



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ



ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ
ΙΝΣΤΙΤΟΥΤΟ ΧΗΜΙΚΗΣ ΒΙΟΛΟΓΙΑΣ

**ΔΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΒΙΟΕΠΙΧΕΙΡΕΙΝ**



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Εξερεύνηση Βάσεων Δεδομένων με τα Εργαλεία
Χημειοπληροφορικής Enalos**

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΑΝΑΓΙΩΤΗΣ ΓΕΩΡΓΙΑΔΗΣ, ΕΡΕΥΝΗΤΗΣ Β΄

ΤΕΧΝΙΚΟΙ ΣΥΜΒΟΥΛΟΙ

ΑΝΔΡΕΑΣ ΑΦΑΝΤΙΤΗΣ: ΔΙΕΥΘΥΝΩΝ ΣΥΜΒΟΥΛΟΣ ΣΤΗΝ NOVAMECHANICS LTD

ΓΕΩΡΓΙΑ ΜΕΛΑΓΡΑΚΗ: ΕΠΙΚ. ΚΑΘΗΓΗΤΡΙΑ ΣΣΕ

ΛΕΚΑ ΓΙΟΥΓΚΕΝΑ

A.M.: 00077

ΑΘΗΝΑ, 2022



UNIVERSITY OF THESSALY
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF BIOCHEMISTRY AND BIOTECHNOLOGY

NATIONAL HELLENIC RESEARCH FOUNDATION
INSTITUTE OF CHEMICAL BIOLOGY



INTERINSTITUTIONAL PROGRAM OF POSTGRADUATE STUDIES
IN BIOENTREPRENEURSHIP



MASTER THESIS

Database Exploration with Enalos Cheminformatics Tools

SUPERVISOR: PANAGIOTIS GEORGIADIS, SENIOR RESEARCHER

TECHNICAL ADVISORS

ANDREAS AFANTITIS: MANAGING DIRECTORS AT NOVA
MECHANICS

GEORGIA MELAGRAKI: ASSISTANT PROFESSOR AT HELLENIC
MILITARY ACADEMY

LEKA GIOUGKENA

A.M.: 00077

2

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στο

ΒΙΟΕΠΙΧΕΙΡΕΙΝ

που απονέμει το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, σε συνεργασία με την εταιρεία NovaMechanics Ltd.

Εγκρίθηκε την από την τριμελή εξεταστική επιτροπή:

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ	ΥΠΟΓΡΑΦΗ
Παναγιώτης Γεωργιάδης	Ερευνητής Β'	
Βασιλική Πλέτσα	Ερευνήτρια Β'	
Παναγιώτης Ζουμπουλάκης	Αν. Καθηγητής, Παν.Δυτικής Αττικής	

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης που απονέμει το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας σε συνεργασία με το Εθνικό Ίδρυμα Ερευνών κατά το ακαδημαϊκό έτος 2020-2022 υπό την επίβλεψη του Ερευνητή Β΄, κυρίου Παναγιώτη Γεωργιάδη.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Παναγιώτη Γεωργιάδη για τις γνώσεις που μου μετέδωσε κατά την διάρκεια των σπουδών μου καθώς και για την βοήθεια του κατά την εκπόνηση της διπλωματικής μου εργασίας. Ακόμη, επιθυμώ να ευχαριστήσω τα μέλη της τριμελούς επιτροπής κ. Βασιλική Πλέτσα και κ. Παναγιώτη Ζουμπουλάκη για το χρόνο που αφιέρωσαν ως μέλη της τριμελούς επιτροπής και φυσικά όλους τους διδάσκοντες του προγράμματος σπουδών για τις γνώσεις και τη βοήθεια που μου παρείχαν καθόλη τη διάρκεια της φοίτησής μου.

Επίσης, θα ήθελα να ευχαριστήσω θερμά τον Ανδρέα Αφαντίτη για την εξαιρετική συνεργασία που είχαμε στα πλαίσια της διπλωματικής μου εργασίας και την κ. Γεωργία Μελαγράκη για την βοήθεια και την καθοδήγηση που μου παρείχε όποτε την χρειάστηκα καθώς και για την καθοδήγηση της κατά την συγγραφή της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την στήριξη και την υπομονή που έδειξαν σε όλη την διάρκεια των σπουδών μου.

Περιεχόμενα

Περίληψη	7
Abstract	8
Σκοπός	9
Ενότητα 1. Θεωρητικό Μέρος - Εισαγωγή στην Χημειοπληροφορική	11
1.1 Χημειοπληροφορική	11
1.2 Μοριακή Απεικόνιση.....	12
1.3 Κωδικοποίηση SMILES, SMARTS, InChI και InChIkey.....	13
1.4 Μοριακά Αποτυπώματα (Molecular Fingerprints).....	15
1.5 Μοριακοί Δείκτες (Molecular Descriptors).....	19
1.6 Μοριακή ομοιότητα και ομοιότητα Tanimoto.....	20
1.7 Quantitative Structure Activity Relationship (QSAR)	22
1.7.1 Μοντέλα QSAR	22
1.7.2 Επικύρωση των μοντέλων QSAR/QSPR.....	24
1.8 Μηχανική μάθηση (Machine Learning, ML)	26
1.9 Εικονική διαλογή (Virtual screening, VS).....	28
1.10 Χρήσιμα εργαλεία Χημειοπληροφορικής.....	29
1.10.1 Βάσεις Δεδομένων	29
1.10.2 Χρήσιμα Εργαλεία Λογισμικού στην Χημειοπληροφορική	34
Ενότητα 2. Εργαλεία Χημειοπληροφορικής Enalos	36
2.1 Εισαγωγή.....	36
2.2 Enalos Suite	37
2.3 Λειτουργίες του Enalos Suite	39
2.3.1 Λειτουργία CIR.....	39
2.3.2 Λειτουργίες του PubChem.....	40
2.3.3 Λειτουργία του UniChem.....	41
2.3.4 Έλεγχος Καινοτομίας	42
2.3.5 Μοριακοί δείκτες του Enalos Suite.....	42
2.4 Μοντέλα πρόβλεψης Enalos Cloud & Enalos Suite	42
2.4.1 Μοντέλο κυτταροτοξικότητας MouseTox.....	43
2.4.2 Μοντέλο αναστολής K562.....	46
2.4.3 Μοντέλο αναστολής του παράγοντα TNF (Tumor Necrosis Factor).....	48
2.4.4 Λοιπά μοντέλα του Enalos Suite.....	49
Ενότητα 3. Μελέτη περίπτωσης - Μεθοδολογία	50
3.1 Εισαγωγή.....	50

3.2 Παράγοντας νέκρωσης όγκων (Tumor Necrosis Factor, TNF).....	51
3.3 Αναστολείς TNF.....	52
3.4 Ανάκτηση Δεδομένων & έλεγχος ομοιότητας.....	52
3.5 Υπολογισμός αναστολής & τοξικότητας	55
3.6 Εμπορική διαθεσιμότητα & κάλυψη διπλωμάτων ευρεσιτεχνίας.....	55
Ενότητα 4. Αποτελέσματα.....	55
4.1 Αποτελέσματα ελέγχου ομοιότητας μέσω του συντελεστή Tanimoto.....	55
4.2 Αποτελέσματα αναστολής & τοξικότητας	56
4.3 Αποτελέσματα εμπορικής διαθεσιμότητας και διπλωμάτων ευρεσιτεχνίας.....	57
4.4 Σχολιασμός Αποτελεσμάτων	65
Συμπεράσματα	69
Αναφορές-Πηγές.....	71

Περίληψη

Οι υπολογιστικές προσεγγίσεις του τομέα της χημειοπληροφορικής στις μέρες μας, αποτελούν αναπόσπαστο μέρος της έρευνας και της ανακάλυψης φαρμάκων. Τα επιτεύγματα της ψηφιακής εποχής έχουν περιορίσει σημαντικά προβλήματα που σχετίζονται με την ερμηνεία πειραματικών αποτελεσμάτων και την ανάλυση δεδομένων ενώ παράλληλα έχουν ελαχιστοποιηθεί σε μεγάλο βαθμό δαπανηρές και χρονοβόρες πειραματικές διαδικασίες καθώς και δοκιμές σε ζώα.

