain diseases efficiently, an approach relying on poly-pharmacy is sometimes used. Poly-pharmacy consists in the administration of two or more drugs in order to treat the symptoms of a condition [1]. This approach is often used among the elderly [2]. Therefore, it is necessary to consider all the risks associated with the use of multiple drugs at once such as higher mortality, falls, cognitive impairment, etc... Such side effects may occur between combinations of drugs, and do not always rely on side effects observed on individual drugs. The abundance of side effects is even an argument against poly-pharmacy approaches [3], specifically targeting the elderly who represent a vulnerable population more susceptible to negative poly-pharmacy consequences. However, it remains a form of treatment for cases of multi-morbidity.

Pharmaceutical compounds available on the market are typically evaluated through clinical trials designed for specific diseases. As a result, their individual pharmacodynamic and pharmacokinetic properties, as well as their side effect profiles, are generally well characterized. Each drug can interact with one or several molecular targets, most commonly proteins. To modulate the therapeutic efficacy or mitigate the adverse effects of a treatment, clinicians may prescribe drug combinations. A well-known example is antiretroviral therapy for HIV infection, which commonly involves the co-administration of three or more drugs. In this context, many clinical trials were conducted to determine the best drug combination that maximizes efficacy while minimizing adverse drug interactions and side effects. In contrast, the vast majority of potential drug combinations remain poorly characterized. This is mainly due to the vast number of possible rare combinations which turns clinical trials expensive and difficult to conduct.

Some studies already try to understand and demonstrate the presence of a given side effect in certain drug combinations. However, as the number of possible combinations appears to be impossible to determine, there is a lack of knowledge to predict possible side effects for a pair of drugs which has not been studied and tested thoroughly before. This gap in the medical knowledge prevents the combination of certain drugs for a treatment as their combined negative effect is unknown, making the prediction of side effects in combined drugs a useful tool.

Drugs typically act by targeting one or more proteins. When multiple drugs are administered concurrently, the effect of one compound on its target can be either amplified or weakened as a result of interactions involving other proteins affected by the co-administered drugs. Such drug–drug interactions may also give rise to unexpected adverse effects, which can threaten the life of the patient. Therefore, a systematic analysis of the interactions between proteins targeted by different drugs represents an important area of research, with significant implications for both drug safety and therapeutic efficacy.

This study aims at presenting the graph convolutional network Decagon [4] which predicts potential side effects for random drug combinations from the combination of protein-protein interaction and drug-protein interaction networks. In order to understand the functioning of this model, this study shall determine and illustrate underlying rules of the model which allow for good predictions.

Strategy

In order to predict side effects in drug combinations, graph neural networks have been implemented such as Decagon in 2018 [4] which became the basis for further research on the topic.

Decagon is an end-to-end graph convolutional network created in order to predict potential side effects in specific drug combinations. The model relies on an important dataset acquired from various databases examining drug-drug interactions, drug-protein interactions as well as protein-protein interactions in order to identify patterns for predicting recurring and similar side effects. The dataset as well as the implementation of Decagon are hosted on Stanford Network Analysis Project.

Decagon is a graph convolutional network. Such a model is able to make local predictions within data structured as a graph. In the case of this study, the graphs taken into account are of various natures and are combined in order to render a multi-modal graph with 2 types of nodes (drugs and proteins) and 3 types of edges (protein-protein interaction, drug-protein interaction and drug-drug interaction). The goal of this study is to characterize the emergence of combined side effects resulting from the concurrent administration of two drugs, based on the interactions among the proteins they target. For each type of combined side effect, we aim to analyze and describe the structural properties of the corresponding protein–protein interaction (PPI) networks that arise from the union of drug target profiles. This approach allows to assess whether the occurrence of drug combination side effects can be significantly predicted by examining the topological and functional characteristics of the protein interaction networks associated with their targets.

To address this question, the properties of PPI networks constructed from the interactions among proteins targeted by drugs that share a similar combined side effect or that are chosen at random are inspected. These empirical networks are then compared with model networks that possess similar global properties but are generated according to defined construction rules. By iteratively refining these rules based on our empirical findings, we aim to reproduce the structural characteristics of the original networks, thereby identifying the underlying mechanisms that may explain the observed drug combination side effects. The implementation of the various codes used for this study is available on Github.

1 | Dataset Description

1.1 - Dataset Source

The data used for this study is the same as the one present on the Stanford Network Analysis Project page. The data consists in 6 csv files each containing one type of information:

- **bio-decagon-ppi.csv:** This file provides every protein-protein interaction and contains 715,612 edges.
- **bio-decagon-combo.csv:** This file documents every known polypharmacy side effect for a given pair of drugs. It contains 4,649,441 drug combinations associated with a specific side effect.
- **bio-decagon-mono.csv:** This file contains the side effects associated with the drugs when taken individually. It contains 174,977 rows of data.
- **bio-decagon-targets.csv:** This file presents drug-protein interactions with 18,690 edges representing the targeted proteins for each drug.
- **bio-decagon-targets-all.csv:** This file is an extended version of the one above with a total of 131,034 edges.
- **bio-decagon-effectcategories.csv:** This file contains a list of side effect categories with the name of the effect and an associated code for a total of 562 identified effects.

1.2 - Data Analysis

The dataset examines various types of side effects. However, not all of them are represented in equal proportions. For instance, certain diseases — such as monogenic disorders — are underrepresented in the dataset. This under-representation may be due to the longer time typically required to diagnose these conditions, compared to other diseases like cardiovascular disorders, which are more frequently identified and therefore more prevalent in the dataset (Figure 1.1).

