National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

Network Structure vs Chemical Information in Drug-Drug Interaction Prediction

Diploma Thesis

of

GEORGE D. KEFALAS

Supervisor: Andreas-Georgios Stafylopatis
Professor, NTUA

Co-Supervisor: Dimitrios Vogiatzis
Researcher, NCSR "Demokritos"

Athens, July 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

Network Structure vs Chemical Information in Drug-Drug Interaction Prediction

Diploma Thesis

of

GEORGE D. KEFALAS

Supervisor: Andreas-Georgios Stafylopatis
Professor, NTUA

Co-Supervisor: Dimitrios Vogiatzis
Researcher, NCSR "Demokritos"

Approved by the examination committee on 12th July 2022.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Andreas-Georgios Stafylopatis	Stefanos Kollias	Giorgos Stamou
Professor, NTUA	Professor, NTUA	Professor, NTUA

Athens, July 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

Copyright © – All rights reserved.
George D. Kefalas, 2022.

The copying, storage and distribution of this diploma thesis, exall or part of it, is prohibited for commercial purposes. Reprinting, storage and distribution for non - profit, educational or of a research nature is allowed, provided that the source is indicated and that this message is retained.

The content of this thesis does not necessarily reflect the views of the Department, the Supervisor, or the committee that approved it.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
George D. Kefalas

12th July 2022

Abstract

This thesis compares network information against chemical information for the problem of drug interaction prediction. Drug interactions can be studied as a network, with the drugs represented as nodes, and interactions as edges. There is also the additional information of the chemical formula of each drug. We apply two embedding mechanisms, mol2vec and node2vec, on the problem of predicting Drug-Drug Interactions (DDIs). These mechanisms, respectively, convert drugs into vectors using the chemical information of the underlying chemical compound and the network information from the graph of drug interactions. Our goal is to compare Single Link Prediction models that are based on each embedding method by exploring the topological features of the drug interactions graph that make each approach more efficient in making correct predictions. We base our experiments on the DrugBank data set and use various computational chemistry tools such RDKit and PubChem, along with NetworkX, in order to create the chemical and structural embeddings for each drug.

Keywords

Drug-Drug Interaction Prediction, DrugBank, Network Embeddings, Chemical Embeddings, Graph Topological Features, mol2vec, node2vec

Περίληψη

Σε αυτή την εργασία μελετάμε τις αλληλεπιδράσεις μεταξύ φαρμακευτικών ουσιών. Πιο συγκεκριμένα, εξερευνούμε τη πληροφορία που μπορεί να αντλήσει κάποιος μέσα από τον γράφο γνώσης που προκύπτει αν θεωρήσουμε κάθε φάρμακο ως κόμβο και κάθε αλληλεπίδραση δύο φαρμάκων ως ακμή μεταξύ των αντίστοιχων κόμβων — τη γνώση αυτή ονομάζουμε δικτυακή ή γραφική πληροφορία. Παράλληλα αξιοποιούμε τον χημικό τύπο από τη δραστική ουσία ενός φαρμάκου για να αποτυπώσουμε τις χημικές ιδιότητές του και να εξάγουμε αυτό που ονομάζουμε στη συνέχεια ως χημική πληροφορία. Στόχος μας είναι να συγκρίνουμε τη δικτυακή πληροφορία με την χημική πληροφορία πάνω στο πρόβλημα της εκτίμησης αλληλεπιδράσεων μεταξύ φαρμάκων. Για να το κάνουμε αυτό, επιστρατεύουμε δύο μεθόδους που εμπνεύστηκαν από τον χώρο της μηχανικής μάθησης, και συγκεκριμένα από τον τομέα επεξεργασίας φυσικής γλώσσας: το mol2vec και το node2vec. Οι μέθοδοι αυτοί μας βοηθούν να παράξουμε διανυσματικές αναπαραστάσεις — τόσο βάσει της χημικής πληροφορίας, όσο και της δικτυακής πληροφορίας — από τα φάρμακα που έχουμε στη διάθεσή μας. Αξιοποιούμε αυτές τις αναπαραστάσεις για να αναπτύξουμε και να εκπαιδεύσουμε ταξινομητές βασισμένους σε νευρωνικά δίκτυα οι οποίοι με δεδομένο ένα ζεύγος φαρμάκων ως είσοδο, δίνουν ως έξοδο μια εκτίμηση για το αν τα φάρμακα αυτά αλληλεπιδρούν. Μέσω της σύγκρισης των αποτελεσμάτων των ταξινομητών μας σε ένα κατάλληλο σύνολο ελέγχου κατορθώνουμε να παράξουμε χρήσιμα συμπεράσματα για τα γραφοθεωρητικά χαρακτηριστικά που προσδιορίζουν για ένα φάρμακο του γράφου γνώσης (γράφος γνωστών αλληλεπιδράσεων) αν είναι προτιμότερο να μελετηθεί βάσει της χημικής ή της δικτυακής πληροφορίας που υπάρχει διαθέσιμη για αυτό. Για τα πειράματά μας χρησιμοποιούμε τη βάση δεδομένων DrugBank και στη πορεία της εξόρυξης και προετοιμασίας των δεδομένων μας επιστρατεύουμε διάφορα εργαλεία της υπολογιστικής χημείας, όπως τα RDKit και PubChem καθώς και τη βιβλιοθήκη NetworkX για να επεξεργαστούμε υπολογιστικά γράφους δεδομένων.

Λέξεις Κλειδιά

Αλληλεπιδράσεις Φαρμάκων, Δικτυακές Αναπραστάσεις Φαρμάκων, Χημικές Αναπαραστάσεις Φαρμάκων, Γραφοθεωρητικά Χαρακτηριστικά Αλληλεπιδράσεων Φαρμάκων, DrugBank, mol2vec, node2vec

to my parents

Acknowledgements

This endeavor would not have been possible without the invaluable guidance and inexhaustible patience of my co-supervisor, Dr. Dimitrios Vogiatzis. I would also like to express my deepest appreciation to my friends and family for their support and encouragement through the course of my studies.

> Humans cannot create anything out of nothingness. Humans cannot accomplish anything without holding onto something. After all, humans are not gods.
>
> *Kaworu Nagisa*

Athens, July 2022

George D. Kefalas

Table of Contents

List of Figures

List of Tables

Σύνοψη της Εργασίας

0.1 Εισαγωγή

Στόχος της παρούσας εργασίας είναι η μελέτη των αλληλεπιδράσεων μεταξύ φαρμάκων και, πιο συγκεκριμένα, η σύγκριση της χημικής πληροφορίας που μπορούμε να εξάγουμε από τη μοριακή δομή της δραστικής ουσίας των φαρμάκων με τη δομική (ή δικτυακή, ή γραφική) πληροφορία την οποία εξάγουμε από τον γράφο που προκύπτει αν θεωρήσουμε τα διάφορα φάρμακα ως κόμβους ο οποίοι συνδέονται μεταξύ τους με μια ακμή για κάθε γνωστή αλληλεπίδραση που έχει ανακαλυφθεί. Για να κάνουμε αυτή τη σύγκριση, επιστρατεύουμε τη βάση δεδομένων DrugBank [1] για να εξορύξουμε φάρμακα και αλληλεπιδράσεις ώστε να κατασκευάσουμε τον γράφο γνώσης μας. Ακολούθως, χρησιμοποιούμε δύο μεθόδους που βασίζονται στον αλγόριθμο word2vec [2] από τον χώρο της Επεξεργασίας Φυσικής Γλώσσας: το mol2vec [3] και το node2vec [4]. Οι μέθοδοι αυτοί θα μας βοηθήσουν να κωδικοποιήσουμε κάθε φάρμακο από τον γράφο γνώσης μας σε μια διανυσματική αναπαράσταση από πραγματικούς αριθμούς βάσει της χημικής και της δικτυακής πληροφορίας αντίστοιχα.

Δημιουργούμε τρεις ταξινομητές βασισμένους σε ένα νευρωνικό δίκτυο δύο επιπέδων οι οποίοι παίρνουν ως είσοδο τις διανυσματικές αναπαραστάσεις από: τη χημική πληροφορία, τη δικτυακή πληροφορία και τον συνδυασμό χημικής και δικτυακής πληροφορίας αντίστοιχα. Αφού εκπαιδεύσουμε τα συστήματα αυτά, εξετάζουμε την απόδοσή τους σε ένα σύνολο δειγμάτων ελέγχου που απομονώνουμε από τον γράφο γνώσης μας. Πιο συγκεκριμένα, εξετάζουμε τις επιλογές του κάθε ταξινομητή και μελετάμε διάφορα γραφοθεωρητικά χαρακτηριστικά που τους αντιστοιχούν ώστε να εξάγουμε συμπεράσματα σχετικά με το που υπερτερεί ή υστερεί κάθε μοντέλο. Με αυτό τον τρόπο κατορθώνουμε να περιγράψουμε και να συγκρίνουμε τη συμπεριφορά ενός συστήματος εκτίμησης αλληλεπιδράσεων φαρμάκων που βασίζεται στη χημική γνώση με ένα αντίστοιχο σύστημα που χρησιμοποιεί τη δικτυακή πληροφορία. Ονομάζουμε τον συνδυασμό της χημικής και της δικτυακής πληροφορίας ως *υβριδική πληροφορία*, και το αντίστοιχο μοντέλο ως *υβριδικό ταξινομητή*.

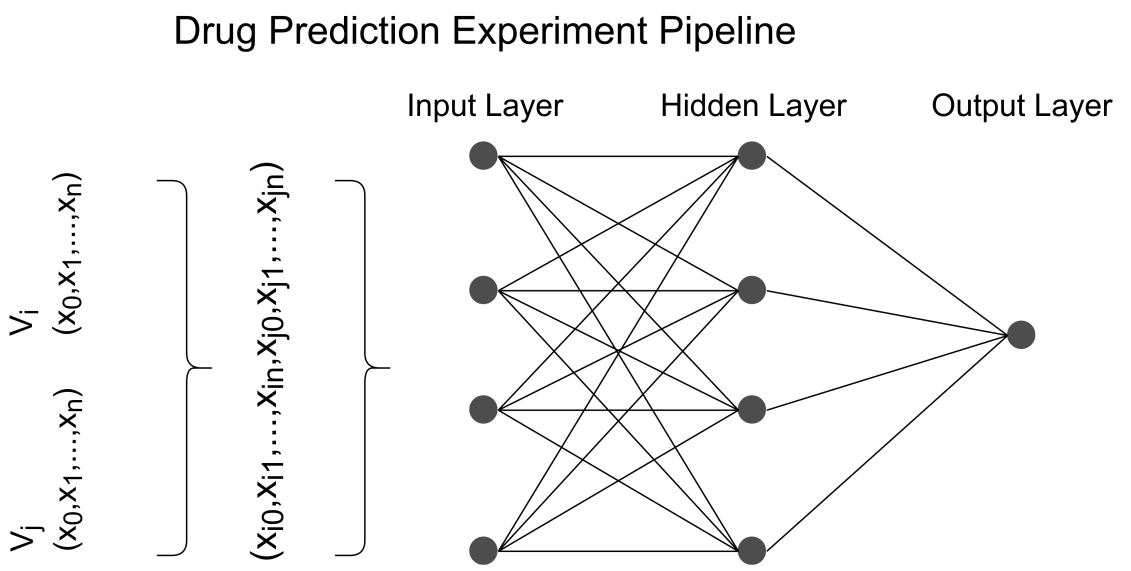


Figure 1. *Δείγματα εκπαίδευσης που προκύπτουν από τις αλληλεπιδράσεις φαρμάκων, με τη μορφή σύζευξης διανυσματικών αναπαραστάσεων, χρησιμοποιούνται για την εκπαίδευση του νευρωνικού δικτύου.*

0.2 Μεθοδολογία

Από το αποθετήριο της DrugBank εξάγουμε συνολικά 14,624 φάρμακα και 1,389,184 μοναδικές αλληπιδράσεις μεταξύ αυτών των φαρμάκων. Από τα δεομένα αυτά κρατάμε εκείνα μόνο τα φάρμακα που παρουσιάσουν τουλάχιστον μία αλληλεπίδραση με κάποιο άλλο φάρμακο, κρατώνας έτσι 3,753 φάρμακα και 1,207,953 αλληλεπιδράσεις — οι αλληλεπιδράσεις μειώθηκαν διότι χρειάστηκε να αφαιρέσουμε από τη γνώση ένα μικρό σύνολο φαρμάκων το οποίο δεν είναι συμβατό με τα εργαλεία που περιγράφουμε στη συνέχεια. Τον τελικό γράφο που προκύπτει ονομάζουμε εφεξής *πλήρη γράφο*.

Για να εξάγουμε τις χημικές αναπαραστάσεις από κάθε φάρμακο, πρέπει πρώτα να βρούμε τον χημικό τύπο του και στη συνέχεια να ανακαλύψουμε τη πληροφορία σχετικά με τη μοριακή δομή της δραστικής του ουσίας. Για τον σκοπό αυτό χρησιμοποιούμε την υπηρεσία PubChem[5], ένα από τα εργαλεία της οποίας δέχεται ως είσοδο το μοναδικό αναγνωριστικό κάθε φαρμάκου από το αποθετήριο της DrugBank και επιστρέφει τον χημικό τύπο του με τη μορφή SMILES [6]. Στη συνέχεια αξιοποιούμε τη βιβλιοθήκη RDKit [7] της Python ώστε να μετατρέψουμε αυτή τη μορφή χημικού τύπου σε μια δομή που ονομάζεται MOL και ουσιαστικά περιέχει κωδικοποιημένη σε πίνακες γειτνίασης τη δομή του χημικού μορίου. Στο σημείο αυτό, είμαστε έτοιμοι να εφαρμόσουμε ένα προεκπαιδευμένο μοντέλο mol2vec[1] για να παράξουμε ένα διάνυσμα 300 διαστάσεων ως τη χημική αναπαράσταση για κάθε φάρμακο στον *πλήρη γράφο*.

Ως προς τις δικτυακές (ή γραφικές) αναπαραστάσεις, δεν χρειαζόμαστε προεπεξεργασία των δεδομένων μας. Χρησιμοποιούμε απευθείας τον αλγόριθμο του node2vec, ο οποίος εκπαιδεύεται πρώτα σε έναν γράφο γνώσης και στη συνέχεια χρησιμοποιείται για να αντιστοιχίσει κάθε κόμβο του σε ένα διάνυσμα 128 διαστάσεων — τη δικτυακή αναπαράσταση.