Στην πρώτη ενότητα της παρούσας διπλωματικής εργασίας επεξηγούνται βασικές έννοιες του τομέα της χημειοπληροφορικής όπως είναι η μοριακή απεικόνιση, τα μοριακά αποτυπώματα και οι μοριακοί δείκτες καθώς και η έννοια της ομοιότητας μεταξύ των μορίων. Αναλύονται επίσης χρήσιμα εργαλεία χημειοπληροφορικής καθώς και σημαντικοί τομείς της που έχουν συμβάλει στην ανακάλυψη των φαρμάκων όπως είναι η μηχανική μάθηση και η εικονική διαλογή. Στην δεύτερη ενότητα αναλύονται τα εργαλεία χημειοπληροφορικής Enalos τα οποία έχουν σχεδιαστεί από την εταιρεία NovaMechanics Ltd. Συγκεκριμένα, γίνεται αναφορά των βασικών λειτουργιών του Enalos Suite και του Enalos Cloud καθώς και όλων των μοντέλων πρόβλεψης που βοηθούν, μέσω της μοριακής μοντελοποίησης, τον χρήστη να έχει απευθείας πρόσβαση σε πολλαπλές βάσεις δεδομένων για εξόρυξη και χειρισμό δεδομένων. Για την κατανόηση της χρησιμότητας και της λειτουργίας των εργαλείων χημειοπληροφορικής Enalos, στην τρίτη ενότητα αναφέρεται μια μελέτη περίπτωσης με στόχο την ανακάλυψη φαρμάκων τα οποία αναστέλλουν τον παράγοντα νέκρωσης όγκων TNF. Στην τέταρτη ενότητα παρατίθενται τα αποτελέσματα αυτής της μελέτης και τέλος ακολουθούν σχολιασμός των αποτελεσμάτων και μελλοντικές προοπτικές με βάση τις πλέον υποσχόμενες ενώσεις που προέκυψαν από την μελέτη.

Λέξεις-Κλειδιά: *Χημειοπληροφορική, Βάσεις δεδομένων, Μοντέλο πρόβλεψης, ομοιότητα Tanimoto, παράγοντας TNF, Enalos Suite*

Abstract

Nowadays, the computational approaches within a cheminformatics context are an integral part of drug research and discovery. The achievements of the digital age have addressed the problems related to the interpretation of experimental results and data analysis and at the same time have greatly minimized costly and time-consuming experimental procedures as well as animal testing.

The first section of this dissertation reports main concepts in the field of cheminformatics such as molecular representation, molecular fingerprints, molecular descriptors, and the concept of similarity between molecules. Useful tools of cheminformatics are also analyzed and important areas of chemistry that have contributed to the discovery of drugs such as machine learning and virtual screening. The second section describes the Enalos cheminformatics tools designed by NovaMechanics Ltd. Specifically, the main functions of the Enalos Suite and the Enalos Cloud are reported, as well as all the predictive models that allow a user to have direct access to validated predictive models as well as multiple databases for data mining and data handling. In order to understand the usefulness and function of Enalos cheminformatics tools, a case study targeting the drug discovery of TNF inhibitors is reported in the third section and future perspectives based on the most promising compounds that emerged from this study are presented in the fourth section.

Keywords: Chemoinformatics, Databases, Predictive model, Tanimoto similarity, TNF factor, Enalos Suite

Σκοπός

Ζώντας στον 21ο αιώνα, στην εποχή των μεγάλων δεδομένων και της προηγμένης ανάλυσης, θεωρείται απαραίτητη η δυνατότητα διαχείρισης πληροφοριών που παρέχονται από πληθώρα ερευνών και ανασκοπήσεων σε όλους τους τομείς συμπεριλαμβανομένου και του τομέα ανάπτυξης των φαρμάκων. Έτσι, μέσω της χρήσης και της ανάπτυξης διάφορων εργαλείων χημειοπληροφορικής επιτυγχάνεται η βέλτιστη και ταχύτερη επεξεργασία χημικών δεδομένων. Σκοπός της παρούσας διπλωματικής εργασίας είναι η αξιοποίηση των χημικών βάσεων δεδομένων με την βοήθεια των εργαλείων χημειοπληροφορικής Enalos τα οποία προσφέρουν ένα φάσμα δυνατοτήτων τόσο για την αξιοποίηση όσο και ανάλυση δεδομένων.

Κατάλογος συντομογραφιών

ECFP	Extended Connectivity Fingerprint (Αποτύπωμα εκτεταμένης σύνδεσης)
FCFP	Functional Class Fingerprint (Αποτύπωμα λειτουργικής κλάσης)
InChI	International Chemical Identifier (Διεθνές χημικό αναγνωριστικό)
IUPAC	International Union of Pure and Applied Chemistry (Διεθνής Ένωση Καθαρής και Εφαρμοσμένης Χημείας)
SMILES	Simplified Molecular Input Line Entry System (Σύστημα απλοποιημένης μοριακής γραμμικής γραφής)
SMARTS	SMILES Arbitrary Target Specification
QSAR	Quantitative Structure-Activity Relationship (ποσοτική σχέση δομής-δραστικότητας)
QSPR	Quantitative Structure-Property Relationship (ποσοτική σχέση δομής-ιδιότητας)
TNF	Tumor Necrosis Factor

Ενότητα 1. Θεωρητικό Μέρος - Εισαγωγή στην Χημειοπληροφορική

1.1 Χημειοπληροφορική

Η επιστήμη της Χημείας από τα πρώτα κι όλας χρόνια βασιζόταν σε δεδομένα που προκύπταν από τις παρατηρήσεις πειραμάτων. Η θεωρητική Χημεία με το πέρασμα των χρόνων, κατάφερε να φτάσει σε τέτοιο βαθμό που να μπορεί σε μερικές περιπτώσεις να εξάγει δεδομένα με υπολογιστικούς τρόπους. Ακόμα, αρκετά χημικά φαινόμενα είναι ιδιαίτερα περίπλοκα για να μπορέσουν να ερμηνευτούν με παραδοσιακούς υπολογιστικούς τρόπους. Έτσι από την δεκαετία του 1960 άρχισαν να γίνονται προσπάθειες ώστε να χρησιμοποιηθεί η υπολογιστική ισχύς των ηλεκτρονικών υπολογιστών για την μοντελοποίηση και αποσαφήνιση των χημικών φαινομένων. Αυτό είχε ως αποτέλεσμα την δημιουργία της χημειοπληροφορικής (Gasteiger, 2016). Ο τομέας της χημειοπληροφορικής έλαβε το όνομα του το 1996 από τον Frank Brown και έχει επηρεάσει σημαντικά την ανακάλυψη των φαρμάκων (Bender & Nathan 2018). Όμως, η επιρροή της χημειοπληροφορικής είναι ευρεία και σε άλλους τομείς. Ο όρος χημειοπληροφορική συγχωνεύει τρεις διαφορετικές επιστήμες, τη Χημεία, την Πληροφορική και τα Μαθηματικά, σε ένα κοινό διεπιστημονικό πεδίο. Ο επικρατέστερος ορισμός του όρου χημειοπληροφορική είναι ο ακόλουθος: πρόκειται για το πεδίο το οποίο εφαρμόζει μεθόδους της Πληροφορικής για να επιλύσει χημικά προβλήματα (Gasteiger & Engel, 2003). Το πεδίο της χημειοπληροφορικής εξαπλώθηκε γρήγορα, παράλληλα με την αλματώδη ανάπτυξη της πληροφορικής και με τον ολοένα και αυξανόμενο όγκο των διαθέσιμων πληροφοριών, ιδιαίτερα στον τομέα της αναζήτησης και της ανάκτησης σε συστήματα πληροφοριών χημικών δομών. Η τελευταία δεκαετία έχει φέρει κοντά στο επιστημονικό κοινό, μεγάλο εύρος χημικών δεδομένων με βάσεις δεδομένων όπως η PubChem, η ChEMBL και πολλές άλλες. Μία από τις κύριες πρακτικές εφαρμογές της χημειοπληροφορικής είναι η δημιουργία μοντέλων πρόβλεψης για διάφορες ιδιότητες όπως φυσικοχημικές ιδιότητες, δραστικότητα, ή τις αλληλεπιδράσεις των ενώσεων-στόχου. Εκτός από τις επιθυμητές βιολογικές δραστικότητες, τα μοντέλα χημειοπληροφορικής μπορούν επίσης να προβλέψουν ανεπιθύμητες ιδιότητες ενώσεων, ή ιδιότητες που σχετίζονται με απορρόφηση, κατανομή, μεταβολισμό, απέκκριση και τοξικότητα μιας ένωσης (Andreas and Nathan 2018).

1.2 Μοριακή Απεικόνιση

Η μεθοδολογία που χρησιμοποιείται για την δημιουργία της αναπαράστασης των μορίων επηρεάζει άμεσα την ποσότητα και τον τύπο των χημικών πληροφοριών που θα διατηρηθούν. Για να γίνει καλύτερα αντιληπτή η έννοια της χημικής δομής δημιουργήθηκε μια διαδικασία απεικόνισης των προς μελέτη μορίων. Μπορούν να χρησιμοποιηθούν μοριακές αναπαραστάσεις διαφορετικών επιπέδων πολυπλοκότητας ανάλογα με την χημική πληροφορία που θέλουμε να πάρουμε από μία ένωση (Grisoni et al., 2018):

-Μηδενικής διάστασης (0D): Η απλούστερη κατηγορία μοριακής απεικόνισης είναι ο χημικός τύπος μιας ένωσης, που απεικονίζει τα ποιοτικά (το σύνολο των χημικών στοιχείων της ένωσης) και ποσοτικά (τον αριθμό των ατόμων κάθε στοιχείου στο μόριο της ένωσης) χαρακτηριστικά της. Παράδειγμα αποτελεί η ιβουπροφίνη με χημικό τύπο $C_{13}H_{18}O_2$, ο οποίος μας δείχνει την παρουσία 13 ατόμων άνθρακα, 18 ατόμων υδρογόνου και 2 ατόμων οξυγόνου στο μόριο της ιβουπροφίνης. Αυτός ο τρόπος απεικόνισης είναι ανεξάρτητος από οποιαδήποτε γνώση σχετικά με τη μοριακή δομή και τη συνδεσιμότητα των ατόμων της ιβουπροφίνης. Η μηδενικής διάστασης απεικόνιση (0D) χημικών μορίων είναι πολύ απλή στον υπολογισμό και στην ερμηνεία της, αλλά παρέχει περιορισμένο αριθμό πληροφοριών.