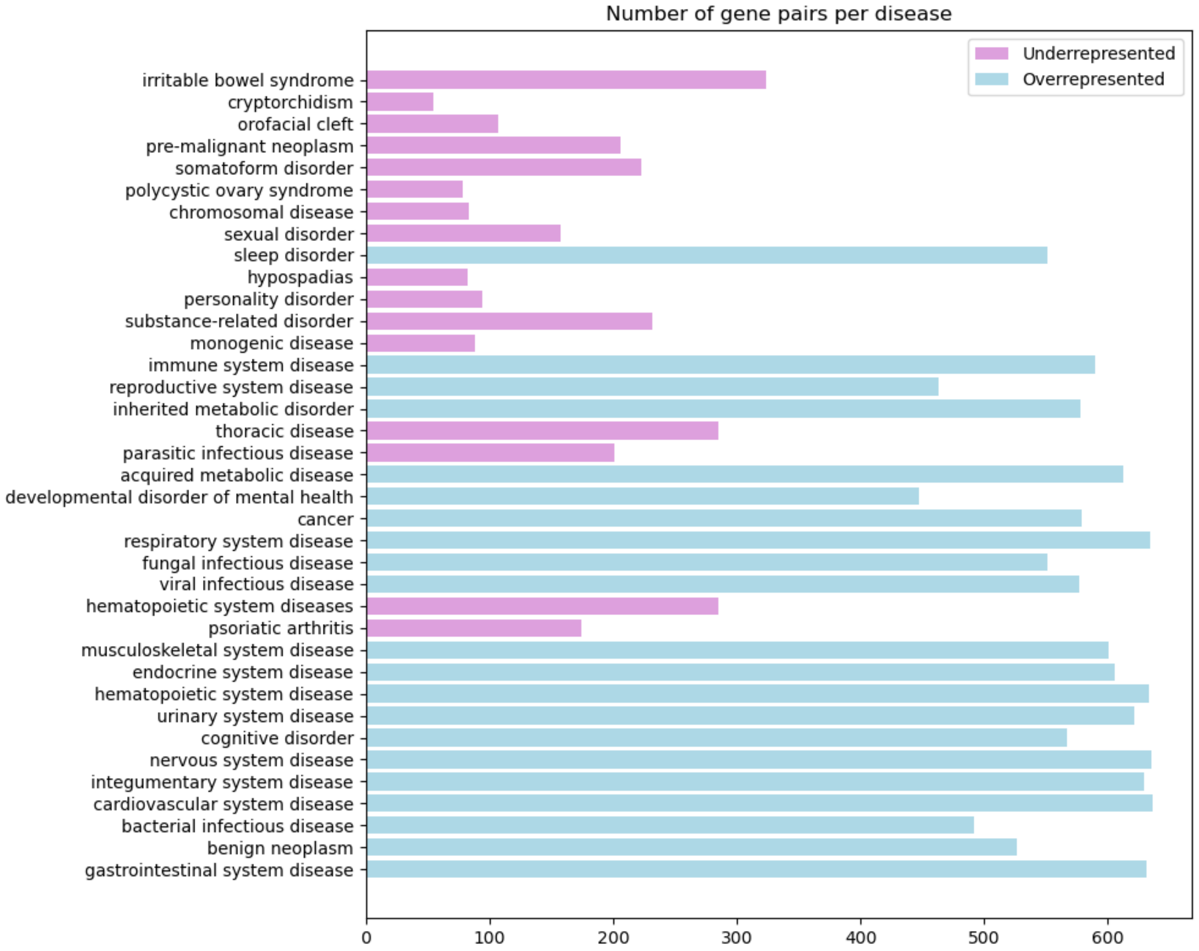


Figure 1.1: Number of couple of drugs depending on the combined side effect

1.3 - Data subsets considered for analysis

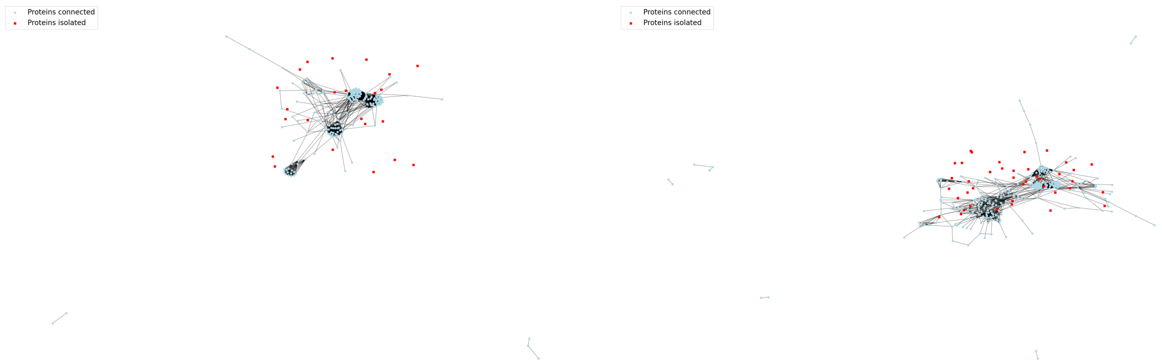
In order to isolate specific rules for the modeling of Decagon's predictions, various graphs have been examined from the provided dataset. For a matter of readability, the figures in this study present smaller portions of studied graphs but the background computation considered larger amounts of data for a more precise result.