[1] https://github.com/samoturk/mol2vec/tree/master/examples/models

Εδώ, ωστόσο, υπάρχει ένα σημαντικό πρόβλημα: δεν μπορούμε να εκπαιδεύσουμε ένα μοντέλο node2vec στον *πλήρη γράφο*. Πιο συγκεκριμένα, ο αλγόριθμος του node2vec εξετάζει τη γειτονιά κάθε κόμβου στον γράφο, διατρέχοντας τυχαία μονοπάτια, ώστε να κατασκευάσει διανύσματα τα οποία βρίσκονται σε μικρότερη απόσταση μεταξύ τους (π.χ. ευκλείδια απόσταση) όταν αντιστοιχούν σε κόμβους του γράφου με όμοια γειτονιά. Αυτό σημαίνει ότι εκπαιδεύοντας το μοντέλο στον *πλήρη γράφο* θα πάρουμε διανύσματα τα οποία θα έχουν ενσωματωμένη μέσα τους, σε έναν σημαντικό βαθμό, τη πληροφορία από όλες τις αλληλεπιδράσεις του γράφου. Κάτι τέτοιο θα ήταν απαγορευτικό για τα πειράματα που ακολουθούν, καθώς δεν θα είχαμε τη δυνατότητα να χωρίσουμε τον *πλήρη γράφο* σε σύνολα εκπαίδευσης, επαλήθευσης και ελέγχου. Για να λύσουμε αυτό το πρόβλημα εφαρμόζουμε δειγματοληψία, κρατώντας όλους τους κόμβους και επιλέγοντας τυχαία να διατηρήσουμε μόνο το 1% των ακμών του *πλήρη γράφου*. Το αποτέλεσμα θα ονομάζουμε *γράφο δειγματοληψίας*, και πρόκειται για έναν γράφο με 12,080 ακμές. Όπως γίνεται αντιληπτό στη συνέχεια, ο γράφος που προέκυψε, παρά την ένταση της δειγματοληψίας, κατορθώνει να εκπαιδεύσει τον ταξινομητή που βασίζεται στη δικτυακή πληροφορία με εξαιρετική επιτυχία. Μάλιστα, μέσω της δειγματοληψίας που κάνουμε κατορθώνουμε να εξασφαλίσουμε την εξής σημαντική ιδιότητα για τις δικτυακές αναπαραστάσεις μας: τα διανύσματα προκύπτουν όχι από τις άμεσες αλληλεπιδράσεις μεταξύ των φαρμάκων, αλλά από τα υπόλοιπα γραφοθεωρητικά χαρακτηριστικά που περιγράφουν μια γειτονιά του γράφου — ακριβώς αυτού του είδους τα χαρακτηριστικά θα προσπαθήσουμε στη πορεία να μελετήσουμε και να συγκρίνουμε για να ανακαλύψουμε περιοχές του *πλήρη γράφου* οι οποίες θα καταστούν τη χημική πληροφορία ανώτερη της δικτυακής πληροφορίας (ή το αντίθετο) για την πρόβλεψη αλληλεπιδράσεων μεταξύ φαρμάκων.

Στο σημείο αυτό μπορούμε να εστιάσουμε στα δείγματα εκπαίδευσης και τον τρόπο που χρησιμοποιούνται από τους ταξινομητές μας. Ειδικότερα, θέλουμε να αναπτύξουμε συστήματα τα οποία θα δέχονται ως είσοδο ένα ζεύγος φαρμάκων, και ως έξοδο να έχουν δύο δυνατές τιμές οι οποίες αντιπροσωπεύουν την ύπαρξη ή μη μιας αλληλεπίδρασης μεταξύ των φαρμάκων εισόδου. Επομένως, επιλέγουμε να δίνουμε ως είσοδο στα νευρωνικά μας μοντέλα τη διανυσματική σύζευξη από τις αναπαραστάσεις των φαρμάκων, όπως φαίνεται στο διάγραμμα 1. Το νευρωνικό δίκτυο του σχήματος αποτυπώνει τη δομή του μοντέλου που χρησιμοποιούμε για τους τρεις ταξινομητές με βάση τη χημική, τη δικτυακή, και την υβριδική πληροφορία. Προτού προχωρήσουμε, όμως, με την εκπαίδευση των μοντέλων μας πρέπει να λύσουμε ένα ακόμη πρόβλημα που προκύπτει από το σύνολο των δεδομένων που έχουμε στη διάθεσή μας: δεν γνωρίζουμε κανένα αρνητικό δείγμα. Αναλυτικότερα, θεωρούμε ως θετικό δείγμα εκπαίδευσης ένα συζευγμένο διάνυσμα που προέκυψε από μια πραγματική ακμή του *πλήρη γράφου*, δεν έχουμε όμως αντίστοιχες ¨μη πραγματικές¨ ακμές στον γράφο για να μας δώσουν τα αρνητικά δείγματα. Χρειάζεται, συνεπώς, να κάνουμε άλλη μια δειγματοληψία ώστε να παράξουμε νέες ¨αρνητικές¨ ακμές στον γράφο μας για να παράξουμε αρνητικά δείγματα εκπαίδευσης από την συνένωση των αντίστοιχων διανυσμάτων — φροντίζουμε οι τυχαίες ακμές που παράγουμε να μην υπάρχουν ήδη στον γράφο ως πραγματικές αλληλεπιδράσεις.

Έχοντας ετοιμάσει τα θετικά και τα αρνητικά δείγματα για τον *πλήρη γράφο*, είμαστε έτοιμοι να εκπαιδεύσουμε και, έπειτα, να εξετάσουμε την επίδοση των ταξινομητών μας.

Χωρίζουμε τα δεδομένα μας σε σύνολα εκπαίδευσης, επαλήθευσης και ελέγχου με ποσοστά 65%/5%/30%.

0.3 Πειραματικά Αποτελέσματα

Class	Classifier	Precision	Recall	F1 Score
Negative Samples (Class 0)	mol2vec	0.91	0.89	0.90
	node2vec	0.90	0.94	0.92
	Hybrid	**0.96**	**0.94**	**0.95**
Positive Samples (Class 1)	mol2vec	0.88	0.91	0.89
	node2vec	**0.96**	0.93	0.94
	Hybrid	0.93	**0.95**	**0.94**

(a) *Classification Report για τους τρεις ταξινομητές*

Category	mol2vec	node2vec	Hybrid
Experimental	87	94	96
Approved	85	93	94
Investigational	87	94	95
Vet Approved	88	94	97
Withdrawn	87	96	97
Illicit	87	95	96
Nutraceutical	91	90	97

(b) *Σύγκριση της ποσοστιαίας ακρίβειας των προβλέψεων κάθε μοντέλου για κάθε κατηγορία φαρμάκων*

Figure 2. *Αξιολόγηση της Επίδοσης των Ταξινομητών*

Στην ενότητα αυτή παραθέτουμε τα σημαντικότερα αποτελέσματα από τα πειράματα μας. Ο πίνακας 2α περιέχει το classification report που προκύπτει από την αξιολόγηση των μοντέλων μας στο σύνολο ελέγχου — οι αρνητικές αλληλεπιδράσεις αποτυπώνονται με τη κλάση 0 και οι θετικές αλληλεπιδράσεις από την κλάση 1. Στο παράρτημα A παραθέτουμε τα λεπτομερή στοιχεία σχετικά με τα δεδομένα εκπαίδευσης, επαλήθευσης και ελέγχου για τους τρεις ταξινομητές, ενώ στο παράρτημα B παραθέτουμε τα πλήρη δεδομένα σχετικά με τον πίνακα 2β, παρουσιάζοντας μια αναλυτικότερη αξιολόγηση ως προς την ακρίβεια κάθε μοντέλου για κάθε μοναδική κατηγορία φαρμάκων. Η κατηγοριοποίηση των φαρμάκων δίνεται από τη DrugBank και εξυπηρετεί ώστε να κάνουμε ειδικότερες συγκρίσεις στο σύνολο των φαρμάκων, όπως για παράδειγμα να μελετήσουμε την αξιοπιστία της χημικής και της δικτυακής πληροφορίας ως προς την αναζήτηση αλληλεπιδράσεων όταν εμπλέκονται πειραματικά, ή κτηνιατρικά φάρμακα.

Μπορούμε να διακρίνουμε ότι ο υβριδικός ταξινομητής έχει σημαντικά καλύτερες επιδόσεις σε όλες σχεδόν τις συγκρίσεις με τα άλλα δύο συστήματα. Το γεγονός αυτό υποδεικνύει ότι υπάρχει μη επικαλυπτόμενη γνώση από τις δύο προσσεγίσεις (χημική και δικτυακή πληροφορία) η οποία είναι ωφέλιμο να συνδυαστεί — παρά το γεγονός ότι μεγαλώνει σημαντικά το μήκος της υβριδικής διανυσματικής αναπαράστασης και, κατά συνεπεια, δυσκολεύει η εκπαίδευση των νευρωνικών μας δικτύων με τις εισόδους μεγαλύτερων διαστάσεων. Μόνη εξαίρεση στην υπεροχή του υβριδικού ταξινομητή αποτελεί το μεγαλύτερο precision του node2vec μοντέλου ως πος τα θετικά δείγματα, κάτι που μας οδηγεί στο συμπέρασμα ότι η δικτυακή πληροφορία είναι ισχυρότερη στην αναγνώριση των αρνητικών δειγμάτων και την αποφυγή των false positive προβλέψεων.

Ο πίνακας 3α παρουσιάζει τη σύγκριση της μέσης τιμής από διάφορα χαρακτηριστικά των δειγμάτων ελέγχου, και ο πίνακας 3β περιέχει την απόκλιση Kullback-Leibler ανάμεσα στις κατανομές που ορίζονται από τις σωστές και τις λανθασμένες εκτιμήσεις των ταξινομητών μας. Ο όρος *average mean degree* αναφέρεται στη μέση τιμή από το σύνολο που προκύπτει

	Metric	mol2vec	node2vec	Hybrid
False Prediction	Average Min Degree	518.06	470.07	517.01
	Average Max Degree	1137.02	1143.09	1163.44
	Average Mean Degree	827.54	806.58	840.22
	Average Core Difference	237.17	266.10	242.34
	Average Betweenness Centrality	4.57e-06	6.99e-06	6.76e-06
Correct Prediction	Average Min Degree	519.66	593.45	512.09
	Average Max Degree	1083.25	1143.96	1078.67
	Average Mean Degree	801.45	868.713	795.38
	Average Core Difference	228.50	196.01	231.80
	Average Betweenness Centrality	1.39e-06	1.29e-06	1.44e-06

(a) *Σύγκριση γραφοθεωρητικών χαρακτηριστικών ως προς συνολα δειγμάτων ελέγχου που ορίζονται από την επιλογή (σωστή ή λανθασμένη) κάθε ταξινομητή*

Classifier	Correct	False
Average Degree	0.018	0.047
Min Degree	0.019	**0.199**
Max Degree	0.010	**0.110**
Core Difference	0.013	**0.103**
Betweenness C.	**0.109**	0.071

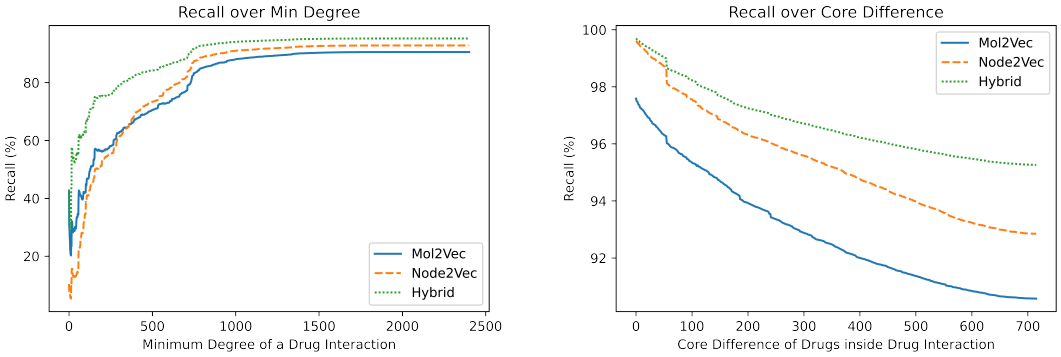
(b) *Απόκλιση KL ανάμεσα στις κατανομές δειγμάτων των ταξινομητών mol2vec και node2vec (τιμές μεγαλύτερες από* 0.1 *εμφανίζονται με έντονη γραφή)*

Figure 3. *Χαρακτηριστικά αποτελέσματα ανά κατηγορία πρόβλεψης (Σωστή ή Λανθασμένη)*

αν υπολογίσουμε το μέσο όρο από τους βαθμούς των κόβων που απαρτίζουν κάθε ζεύγος πιθανών αλληλεπιδράσεων (δειγμάτων) στο σύνολο ελέγχου. Όμοια, ο όρος *average min degree* περιγράφει τη μέση τιμή από τον ελάχιστο βαθμό κάθε ζεύγους στα δείγματα ελέγχου και ο όρος *average max degree* αναφέρεται στη μέση τιμή από τους μέγιστους βαθμούς. Ακόμη, η τιμή του *betweenness centrality* υπολογίζεται μόνο για τις θετικές αλληλεπιδράσεις (δείγματα που αποτυπώνουν πραγματική αλληλεπίδραση μεταξύ φαρμάκων), ενώ ο όρος *core difference* αναφέρεται στην απόλυτη διαφορά των k-cores ανάμεσα στους δύο κόμβους που απαρτίζουν ένα δείγμα αλληλεπίδρασης.

Ακόμη, τα διαγράμματα 4α και 4β αποτυπώνουν τη διαμόρφωση της μέσης τιμής του recall ως προς τον ελάχιστο βαθμό (min degree) και το core difference των δειγμάτων. Πιο συγκεκριμένα, σε αυτά τα διαγράμματα, ο άξονας x καθορίζει ποια δείγματα ελέγχου συμμετέχουν στη διαμόρφωση της τιμής του recall: επιλέγουμε όλα τα δείγματα που έχουν το υπό έλεγχο χαρακτηριστικό (ελάχιστο βαθμό ή core difference) να είναι μικρότερο ή ίσο της τιμής που έχει ο άξονας x για κάθε σημείο της γραφικής παράστασης — επομένως για $x = 0$ δεν συμμετέχει κανένα δείγμα, ενώ για τη μέγιστη τιμή του x (δεξιότερο σημείο της γραφικής παράστασης) συμμετέχουν όλα τα δείγματα.

Το διάγραμμα 4α μας βοηθά να εξάγουμε ένα πολύ χρήσιμο συμπέρασμα: τα δείγματα στα οποία το φάρμακο με τις λιγότερες αλληλεπιδράσεις (κόμβος του γράφου με τον μικρότερο βαθμό στα άκρα μιας ακμής) έχει αριθμό αλληλεπιδράσεων μικρότερο από ένα



(a) *Recall over Min Degree* **(b)** *Recall over Core Difference*

Figure 4. *Διαγράμματα με τη μετρική recall για μοντέλα που βασίζονται στις αναπαραστάσεις της χημικής, της δικτυακής και της υβριδικής πληροφορίας.*

κατώφλι — στο διάγραμμα διακρίνουμε το κατώφλι αυτό να βρίσκεται κοντά στον αριθμό 350 — συνιστά καλύτερο υποψήφιο για τις μεθόδους που βασίζονται στη χημική πληροφορία σε σύγκριση με τις μεθόδους που εκμεταλλεύονται τη δικτυακή πληροφορία. Πρόκειται για μια λογική υπόθεση, καθώς φάρμακα τα οποία έχουν μικρή παρουσία σε έναν γράφο είναι λογικό να μην μπορούν να αξιοποιηθούν από δικτυακές προσεγγίσεις. Ωστόσο, το γεγονός ότι μπορούμε να επιβεβαιώσουμε την υπόθεση πειραματικά και, ακόμα περισσότερο, να ορίσουμε και ένα κατώφλι για τον αριθμό των αλληλεπιδράσεων είναι ιδιαίτερα σημαντικό για την εργασία μας.

Μελετώντας τα διαγράμματα 4α και 4β μπορούμε να εκτιμήσουμε ότι τα μοντέλα μας κάνουν πιο αξιόπιστες προβλέψεις για δείγματα που παρουσιάζουν χαμηλό core difference. Επιπρόσθετα, για χαμηλότερες τιμές στο core difference των δειγμάτων ο ταξινομητής node2vec πετυχαίνει παρόμοια επίδοση με τον υβρδικό ταξινομητή, γεφυρώνοντας το χάσμα που κερδίζαμε συνδυάζοντας τη χημική με τη δικτυακή πληροφορία.

0.4 Συμπεράσματα

Έχοντας ολοκληρώσει τα στάδια της εξόρυξης και επεξεργασίας των δεδομένων μας ώστε να παράξουμε τις χημικές και τις δικτυακές αναπαραστάσεις για τα φάρμακα της DrugBank, την εκπαίδευση των ταξινομητών μας καθώς και τους κύκλους πειραμάτων αναφορικά με την ανάλυση των επιλογών των μοντέλων μας ως προς τα διάφορα χαρακτηριστικά του *πλήρη γράφου*, μπορούμε να συνοψίσουμε σε αυτή την ενότητα τα σημαντικότερα συμπεράσματα που εξάγουμε από την όλη πορεία μας.