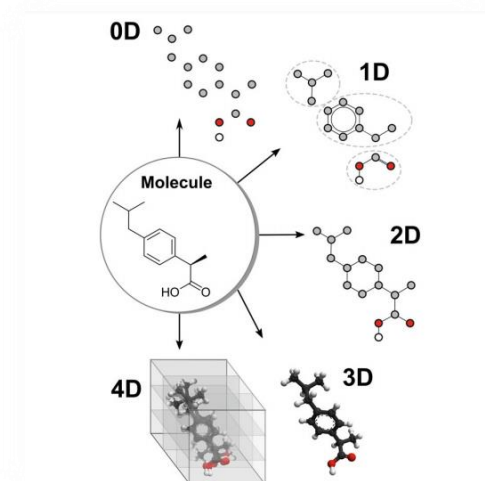
-Μονοδιάστατη (1D): Τα μόρια αναπαρίστανται ανά τμηματικές δομές (substructures) ενδιαφέροντος, όπως είναι τα μοριακά θραύσματα, οι λειτουργικές ομάδες, ή δομές υποκατάστασης. Αυτή η απεικόνιση δεν προϋποθέτει την ακριβή γνώση της χημικής δομής.

- Δισδιάστατη (2D): Αυτή η απεικόνιση εξετάζει τον τρόπο σύνδεσης των ατόμων, όσον αφορά την παρουσία και τη φύση των χημικών δεσμών. Συνήθως το μόριο παρουσιάζεται σαν ένα γράφημα του οποίου οι ακμές είναι οι δεσμοί και οι κορυφές είναι τα άτομα. Συχνά, λαμβάνονται υπόψη συγκεκριμένες χημικές ιδιότητες των ατόμων, π.χ. μάζα, πολικότητα κλπ.

- Τρισδιάστατη (3D): Η τρισδιάστατη αναπαράσταση απεικονίζει ένα μόριο ως γεωμετρικό αντικείμενο στον χώρο και, εκτός από τη φύση και τη συνδεσιμότητα των ατόμων, δίνει πληροφορίες και για τη διαμόρφωσή τους στον χώρο. Η τρισδιάστατη απεικόνιση χημικών μορίων παρέχει μεγάλο περιεχόμενο πληροφοριών και μπορεί να είναι ιδιαίτερα χρήσιμη για τη μοντελοποίηση φαρμακευτικών και βιολογικών ιδιοτήτων.

- Τετραδιάστατη (4D): Εκτός από τη μοριακή γεωμετρία, μπορεί να εισαχθεί μια «τέταρτη διάσταση», που συνήθως στοχεύει στον ποσοτικό προσδιορισμό και τον

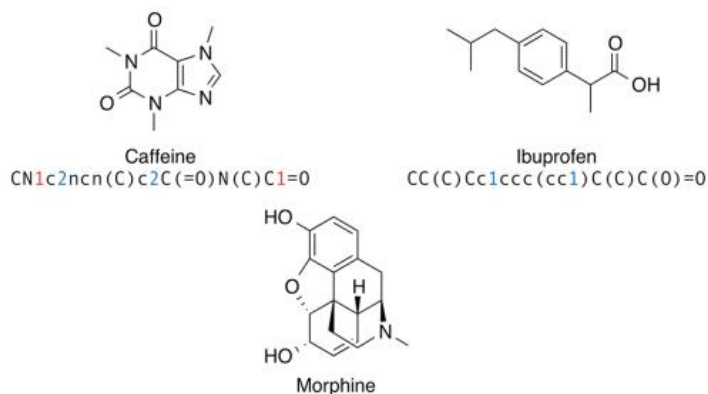
χαρακτηρισμό των αλληλεπιδράσεων μεταξύ ενός μορίου και μιας ενεργής θέσης ενός υποδοχέα (Grisoni et al., 2018).



Εικόνα 1.1. Παράδειγμα διαφορετικών μοριακών αναπαραστάσεων της ίδιας δομής (ιβουπροφαίνη) (Grisoni F. et al, 2018)

1.3 Κωδικοποίηση SMILES, SMARTS, InChI και InChIkey

Συνήθως τα μόρια μοντελοποιούνται με την μορφή γραφικών παραστάσεων, οι οποίες στην επιστήμη της χημείας ονομάζονται δομές Lewis. Στις δομές αυτές τα άτομα χαρακτηρίζονται ως κόμβοι και τα άκρα αυτών είναι οι δεσμοί μεταξύ των ατόμων. Θα μπορούσε επομένως να φανταστεί κανείς ότι έχει ένα μοντέλο όπως τα SMILES που διαβάζει και εξάγει γραφήματα. Τα SMILES (Simplified Molecular Input Line Entry System) προτάθηκαν από τον Weininger το 1987 και χρησιμοποιούνται για την αναπαράσταση της αρχιτεκτονικής ενός μορίου (Segler et al., 2017). Το σύστημα απλοποιημένης μοριακής γραμμικής γραφής (SMILES) είναι μια προδιαγραφή με τη μορφή μιας γραμμικής σημειογραφίας για την περιγραφή της δομής των χημικών ειδών χρησιμοποιώντας σύντομες συμβολοσειρές ASCII (American Standard Code for Information Interchange). Περιγράφει μόρια χρησιμοποιώντας ένα αλφάβητο για κάθε χαρακτήρα (Segler et al., 2017). Επίσης, τα SMILES παρέχουν πιο περίπλοκη πληροφορία σε σχέση με τον χημικό τύπο μιας ένωσης ενώ παράλληλα καταλαμβάνει 50-70% λιγότερο χώρο από άλλες μεθόδους μοριακής απεικόνισης.



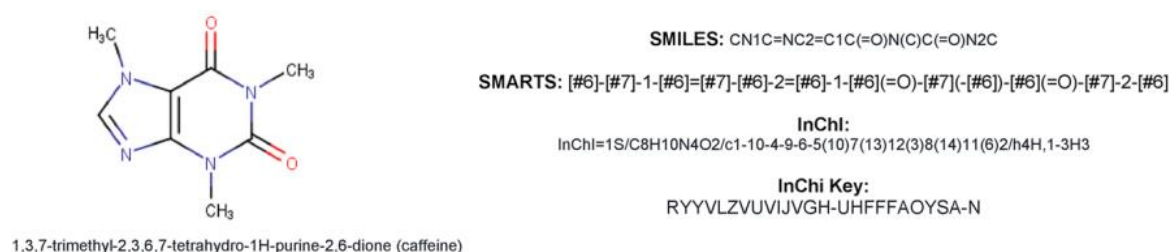
Εικόνα 1.2. Παραδείγματα μορίων και της αναπαράστασης των SMILES τους (Segler et al., 2018)

Συγκεκριμένα, για την κωδικοποίηση των μορίων σε SMILES χρησιμοποιούνται 5 βασικοί κανόνες:

1. Τα άτομα αναπαρίστανται από τα ατομικά τους γράμματα (πχ N για το άζωτο, O για το οξυγόνο). Τα ασθενή άτομα υδρογόνου δεν αναπαρίστανται αναλυτικά.
2. Τα γειτονικά άτομα τοποθετούνται το ένα δίπλα στο άλλο και οι χημικοί δεσμοί συμβολίζονται ως εξής: με '-' για τους απλούς δεσμούς, με '=' για τους διπλούς, με '#' για τους τριπλούς και με ':' για τους αρωματικούς. Οι απλοί και οι αρωματικοί δεσμοί συνήθως παραλείπονται.
3. Ό,τι περικλείεται σε παρένθεση υποδηλώνει διακλάδωση στην χημική δομή του μορίου.
4. Για την γραμμική αναπαράσταση κυκλικών δομών διαχωρίζεται ένας δεσμός από κάθε δακτύλιο και τα 2 άτομα αυτού του δεσμού ακολουθούνται από το ίδιο αριθμητικό ψηφίο.
5. Τα άτομα στους αρωματικούς δακτυλίους αναπαρίστανται με πεζά γράμματα, γεγονός που προκαλεί προβλήματα μερικές φορές στην αντίληψη των αρωματικών δομών.