Each sub-graph analyzed corresponds to a specific condition defined by the type of side effect resulting from the combination of two drugs. For clarity, only few drugs at a time were selected from all drugs meeting that condition. For each selected drug, the set of targeted proteins is compiled, as well as the proteins that interact with these targets. The resulting individual networks were then combined to generate a global sub-graph, providing a comprehensive view that includes: which combination of drugs respond to the specific condition, the proteins targeted by the selected drugs, and the associated protein–protein interaction networks. In this study, several types of sub-graphs were generated based on distinct selection criteria. The first type was constructed from random selections of drugs, whereas the second type focuses on drugs that, when combined, produce similar side effects associated with cancer.

2 | Graph Properties Analysis

2.1 - Protein–Protein Interaction Graph

When selecting the drugs of interest, the proteins targeted by these drugs are collected, as well as the potential protein–protein interactions among them. The protein interactions are then visualized for drugs chosen either randomly or based on the presence of side effects associated with cancer (Figure 2.1).



(a) Drugs chosen with cancer side effect

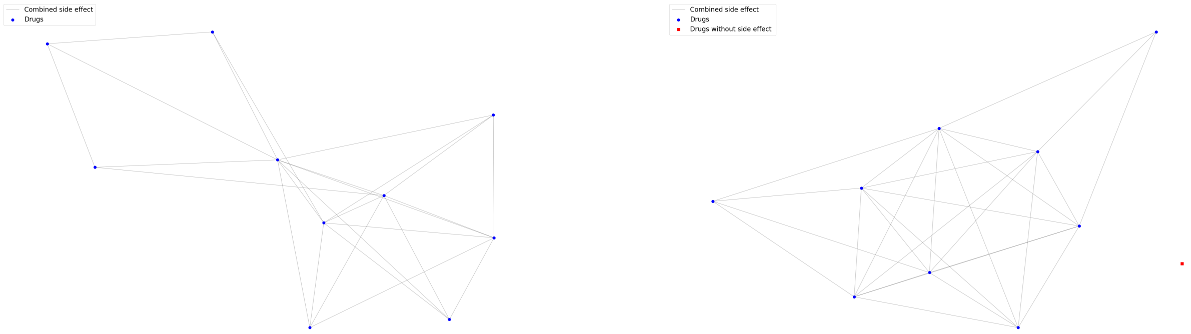
(b) Drugs chosen randomly

Figure 2.1: Protein–protein network of the selected drugs.

It can be observed that when drugs are chosen randomly, the protein network contains more isolated interaction hubs. In contrast, selecting drugs that share a common side effect appears to promote interactions among the proteins targeted by the chosen drug pairs.

2.2 - Drug Interaction Graph

Drugs, whether chosen randomly or according to a specific criterion, can also be represented as a graph, where edges between two drugs indicate the presence of a side effect when the drugs are combined (Figure 2.2).



(a) Drugs chosen with cancer side effect

(b) Drugs chosen randomly

Figure 2.2: Combined side effect between the selected drugs.

To better visualize the interactions, the graph is limited to ten selected drugs. Here it can be observed that when drugs are selected randomly, some have no side effects in combination with the other chosen drugs. This results comes from the fact that not every drug needs to give rise to a side effect when combined with another one. Moreover, it is also important to note that research on poly-pharmacy is incomplete and that not every drug pair in existence has been tested yet. The dataset understudy contains a restricted amount of drugs and therefore a fixed number of drug combinations. Only a few of those combinations present side effects. However, it is interesting to note that drugs which share no known side effect can still target similar proteins and proteins that interact between each other (Table 2.1).

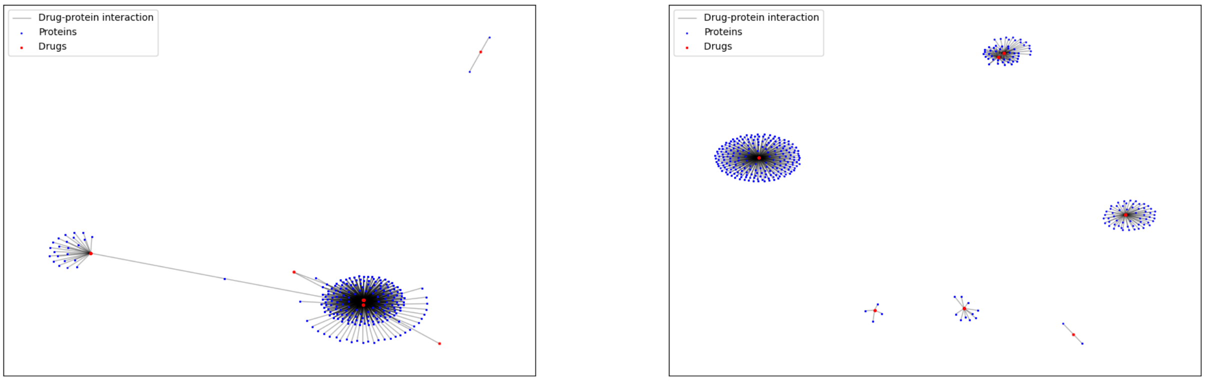
Total drugs	1774
Total possible drug pairs	1572651
Pairs triggering side effects	14247
Pairs triggering no known side effect	1558404

Table 2.1: Available data about drugs

2.3 - Protein–Drug Interaction Graph

Each drug selected targets one or several specific proteins. Through protein–protein interactions, a drug can influence the effect of another drug by modulating the effect of the protein targeted by that drug. Such modulation may occur either directly, via immediate interactions between the targeted proteins, or indirectly, through intermediary proteins within the network. To illustrate this phenomenon, a graph representing the drugs alongside their direct protein targets is represented, highlighting potential pathways for drug–drug interactions mediated by protein networks (Figure 2.3).