Καταρχάς, οφείλουμε να αναγνωρίσουμε ότι το πρόβλημα της αναζήτησης φαρμακευτικών αλληλεπιδράσεων εκτείνεται σε πολύ μεγαλύτερα βάθη από εκείνα που είχαμε την ευκαιρία να εξερευνήσουμε. Χιλιάδες φάρμακα για τα οποία δεν έχουμε ακόμα γνωστές αλληλεπιδράσεις περιμένου ώστε να συμμετέχουν και αυτά στον γράφο γνώσης, μαζί με τις συνδέσεις τους με άλλα φάρμακα που αντίστοιχα αναμένουν την ανακάλυψή τους. Μπορούμε ωστόσο να εκτιμήσουμε τη ποιότητα των δεδομένων της DrugBank, καθώς το αποθετήριο φαρμάκων και αλληλεπιδράσεων στο οποίο βασιστήκαμε αποδείχθηκε από την ανάλυσή μας

εκτενές και ιδιαίτερα χρήσιμο για τα πειράματά μας.

Αξίζει επίσης να σταθούμε στις εξαιρετικές επιδόσεις του αλγορίθμου node2vec. Πιο συγκεκριμένα, το γεγονός ότι παρά την αφαίρεση του 99% των ακμών από τον *πλήρη γράφο* ο αλγόριθμος κατάφερε να παράξει διανυσματικές αναπαραστάσεις που οδήγησαν σε εξαιρετικά επίπεδα μάθησης τα νευρωνικά μας υποδεικνύει τη δύναμη του node2vec να ανακαλύπτει τη πληροφορία από τη γειτονιά κάθε κόμβου ενός γράφου, ακόμα κι αν λείπει η συντριπτική πλειονότητα της αρχικής πληροφορίας, και να οδηγεί σε αξιόπιστες εκτιμήσεις για την ύπαρξη ή μη πιθανών ακμών.

Σε αντίθεση με τις διανυσματικές αναπαραστάσεις, για τις οποίες γνωρίζαμε εκ των προτέρων από τη σχετική βιβλιογραφία ότι έχουν ήδη εφαρμοστεί επιτυχημένα σε συστήματα που προβλέπουν αλληλεπιδράσεις μεταξυ φαρμάκων, δεν είχαμε κάποια ένδειξη προτού εκτελέσουμε τα πειράματα μας για την αντίστοιχη χρησιμότητα των χημικών αναπαραστάσεων μέσα από τον αλγόριθμο mol2vec. Παρόλαυτά, ο σχετικός ταξινομητής κατάφερε να πετύχει εξίσου σημαντικές επιδόσεις στην αναγνώριση θετικών και αρνητικών αλληλεπιδράσεων.

Το κύριο ζητούμενο της εργασίας αυτής, βέβαια, δεν είναι η εξατομικευμένη μελέτη κάθε συστήματος, αλλά η σύγκριση της χημικής με τη δικτυακή πληροφορία για τον σκοπό της πρόβλεψης αλληλεπιδράσεων φαρμάκων. Ως προς τον στόχο αυτό, μπορέσαμε να εξάγουμε ιδιαίτερα χρήσιμα συμπεράσματα, με κυρίοτερα τα εξής:

- Ο ταξινομητής mol2vec παρουσιάζει καλύτερο recall για δείγματα με χαμηλότερο ελάχιστο βαθμό κόμβων, μέχρι ο βαθμός αυτός να ξεπεράσει ένα κατώφλι όπου τότε επικρατεί ο ταξινομητής node2vec. Ακόμη, ο ταξινομητής node2vec επιτυγχάνει πάντα καλύτερες επιδόσεις όταν μελετάται ως προς το core difference των δειγμάτων (αλληλεπιδράσεων).

- Η κατανομή της τιμής του betweenness centrality στα δείγματα για τα οποία έγινε σωστή πρόβλεψη εμφανίζει σημαντικές διαφορές ανάμεσα στα μοντέλα mol2vec και node2vec.

- Χαμηλές τιμές της παραμέτρου core difference καθιστούν τα δείγματα των αλληλεπιδράσεων καλύτερους υποψηφίους για μεθόδους που βασίζονται στη δικτυακή πληροφορία — μάλιστα το μοντέλο node2vec καταφέρνει να φτάσει στην περίπτωση αυτή ακόμη και τις επιδόσεις του υβριδικού μοντέλου.

- Για μεγάλες τιμές της παραμέτρου core difference η χημική πληροφορία μπορεί να συνδυαστεί με τη δικτυακή ώστε να βελτιώσει την επίδοση του ταξινομητή.

Τέλος, γίνεται εμφανές μέσα από τα πειράματά μας ότι το υβριδικό μοντέλο, το οποίο συνδυάζει τη χημική με τη δικτυακή πληροφορία, είναι σαφώς ισχυρότερο από από τα μοντέλα της χημικής και της δικτυακής πληροφορίας ως προς τη σωστή αναγνώριση των θετικών ή αρνητικών αλληλεπιδράσεων.

Κλείνοντας, καταγράφουμε μερικές από τις σημαντικότερες μελλοντικές επεκτάσεις που θα μπορούσαμε να μελετήσουμε ώστε να εξερευνήσουμε περαιτέρω το πρόβλημα της πρόβλεψης αλληλεπιδράσεων μεταξύ φαρμάκων και τη σύγκριση μεταξύ του ρόλου που παίζουν η χημική και η δικτυακή πληροφορία.

- Πρόβλεψη του τύπου της αλληλεπίδρασης: το πρόβλημα που μελετάμε μπορεί να επεκταθεί σε μια πιο σύνθετη έκδοση όπου δεν αρκούμαστε στην αναγνώριση μιας αλλελεπίδρασης μεταξύ δυο φαρμάκων, αλλά προσπαθούμε να εκτιμήσουμε και το είδος της ανάμεσα σε ένα ορισμένο πλήθος από αναγνωρισμένες κατηγορίες — εργασίες όπως η [8] αποτελούν χαρακτηριστικά παραδείγματα αυτής της προσέγγισης. Η μέλετη σε αυτόν τον τομέα θα απαιτεί σαφώς πιο ανεπτυγμένα μοντέλα πρόβλεψης και ίσως οδηγήσει σε πιο σαφή γραφοθεωρητικά στοιχεία αναφορικά με την υπεροχή της γραφικής από τη χημική πληροφορία (και το αντίθετο).

- Αναβάθμιση του τρόπου δειγματοληψίας για αρνητικές αλληλεπιδράσεις: γίνεται σαφές στα πειραματικά μας αποτελέσματα ότι τα μοντέλα μπορούν να αναγνωρίσουν με σημαντικά μεγαλύτερη ευκολία τα αρνητικά δείγματα σε σύγκριση με τα θετικά δείγματα. Θα ήταν χρήσιμο συνεπώς να εξερευνήσουμε καλύτερες μεθόδους για να παράξουμε τις αρνητικές αλληλεπιδράσεις μας με στόχο τη καλύτερη εκπαίδευση των ταξινομητών μας. Μάλιστα, στο χώρο της πρόβλεψης της κατηγορίας των αλληλεπιδράσεων, και όχι απλώς της ύπαρξή τους, έχουν αναπτυχθεί ήδη αξιόλογες μέθοδοι κατασκευής αρνητικών δειγμάτων [9].

Chapter 1

Introduction

1.1 Motivation

The combination of more than one drugs can often improve the outcome of a treatment that would be based on a single one [10]. Also it is very common for patients that suffer from comorbidities to follow a multiple drug scheme. Yet, there can be adverse effects in those combinations, which occasionally might be toxic. While many drug interactions have been discovered, there are potentially new ones that could be predicted with computational methods, before their laboratory investigation [11]. Thus predicting Drug-Drug Interactions (DDIs) is important for the well-being of patients. The problem of drug interaction prediction with computational methods is usually reduced to the problem of link prediction in a network of interactions. Additional features, such as structural, physicochemical and biochemical characteristics of chemical compounds could potentially increase the accuracy of predictions; the usefulness of such properties is already proven in the field of drug discovery [12].

Usually, notable approaches to create systems that predict drug interactions either exploit the knowledge graph of the already known DDIs, or study the chemical features of compounds that are related to the drugs of interest. This is where an important question arises: when is it preferable to choose the network information over the chemical features — or the opposite — in order to study possible interactions in a set of drugs? And further more, is there sufficient gain in combining these sets of characteristics (graph and chemical) for the purpose of creating even more efficient systems for link prediction in the area of drug interactions?

1.2 Field of Study

Our work is focused on single link prediction between drugs. We study the embedding of a drug's chemical formula towards vector representations that encapsulate the properties of chemical compounds. We also study the embedding of the network information that we extract when we consider each drug to be a node in a graph of drug interactions; interactions are denoted as edges in this type of graph. We utilize these embedding methods in creating machine learning models that attempt to predict interactions between drugs. Our interest is not just to isolate those methods and separately

evaluate their efficacy in creating link prediction systems; we delve into the network of drug interactions and compare the two embedding mechanisms in order to discover which are the network properties that make each approach more efficient than the other when it comes to predicting DDIs. For example, we could assume that a drug for which there is plenty of network information (e.g. a high degree node in our graph) is a great candidate for network embedding methods that work on exploiting this kind of knowledge. On the other hand, a drug that may seem isolated in the network of drug interactions (perhaps a newly discovered compound with only a few known interactions), may be more suitable for chemical embedding methods. We study various graph properties, such as the degree, core difference and betweenness centrality of nodes and edges in order to discover parts and characteristics of the interactions graph that will help us compare in detail the chemical and the structural methods of embedding drug information.

1.3 Contribution

In this work we make the following contributions:

- Comparison of two drug interaction prediction methods. One based on network information and and the other on chemical information. node2vec and mol2vec were used for embedding network and chemical information respectively.
- In-depth study of topological features that influence the accuracy of the two link prediction models. In particular, we focus on what renders chemical information more useful than network information when it comes to drug interaction discovery.
- Experimental Evaluation of the link prediction methods on a graph created from DrugBank's [1] drug interactions.

1.4 Thesis Structure

The rest of this thesis is structured as follows:

- Chapter 2.1 provides a literature review regarding notable works in the fields of Link Prediction, Network Analysis and Drug Interaction Prediction. This chapter also provides the reader with the essential definitions behind the concepts discussed in this thesis.
- Chapter 3 explains the details behind our experiments, as well as how we proceed with the data mining operations that are needed in order to prepare our source data for mol2vec and node2vec. Finally, we discuss the way in which we compare network against chemical information for the purpose of DDIs prediction.
- Chapter 4 presents the results of our experiments: cross-experiment study allows for comparisons between embedding types in regard to topological characteristics of the samples, while detailed reports on each classifier's predictions provides in-depth analysis of each embedding model's behavior on the test samples.

- Chapter 5 summarizes the most important observations from our experiments; we also draw our conclusions from the work we did for this thesis. Finally, we suggest ideas for future work on the topic of comparing chemical against network information for the problem of drug interactions prediction.
- Appendix A contains information about the dataset and classification report for each one of our experiments that are based on: mol2vec, node2vec and hybrid embeddings.
- Appendix B provides a detailed evaluation of each one of our classifiers per DrugBank's drug categories (see subsection 3.2.1)
- Appendix C presents a complete report on our hybrid classifier's choices over key graph characteristics.

Chapter 2

Literature & Definitions

2.1 Literature Review

Link Prediction in graphs is a well researched area [13], with many applications in social networks analysis [14, 15] and drug interaction prediction [16], to name just two prominent ones.

Predicting missing (from existing knowledge) or future edges of a network can be based on graph features, upon which matrix and tensor factorization can be applied [17, 15]. Such methods have also been applied to graphs that evolve in time [18]. Node and edge embedding methods are widely used to create features for classification [19]. Lately graph neural networks have become popular in link prediction [20, 21].

Link prediction has also been used for drug interactions, with methods varying from predicting the presence or absence of an interaction, to methods the predict the type of an interaction a in multi-relational network [22, 23]. Moreover, many methods have been proposed that employ additional features apart from graph based features, for instance the usage of chemical information has been proposed in [24]. Also a hybrid method that combines multiple types of information, including structural information has been proposed in [20]. Finally, for multi relational link prediction there are approaches based on graph neural networks [8].

This thesis focuses on the difference between chemical and structural information in DDI prediction, and in particular how the network topological features could render one of these types of information more useful for discovering new interactions.

2.2 Prerequisites

In this section we present the fundamental definitions, algorithms and graph measures that we use in this thesis. We do not include all the concepts that are required for someone to understand this document and for that reason a familiarity with data science and machine learning topics is important for efficiently going through this thesis.

Kullback–Leibler divergence KL divergence [25] is a non-symmetric statistical measure of how one probability distribution p is different from another distribution Q over the same variable x. Denoted by $D_{\mathrm{KL}}(P\|Q)$, for discrete probability distributions P and Q KL

divergence is given by:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{2.1}$$

KL divergence will prove particularly useful in chapter 4, were we will compare the distributions of the graph characteristics of the samples sets that are defined by the choices of our classifiers.

2.2.1 Machine Learning & NLP Concepts

Shallow Neural Networks Feed forward neural network models with only one hidden layer are called Shallow Neural Networks (SNNs). The information stored in the hidden layer after the training of the network is considered to be the projection of the raw data input of the network towards a new representation of features. Essentially, when an input is given to the network, the values that are formed in this hidden layer can be used to produce an embedding for that input.

word2vec Proposed in 2013 by Tomas Mikolov and his team of Google researchers, word2vec [2] is a NLP technique that uses a neural network model in order to capture the relations between words in a large corpus of text. Each unique word is embedded into a fixed-size vector of real numbers; words that are semantically similar should lead to vectors that also satisfy a metric of similarity (e.g. euclidean distance or cosine similarity). These vectors are called word embeddings, and the models that are used to produced them are called Vector Space Models (VSMs). There are two major algorithmic approaches for the word2vec paradigm, based on Shallow Neural Networks:

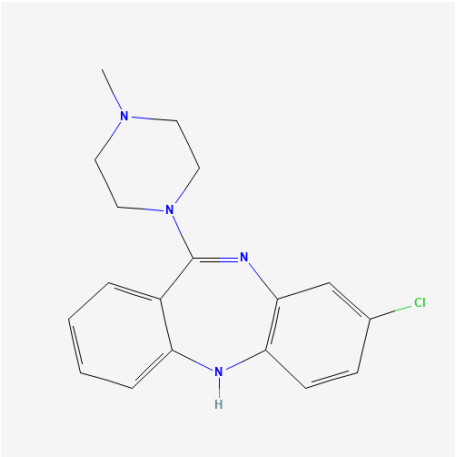
- Continuous skip-gram: the model attempts to predict the most semantically fitting surrounding window of context words when a word is given as an input. The architecture behind the skip-gram approach weighs nearby context words more heavily than more distant context words inside the surrounding window of context words.

- Continuous bag-of-words (CBOW): the model attempts to predict the current word of a sentence from a window of surrounding context words. The order of these context words does not affect prediction of the model — this property is known as the bag-of-words assumption.

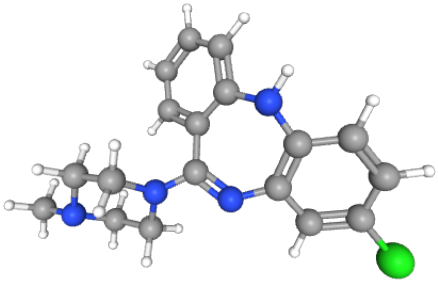
node2vec Inspired by word2vec, node2vec is a machine learning algorithm for producing node embeddings, by mapping the nodes of a graph to a low-dimensional space of features that maximizes the likelihood of preserving their network neighborhoods [4]. To achieve this goal, node2vec simulates biased random walks based on an efficient network-aware search strategy where the nodes appearing in the random walk define neighbourhoods. The search strategy accounts for the relative influence nodes exert in a network.

mol2vec Mol2vec is also a variation of word2vec. This algorithm encodes chemical compounds as vectors through training an unsupervised machine learning approach on a so called corpus of compounds that consists of all available chemical matter [3]. The result vector representations of molecular substructures are close for chemically related substructures, just like vector representations of semantically related words are close in a word2vec model's embeddings.