Μία παραλλαγή των SMILES είναι η κωδικοποίηση SMARTS (SMILES Arbitrary Target Specification). Πρόκειται για μια γλώσσα που αναπτύχθηκε για τον προσδιορισμό μοτίβων σε δομικές υποομάδες που χρησιμοποιούνται για να ταιριάζουν με μόρια και αντιδράσεις. Οι κανόνες που χρησιμοποιούνται για να επιτευχθεί ο προσδιορισμός αυτών των υποομάδων αποτελούν μια επέκταση των κανόνων που χρησιμοποιούνται για τα SMILES. Το πρόβλημα στην χρήση των SMILES εντοπίζεται στο ότι για την ίδια μοριακή απεικόνιση μπορούν να δημιουργηθούν διαφορετικά SMILES που να την περιγράφουν. Έτσι συχνά χρησιμοποιούνται κανονικοποιημένα SMILES προκειμένου να εξασφαλίσουν την

μοναδικότητα ενός μορίου σε μία βάση δεδομένων (Saldívar-González et al., 2020). Στο σημείο αυτό προστίθεται η χρήση των InChI (International Chemical Identifier) και InChIkey. Παρόλο που τα InChI είναι δύσκολο να γίνουν απόλυτα κατανοητά από τον άνθρωπο αποτελούν μια πλήρως εξοπλισμένη, ευέλικτη και τυποποιημένη γραμματοσειρά. Η κωδικοποίηση InChI αναπτύχθηκε με την υποστήριξη της Διεθνούς Ένωσης Καθαρής και Εφαρμοσμένης Χημείας (IUPAC, International Union of Pure and Applied Chemistry) με κύριες συνεισφορές από το Ινστιτούτο NIST (US National Institute of Standards and Technology) και υποστηρίζεται από την InChI Trust, έναν μη κερδοσκοπικό οργανισμό που δημιουργήθηκε για το σκοπό αυτό. Έτσι λοιπόν, τα InChI είναι δωρεάν, ανοιχτού κώδικα και διατηρούνται από έναν μόνο οργανισμό που σημαίνει ότι δεν υπάρχουν ταυτόχρονες, παράλληλες εφαρμογές για την υποστήριξή του. Σε σύγκριση με το SMILES, το InChI χρησιμοποιεί πολύ διαφορετική λογική και αποσκοπεί στην καθιέρωση ενός μοναδικού χαρακτηρισμού για κάθε ένωση, που θα επιτρέπει την καλύτερη οργάνωση των ενώσεων σε βάσεις δεδομένων. Ένα μεγάλο μειονέκτημα των αναπαραστάσεων InChI είναι ότι μπορεί να είναι πραγματικά ογκώδη (με πολλούς χαρακτήρες) για μεγάλα μόρια, γεγονός που τα καθιστά πολλές φορές δύσχρηστα (Bajusz et al., 2017) Για να διορθωθεί αυτή η ανεπάρκεια, αναπτύχθηκε το InChIKey, μια συνεπτυγμένη ψηφιακή αναπαράσταση που προέρχεται από το InChI και αναπτύχθηκε ώστε να γίνονται πιο εύκολα οι διαδικτυακές αναζητήσεις για της χημικές δομές των ενώσεων. Το InChIKey συμπιέζει το περιεχόμενο πληροφοριών μιας αναπαράστασης InChI σε σταθερό μήκος 27 χαρακτήρων στην ακόλουθη μορφή S. R. (Heller et al., 2015):



Εικόνα 1.3. Παράδειγμα των κωδικοποιήσεων SMILES, SMARTS, InChI και InChIkey για την χημική δομή της καφεΐνης (Saldívar-González et al., 2020)

1.4 Μοριακά Αποτυπώματα (Molecular Fingerprints)

Τα μοριακά δακτυλικά αποτυπώματα είναι ένας τρόπος κωδικοποίησης της δομής ενός μορίου. Ο πιο συνηθισμένος τύπος δακτυλικών αποτυπωμάτων είναι μια σειρά

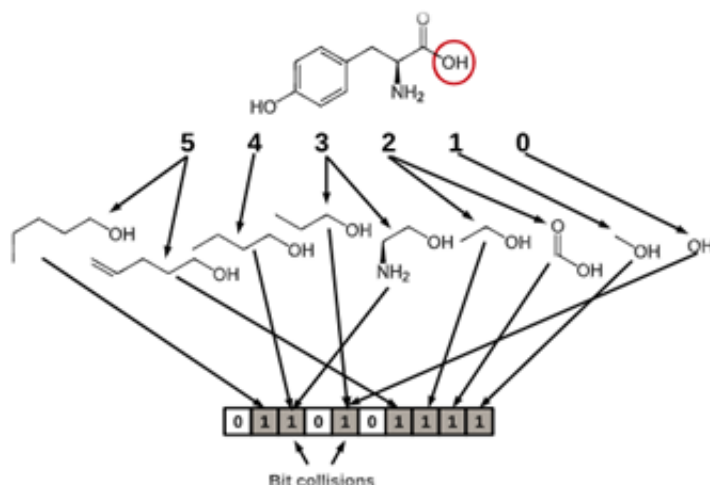
δυναμικών ψηφίων (bits) που αντιπροσωπεύουν την παρουσία ή την απουσία συγκεκριμένων δομών στο μόριο (Bajusz et al., 2017). Ουσιαστικά, οι δομικές υποομάδες αναπαριστώνται ως αλληλουχίες '0' και '1' (bit strings), όπου το '0' συμβολίζει την απουσία της συγκεκριμένης υποομάδας από την δομή του μορίου και αντίστοιχα το '1' συμβολίζει την παρουσία της (Εικόνα 1.5). Αυτή η χαρακτηριστική αλληλουχία από '0' και '1' μιας χημικής δομής ονομάζεται μοριακό αποτύπωμα και τυπικά μπορεί να έχει μήκος από 150–2500 ψηφία (bits). Η σύγκριση των δακτυλικών αποτυπωμάτων μας επιτρέπει να προσδιορίσουμε την ομοιότητα μεταξύ δύο μορίων, να βρούμε αντιστοιχίσεις ομοιοτήτες μεταξύ δύο δομών κλπ. (Gasteiger & Engel, 2003). Τα μοριακά δακτυλικά αποτυπώματα ποικίλλουν πολύ σε μήκος και πολυπλοκότητα. Παρόλο που η 3D μοριακή απεικόνιση είναι ιδιαίτερα διευρυμένη, ο συχνότερος τρόπος κωδικοποίησης χαρακτηριστικών μικρών μορίων σε συμβολοσειρές δακτυλικών αποτυπωμάτων είναι η 2D μοριακή απεικόνιση. Αυτό οφείλεται κυρίως στην ταχύτητα και στην ευκολία υπολογισμού τους.

Τα πιο δημοφιλή μοριακά αποτυπώματα είναι (Muegge & Mukherjee, 2016):

1. Τοπολογικά αποτυπώματα (substructure key-based)
2. Τα αποτυπώματα που βασίζονται σε δομικά κλειδιά υποομάδων (substructure key-based)
3. Κυκλικά αποτυπώματα (circular)
4. Αποτυπώματα φαρμακοφόρου
5. Υβριδικά αποτυπώματα
6. Άλλα αποτυπώματα που εστιάζουν στην κωδικοποίηση αλληλεπιδράσεων πρωτεΐνης-προσδέματος

- **Τοπολογικά αποτυπώματα**

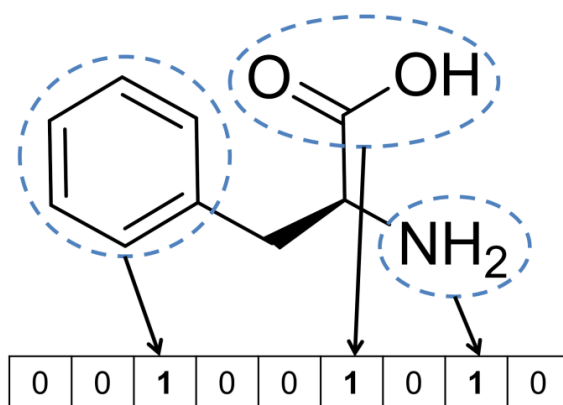
Τα τοπολογικά αποτυπώματα αναλύουν όλες τις υποομάδες ενός μορίου ακολουθώντας συνήθως μια γραμμική διαδρομή μέχρι έναν συγκεκριμένο αριθμό χημικών δεσμών και την συνέχεια κωδικοποιούν τμηματικά καθεμία από αυτές τις διαδρομές για να δημιουργήσουν το αποτύπωμα (Cereto-Massagué et al., 2015). Τα αποτυπώματα Daylight 150 που αποτελούνται από 2048 bits είναι τα πιο σημαντικά αυτής της κατηγορίας κωδικοποιούν όλες τις πιθανές διαδρομές συνδεσιμότητας στο μόριο μέχρι ένα δεδομένο μήκος.