One can observe that when drugs are selected randomly, each drug tends to target proteins that are not directly targeted by other drugs. In contrast, when drugs share a common combined side effect, they appear to have more overlapping protein targets. Consequently, the effect of one drug may influence that of another through direct interactions with the proteins targeted by the latter, highlighting potential mechanisms of drug–drug interactions.



(a) Drugs chosen with cancer side effect

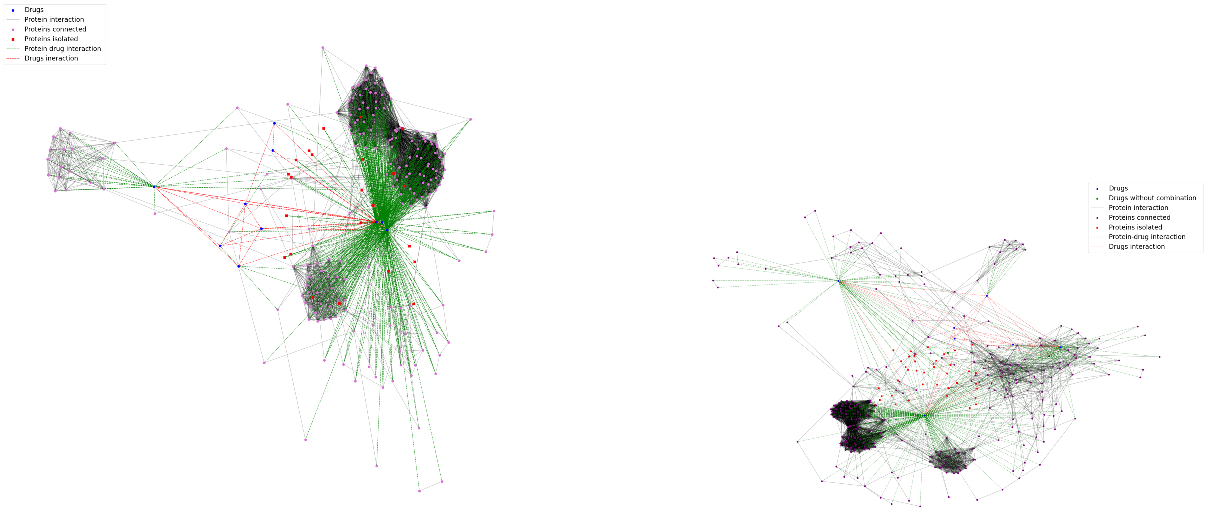
(b) Drugs chosen randomly

Figure 2.3: Direct targets of the drugs

2.4 - Global Graph

Decagon utilizes graphs that integrate all of the previously described networks. Specifically, to understand how protein networks associated with drugs may explain the emergence of combined side effects, Decagon combines three types of networks: the drug–drug side effect network, the protein–protein interaction network, and the drug–target interaction network. It is possible to construct similar integrated graphs to illustrate these relationships.

Figure 2.4 shows that proteins can interact indirectly with multiple drugs. This illustrates the complexity of drug–drug interactions and the challenges in predicting potential side effects between two drugs. Notably, when drugs are selected randomly, there appears to be fewer indirect interactions compared to when drugs are chosen based on a shared combined side effect.



(a) Drugs chosen with cancer side effect

(b) Drugs chosen randomly

Figure 2.4: General network

2.5 - Statistics

To study the general graph, 6 features are gathered and summarized in Table 2.2:

- **Number of nodes:** Represents the size of the graph and helps normalize the metrics when comparing two graphs.
- **Number of edges:** Is essential for understanding network topology and normalization.
- **Density of the graph:** Defined as the ratio of the number of edges present in the graph to the total number of possible edges between all nodes. Higher density indicates more interactions between nodes, while lower density suggests a sparser network.
- **Average degree:** This indicates the mean number of connections per node. A higher average degree suggests that nodes are more interconnected.
- **Minimal degree:** It identifies the least connected node.
- **Maximal degree:** It identifies the most highly connected node and highlights potential hubs in the network.

Number of nodes	19081
Number of edges	715612
Density	0.0039
Average degree	75.0078
Maximal degree	2519
Minimal degree	0

Table 2.2: Features extracted from the general graph

To investigate how these network features differ between randomly selected drugs and drugs sharing a common combined effect, 50 graphs were generated for each condition and statistical analyses were conducted on them(Figure 2.5).