For example, figures 2.1a and 2.1b show the 2D and 3D representations for the chemical compound of Clozapine: the first atypical antipsychotic approved for treatment of schizophrenia. These figures were generated by the PubChem website [1]. The purpose of mol2vec is to embedd the chemical structure of compounds like clozapine into a vector of real numbers.



(a) *2D Structure of the chemical compound of Clozapine*



(b) *3D Conformer of the chemical compound of Clozapine*

Figure 2.1. *Representations of Clozapine from PubChem*

2.2.2 Computational Chemistry Definitions

SMILES Notation SMILES (Simplified Molecular Input Line Entry System) is chemical notation system based on principles of molecular graph theory [6]. Given a chemical graph from which hydrogen atoms have been removed and cycles have been broken in order to turn the graph into a spanning tree T, SMILES is a string that we obtain by concatenating the symbol nodes that we encounter in a depth-first traversal of T. For the example of Clozapine, the corresponding molecular formula and canonical SMILES notations are respectively:

$$C_{18}H_{19}ClN_4$$

and

$$CN1CCN(CC1)C2 = NC3 = C(C = CC(= C3)Cl)NC4 = CC = CC = C42$$

MOL Data Structure RDKit's [7] MOL data structure, is essentially an MDL Molfile; a file format for holding information about the atoms, bonds, connectivity and coordinates

[1] https://pubchem.ncbi.nlm.nih.gov/compound/135398737

of a molecule. Stored in what is called a "molfile ", this data structure holds some header information, the Connection Table (CT) containing atom info and the bond connections and types of a chemical compound, followed by sections for more complex information. We use MOL data structures as input for the mol2vec model in order to produce the chemical embeddings of drug chemical compounds.

2.2.3 Graph Metrics

The following paragraphs briefly present the most important graph metrics that we use in both our Link Prediction experiments and comparisons between chemical and network embeddings in regard of their usefulness in DDIs prediction.

Node Degree & Degree Centrality The degree of a node in the graph denotes the number of edges that are connected to it. Also, the Degree Centrality of a node is equal to its degree; this is the simplest one of the centrality measures that we use. We only focus on the variations of these metrics for unweighted and undirected graphs since our graph of drug interactions fits this category.

Clustering Coefficient The local clustering coefficient ([26], [27]) for a node v_i of an undirected graph $G = (V, E)$ is

$$C_i = \frac{2\left|\left\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\right\}\right|}{k_i\,(k_i - 1)} \tag{2.2}$$

where we consider N_i to be the neighbourhood of v_i, defined as the set of its immediately connected neighbours (nodes):

$$N_i = \left\{v_j : e_{ij} \in E \vee e_{ji} \in E\right\} \tag{2.3}$$

and k_i is the size of N_i. The clustering coefficient essentially quantifies how close are a node's neighbours to being a clique (complete graph). The average clustering coefficient of a graph, is given by:

$$C = \frac{1}{n}\sum c_v \tag{2.4}$$

where n is the number of nodes in G.

Eigenvector Centrality Eigenvector centrality ([28]) computes the centrality for a node based on the centrality of its neighbors. The eigenvector centrality for node i is:

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{2.5}$$

where $\mathbf{A}$ is the adjacency matrix of the graph G with eigenvalue λ. The Perron–Frobenius theorem ([29]) establishes that there is a unique and positive solution for this equation if λ is the largest eigenvalue associated with the eigenvector of the adjacency matrix $\mathbf{A}$.

In other words, Eigenvector Centrality quantifies the transitive influence of nodes on the graph. Edges originating from high-scoring nodes contribute more to the eigenvector

centrality of a node than edges from low-scoring nodes. A high eigenvector centrality value means that a node is connected to many nodes who themselves have a high eigenvector centrality.

Closeness Centrality The closeness centrality [30] of a node v is given by the reciprocal of the sum of the shortest path distances from v to all the other $n-1$ nodes of the graph. Since the sum of these distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum the $n-1$ possible distances.

$$C(v) = \frac{n-1}{\sum_{u=1}^{n-1} d(u,v)} \tag{2.6}$$

Closeness centrality is a way of identifying nodes that have the capacity to convey information very efficiently through a network.The closeness centrality of a node denotes its average farness (inverse distance) to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes of the graph.

Node Betweenness Centrality In graph theory, betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

Defined in the NetworkX documentation [2]: betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v :

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)} \tag{2.7}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t \mid v)$ is the number of those paths passing through some node v other than s,t. If $s = t, \sigma(s,t) = 1$, and if $v \in s,t, \sigma(s,t \mid v) = 0$ [31].

Edge Betweenness Centrality In chapter 4 we use the edge betweenness centrality, which, for an edge e of the graph, is calculated by the sum of the fraction of all-pairs of the shortest paths that pass through e :

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t \mid e)}{\sigma(s,t)} \tag{2.8}$$

where V is the set of nodes, sigma(s, t)^2 is the number of shortest (s,t)-paths, and $\sigma(s,t \mid e)$ is the number of those paths passing through edge e [30].

[2] https://networkx.org/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.betweenness_centrality.html

K-Cores Given a graph G and an integer K, the K-cores of the graph are connected components that are left after all nodes that have degree less than k have been removed. Also, the k-shell is the set of vertices of a graph that are part of the k-core set but not part of the (k+1)-core set. For each k-shell set, we assign k as the k-core value to each node of that set. In our experiments we will refer to the *Core Difference* of our drug interactions graph's edges as the absolute difference of the k-core values of the nodes that are denoted by that edge of the graph (where the edge also expresses an interaction between the underlying drugs).

Chapter 3

Methodology

This chapter describes all of our link prediction experiments, along with the steps we follow in order to prepare our data for these experiments. Although the primary goal of this thesis is to explore topological features that render network embeddings of drugs more useful than chemical embeddings, and vice versa, for the problem of DDIs prediction, it is crucial that we prepare a complete set of experiments that will form the basis for our comparisons between the embedding methods. At the same time, these experiments constitute a case study for the application of mol2vec and node2vec on the DrugBank dataset for predicting drug interactions.

3.1 Experiment Design

For each drug, we create two embedding vectors: a network and a chemical based representation. Next, we use three neural network classifiers to predict DDIs: a classifier based on graph (network) embeddings, another classifier for chemical embeddings, and a hybrid classifier that is based on both network and chemical embeddings. We evaluate the performance of each classifier in chapter 4, and proceed to compare their DDI prediction behavior against various network topological features on the network of drug interactions.

The embedding mechanisms we use on drugs are inspired by Natural Language Processing techniques and are both variations of word2vec [2]. The result in both cases is a vector that encapsulates information derived from either network or chemical properties. In order to create DDI prediction models, we train the neural network classifiers to accept as an input the vector embedding from pairs of drugs, and predict whether an interaction exists between them.

3.1.1 Chemical Embeddings

We discussed in 2.2.1 that mol2vec creates vector embeddings from drugs, by traversing the molecular structure of a drug's underlying chemical compound, in a similar fashion to word2vec's traversal of a sentence of words. Our set of unique chemical compounds — which is presented in 3.2.1 — consists only of a few thousand chemicals; a number which is sufficient when we explore drugs and their interactions, but is not enough to train an efficient mol2vec model. Therefore, in our experiments we use a pre-trained

model of mol2vec to produce the chemical embeddings of the drugs in our dataset[1].

3.1.2 Network Embeddings

After harvesting DrugBank's drug interaction graph and then extracting a sample subgraph (see 3.2.2), we train a node2vec model on the sampled graph and use it to produce the structural embeddings of all the drugs in the dataset. The embedding mechanism of node2vec is presented in 2.2.1.

3.1.3 Training Samples

Since the experiments focus on predicting drug interactions, we consider two categories of samples: positive and negative. Positive samples refer to observed interactions between drugs, and negative samples refer to interactions that have not been observed so far, i.e. a closed world assumption. The input to a classifier is a pair of drugs, represented as a vector, for all of our models. The output is the presence or absence of an interaction, denoted by classes 1 and 0 respectively. Figure 3.1 displays how drugs v_i and v_j can form a training sample for our neural network classifiers, through the concatenation of their embedding vectors; the desired result from the classifier should be 0, if this is a negative interaction (meaning that these drugs do not interact), or 1, in the opposite scenario.

Note that each DDI leads to two different vector concatenations depending on the order that the concatenation is done; in the following experiments we included both options to generate positive samples. Acquisition of positive samples was straightforward. However, to create negative interactions it was assumed that unobserved interactions do not exist, and thus the graph was sampled for pairs of nodes that do not form edges. We chose to created balanced sets for training and evaluating the classifiers by sampling a number of negative samples equal the number of the positive samples.

3.1.4 Experiments & Evaluation

In total, we evaluated three classifiers based on the same neural network architecture of a feed forward model with two hidden layers.

- A mol2vec classifier based on mol2vec embeddings.
- A node2vec classifier based on node2vec embeddings.
- A hybrid classifier based on both mol2vec and node2vec embeddings.

The aim is to compare chemical and structural information when used for DDI prediction. For this purpose, along with commonly used metrics in classification, we also use in chapter 4 network topological features such as node degree, k-core values and betweenness centrality, in order to identify which topology characteristics make each approach more efficient in the problem of link prediction.

[1] https://github.com/samoturk/mol2vec/tree/master/examples/models.

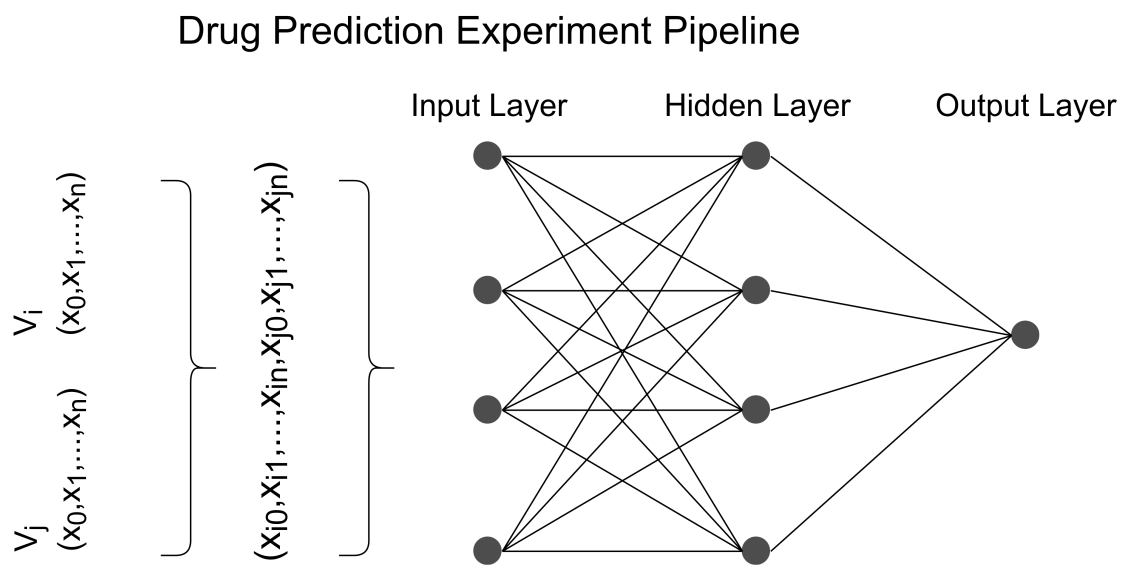


Figure 3.1. *DDIs, in the form of concatenated vectors of drug embeddings, are used as training samples for our classifiers.*

3.2 Data Harvesting

3.2.1 Drug Interactions Graph

We extract our dataset from DrugBank [1] (version 5.1.9), a drug interactions repository that is human curated. The dataset includes 14,624 drug entries and 1,389,184 unique DDIs. We keep only the drugs that have at least one known interaction, and we also exclude (a relatively small number of a few hundred) drugs that have chemical compounds that are incompatible with the tools that we use for our data mining processes (i.e. PubChem) and are described in this section. After the data cleaning, we obtain a set that comprises of 3,753 nodes and 1,207,953 edges, and it will be referred as the *full graph*. Table 3.1 contains the DrugBank entries that have the most interactions in the dataset; interactions are denoted in the form of tuples: (id_1, id_2).

Drug Name	# Interactions	DrugBank Categories	DrugBank ID
Quinidine	2402	approved, investigational	DB00908
Clozapine	2379	approved	DB00363
Chlorpromazine	2369	approved, investigational vet approved	DB00477
Amitriptyline	2298	approved	DB00321
Imipramine	2287	approved	DB00458
Doxepin	2236	approved, investigational	DB01142
Clomipramine	2217	approved, investigational, vet approved	DB01242
Haloperidol	2214	approved	DB00502
Methylene blue	2203	approved, investigational	DB09241
Nefazodone	2196	approved, withdrawn	DB01149

Table 3.1. *DrugBank drugs with the most interactions. DrugBank ID refers to the dataset's unique identifier for each drug.*

Graph	Nodes	Edges	Avg. Degree	Clustering Co.	Conn. Components	Diameter
Full	3,753	1,207,953	643.73	0.621	1	5
Sampled	3,753	12,080	6.44	0.005	799	∞

Table 3.2. *Basic Properties for full graph and sampled graph*

Graph	Degree C.	Eigenvector C.	Closeness C.	Betweenness C.
Full	0.1715	0.012	0.5030	0.00027
Sampled	0.0017	0.011	0.1556	0.00049

Table 3.3. *Average Centrality Measures for full graph and sampled graph*

When we refer to structural information regarding DDIs, it is critical to make the distinction between direct edges which denote drug interactions and information that derives from the rest of the properties the two interacting drugs (nodes) share. There is no point in creating node2vec embeddings over a graph that holds all the direct edges between interacting drugs, because that information — which is essentially the train and test set of the following experiments — will infiltrate in the structural embeddings. What we do instead, is take a small sample of the *full graph* that contains the same number of nodes and only 1% of its edges. We will refer to the result of this process as *sampled graph*. We train the node2vec model on the *sampled graph* and then use it to create the structural embeddings for our experiments.

3.2.2 Graph Sampling

Sampling 1% of the edges of the *full graph*, serves two purposes. First, it showcases the ability of node2vec to capture graph properties of the original graph, from a subset of edges that is smaller by two orders of magnitude; this fact is presented in detail in section 4. The second and most important purpose is the removal of direct links between interacting drugs. The graph information that we wish to use in our experiments needs to depend on properties between interacting drugs minimizing the effect of the edge that directly links the corresponding graph nodes. Closeness centrality in table 3.3 confirms that the sampled graph has been stripped off most of the direct edges, raising the average distance between nodes of the graph and, therefore, setting a level of difficulty for our model to identify interacting drugs through graph information that does not include the direct edges. Figure 3.3a provides a more detailed view on closeness centrality [32] for both graphs; for the case of more than one connected components we use the Wasserman and Faust formula [33].

Table 3.2 depicts some basic properties for the full and the sampled graph, and table 3.3 contains the average centrality measures. Also, DrugBank labels drugs with the categories shown in table 3.4; note that some drugs belong to more than one category. Interestingly, drugs that belong to different categories also seem to differentiate on a graph/topological perspective (avg. node degrees and std. of node degrees).