Εικόνα 1.4. Μια αναπαράσταση ενός υποθετικού τοπολογικού δακτυλικού αποτυπώματος 10 bit βασισμένο σε γραμμικές διαδρομές (Cereto-Massagué et al., 2015)

- Τα αποτυπώματα που βασίζονται σε δομικά κλειδιά υποομάδων (**substructure key-based**)

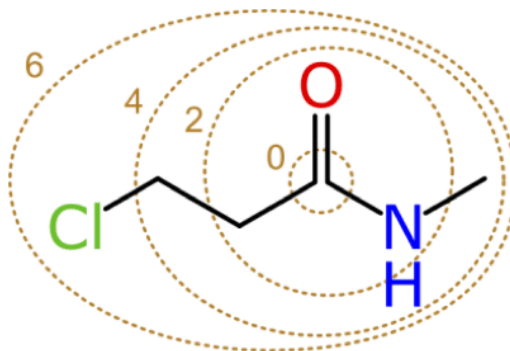
Στα δακτυλικά αποτυπώματα τα οποία είναι βασισμένα σε κλειδιά, τα δυαδικά ψηφία ορίζονται σύμφωνα με την παρουσία ή την απουσία προκαθορισμένων υπομονάδων (structural keys). Το μήκος του δακτυλικού αποτυπώματος καθορίζεται από τον αριθμό των δομικών υπομονάδων και κάθε bit αντιστοιχεί σε μια συγκεκριμένη υπομονάδα. Τα MACCS (for Molecular Access System) είναι το πιο γνωστό παράδειγμα δακτυλικού αποτυπώματος που βασίζεται σε κλειδιά. Αναπτύχθηκε από την MDL (Molecular Design Limited, τώρα θυγατρική της BIOVIA106) και έχει δύο παραλλαγές: η μία περιέχει 166 κλειδιά και είναι η πιο συχνά χρησιμοποιούμενη, καθώς έχει μικρό μέγεθος (166 bits) και εμπεριέχει τα περισσότερα χημικά χαρακτηριστικά ενδιαφέροντος, ενώ η άλλη περιέχει 960 κλειδιά. Επίσης, το δακτυλικό αποτύπωμα PubChem αποτελείται από 881 δομικά κλειδιά που κωδικοποιούν αρκετά διαφορετικά χαρακτηριστικά. Περαιτέρω παραδείγματα δακτυλικών αποτυπωμάτων βασισμένων σε κλειδιά αποτελούν τα τροποποιημένα δακτυλικά αποτυπώματα BCI (Barnard Chemical Information Ltd) με 1052 δομικά κλειδιά (Bajusz et al., 2017; Barnard & Downs, 1997).



Εικόνα 1.5. Παράδειγμα ενός δακτυλικού αποτυπώματος που βασίζεται σε δομικά κλειδιά υπομονάδων. Τα bit ορίζονται σύμφωνα με τις υποδομές που υπάρχουν στο μόριο. (1 ή "on" εάν υπάρχει η δεδομένη υποδομή και 0 ή "off" εάν απουσιάζει). Έτσι, κάθε θέση bit αντιστοιχεί σε μια συγκεκριμένη υποδομή (Bajusz et al., 2017)

- **Κυκλικά αποτυπώματα (circular)**

Αντί για διαδρομές, τα κυκλικά δακτυλικά αποτυπώματα κωδικοποιούν το περιβάλλον ενός ατόμου που ξεκινούν από το κεντρικό άτομο ενός μορίου και επεκτείνονται μέχρι μια συγκεκριμένη διάμετρο. Τα αποτυπώματα Molprint2D και τα αποτυπώματα εκτεταμένης συνδεσιμότητας (Extended Connectivity – ECFP) αποτελούν δύο συνήθη παραδείγματα αυτής της κατηγορίας (Bajusz et al., 2017). Συνήθως απαντώνται με την μορφή ECFP2 ή ECFP4, όπου το νούμερο στο τέλος υποδηλώνει την ακτίνα σε Angstrom (10-10 m) η οποία καθορίζει την κυκλική περιοχή γύρω από το κάθε άτομο. Ακόμα, μια παραλλαγή των ECFP αποτυπωμάτων είναι τα αποτυπώματα λειτουργικής κλάσης (Functional Class - FCFP), τα οποία αντί να καταγράφουν το περιβάλλον γύρω από ένα άτομο, όπως τα ECFP, καταγράφουν τον λειτουργικό ρόλο του ατόμου. Έτσι διαφορετικά άτομα με παρόμοιο λειτουργικό ρόλο δεν μπορούν να γίνουν διακριτά από αυτό το αποτύπωμα (Cereto-Massagué et al., 2015).



Εικόνα 1.6. Παράδειγμα κυκλικού αποτυπώματος. Τα νούμερα καθορίζουν την ακτίνα σε Angstrom που ελέγχεται κάθε φορά (ανακτήθηκε από collaborativedrug.com)

- **Αποτυπώματα φαρμακοφόρου**

Τα φαρμακοφόρα αποτυπώματα συνήθως εμπεριέχουν την πληροφορία από μία λίστα χαρακτηριστικών που παρουσιάζει το μόριο, κατά παρόμοιο τρόπο με τα αποτυπώματα που είναι βασισμένα σε δομικά κλειδιά υποομάδων. Η διαφορά είναι ότι λαμβάνεται υπόψη και η απόσταση που έχουν αυτά τα χαρακτηριστικά μεταξύ τους και έτσι με αυτόν τον τρόπο μεταφέρεται και 3D πληροφορία στο αποτύπωμα (Cereto-Massagué et al., 2015)

- **Υβριδικά αποτυπώματα**

Τα υβριδικά αποτυπώματα συνδυάζουν τα ίδια δυαδικά ψηφία χρησιμοποιώντας διαφορετικές προσεγγίσεις. Τέλος, υπάρχουν και άλλοι τύποι μοριακών αποτυπωμάτων που χρησιμοποιούν εντελώς διαφορετικές προσεγγίσεις. Για παράδειγμα, τα LINGO (O'Boyle et al., 2011) και SMIfp είναι αποτυπώματα που ο υπολογισμός τους γίνεται με βάση τα SMILES των μορίων. Τα αποτυπώματα αλληλεπίδρασης πρωτεΐνης-προσδέτη (PLIF), όπως υποδηλώνει το όνομά τους, κωδικοποιούν πληροφορίες σχετικά με αλληλεπιδράσεις πρωτεΐνης-προσδέτη, όπως δεσμοί υδρογόνου, ιοντικές αλληλεπιδράσεις [Chemical Computing Group Inc., Molecular operating environment (MOE), 546 (2013)].

1.5 Μοριακοί Δείκτες (Molecular Descriptors)

Σύμφωνα με τις παραπάνω απεικονίσεις μπορούμε να υπολογίσουμε διαφορετικούς μοριακούς δείκτες (molecular descriptors), ανάλογα με την πληροφορία που θέλουμε να εξάγουμε από το μόριο. Οι μοριακοί δείκτες είναι αριθμητικά χαρακτηριστικά τα οποία εξάγονται από την χημική δομή ενός μορίου και παρέχουν πληροφορίες για το συγκεκριμένο μόριο. Μπορούν να είναι μονοδιάστατοι (0D ή 1D), δισδιάστατοι (2D), τρισδιάστατοι (3D) ή τετραδιάστατοι (4D). Οι μονοδιάστατοι δείκτες είναι ο απλούστερος τύπος μοριακών δεικτών και παρέχουν μια συγκεντρωτική πληροφορία για το μόριο σύμφωνα με τον χημικό του τύπο. Τέτοιοι δείκτες είναι το μοριακό βάρος, ο αριθμός και ο τύπος των ατόμων καθώς και ο αριθμός των δεσμών στο μόριο. Παρόλο που είναι εύκολο να υπολογιστούν, οι μονοδιάστατοι δείκτες αντιμετωπίζουν ορισμένα προβλήματα εκφυλισμού, δηλαδή διαφορετικά μόρια λαμβάνουν ίδιες τιμές για τον ίδιο δείκτη. Για τον λόγο αυτό, οι μονοδιάστατοι δείκτες συνήθως χρησιμοποιούνται σε συνδυασμό με μοριακούς δείκτες μεγαλύτερων διαστάσεων. Οι δισδιάστατοι μοριακοί δείκτες βασίζονται στην τοπολογία της δομής του μορίου, όπως είναι οι χαρακτηριστικές ομάδες ή τα μοριακά θραύσματα. Οι τρισδιάστατοι δείκτες εξάγουν πληροφορίες από την απεικόνιση των 3D

συντεταγμένων του μορίου, ως εκ τούτου βασίζονται στην γεωμετρία του. Γνωστοί 3D δείκτες περιλαμβάνουν σταθερές ατόμων υποκατάστασης, δείκτες αυτοσυσχέτισης, δείκτες του λόγου επιφάνειας-όγκου και κβαντοχημικούς δείκτες. Τέλος, οι τετραδιάστατοι μοριακοί δείκτες αποτελούν μια επέκταση των τρισδιάστατων, όπου λαμβάνουν υπόψη πολλαπλές δομικές διαμορφώσεις ταυτόχρονα (Lo et al., 2018; Schneider et al., 2019). Ο υπολογισμός των μοριακών δεικτών προϋποθέτει την ύπαρξη μιας συγκεκριμένης κωδικοποίησης για το κάθε μόριο, προκειμένου ένα υπολογιστικό σύστημα να μπορεί να διακρίνει από ποια στοιχεία αποτελείται το εκάστοτε μόριο. Υπάρχουν διαφορετικές μεθοδολογίες και λογισμικά εργαλεία που μπορούν να εφαρμοστούν για τον προσδιορισμό των μοριακών δεικτών τα οποία χρησιμοποιούνται στην οργανική και κβαντική χημεία κ.λπ. Το Εθνικό Κέντρο Τοξικολογικής Έρευνας του FDA (Food and Drug administration, USA) σχεδίασε και κυκλοφόρησε το Mold2, ένα δωρεάν διαθέσιμο λογισμικό, το οποίο μπορεί να χρησιμοποιηθεί για τον υπολογισμό 777 σημαντικών μοριακών δεικτών που κωδικοποιούν δισδιάστατες πληροφορίες δομών χημικών ενώσεων, συμπεριλαμβανομένων τοπολογικών, γεωμετρικών και δομικών χαρακτηριστικών τους (Varsou et al., 2018).