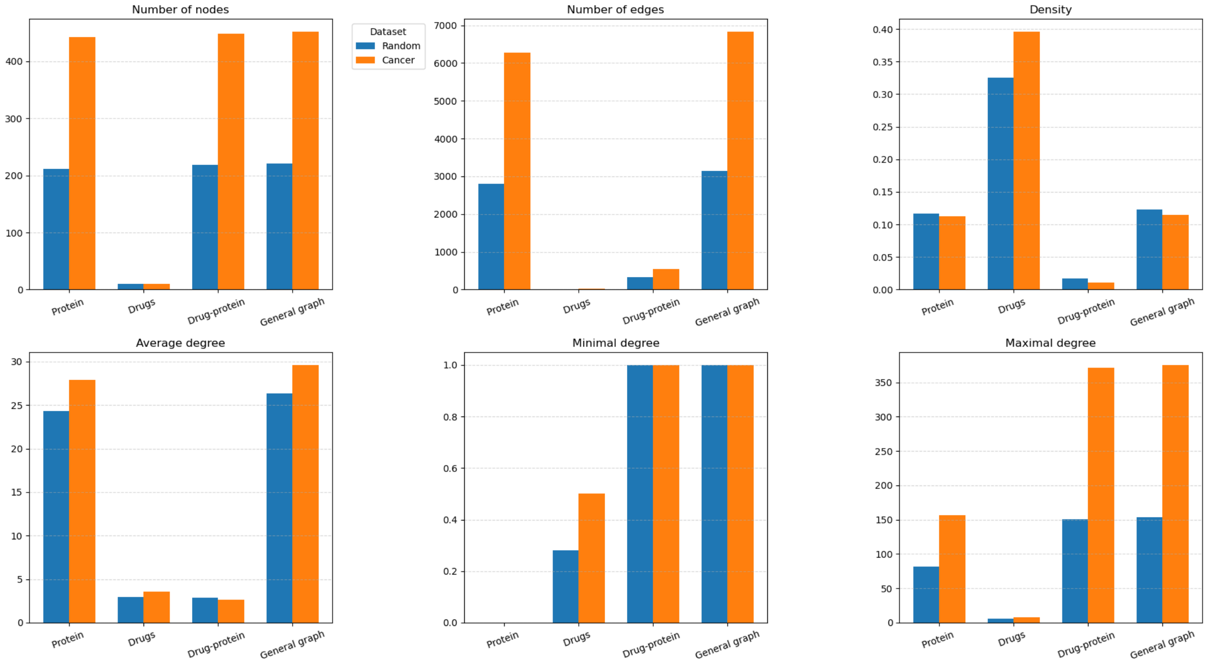


Figure 2.5: Statistical analyses of network features depending on the drugs selected

It can be observed that when drugs sharing a similar combined side effect are selected, the resulting networks contain a greater number of proteins compared to those obtained from random drug selection. This suggests that drugs which tend to trigger cancer when combined may affect a larger set of proteins. This observation is consistent with biological knowledge, as cancer is often the result of disruptions in multiple protein interactions and signaling pathways.

Moreover, the number of edges in the global network is primarily driven by protein–protein interactions. This indicates that when a drug targets a protein, it can also indirectly influence many other proteins through interaction cascades. This effect is more pronounced when the selected drugs are associated with cancer, suggesting that cancer-related drugs tend to target proteins that are highly connected within the network—so-called hub proteins.

The density of the graph is similar between both selected conditions, indicating that while the total number of proteins and interactions increases for cancer-associated drugs, the overall level of network connectivity remains comparable. In contrast, randomly chosen drugs tend to have fewer connections overall.

Direct interactions between drugs and proteins, as well as protein–protein interactions, are relatively rare. Therefore, each observed connection is likely to be meaningful. The average degree is comparable between the two conditions, with proteins showing on average around 25 interactions with other proteins. Some proteins, however, remain isolated with no detectable interactions.

Finally, the maximal degree is higher in the condition involving cancer-associated drugs. This increase results from the presence of drugs that interact directly with multiple proteins, acting as central hubs within the network and potentially playing key roles in mediating drug–drug interactions.

Another analysis was conducted by taking into account 15 sets of 80 drugs randomly selected, with each set having unique drugs taken from the global graph (Figure 2.6). Those sets allow the computation and comparison of average shared proteins and average protein-protein interaction per set. Shared proteins corresponds to the number of proteins interacting with at least two drugs in the set. The protein-protein interaction corresponds to the interaction between proteins for pairs of drugs targeting different proteins. The average is calculated based on the interaction for pairs of drugs within a given set.

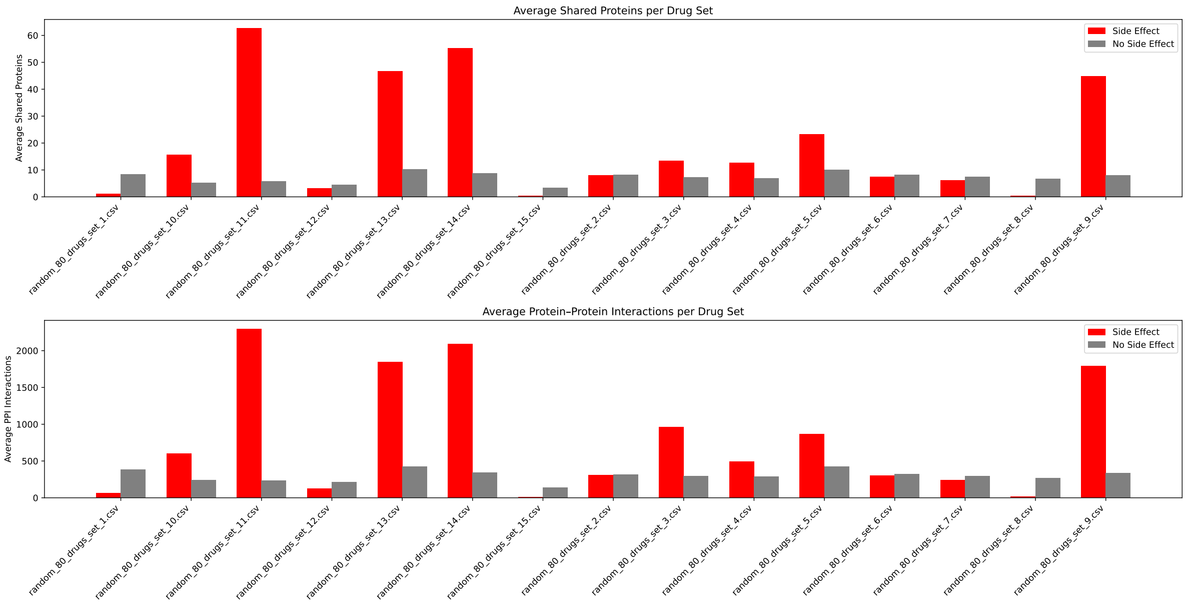


Figure 2.6: Statistical analyses of 80-random-drug sets

First, a link between shared proteins and protein interaction can be identified. For every set of drugs with a high number of shared proteins, a high level of protein interaction can also be observed. Moreover, most of the interaction seem to happen for drug pairs triggering at least one side effect. This may indicate that in order to have a side effect in a drug pair, there needs to be a minimal level of interaction between targeted proteins. Eventually, the drug pairs triggering no known side effect seem to present a constant amount of shared proteins and protein-protein interaction. However, those numbers representing averages, it is possible that specific situations concerning drug pairs with no known side effect may have been occulted.