The average node degree of the *full graph* reflects a high density of drug-interactions,

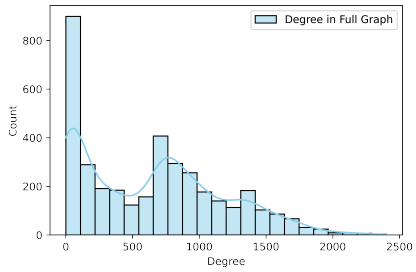
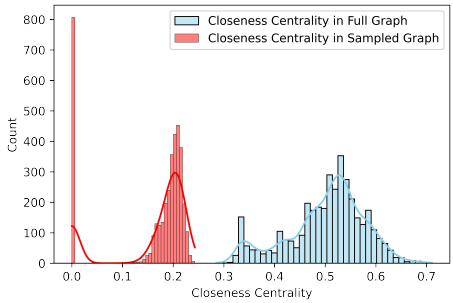


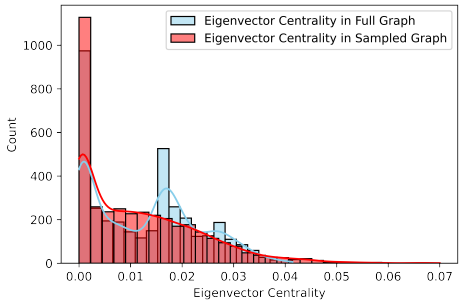
Figure 3.2. *Histogram and KDE for Node Degrees in full graph*

Category	Frequency	Degree (mean, std.)
Approved	2,179	787.54 ± 540.90
Investigational	1,585	635.09 ± 541.51
Experimental	853	480.82 ± 447.84
Vet Approved	309	731.11 ± 578.32
Withdrawn	190	871.08 ± 492.24
Illicit	123	870.91 ± 503.60
Nutraceutical	63	282.95 ± 337.05

Table 3.4. *Full Graph's Drug Categories*



(a) *Histogram and KDE of Closeness Centrality in full graph and sampled graph*



(b) *Histogram and KDE of Eigenvector Centrality in full graph and sampled graph*

Figure 3.3. *Closeness Centrality & Eigenvector Centrality in full graph and sampled graph*

thus an average drugs interacts with 600 other drugs. Also, figure 3.2 reveals that there is a considerable number of drugs that have only few known interactions, as well as many drugs with thousands of identified interactions.

Although reducing closeness centrality in the *sampled graph* is one of our goals, we also need to keep enough edges to train node2vec embeddings to be useful for the classification. Thus we considered the eigenvector centrality of the *full graph* and the *sampled graph*. Edges originating from high-scoring nodes contribute more to the score of a node than connections from low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores [34]. Figure 3.3b shows that sampling the full graph did not result in big difference at nodes' eigenvector centrality. We can say that nodes in the full graph that not only have many connections to other nodes, but are also connected with other nodes of importance — in the sense of eigenvector centrality — continue to hold this characteristic in the sampled graph.

3.2.3 Data pre-processing

node2vec embeddings Starting from DrugBank's drug IDs, we use various tools in order to obtain the chemical and the structural embeddings (see Figure 3.4). With the NetworkX library [35] we create a graph data structure out of DrugBank's data — the *full graph* — and we also use the implemented sampling methods to create the *sampled*

graph. We train node2vec on the *sampled graph*, setting the algorithm to embed nodes to vectors of 128 dimensions. Once the node2vec [4] model is trained, we apply it on each drug (node) to produce the corresponding node2vec (structural) embedding.

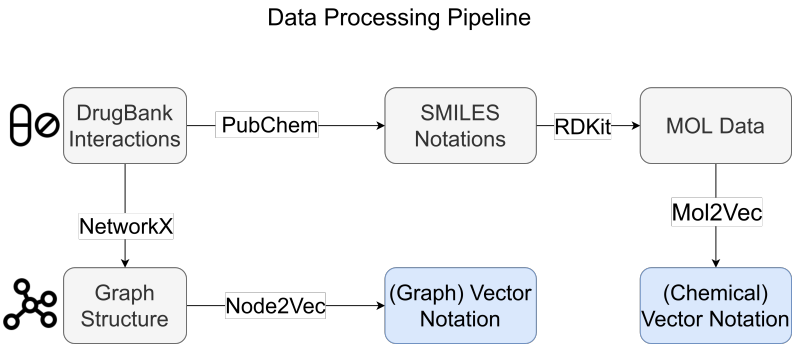


Figure 3.4. *Data Processing Pipeline: steps between DrugBank data and final embeddings for each drug*

mol2vec embeddings To compute the mol2vec (chemical) embeddings, we must first use PubChem's services and acquire the isomeric SMILES notation for each drug [5]. PubChem is a large repository of chemical information that is free to access, and also offers various tools for searching and studying chemical compunds; specifically, we utilize the PubChemPy [2] library for Python. We use RDKit [7] to convert each SMILES entry to a MOL data structure [36], a widely-used chemical structure file format in which adjacent lists and adjacent matrices are mostly used to describe a chemical compound's structure. Finally, we apply mol2vec [3] on the MOL data structures, using a pre-trained model, to embed drugs to vectors of 300 dimensions.

At this point, we are ready to proceed with our experiments. We have completed the data harvest of DrugBank's contents, and we have also created a unique network (or structural) embedding as well as a unique chemical embedding for each drug. We used the *sampled graph* to train the node2vec model that produced our network embeddings, however these embeddings also refer to the *full graph*, which is the only graph that we should be concerned about from this point and forward; the *sampled graph* has fulfilled its purpose.

[2] https://pubchempy.readthedocs.io/en/latest

Chapter 4

Experimental Results

4.1 Classifier Evaluation

Table 4.1a displays the classification report of the three models on the test set; negative interactions are denoted by class 0 and positive interactions are denoted by class 1. We used a 65/5/30 split on all samples (positive and negative) of the *Full Graph* for training, validation and testing — we applied a random split, with the exception that all of the edges in the *sampled graph* (which are also present in the *full graph*) ended up on the training set. Also, we make sure that both variations of an interaction sample (regarding the order of the concatenation of the corresponding drug embeddings) are always included in the same set. Appendix A contains detailed information about the data used for training, evaluating and testing the classifiers, as well as their detailed classification reports. Appendix B provides the complete data behind table 4.1b, presenting a detailed evaluation of each classifier's accuracy for each drug category.

Class	Classifier	Precision	Recall	F1 Score
Negative Samples (Class 0)	mol2vec	0.91	0.89	0.90
	node2vec	0.90	0.94	0.92
	Hybrid	**0.96**	**0.94**	**0.95**
Positive Samples (Class 1)	mol2vec	0.88	0.91	0.89
	node2vec	**0.96**	0.93	0.94
	Hybrid	0.93	**0.95**	**0.94**

(a) *Classification Report*

Category	mol2vec	node2vec	Hybrid
Experimental	87	94	96
Approved	85	93	94
Investigational	87	94	95
Vet Approved	88	94	97
Withdrawn	87	96	97
Illicit	87	95	96
Nutraceutical	91	90	97

(b) *Accuracy percentage comparison for each drug category*

Figure 4.1. *Classification Report & Accuracy Comparison per Drug Category*

Table 4.2a compares the average value of useful graph properties of the test samples, and table 4.2b provides a comparison between the Kullback-Leibler divergence of the distributions of test sample properties for the correct and false predictions of mol2vec and node2vec classifiers. *Average mean degree* refers to the average of all means of the node degrees of the test samples; similarly *average min degree* denotes the average of all minimum degrees between all test interactions, and *average max degree* denotes the average maximum degree. Also, *betweenness* centrality in this section is calculated only for positive samples, and refers to the corresponding edge betweenness centrality (and not the node betweenness centrality that is reported in section 3.2).

	Metric	mol2vec	node2vec	Hybrid
False Prediction	Average Min Degree	518.06	470.07	517.01
	Average Max Degree	1137.02	1143.09	1163.44
	Average Mean Degree	827.54	806.58	840.22
	Average Core Difference	237.17	266.10	242.34
	Average Betweenness Centrality	4.57e-06	6.99e-06	6.76e-06
Correct Prediction	Average Min Degree	519.66	593.45	512.09
	Average Max Degree	1083.25	1143.96	1078.67
	Average Mean Degree	801.45	868.713	795.38
	Average Core Difference	228.50	196.01	231.80
	Average Betweenness Centrality	1.39e-06	1.29e-06	1.44e-06

(a) *Comparison of average graph properties of interactions grouped by each model's prediction*

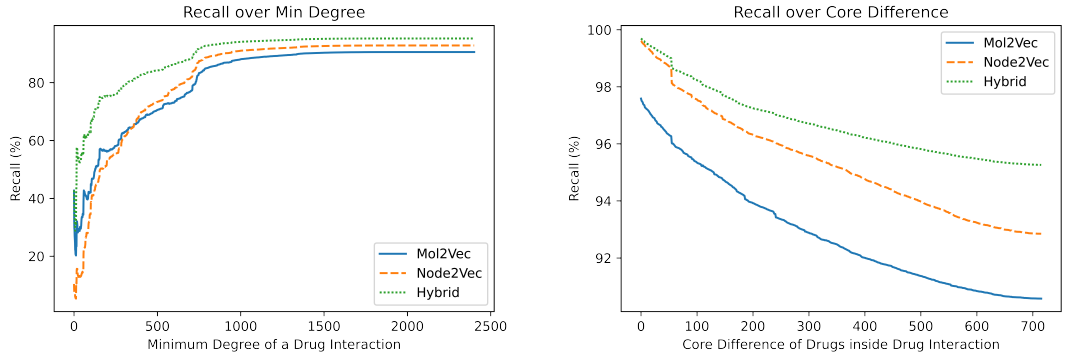
Classifier	Correct	False
Average Degree	0.018	0.047
Min Degree	0.019	**0.199**
Max Degree	0.010	**0.110**
Core Difference	0.013	**0.103**
Betweenness C.	**0.109**	0.071

(b) *KL Divergence between mol2vec and node2vec sample distributions (values over* 0.1 *in bold)*

Figure 4.2. *Result Metrics by Classifier Choice (Correct or False)*

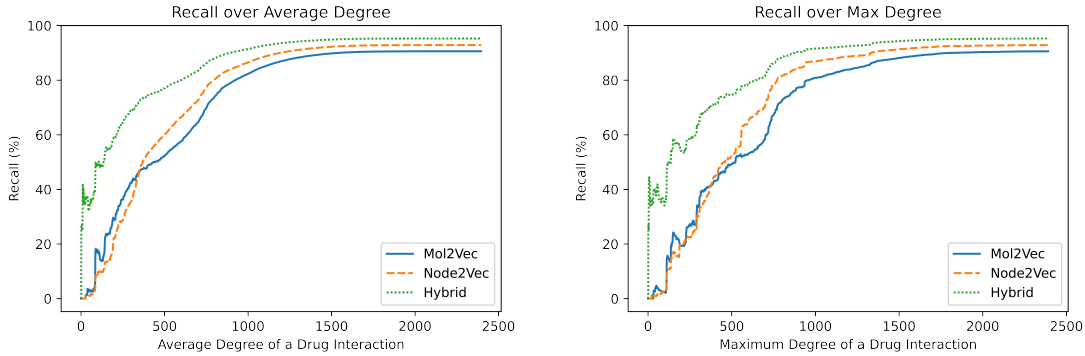
Note that all of the figures in this section do not plot a function directly on a variable of the corresponding samples characteristic, but show the classification value (accuracy or recall) over the set that includes all samples that have a characteristic value that is equal or less than the variable of x-axis for each point of the plot. This means that the value of 0 of the x-axis refers to an empty set of samples, where the greatest shown value of the x-axis refers to all the samples of the test set.

The higher performance of the hybrid classifier compared to the other models in the classification report suggests that there is knowledge on DDIs that is unique for both the chemical and the structural embedding methods. Combining the embeddings to train the hybrid model — trading this abundance of information with higher vector dimensions that are known to hinder the learning capabilities of neural networks — leads to a better predictor. The only exception here is the higher precision on positive interactions for the node2vec classifier; hinting that the node2vec classifier shows a greater ability to identify negative interactions properly and maintain a lower number of false positives choices through the evaluation phase. Figure 4.3a confirms our assumption that drugs with few known interactions (possibly newly discovered compounds) make better candidates for chemical based predictors when recall is more important than accuracy; identifying more true interactions by trading some false positives may be a good trade off for a chemical researcher. The figure even sets a threshold value at a Min Degree of 350 where node2vec begins to perform better than mol2vec in terms of recall.



(a) *Recall over Min Degree* **(b)** *Recall over Core Difference*

Figure 4.3. *Recall plots for mol2vec, node2vec and hybrid Classifiers for min degree and core difference of sample interactions*

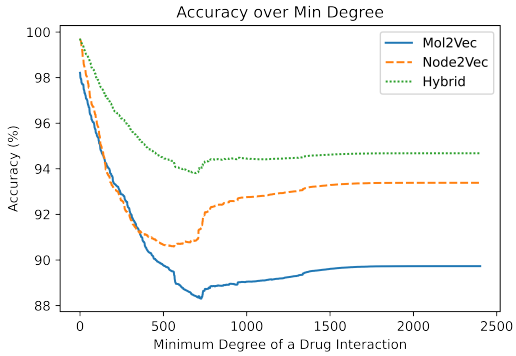


(a) *Recall over Avg. Degree* **(b)** *Recall over Max Degree*

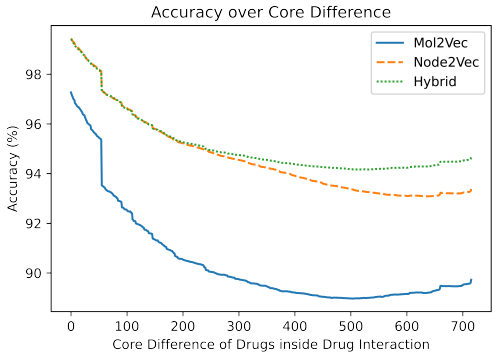
Figure 4.4. *Recall plots for mol2vec, node2vec and hybrid Classifiers for avg. and max degree of sample interactions*

Figures 4.3b ans 4.5b show that for low core difference values of sample interactions the node2vec classifier performs as well as the hybrid classifier in terms of recall and accuracy. The fact that node2vec classifier reaches hybrid model's efficiency means that chemical embeddings, when it comes for test interactions with low core difference, show no unique knowledge to add to structure embeddings' learning capabilities. Also, all models seem to perform better for lower values of core difference.

The sections that follow will focus on each classifier separately as we attempt to explore the topological characteristics of the samples that lead to correct and false estimations of our classifiers.