1.6 Μοριακή ομοιότητα και ομοιότητα Tanimoto

Η μοριακή ομοιότητα αποτελεί μια βασική έννοια-κλειδί για την ανακάλυψη φαρμάκων και χρησιμοποιείται συνήθως κατά την ανακάλυψη και το σχεδιασμό νέων μορίων. Ορίζεται ως η ομοιότητα μεταξύ χημικών στοιχείων, μορίων ή χημικών ενώσεων σε σχέση με τις δομικές ή με τις λειτουργικές τους ιδιότητες. Συγκεκριμένα, η μοριακή ομοιότητα βασίζεται στην ιδέα ότι δύο δομικά παρόμοια μόρια συχνά μοιράζονται παρόμοιες φυσικές ιδιότητες και βιολογική λειτουργία (Kumar & Zhang, 2018). Προκειμένου να διεξαχθεί η αναζήτηση ομοιότητας, υπάρχουν τρία σημαντικά στοιχεία των μέτρων ομοιότητας που πρέπει να ληφθούν υπόψη. Πρώτον, η αναπαράσταση της δομής (χρησιμοποιείται για τον χαρακτηρισμό μορίων), δεύτερον, το σχήμα στάθμισης, weighting scheme (χρησιμοποιείται για τον προσδιορισμό της σχετικής σημασίας τιμής της αναπαράστασης) και τέλος είναι ο συντελεστής ομοιότητας (Willett, 2006). Η μοριακή αναπαράσταση είναι πάντα το πρώτο πράγμα που πρέπει να λαμβάνεται υπόψη πριν την πραγματοποίηση οποιασδήποτε αναζήτησης ομοιότητας στη μοριακή δομή. Μια μοριακή δομή είναι ένα σημαντικό στοιχείο στον τομέα της αναζήτησης ομοιότητας καθώς φέρνει την περιγραφή του τύπου, της διάταξης, της θέσης και της κατεύθυνσης των δεσμών που συνδέουν το άτομο μέσα σε ένα μόριο. Εκτός αυτού, η μοριακή δομή είναι η θεμελιώδης μονάδα

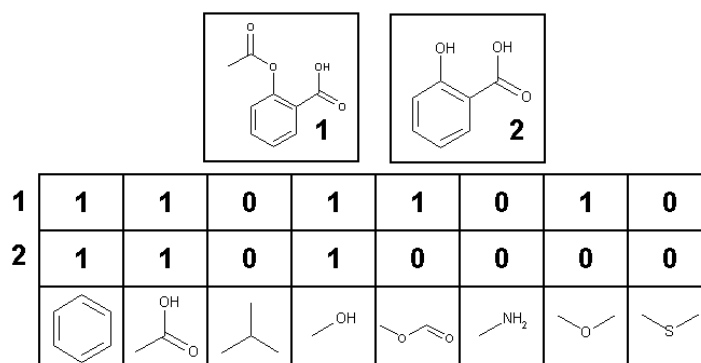
στη μελέτη των χημικών πληροφοριών του μορίου (<https://en.wiktionary.org>). Μια μοριακή δομή μπορεί να εκφραστεί σε 1D, 2D ή 3D (Shin et al., 2015). Πολλοί ερευνητές έχουν εργαστεί με τη δισδιάστατη μοριακή δομή αντί για την τρισδιάστατη δομή στη μελέτη της αναζήτησης ομοιότητας μοριακής δομής (Flower, 1998; Willett, 2009). Αυτό οφείλεται στο γεγονός ότι η δισδιάστατη δομή είναι πιο απλή, εύκολη στην εργασία και όχι τόσο περίπλοκη όσο η 3D. Ο συντελεστής ομοιότητας παίζει σημαντικό ρόλο στο πλαίσιο της αναζήτησης ομοιότητας, καθώς χρησιμοποιείται για να ποσοτικοποιήσει τον βαθμό ομοιότητας μεταξύ των δύο δομών. Η ομοιότητα μεταξύ μορίων μπορεί είτε να αξιολογηθεί μεταξύ των δομών τους είτε με βάση τις ιδιότητες που εμφανίζουν τα μόρια αυτά. Η πιο συχνά χρησιμοποιούμενη μέθοδος για τη μελέτη της ομοιότητας δομής μεταξύ μορίων βασίζεται σε μοριακά αποτυπώματα των ενώσεων και κυρίως στα 2D αποτυπώματα. Οι μετρήσεις που χρησιμοποιούνται συχνά σε αυτή τη μέθοδο είναι απλές μετρήσεις απόστασης με συντελεστές απόστασης όπως αυτός της Ευκλείδειας απόστασης και συντελεστές συσχέτισης όπως ο συντελεστής Tanimoto, ο Dice και ο Cosine (Nikolova & Jaworska, 2004). Όσον αφορά τους συντελεστές απόστασης, η απόσταση χρησιμοποιείται για να ποσοτικοποιήσει τον βαθμό διαφοράς μεταξύ δύο μορίων. Όσο μικρότερη είναι η τιμή της απόστασης, τόσο μεγαλύτερος είναι ο βαθμός ομοιότητας και το αντίστροφο (Holliday et al., 2002).

Ο συντελεστής Tanimoto (T_c) είναι το πιο δημοφιλές και ευρέως χρησιμοποιούμενο μέτρο ομοιότητας μεταξύ δύο μορίων. Ο συντελεστής Tanimoto ποσοτικοποιεί την αλληλοεπικάλυψη των χαρακτηριστικών δύο μορίων, ως τον λόγο του πλήθους των κοινών χαρακτηριστικών προς το σύνολο των χαρακτηριστικών σε κάθε μοριακό αποτύπωμα και εκφράζεται με την ακόλουθη εξίσωση:

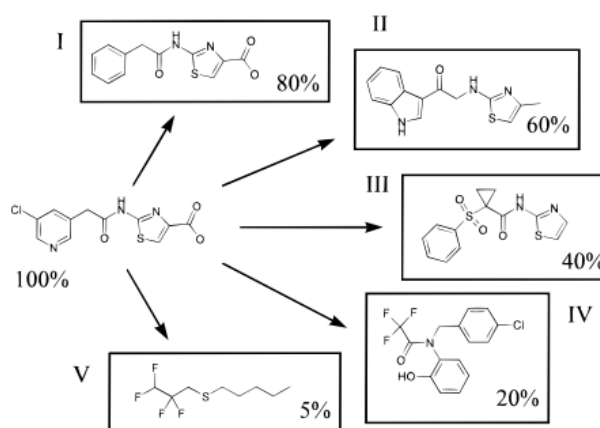
$$T_c = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

όπου το N_a είναι ο αριθμός των bit που είναι '1' στο αποτύπωμα του ενός μορίου a , το N_b είναι ο αριθμός των bit που είναι '1' στο αποτύπωμα του άλλου μορίου b και το N_{ab} είναι ο αριθμός των bit που είναι '1' και στα δύο αποτυπώματα ταυτόχρονα. Η τιμή του συντελεστή Tanimoto κυμαίνεται από το 0 έως το 1 και μπορεί να ερμηνευτεί ως το ποσοστό των κοινών στοιχείων δύο ενώσεων (Tovar et al., 2007; Vogt & Bajorath, 2020). Δύο ενώσεις θεωρούνται παρόμοιες όταν ο συντελεστής Tanimoto είναι μεγαλύτερος από 0.85 ή 85%, αλλά αυτό δεν συνεπάγεται ότι οι ενώσεις αυτές έχουν και παρόμοια βιολογική δράση. Tanimoto ομοιότητα με τιμή 1 σημαίνει ότι τα

δύο μόρια έχουν ακριβώς τα ίδια αποτυπώματα ενώ Tanimoto ομοιότητα με τιμή 0 υποδηλώνει απουσία ομοιότητας (Maggiore et al., 2014). Στην Εικόνα 1.7 παρουσιάζονται τα αποτυπώματα δύο ενώσεων με τον συντελεστή Tanimoto.