Average shared proteins (side-effect combos)	16.36
Average PPI (side-effect combos)	635.90
Average shared proteins (no-side-effect combos)	7.40
Average PPI (no-side-effect combos)	308.92

Table 2.3: Features extracted from the general graph

It is interesting to note that this greater presence of PPI and shared proteins for drug pairs presenting a side effect is also present when analyzing the general graph taken into account by Decagon (Table 2.3). This analysis seems to illustrate one of the possible mechanism for the apparition of side effects during drug combination. Of course, the type of side effect arising from a certain drug combination is linked to the specific target proteins. This mechanism is simplified in this analysis and tested on a larger scale. Decagon presents the benefit of being able to take into account parameters about individual proteins, making the prediction task even more precise.

3 | Comparison with models of graphs

To better understand the phenomenon which explains the graphs obtained previously, it is possible to use data extracted from them (number of nodes and edges for instance) to model new ones which should resemble them following a set of rules.

3.1 - Heterogeneous Random model

The first model used to reproduce Decagon's graph is a random model relying on some fixed information. This random model is not an Erdös–Rényi model which would present a uniform distribution of edges and only one type of nodes. In the considered random model, two types of nodes and three types of edges are created, but the edges distribution is random within a certain range.

A first generation is computed relying on a rather small number of nodes and edges:

```
Nodes count:
    Drugs: 10
    Proteins: 50

Edge counts chosen by the model:
  Drug-Drug edges: 5
  Drug-Protein edges: 118
  Protein-Protein edges: 1235

Average shared proteins between linked drugs: 3.00
Average protein-protein interactions between linked drugs: 127.00
```

Figure 3.1 presents the generated graph resulting from this first configuration. A comparison can be established with results from the general graph in Table 2.3. For instance, the ratio between the average number of shared proteins and the average PPI is of about 39 for pairs of drugs presenting side effects. This ratio is rather similar for the generated random graph as it reaches 42. Of course, other generations present a ratio either much higher or much smaller because of the random nature of the generation. In addition, the average protein interaction is usually a high number as proteins interacts with several others at a time and raise this average quickly.

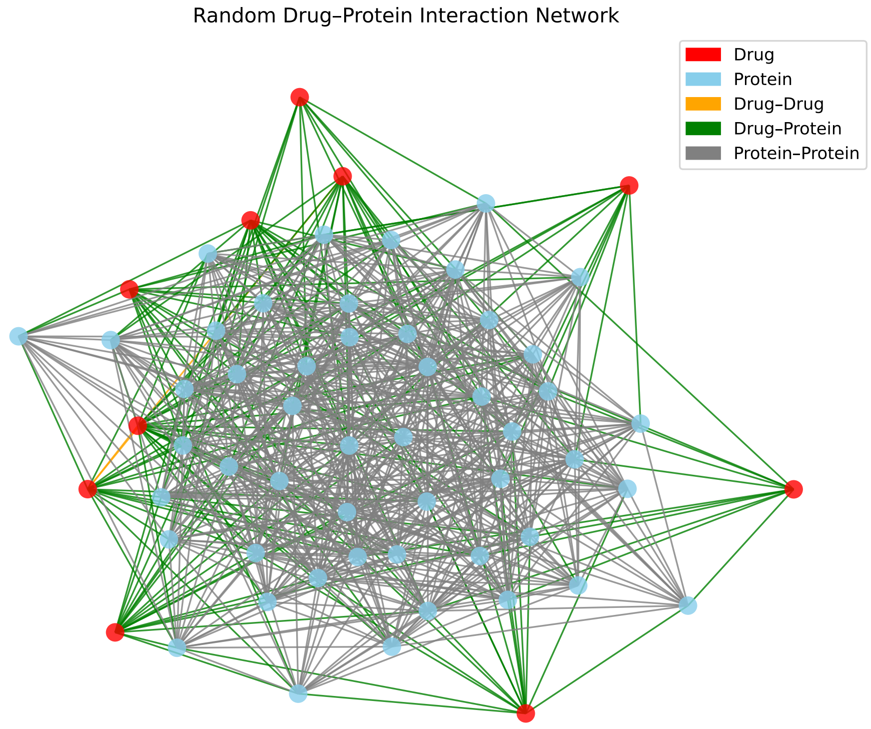


Figure 3.1: Example of random graph generated

Another random model was computed with more drugs and proteins nodes. In order to have a minimally accurate model, the random number of generated edges is chosen within a range of fixed values which follow a certain power law: there needs to be more drug-protein interaction than drug drug interaction, and more PPI than drug-protein interaction.

```
Edge and Node counts:
    Drug-Drug edges: 14257
    Drug-Protein edges: 72711
    Protein-Protein edges: 92907
    Number of nodes: 20855
    Number of edges: 179875

Graph Statistics
    Density: 0.0008
    Average degree: 17.25
    Maximum degree: 80
    Minimum degree: 2

Drug Interaction Metrics
    Average shared proteins between linked drugs: 0.09
    Average protein-protein interactions between linked drugs: 0.86
```

One thing which can be noted right away is that the number of drug-protein edges and protein-protein edges is rather close compared to those from the previous random graph. This could explain the very small density of 0.0008 in the graph, compared to that of Table 2.2, which is also very small but not on the same scale. Moreover, the ratio between shared proteins and PPI disappears for this generation and even presents the opposite of previous observations on the high PPI for pairs of drugs creating side effect. Finally, the distribution of degrees within the graph seems to be very homogeneous, ranging from 2 to 80 with an average degree of 17. This degree distribution is completely erroneous when considering the real degree distribution from the general graph in Table 2.2.