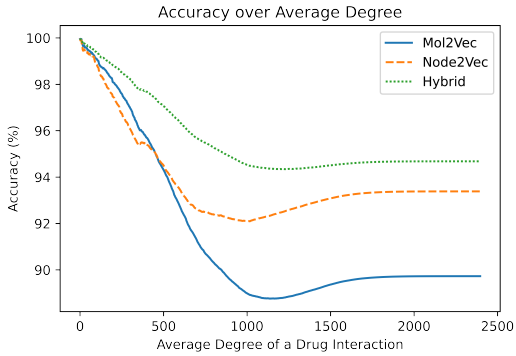


(a) *Accuracy over Min Degree*

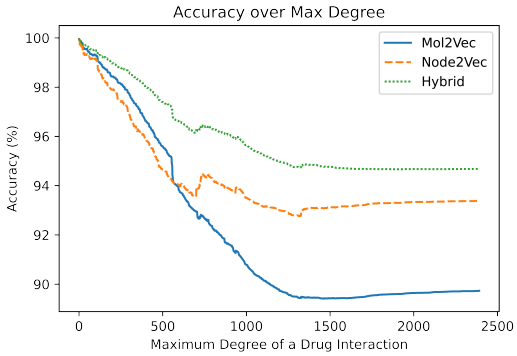


(b) *Accuracy over Core Difference*

Figure 4.5. *Accuracy plots for mol2vec, node2vec and hybrid Classifiers for min degree and core difference of sample interactions*



(a) *Accuracy over Avg. Degree*



(b) *Accuracy over Max Degree*

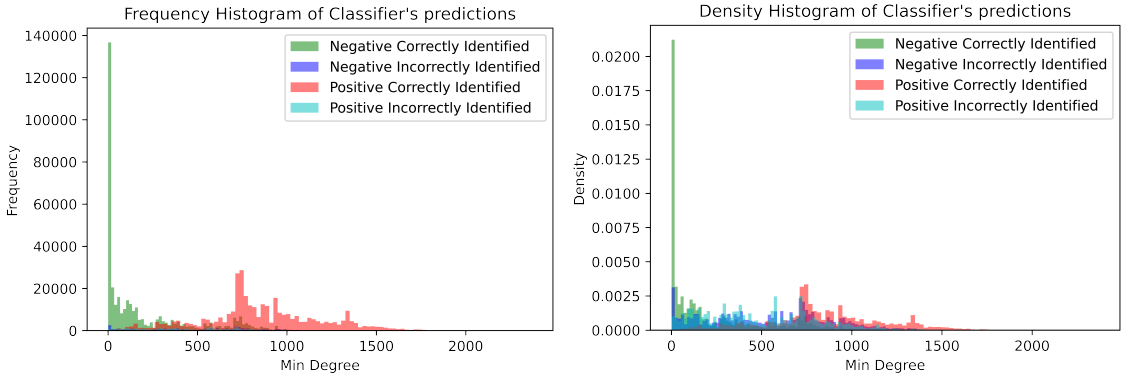
Figure 4.6. *Accuracy plots for mol2vec, node2vec and hybrid Classifiers for avg. and max degree of sample interactions*

4.2 In-depth Study of mol2vec Model's Performance

We consider the following prediction categories:

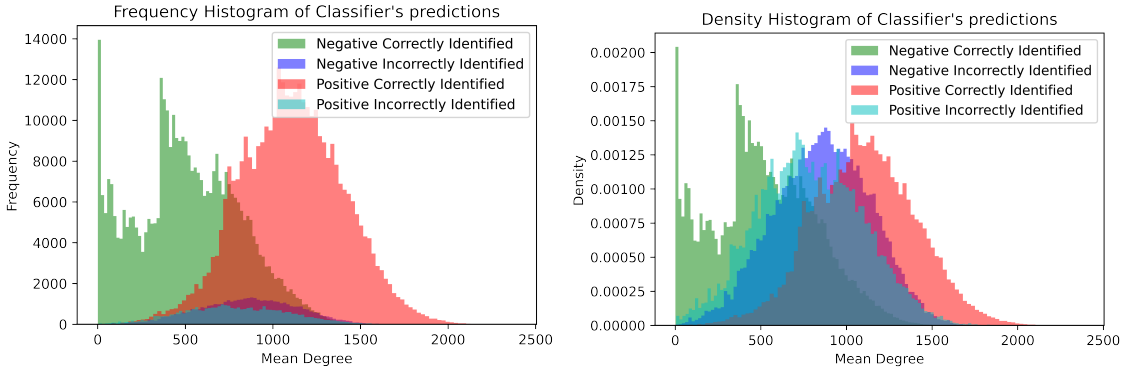
- Negative Samples Correctly Identified (NSCI)
- Negative Samples Incorrectly Identified (NSII)
- Positive Samples Correctly Identified (PSCI)
- Positive Samples Incorrectly Identified (PSII)

The following figures study mol2vec classifier's choices by splitting the test sample set by prediction category. The aim of this categorization is to provide a topological insight behind the model's weaknesses and strengths when it comes to DDIs prediction.



(a) *Frequency Histogram over Min. Degree* **(b)** *Density Histogram over Min. Degree*

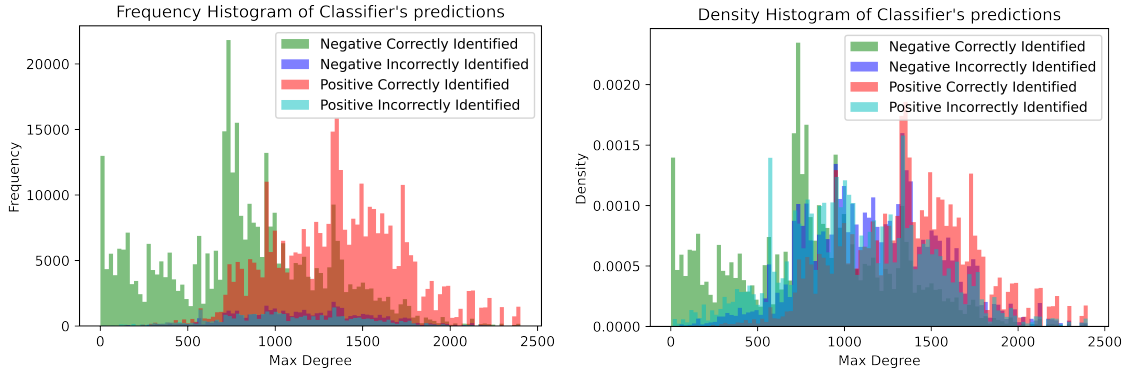
Figure 4.7. *Frequency and Density Histograms for mol2vec's sample distributions' Min. Degree per prediction type*



(a) *Frequency Histogram over Avg. Degree* **(b)** *Density Histogram over Avg. Degree*

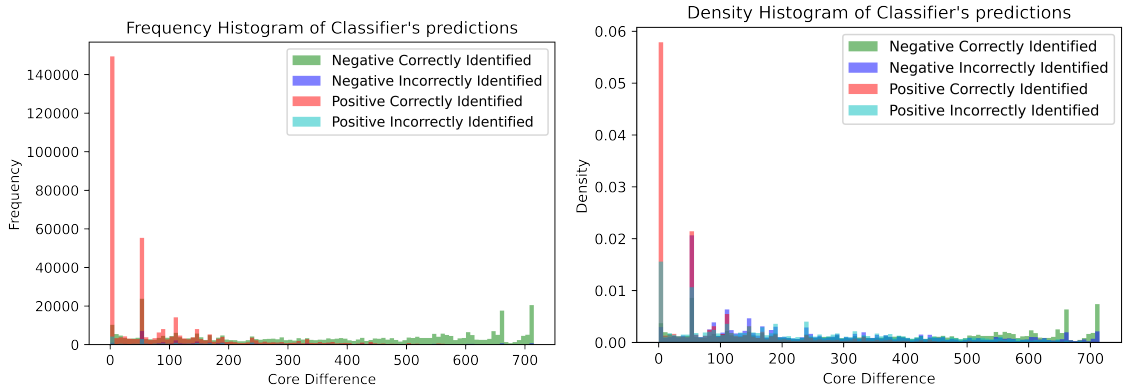
Figure 4.8. *Frequency and Density Histograms for mol2vec's sample distributions' Min. Degree per prediction type*

Figures 4.7a and 4.7b present the frequency and density histograms respectively for all the test samples (edges); minimum degree means that for each edge, only the node



(a) *Frequency Histogram over Max. Degree* **(b)** *Density Histogram over Max. Degree*

Figure 4.9. *Frequency and Density Histograms for mol2vec's sample distributions' Max. Degree per prediction type*



(a) *Frequency Histogram over Core Difference* **(b)** *Density Histogram over Core Difference*

Figure 4.10. *Frequency and Density Histograms for mol2vec's sample distributions' Core Difference per prediction type*

with the minimum degree was taken into account. Similarly, figures 4.8a and 4.8b study the average degree (mean value of the degrees of the two nodes for each edge/sample of the test set) and figures 4.9a and 4.9b report on the maximum degree. Core Difference (see chapter 2.1 for definition) is studied in figures 4.10a and 4.10b

Frequency Histograms are easy to read for the categories where our models' predictions are correct since the high accuracy of our classifiers in the test set means that most of the samples in these histograms will represent a correct prediction. However, the scarcity of false predictions led us to also present the density histograms, where it is easier to study the regions for each characteristic that false predictions occur.

All histograms over node degrees (minimum, average or maximum for each sample interaction) show that correctly identified negative samples are gathered on the left side of the plot and correctly identified positive samples are gathered on the right side; the latter observation is consistent with our previous notes on how all of our models' efficiency rises as the drugs that are evaluated correspond to nodes of the *full graph* with higher degrees. However, the fact that NSCI gather at the left of our plots suggests the — not obvious —

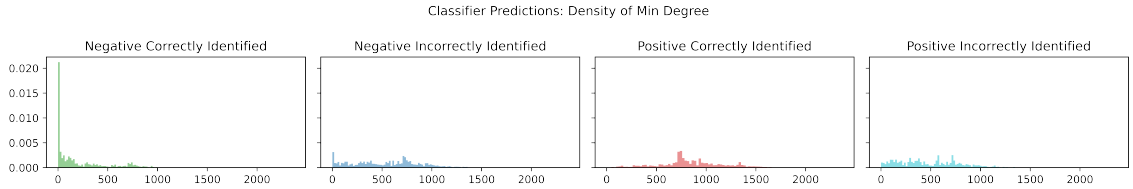


Figure 4.11. *Density Histograms over Min. Degree for each prediction category of mol2vec test samples*

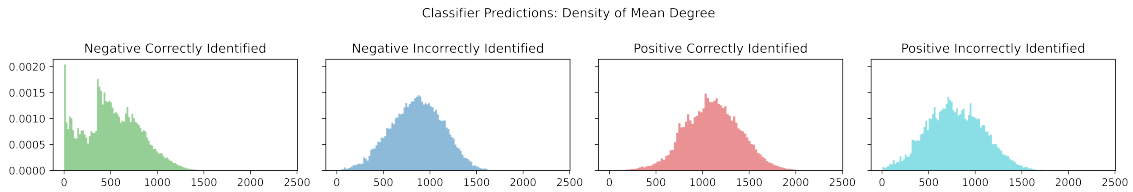


Figure 4.12. *Density Histograms over Avg. Degree for each prediction category of mol2vec test samples*

idea that it is easier for mol2vec model to understand that a sample interaction is fake when the nodes involved have small degrees. This observation provides proof over the guess we made in section 1.2: drugs that do not have a strong presence in the interaction graph are better candidates for methods that are based on the chemical information. Also, we observe that plotting the density over max. degree provides less distinguishable distributions in figure 4.9b than the other density plots.

Interestingly, figure 4.10b follows the opposite pattern from the previous figures regarding the various degree characteristics. The fact that NSCI are on the right side and PSCI on the left side suggests that both great and minor values of core difference, which in turn translates to nodes in a sample interaction with either very distant or very close k-core shells, correlate with better accuracy scores for our model.

Figures 4.11, 4.12, 4.13, 4.14 separate the sample distributions in autonomous plots to produce a more clear image from the mixed category density histograms

4.3 In-depth Study of node2vec Model's Performance

Similar to our study of mol2vec model's test samples' distributions per prediction category, we proceed to analyze node2vec model's decisions.

Figures 4.15a and 4.15b present the frequency and density histograms respectively

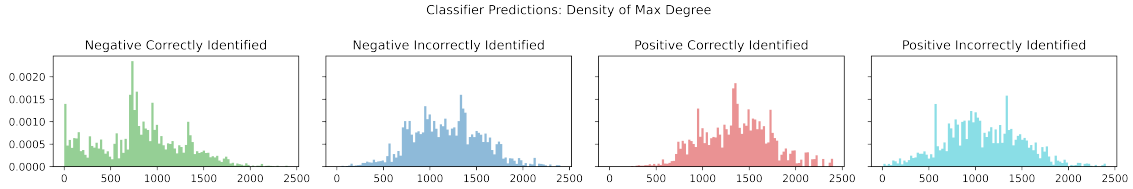


Figure 4.13. *Density Histograms over Max. Degree for each prediction category of mol2vec test samples*

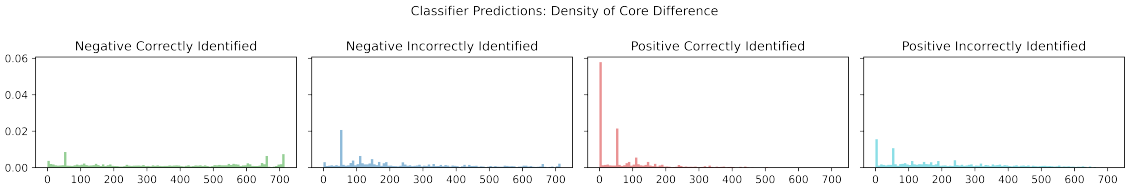
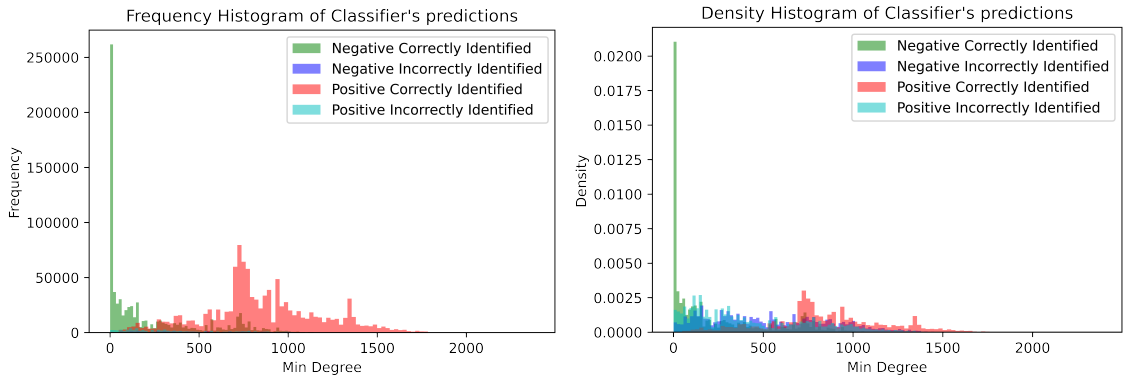


Figure 4.14. *Density Histograms over Core Difference for each prediction category of mol2vec test samples*



(a) *Frequency Histogram over Min. Degree*

(b) *Density Histogram over Min. Degree*

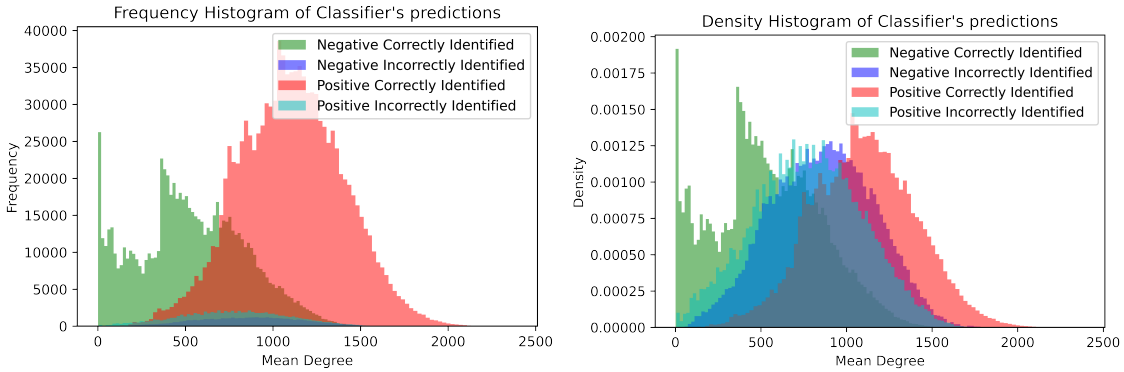
Figure 4.15. *Frequency and Density Histograms for node2vec's sample distributions' Min. Degree per prediction type*

for all the test samples (edges). Similarly, figures 4.16a and 4.16b study the average degree and figures 4.17a and 4.17b report on the maximum degree. Core Difference is studied in figures 4.18a and 4.18b.

The pattern of NSCI on the left and PSCI on the right is also present in node2vec's classifier when it comes to density of degree characteristics; the reversed pattern is also consistent for figure 4.18b.