Εικόνα 1.7. Ο συντελεστής Tanimoto από τα αποτυπώματα των ενώσεων 1 & 2 υπολογίζεται ως εξής: $N_a=5$, $N_b=3$ και $N_{ab}=3$, οπότε από την εξίσωση (1) προκύπτει ότι $T_c=3/(5+3-3) = 0,6$ (<https://www.gsitechnology.com>)



Εικόνα 1.8. Μοριακή ομοιότητα πέντε ενώσεων (I-V) σε σχέση με μια συγκεκριμένη ένωση με βάση τον συντελεστή (Flower, 1998)

1.7 Quantitative Structure Activity Relationship (QSAR)

1.7.1 Μοντέλα QSAR

Μερικούς από τους τομείς στους οποίους εφαρμόζονται τα εργαλεία της χημειοπληροφορικής είναι η εξαγωγή δεδομένων από χημικές δομές, η αναζήτηση χημικών ενώσεων σε βάσεις δεδομένων, καθώς και η ανακάλυψη νέων φαρμάκων. Σημαντικό ρόλο στην ανάπτυξη της χημειοπληροφορικής στον φαρμακευτικό τομέα έχει παίξει και η Μηχανική Μάθηση. Με την αξιοποίηση των αλγορίθμων της Μηχανικής Μάθησης μπορούν να προβλεφθούν και να μοντελοποιηθούν διάφορες

ιδιότητες των φαρμάκων, όπως είναι η τοξικότητα, η αλληλεπίδραση με άλλα φάρμακα και η καρκινογένεση. Η μοντελοποίηση αυτή γίνεται συνήθως με μοντέλα που εξετάζουν την ποσοτική σχέση δομής-δραστικότητας (QSAR) τα οποία βρίσκουν ευρεία εφαρμογή στην ανακάλυψη νέων φαρμάκων, καθώς μέσα από αυτά μπορεί και προβλέπεται η βιολογική συμπεριφορά των ενώσεων μετά από πιθανές χημικές τροποποιήσεις. Με την βοήθεια των μοντέλων QSAR μπορούν να αναπτυχθούν προγράμματα τεχνητής νοημοσύνης τα οποία να προβλέπουν *in silico* (μέσω υπολογιστή) με ακρίβεια το πώς κάποιες χημικές τροποποιήσεις επηρεάζουν την βιολογική συμπεριφορά των φαρμάκων. Μέσα από αυτές τις τεχνικές μπορούν να μοντελοποιηθούν αρκετές φυσικοχημικές ιδιότητες των φαρμάκων, όπως είναι η τοξικότητα, ο μεταβολισμός, η αλληλεπίδραση με άλλα φάρμακα και η καρκινογένεση. Ουσιαστικά, ένα μοντέλο Quantitative Structure Property Relationship (QSPR) περιγράφει μια μαθηματική σχέση μεταξύ των δομικών χαρακτηριστικών και μιας ιδιότητας ενός συνόλου χημικών ουσιών. Η χρήση μοντέλων QSPR για τον έλεγχο βάσεων δεδομένων χημικών ή εικονικών βιβλιοθηκών πριν από τη σύνθεσή τους είναι πολλά υποσχόμενη και φαίνεται εξίσου ελκυστική για τους παραγωγούς χημικών προϊόντων, τις φαρμακευτικές εταιρείες και τις κρατικές υπηρεσίες. Δεδομένου των αυξανόμενων μεγεθών των βάσεων δεδομένων που προκύπτουν από τη συνδυαστική σύνθεση καθώς και των ρυθμιστικών και κοινωνικών πιέσεων για έγκαιρη αξιολόγηση των κινδύνων για την υγεία και το περιβάλλον, η ανάγκη για αξιόπιστα μοντέλα QSPR είναι επιτακτική. Μέσα σε αυτά τα πλαίσια, για να είναι αξιόπιστα και προβλέψιμα, τα μοντέλα QSPR θα πρέπει:

- να είναι στατιστικά σημαντικά και ισχυρά,
- να επικυρώνονται κάνοντας ακριβείς προβλέψεις για εξωτερικά σύνολα δεδομένων που δεν χρησιμοποιήθηκαν στην ανάπτυξη του μοντέλου, και
- να έχουν καθορισμένα όρια εφαρμογής.

Τα δύο βασικά στάδια στην ανάπτυξη ενός μοντέλου QSPR είναι:

1. η προετοιμασία δεδομένων, η οποία περιλαμβάνει (i) την συλλογή και τον καθαρισμό των δεδομένων που έχουμε διαθέσιμα για την αξιολόγηση του στόχου μας, (ii) τον υπολογισμό των μοριακών δεικτών που έχουν αποδεκτές ιδιότητες προς τον εκάστοτε στόχο και (iii) την συγχώνευση των τιμών των ιδιοτήτων και των δεικτών σε μια διαχειρίσιμη βάση δεδομένων SPR (Subscriber Profile Repository) και

2. η δημιουργία μοντέλου, που συνεπάγεται τη δημιουργία στατιστικά σημαντικών σχέσεων μεταξύ της ιδιότητας του στόχου και των τιμών των δεικτών.

Σημαντικό είναι να αναφερθεί ότι, παρά την υψηλή τους ακρίβεια προσαρμογής και την προφανή μηχανιστική τους έλξη, ορισμένα δημοσιευμένα μοντέλα QSPR αποτυγχάνουν σε αυστηρές δοκιμές επικύρωσης και, ως εκ τούτου, μπορεί να μην έχουν πρακτική χρησιμότητα ως αξιόπιστα εργαλεία διαλογής. Η επιστημονική κοινότητα προτείνει ότι μόνο επικυρωμένα μοντέλα QSPR μπορούν να προσφέρουν μια ουσιαστική μηχανιστική ερμηνεία, ειδικά στο πλαίσιο του σχεδιασμού ή της ανακάλυψης νέων χημικών παραγόντων με επιθυμητές ιδιότητες. Δεδομένου ότι η πραγματική χρησιμότητα ενός μοντέλου QSPR έγκειται στην ικανότητά του να προβλέπει με ακρίβεια την μοντελοποιημένη ιδιότητα για νέα χημικά, πρέπει να εξακριβωθεί μια ρεαλιστική εκτίμηση της πραγματικής προγνωστικής ισχύος του μοντέλου. Αυτό συνιστά τα ακόλουθα δύο πρόσθετα βήματα ανάπτυξης ενός τέτοιου μοντέλου:

3. Επικύρωση ενός μοντέλου QSPR, η οποία συνεπάγεται ποσοτική αξιολόγηση της αξιοπιστίας του μοντέλου και της προγνωστικής του ισχύος, και
4. Ορισμός του πεδίου εφαρμογής του μοντέλου στο χώρο των μοριακών δεικτών που χρησιμοποιούνται για την εξαγωγή του μοντέλου (Tropsha et al., 2003).

1.7.2 Επικύρωση των μοντέλων QSAR/QSPR

Οι ποσοτικές σχέσεις δομής-δραστηριότητας και δομής-ιδιότητας (QSARs/QSPRs), όπως είναι επί του παρόντος κατανοητό, έχουν αναπτυχθεί και χρησιμοποιούνται για σχεδόν 50 χρόνια, ξεκινώντας με τη θεμελιώδη εργασία του Corwin Hansch και των συνεργατών του για τα φυτοφάρμακα το 1962 (Hansch et al., 1962). Δύο κύριες μέθοδοι χρησιμοποιούνται για τον προσδιορισμό της αξιοπιστίας των μοντέλων αυτών: η εσωτερική διασταυρούμενη επικύρωση και η εξωτερική επικύρωση με ένα δοκιμαστικό σύνολο ενώσεων. Είναι γενικά αποδεκτό πλέον ότι μόνο τα QSAR και QSPR που έχουν εξωτερικά επικυρωθεί κατάλληλα μπορούν να θεωρείται αξιόπιστο τόσο για επιστημονικούς όσο και για ρυθμιστικούς σκοπούς. Τον Μάρτιο του 2002 πραγματοποιήθηκε μια συνάντηση εμπειρογνομώνων QSAR/QSPR στο Setúbal της Πορτογαλίας, για να διατυπωθεί ένα σύνολο κατευθυντήριων γραμμών για την επικύρωση των QSAR/QSPR, ιδίως για ρυθμιστικούς σκοπούς (Jaworska et al., 2003).

Συντάχθηκαν έξι κατευθυντήριες γραμμές, οι οποίες υιοθετήθηκαν αργότερα από τον

Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ) (<http://www.oecd.org>) και τροποποιήθηκαν σε πέντε.