In other words, it is possible to identify some important rules to model this multi modal graph: degree distribution should allow for nodes with higher degrees, mostly because of drug-protein or protein-protein interaction, which impacts the final density of the graph. Moreover, the ratio between shared proteins and PPI should be around 40, with an important average of protein-protein interaction. Another model can be used to try to model a better degree distribution: a Barabási model.

3.2 - Barabási model

The Barabási–Albert (BA) model is used to describe networks characterized by a power-law degree distribution. In such networks, a small number of nodes (known as hubs) have a very high number of connections, while most nodes are connected to only a few others. This addition of hubs could emulate key proteins from the graph which possess a very high degree and interact with many other proteins.

To construct such graphs, the model begins with a small set of connected nodes. At each time step, a new node is added to the network and forms links with a fixed number of existing nodes. The probability that a new node connects to an existing one is proportional to the degree of that existing node—meaning that nodes with many connections are more likely to attract new links. Since hubs were observed in the previous network visualizations, and their number increases when focusing on drug combinations associated with cancer-related side effects, the choice was made to apply the Barabási–Albert model to reconstruct the network structure. Statistical analysis were conducted on the properties of this graph generated 50 times randomly (Figure 3.2).

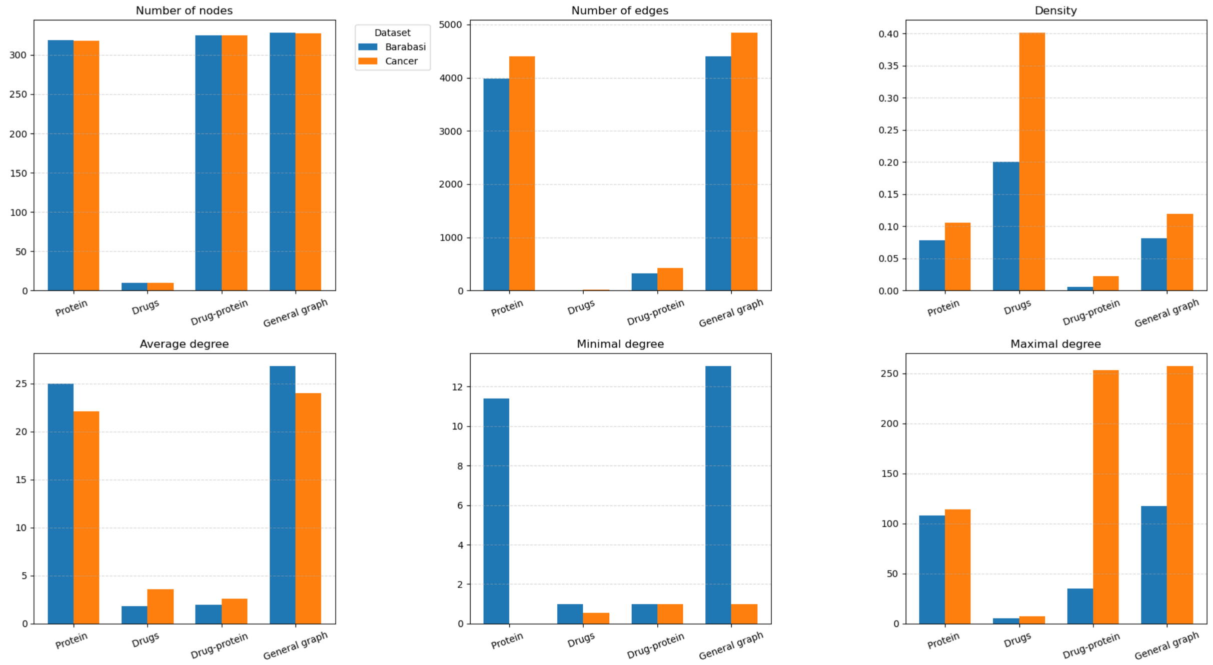


Figure 3.2: Properties of graphs from drugs selected with combined side effect of hypertension with or without anxiety

Here, one can observe that each graph contains the same number of nodes, as expected, since the Barabási–Albert graph backbone is generated based on the statistical properties of the original graph built from the selected drugs that share side effects similar to cancer. However, the Barabási model shows a smaller number of edges, resulting in a significantly lower graph density compared to the original drug network. This indicates that drugs associated with cancer tend to form many diverse connections in the real network. In contrast, the Barabási–Albert model displays fewer drug–protein interactions, suggesting that, in the actual network, drugs act more as hubs than the proteins they target.

4 | Co-occurrence between side effects

Some pairs of drugs can induce different side effects when combined. Studying the co-occurrence of these side effects can provide insights into how specific drug combinations trigger particular adverse reactions. The main article studied [4], deserves a part of its analysis to that of co-occurrences of certain side effects, whether this co-occurrence is over or under represented within the dataset. Therefore, this section focuses on drug pairs that, when combined, are associated with both hypertension and anxiety or fever which are effects identified by Decagon's team.