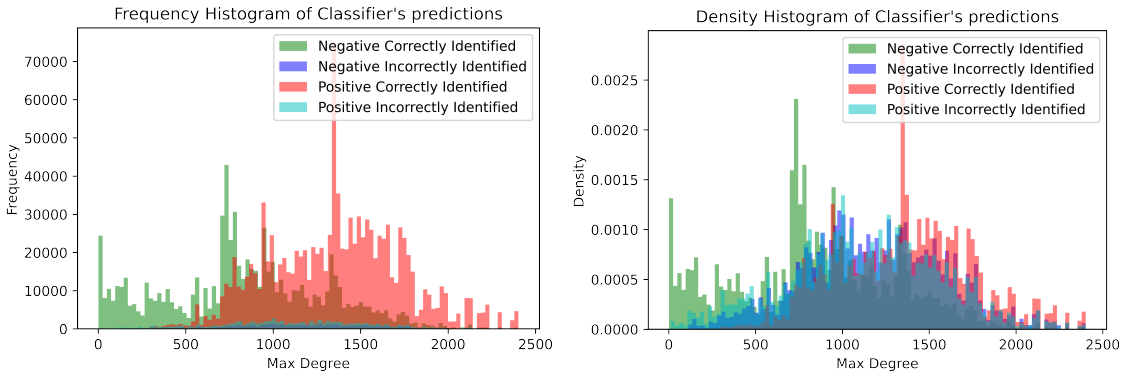
Comparing the separated density plots over min. degree for mol2vec and node2vec classifiers (figures 4.11 and 4.19) we observe almost the same pattern for all prediction categories' distributions. The only noticeable difference can be found positive incorrectly identified samples' distribution, where node2vec shows greater density values at the left side of the x-axis; validating the logical assumption that lack of network information for one of the two drugs that are studied for interactions (as in nodes that have few connections and lead to sample interactions with low min. degree) constitutes a weakness for graph-based models.

Separated density plots over avg. and max. degree also seem to be similar between mol2vec and node2vec. When it comes to core difference, however, figures 4.14 and 4.22 suggest that mol2vec is more vulnerable to handling positive samples that have small values of core difference; PSII are gathered with greater density on the left side of x-axis, something that is not happening on the same scale for the node2vec classifier's plot.



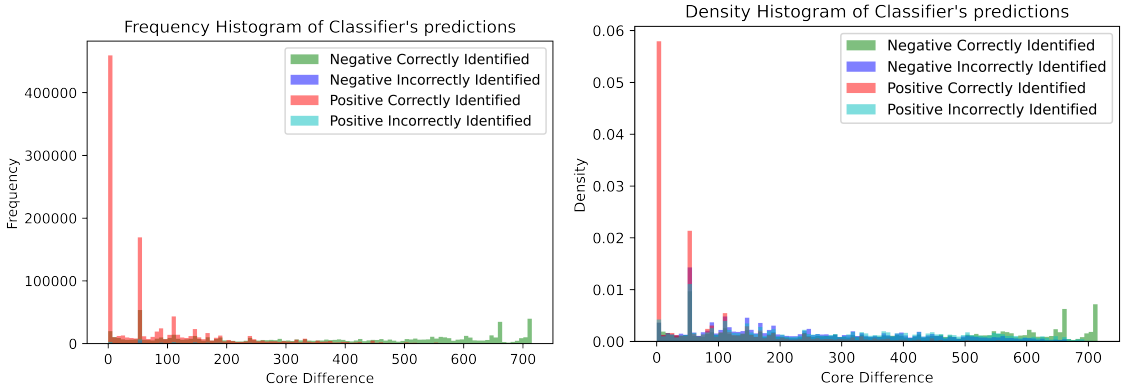
(a) *Frequency Histogram over Avg. Degree* **(b)** *Density Histogram over Avg. Degree*

Figure 4.16. *Frequency and Density Histograms for node2vec's sample distributions' Min. Degree per prediction type*



(a) *Frequency Histogram over Max. Degree* **(b)** *Density Histogram over Max. Degree*

Figure 4.17. *Frequency and Density Histograms for node2vec's sample distributions' Max. Degree per prediction type*



(a) *Frequency Histogram over Core Difference* **(b)** *Density Histogram over Core Difference*

Figure 4.18. *Frequency and Density Histograms for node2vec's sample distributions' Core Difference per prediction type*

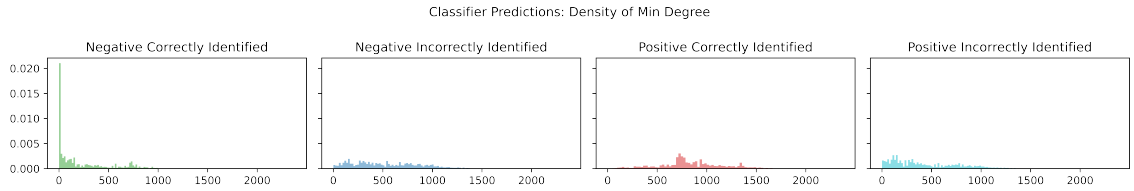


Figure 4.19. *Density Histograms over Min. Degree for each prediction category of node2vec test samples*

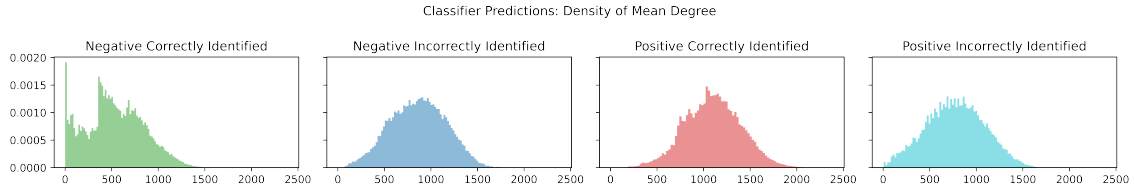


Figure 4.20. *Density Histograms over Avg. Degree for each prediction category of node2vec test samples*

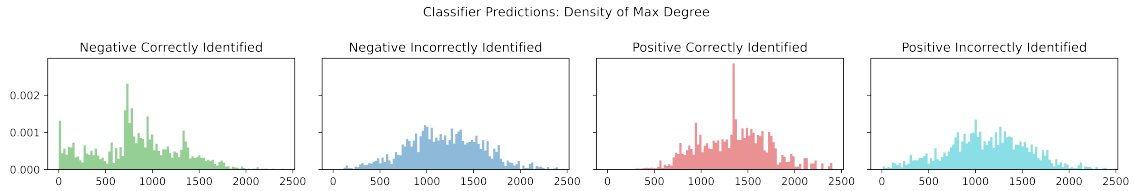


Figure 4.21. *Density Histograms over Max. Degree for each prediction category of node2vec test samples*

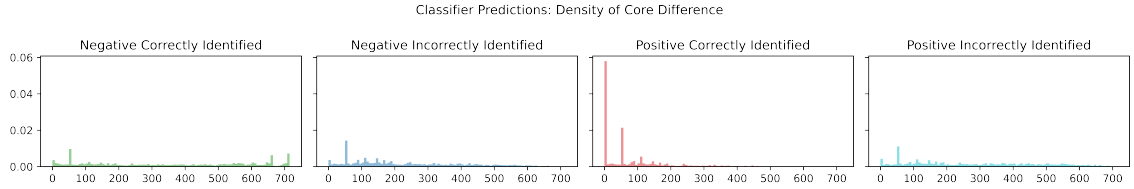


Figure 4.22. *Density Histograms over Core Difference for each prediction category of node2vec test samples*

Chapter 5

Conclusions

In this chapter we follow the order of the various tasks we had to complete in order to set up and evaluate our experiments as we summarize our observations and make inferences.

5.1 Drug Interaction Prediction Experiments on DrugBank

Without doubt, the challenge of predicting drug interactions extends to far greater depths than those we had the pleasure of exploring in this thesis; there are thousands of already existing drugs that we did not include in our experiments because we have no known interactions for them at the moment, and even greater numbers of chemical compounds will be proposed and tested as new drugs in the near future. On this premise we can say with certainty that the size, diversity and real-life resemblance of the knowledge graph need to be the top priority for designing useful and valid experiments for this subject. DrugBank proved to be an excellent data set for our work and appendix B validates that it is not only possible to study drug interactions in general, but also focus on specific categories of drugs (e.g. approved, experimental).

5.2 Structural Information

The whole process of creating our *sample graph* in chapter 3, using it to train node2vec before utilizing it to produce our structural embeddings, truly signifies the power of network information, as well as the efficiency of the node2vec paradigm. We removed 99% of the *full graph's* edges, and the result vectors in the *sample graph* where still able to provide sufficient training for our node2vec classifier and achieve great classification scores and an overall accuracy of 93% on the test set.

5.3 Chemical Information

In contrast with graph based embeddings, for which we were already aware of their capabilities in the area of Link Prediction (see chapter 2.1), we had no guarantee that our approach with chemical embeddings from mol2vec would lead to such rewarding results for the mol2vec classifier. In greater detail, the mol2vec model which produced our

chemical embeddings was not trained on the drugs of our data set (as node2vec did); we used a pre-trained model that was trained on millions of chemical compounds where most of them have no connection with DrugBank compounds. Also, chemical properties that occur from studying the structure of a compound have not been extensively tested like graph properties in the area of DDIs prediction and, therefore, our experiments constitute a notable contribution for establishing this type of information as useful research material for predicting drug interactions.

5.4 Network vs. Chemical Embeddings

Regarding the prediction of positive and negative interactions (i.e. presence or absence of edges of the *full graph*) we have observed the following results:

- Recall for the min-degree criterion is better for mol2vec up to a certain threshold (defined in chapter 4), but then node2vec takes over. Also regarding the core-differences criterion, node2vec is always better than mol2vec.
- The distribution of the betweenness centralities of the pairs of nodes for which the presence or absence of edges were correctly predicted were very dissimilar for the mol2vec versus the node2vec model.
- We have also shown that low core differences between a pair of nodes make structural information based models a better candidate to predict interactions achieving the same accuracy and recall even with the hybrid approach.
- For higher core differences, chemical information can boost the information that is provided by structural information and thus enhance the performance of interaction prediction.

Lastly, by and large the hybrid model that combines structural and chemical information of drugs leads to more efficient predictors of DDIs than using either of the models.

5.5 Future Work

5.5.1 Multi-label Prediction

After comparing structural and chemical approaches for predicting simple interactions between drugs, a question arises about what conclusions we would reach if we expanded our research on the broader field of predicting multi-labeled drug interactions. In this broader area, predicting that two drugs interact is not enough, we also need to define a set of different types of interaction, and then proceed to predict which one applies to the positive sample at hand. In the same manner that we studied drug categories and the efficiency of models in each one, it would be useful to experiment on different types of interactions and research on which of those types render chemical or network information a better resource to base a predictor. Similar work is found in [8]; in this paper the authors

not only base their experiments on multi-label DDIs prediction on network information of an interactions graph, but also expand their data graph by adding nodes regarding proteins and edges that connect them with other proteins or drugs that they may interact with.

5.5.2 Sampling for Negative Interactions

Judging by the classification reports in chapter 4, random generation of negative interactions led to easily identifiable samples; boosting the overall accuracy of our models (mainly the node2vec and hybrid classifiers) without, a similar boost on their recall measures. Further study on techniques for negative samples generation could improve the training set of our experiments and, possibly, allow for better predictors. When it comes to multi-relational link prediction, [9] provides a notable analysis on the importance of negative sampling, as well as useful methods for negative sample generation, such as the "corruption of positive samples" or "nearest neighbor sampling".

Appendixes

A

Classifier Training Data & Classification Reports

In table A.1 we present the dataset split information for each classifier's experiment. Also, tables A.2, A.3, A.4 present the classification reports we produced with python's Scikit-Learn library for mol2vec, node2vec and hybrid DDIs prediction models respectively. [1].

	Class	Train	Val	Test
mol2vec	0	1.502.158	266.288	870.700
	1	1.376.692	241.746	797.468
	Total	2.878.850	508.034	1.668.168
	Class	Train	Val	Test
node2vec	0	1.416.939	250.871	821.458
	1	1.036.784	182.140	1.196.982
	Total	2.453.723	433.011	2.018.440
	Class	Train	Val	Test
hybrid	0	1.417.479	250.615	821.598
	1	1.444.595	254.458	716.853
	Total	2.862.074	505.073	1.538.451

Table A.1. *Data-set split for each classifier*

In chapter 3 we note that each drug interaction inside DrugBank's dataset is responsible for two positive samples for each experiment. More specifically, if drugs d_i and d_j have embeddings m_i and m_j, we can produce the concatenated vectors of m_ij and m_ji. In our dataset splits for all our experiments, we make sure that these pairs of positive samples for each drug interaction are always placed in the same set (train, evaluation or test). This detail is important in order to keep training information from infiltrating our testing set and, possibly, affect our results.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Class	Precision	Recall	F1-score	Support
0	0.91	0.89	0.90	870700
1	0.88	0.91	0.89	797468
accuracy	-	-	0.90	1668168
macro avg	0.90	0.90	0.90	1668168
weighted avg	0.90	0.90	0.90	1668168

Table A.2. *Classification report for mol2vec classifier*

Class	Precision	Recall	F1-score	Support
0	0.90	0.94	0.92	821623
1	0.96	0.93	0.94	1195957
accuracy	-	-	0.93	2017580
macro avg	0.93	0.94	0.93	2017580
weighted avg	0.93	0.93	0.93	2017580

Table A.3. *Classification report for node2vec classifier*

Class	Precision	Recall	F1-score	Support
0	0.96	0.94	0.95	821598
1	0.93	0.95	0.94	716853
accuracy	-	-	0.95	1538451
macro avg	0.95	0.95	0.95	1538451
weighted avg	0.95	0.95	0.95	1538451

Table A.4. *Classification report for hybrid classifier*

B

Classifier Evaluation by DrugBank Label

In order to study each classifier's efficiency per drug category — in chapter 3 we present DrugBank's labels for drugs in the dataset — we isolate each set of corresponding samples in the test set of each experiment. Table B.1 provides a complete report on each classifier's evaluation per drug category; in parenthesis we compute the measurement that refers only to the positive or negative samples of a category.

	Category	Test Samples (Pos./Neg.)	Accuracy (Pos./Neg.) %
mol2vec	Experimental	297423 (118670/178753)	87 (86/88)
	Approved	714569 (361902/352667)	85 (88/83)
	Investigational	210752 (68759/141993)	87 (83/89)
	Vet approved	22510 (7568/14942)	88 (84/90)
	Withdrawn	21322 (12975/8347)	87 (92/80)
	Illicit	7245 (3852/3393)	87 (89/86)
	Nutraceutical	2613 (480/2133)	91 (66/96)
node2vec	Experimental	694346 (356408/337938)	94 (92/96)
	Approved	1750771 (1084681/666090)	93 (93/93)
	Investigational	473938 (206718/267220)	94 (91/97)
	Vet approved	51208 (22983/28225)	94 (91/96)
	Withdrawn	54759 (38899/15860)	96 (96/95)
	Illicit	18005 (11562/6443)	95 (95/95)
	Nutraceutical	5433 (1412/4021)	90 (72/96)
hybrid	Experimental	550343 (213685/336658)	96 (96/96)
	Approved	1316602 (650360/666242)	94 (95/93)
	Investigational	391811 (123579/123579)	95 (94/96)
	Vet approved	41980 (13768/28212)	97 (96/98)
	Withdrawn	39540 (23466/16074)	97 (98/95)
	Illicit	13017 (6859/6158)	96 (96/96)
	Nutraceutical	5024 (839/4185)	97 (90/97)

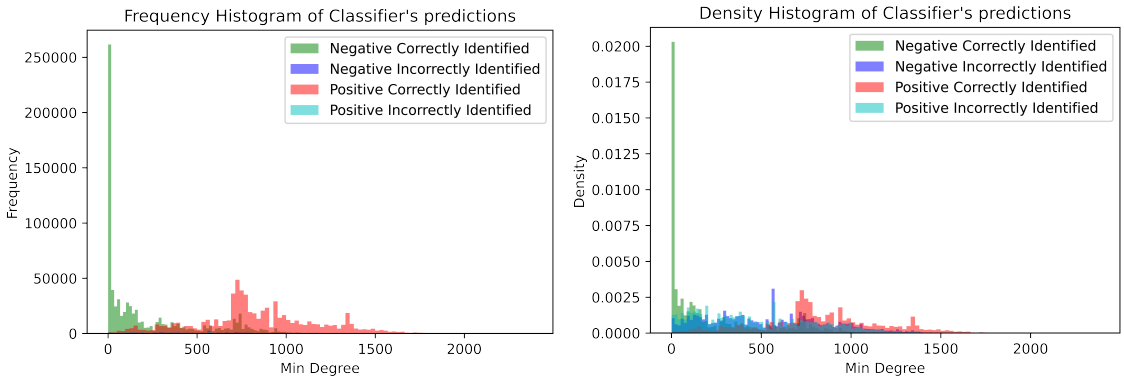
Table B.1. *Classifier test results by DrugBank's drug categories*

It is apparent in this table that the hybrid model shows not only the highest overall accuracy for each drug category, but also better recall for each sample class (positive and negative); validating our understanding in chapter 4 that there is a considerable knowledge gain by combining chemical and network information for the problem of DDIs prediction — even at the cost of larger vector embeddings.

C

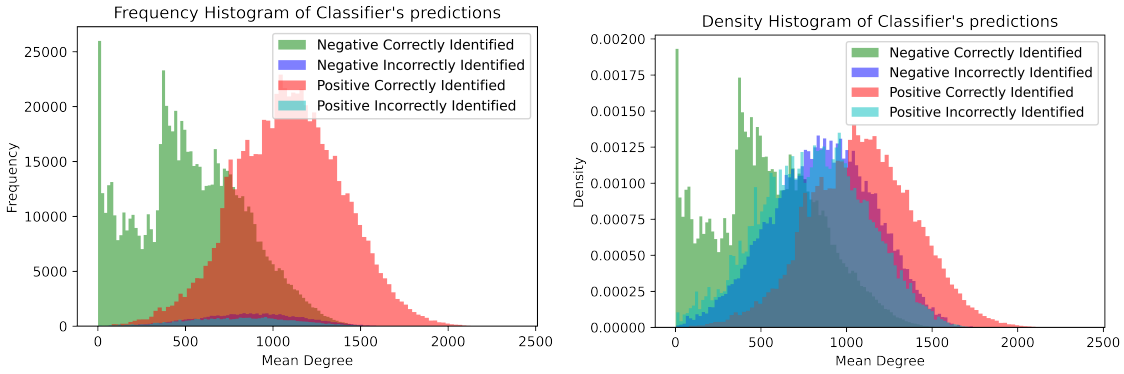
In-depth Study of hybrid Model's Performance

For completeness, we expand chapter 4 by presenting the result data regarding the hybrid model's test samples' distributions per prediction category.



(a) *Frequency Histogram over Min. Degree* **(b)** *Density Histogram over Min. Degree*

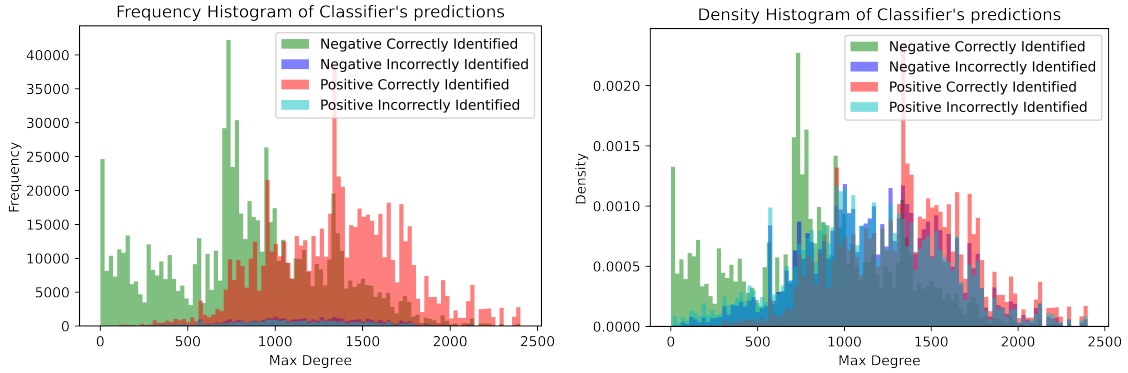
Figure C.1. *Frequency and Density Histograms for hybrid model's sample distributions' Min. Degree per prediction type*



(a) *Frequency Histogram over Avg. Degree* **(b)** *Density Histogram over Avg. Degree*

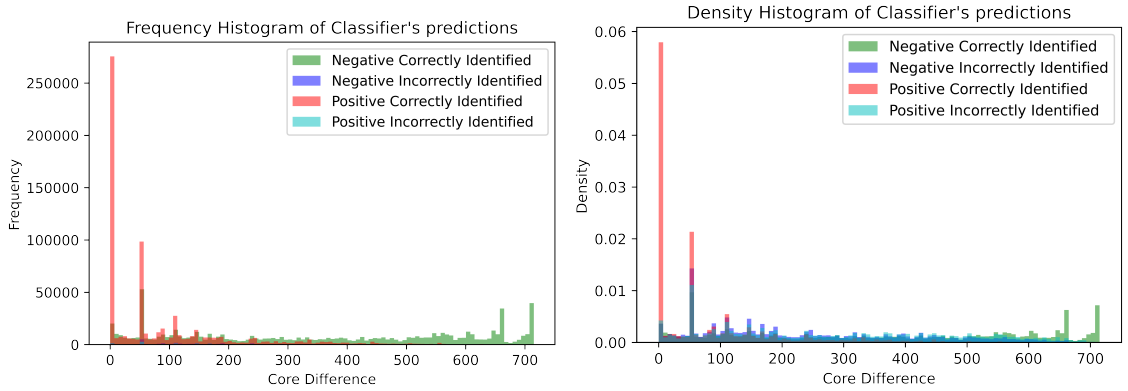
Figure C.2. *Frequency and Density Histograms for hybrid model's sample distributions' Min. Degree per prediction type*

Figures C.1a and C.1b present the frequency and density histograms respectively for all the test samples (edges). Similarly, figures C.2a and C.2b study the average degree



(a) *Frequency Histogram over Max. Degree* **(b)** *Density Histogram over Max. Degree*

Figure C.3. *Frequency and Density Histograms for hybrid model's sample distributions' Max. Degree per prediction type*



(a) *Frequency Histogram over Core Difference* **(b)** *Density Histogram over Core Difference*

Figure C.4. *Frequency and Density Histograms for hybrid model's sample distributions' Core Difference per prediction type*

and figures C.3a and C.3b report on the maximum degree. Core Difference is studied in figures C.4a and C.4b

Also, figures C.5, C.6, C.7, C.8 present the separated distribution plots over minimum, average, maximum degree, as well as the core difference of the test samples.

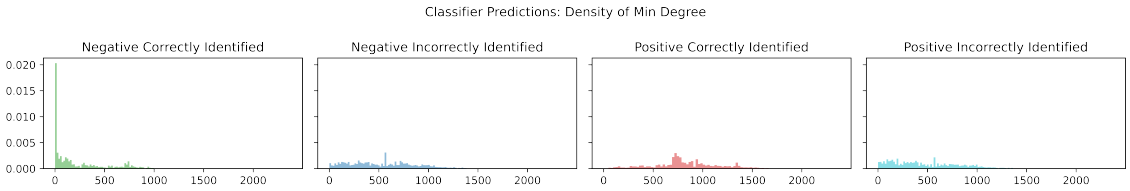


Figure C.5. *Density Histograms over Min. Degree for each prediction category of hybrid model's test samples*

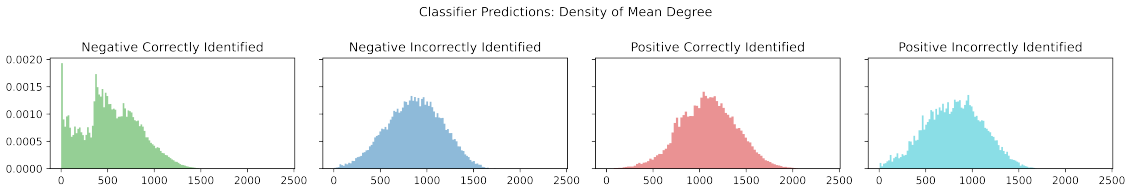


Figure C.6. *Density Histograms over Avg. Degree for each prediction category of hybrid model's test samples*

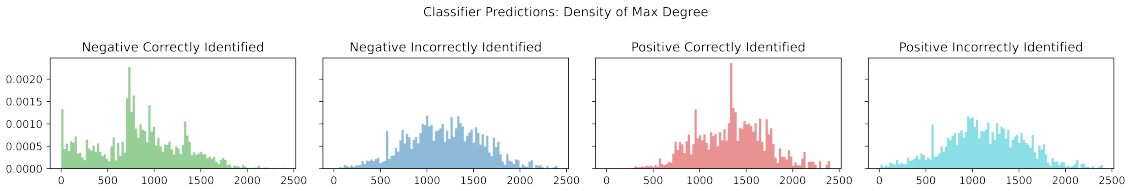


Figure C.7. *Density Histograms over Max. Degree for each prediction category of hybrid model's test samples*

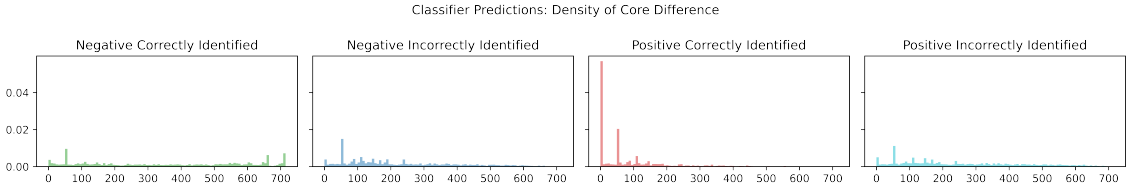


Figure C.8. *Density Histograms over Core Difference for each prediction category of hybrid model's test samples*

Bibliography

[1] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox and Michael Wilson. *DrugBank 5.0: a major update to the DrugBank database for 2018*. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2017.

[2] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. *Efficient estimation of word representations in vector space*. *arXiv preprint arXiv:1301.3781*, 2013.

[3] Sabrina Jaeger, Simone Fulle and Samo Turk. *Mol2vec: unsupervised machine learning approach with chemical intuition*. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

[4] Aditya Grover and Jure Leskovec. *node2vec: Scalable feature learning for networks*. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[5] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu and others. *PubChem 2019 update: improved access to chemical data*. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

[6] David Weininger. *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[7] Greg Landrum and others. *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling*, 2013.

[8] Marinka Zitnik, Monica Agrawal and Jure Leskovec. *Modeling polypharmacy side effects with graph convolutional networks*. *Bioinformatics*, 34(13):i457–i466, 2018.

[9] Bhushan Kotnis and Vivi Nastase. *Analysis of the impact of negative sampling on link prediction in knowledge graphs*. *arXiv preprint arXiv:1708.06816*, 2017.

[10] Adam S Crystal, Alice T Shaw, Lecia V Sequist, Luc Friboulet, Matthew J Niederst, Elizabeth L Lockerman, Rosa L Frias, Justin F Gainor, Arnaud Amzallag, Patricia Greninger and others. *Patient-derived models of acquired resistance can identify effective drug combinations for cancer*. *Science*, 346(6216):1480–1486, 2014.

[11] Feixiong Cheng, István A Kovács and Albert László Barabási. *Network-based prediction of drug combinations*. *Nature communications*, 10(1):1–11, 2019.

[12] Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian and Xiaohong Li. *Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data*. *BMC bioinformatics*, 18(1):1–12, 2017.

[13] Yang Yang, Ryan N Lichtenwalter and Nitesh V Chawla. *Evaluating link prediction methods*. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[14] Herman Yuliansyah, Zulaiha Ali Othman and Azuraliza Abu Bakar. *Taxonomy of Link Prediction for Social Network Analysis: A Review*. *IEEE Access*, 8:183470–183487, 2020.

[15] Peng Wang, BaoWen Xu, YuRong Wu and XiaoYu Zhou. *Link prediction in social networks: the state-of-the-art*. *Science China Information Sciences*, 58(1):1–38, 2015.

[16] Isabel Segura-Bedmar, Paloma Martínez and María Herrero-Zazo. *Lessons learnt from the DDIExtraction-2013 shared task*. *Journal of biomedical informatics*, 51:152–164, 2014.

[17] Aditya Krishna Menon and Charles Elkan. *Link prediction via matrix factorization*. *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer, 2011.

[18] Daniel M Dunlavy, Tamara G Kolda and Evrim Acar. *Temporal link prediction using matrix and tensor factorizations*. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–27, 2011.

[19] Seyed Mehran Kazemi and David Poole. *Simple embedding for link prediction in knowledge graphs*. *Advances in neural information processing systems*, 31, 2018.

[20] Muhan Zhang and Yixin Chen. *Link prediction based on graph neural networks*. *Advances in neural information processing systems*, 31, 2018.

[21] Yahui Long, Min Wu, Yong Liu, Yuan Fang, Chee Keong Kwoh, Jinmiao Chen, Jiawei Luo and Xiaoli Li. *Pre-training graph neural networks for link prediction in biomedical networks*. *Bioinformatics*, 38(8):2254–2262, 2022.

[22] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska and Kayvan Najarian. *Machine learning approaches and databases for prediction of drug-target interaction: a survey paper*. *Briefings in bioinformatics*, 22(1):247–269, 2021.

[23] Andrej Kastrin, Polonca Ferk and Brane Leskošek. *Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning*. *PloS one*, 13(5):e0196865, 2018.

[24] Daiki Koge, Naoaki Ono, Ming Huang, Md Altaf-Ul-Amin and Shigehiko Kanaya. *Embedding of molecular structure using molecular hypergraph variational autoencoder with metric learning*. *Molecular informatics*, 40(2):2000203, 2021.

[25] David JC MacKay, David JC Mac Kay and others. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[26] Jari Saramäki, Mikko Kivelä, Jukka Pekka Onnela, Kimmo Kaski and Janos Kertesz. *Generalizations of the clustering coefficient to weighted complex networks*. *Physical Review E*, 75(2):027105, 2007.

[27] Marcus Kaiser. *Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks*. *New Journal of Physics*, 10(8):083042, 2008.

[28] Phillip Bonacich. *Power and centrality: A family of measures*. *American journal of sociology*, 92(5):1170–1182, 1987.

[29] S Unnikrishna Pillai, Torsten Suel and Seunghun Cha. *The Perron-Frobenius theorem: some of its applications*. *IEEE Signal Processing Magazine*, 22(2):62–75, 2005.

[30] Linton C Freeman. *Centrality in social networks conceptual clarification*. *Social networks*, 1(3):215–239, 1978.

[31] Ulrik Brandes. *On variants of shortest-path betweenness centrality and their generic computation*. *Social networks*, 30(2):136–145, 2008.

[32] Linton C Freeman. *Centrality in social networks conceptual clarification*. *Social networks*, 1(3):215–239, 1978.

[33] Stanley Wasserman, Katherine Faust and others. *Social network analysis: Methods and applications*. 1994.

[34] Phillip Bonacich. *Some unique properties of eigenvector centrality*. *Social networks*, 29(4):555–564, 2007.

[35] Aric Hagberg, Pieter Swart and Daniel S Chult. *Exploring network structure, dynamics, and function using networkx*. 2008.

[36] Wei Zhang, Tingjun Hou, Xuebin Qiao and Xiaojie Xu. *Some basic data structures and algorithms for chemical generic programming*. *Journal of chemical information and computer sciences*, 44(5):1571–1575, 2004.

List of Abbreviations

CBOW	Continuous Bag-of-Words
DDIs	Drug-Drug Interactions
KL Divergence	Kullback–Leibler Divergence
ML	Machine Learning
NN	Neural Network
NLP	Natural Language Processing
NSCI	Negative Samples Correctly Identified
NSII	Negative Samples Incorrectly Identified
PSCI	Positive Samples Correctly Identified
PSII	Positive Samples Incorrectly Identified
SNNs	Shallow Neural Networks
SMILES	Simplified Molecular-Input Line-Entry System
VSMs	Vector Space Models