4.1 - Number of co-occurrence

It is possible to analyze pairs of drugs that are associated with a combined side effect related to hypertension. From this subset, the drug pairs that are also associated with anxiety are selected. It can be observed that 60.6% of the drug pairs linked to hypertension also co-occur with anxiety. This finding is consistent with the observations reported in the reference article. On the contrary, the co-occurrence between hypertension and fever shows a rather weak link between the two side effect. For 8677 pairs producing hypertension, only 5.3% co-occurs with fever, which is also consistent with the observations from the reference article.

4.2 - Results

It is interesting to analyze the properties of the graphs when drugs are selected as having side effect associated with hypertension and hypertension with anxiety as shown in figure4.1.

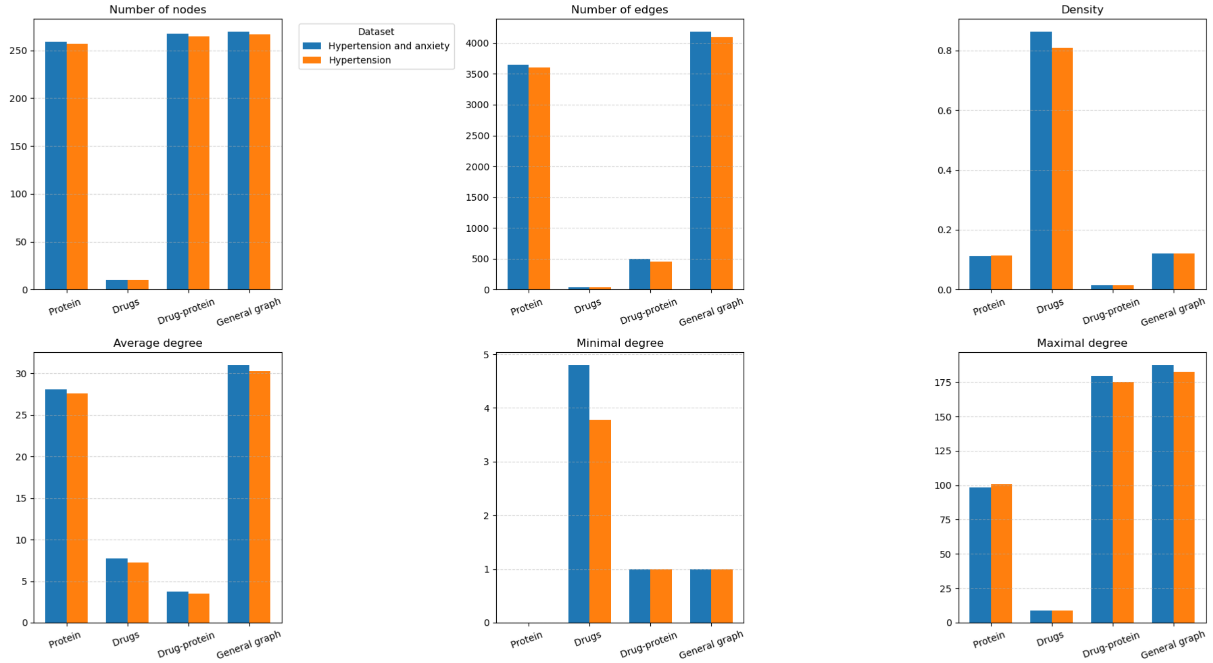


Figure 4.1: Properties of graphs from drugs selected with combined side effect of hypertension with or without anxiety

It can be observed that the number of protein nodes increases slightly when drugs are associated with the co-occurring side effects hypertension and anxiety, compared to anxiety alone. This indicates that co-occurrence tends to increase the number of proteins directly targeted by the drugs. Furthermore, the average degree of the network also increases when selecting drugs that share the combined side effect of hypertension and anxiety. This suggests that the network becomes more connected, which is consistent with the idea that co-occurrence implies a higher level of interaction among the involved proteins.

Conclusion

In conclusion, this study successfully constructed sub-graphs from a selected set of drugs which, when combined, exhibit similar side effects. These sub-graphs were compared with randomly generated graphs to statistically analyze and contrast their structural properties. Using these properties, models based on specific generative rules were generated, and they provided deeper insight into the structural and functional characteristics of the drug–side effect networks.

Furthermore, the analysis was extended to more complex cases involving drugs that share two similar side effects when combined. Future work could explore additional sub-graphs derived from drugs associated with other shared or co-occurring side effects, thereby broadening the understanding of the underlying mechanisms governing drug–side effect relationships.

Bibliography

[1] Nashwa Masnoon et al. "What is polypharmacy? A systematic review of definitions". In: *BMC Geriatrics* 17.1 (Oct. 10, 2017), p. 230. ISSN: 1471-2318. DOI: 10.1186/s12877-017-0621-2. URL: https://doi.org/10.1186/s12877-017-0621-2 (visited on 10/24/2025).

[2] Robert L Maher, Joseph Hanlon, and Emily R Hajjar. "Clinical consequences of polypharmacy in elderly". In: *Expert Opinion on Drug Safety* 13.1 (Jan. 1, 2014). Publisher: Taylor & Francis _eprint: https://doi.org/10.1517/14740338.2013.827660, pp. 57–65. ISSN: 1474-0338. DOI: 10.1517/14740338.2013.827660. URL: https://doi.org/10.1517/14740338.2013.827660 (visited on 10/24/2025).

[3] Turabian Jl and Jose Turabian. "Polypharmacy is an Indicator of Bad Practice and Low Quality in General Medicine". In: *Journal of Quality in Health Care & Economics* 2 (July 19, 2019). DOI: 10.23880/JQHE-16000130.

[4] Monica Agrawal Marinka Zitnik and Jure Leskovec. "Modeling polypharmacy side effects with graph convolutional networks". In: *Bioinformatics* (2018).