Οι οδηγίες είναι ότι ένα έγκυρο QSAR/QSPR πρέπει να έχει:

- (1) ένα καθορισμένο τελικό σημείο/ καθορισμό του στόχου της μελέτης.
- (2) ένα σαφή αλγόριθμο.
- (3) καθορισμένο πεδίο εφαρμογής.
- (4) κατάλληλα μέτρα καλής εφαρμογής, αξιοπιστίας και προβλεψιμότητας.
- (5) μια μηχανιστική ερμηνεία, αν είναι δυνατόν.

Οι κατευθυντήριες γραμμές είναι πλέον γνωστές ως Αρχές του ΟΟΣΑ για την επικύρωση των (Q)SAR, αν και προορίζονται να ισχύουν και για τα QSPR. Ο ΟΟΣΑ έχει επίσης παράσχει μια λίστα ελέγχου για την παροχή καθοδήγησης σχετικά με την ερμηνεία των αρχών (Kearns, 2008). Είναι ενδιαφέρον να σημειωθεί ότι παρόμοιες κατευθυντήριες γραμμές είχαν προταθεί ήδη από το 1973 (Unger & Hansch, 1973). Παρά τα προαναφερθέντα, εξακολουθούν να γίνονται σφάλματα στην ανάπτυξη και χρήση των QSAR και QSPR. Πάνω από 20 κύριοι τύποι σφαλμάτων μπορούν να βρεθούν και να συνεχίσουν να γίνονται, στην ανάπτυξη και χρήση του QSAR/QSPR, και αυτοί παρατίθενται στον Πίνακα 1, μαζί με την αντίστοιχη αρχή του ΟΟΣΑ (Dearden et al., 2009).

Πίνακας 1.1. Τύποι σφαλμάτων κατά την ανάπτυξη και χρήση του QSAR/QSPR (αναπροσαρμογή από πηγή: Dearden et al., 2009 στις 16/01/2022)

Τύποι σφαλμάτων στην ανάπτυξη και χρήση QSAR/QSPR		
	Τύπος σφάλματος	Σχετικές αρχές (ή) του ΟΟΣΑ
1	Παράλειψη να ληφθεί υπόψη η ετερογένεια των δεδομένων	1
2	Χρήση ακατάλληλων δεδομένων εξόδου	1
3	Χρήση συγγραμμικών δεικτών	2,4,5
4	Χρήση ακατανόητων δεικτών	2,5
5	Σφάλμα στις τιμές περιγραφής	2
6	Κακή δυνατότητα μεταφοράς του QSAR/QSPR	2
7	Μη επιβεβαιωμένη παράλειψη σημείων δεδομένων	3
8	Χρήση ανεπαρκών δεδομένων	3
9	Αναδιπλασιασμός των ενώσεων στο σύνολο δεδομένων	3
10	Πολύ στενό εύρος τιμών τελικού σημείου	3
11	Υπερπροσαρμογή δεδομένων	4
12	Χρήση υπερβολικού αριθμού δεικτών σε ένα QSAR/QSPR	4
13	Έλλειψη/ανεπάρκεια στατιστικών στοιχείων	4
14	Λανθασμένος υπολογισμός	4
15	Έλλειψη αυτόματης κλιμάκωσης του μοριακού δείκτη	4

16	Κακή χρήση/ερμηνεία στατιστικών στοιχείων	4
17	Δεν λαμβάνεται υπόψη η κατανομή των υπολειμμάτων	4
18	Ανεπαρκής επιλογή σετ εκπαίδευσης/δοκιμών	4
19	Αποτυχία σωστής επικύρωσης QSAR/QSPR	4
20	Έλλειψη μηχανιστικής ερμηνείας	5

Οι κατευθυντήριες γραμμές του ΟΟΣΑ για τις δοκιμές χημικών ουσιών είναι διεθνώς αποδεκτές ως τυπικές μέθοδοι για δοκιμές ασφάλειας και αξιολόγηση φυτοφαρμάκων, προϊόντων προσωπικής φροντίδας, βιομηχανικών χημικών ουσιών, ακόμη και για να βοηθήσουν στη λήψη αποφάσεων σε περιπτώσεις αντιμετώπισης έκτακτης ανάγκης (Demchuk et al., 2011; Ruiz et al., 2012). Αρκετές μελέτες έχουν γίνει για την δημιουργία QSAR μοντέλων τα οποία να μπορούν να διαχωρίζουν αποτελεσματικά ενώσεις φυτικής προέλευσης από συνθετικές ενώσεις. Η μελέτη των Henkel et al. είναι ίσως από τις πρώτες μελέτες που εξέτασαν τις διαφορές των μοριακών ιδιοτήτων και των δομικών χαρακτηριστικών ανάμεσα σε φυσικά προϊόντα και συνθετικές ενώσεις (Henkel et al., 1999). Οι Hansch και Free-Wilson χρησιμοποίησαν απλές τεχνικές παλινδρόμησης για να συσχετίσουν την δραστικότητα ενώσεων με διάφορα μοτίβα υποομάδων και κάποιες χημικές ιδιότητες, όπως είναι η διαλυτότητα, η υδροφοβικότητα και κάποιοι ηλεκτρονιακοί παράγοντες, δημιουργώντας έτσι μοντέλα ανάλυσης που αποτελούν από τα πρώτα QSAR μοντέλα (Lo et al., 2018). Σε μια άλλη μελέτη των Stahura et al. εντοπίστηκαν κάποιοι μοριακοί δείκτες οι οποίοι μπορούν να διαχωρίσουν ενώσεις φυσικών προϊόντων από συνθετικά μόρια, στηριζόμενοι στο μέγεθος της εντροπίας Shannon (Stahura et al., 2000).

1.8 Μηχανική μάθηση (Machine Learning, ML)

Η μηχανική μάθηση είναι αυτή τη στιγμή ένας από τους πιο σημαντικούς και ταχέως εξελισσόμενους τομείς στο πεδίο της *in silico* ανακάλυψης φαρμάκων (Varnek & Baskin, 2012). Πρόκειται για μια συνένωση της επιστήμης των υπολογιστών και των μαθηματικών, καθώς με την χρήση στατιστικών μεθόδων διερευνάται η σχέση ανάμεσα στα δεδομένα που τροφοδοτούν το σύστημα. Στην συνέχεια μέσα από κατάλληλη επεξεργασία των δεδομένων και μέσα από αλγορίθμους αναπτύσσονται αντίστοιχα μοντέλα πρόβλεψης (Deo, 2015). Κάθε σύνολο δεδομένων (dataset) περιέχει κάποια χαρακτηριστικά (features), τα οποία παρέχουν την πληροφορία. Η ποιότητα των χαρακτηριστικών που θα τροφοδοτήσουν το υπολογιστικό σύστημα παίζει καθοριστικό ρόλο για την ακρίβεια της πρόβλεψης. Διαφορετικοί συνδυασμοί χαρακτηριστικών παράγουν διαφορετικά αποτελέσματα ακρίβειας, οπότε η

διαδικασία επαναλαμβάνεται αρκετές φορές μέχρι να βρεθεί το επιθυμητό αποτέλεσμα (Jayatilake & Ganegoda, 2021). Σε αντίθεση με τα φυσικά μοντέλα που βασίζονται αποκλειστικά σε φυσικές εξισώσεις όπως η κβαντική χημεία ή οι προσομοιώσεις μοριακής δυναμικής, οι προσεγγίσεις μηχανικής μάθησης χρησιμοποιούν αλγόριθμους αναγνώρισης προτύπων για να διακρίνουν μαθηματικές σχέσεις μεταξύ μικρών μορίων και να τις ενισχύσουν για να προβλέψουν χημικές, βιολογικές και φυσικές ιδιότητες των νέων ενώσεων. Επίσης, σε σύγκριση με τα φυσικά μοντέλα, οι τεχνικές μηχανικής μάθησης είναι πιο αποτελεσματικές και μπορούν εύκολα να κλιμακωθούν σε μεγάλα σύνολα δεδομένων χωρίς την ανάγκη εκτεταμένων υπολογιστικών πόρων. Ένας από τους κύριους τομείς εφαρμογής της μηχανικής μάθησης στην ανακάλυψη φαρμάκων είναι να βοηθά τους ερευνητές να κατανοήσουν και να εκμεταλλευτούν τις σχέσεις μεταξύ των χημικών δομών και της βιολογικής δραστηρότητάς τους (Lo et al., 2018). Οι αλγόριθμοι μηχανικής μάθησης χωρίζονται κυρίως σε τέσσερις κατηγορίες: την επιβλεπόμενη μάθηση (Supervised learning), την μη επιβλεπόμενη μάθηση (Unsupervised learning), την μάθηση με ημι-επίβλεψη (Semi-supervised learning) και την ενισχυτική μάθηση (Reinforcement learning), (Mohammed et al., 2016; Sarker, 2021):

- **Επιβλεπόμενη μάθηση:**

Το σύστημα προσπαθεί να δημιουργήσει μια συνάρτηση η οποία θα μπορέσει να αντιστοιχίσει μία γνωστή είσοδο σε μία γνωστή έξοδο. Για το λόγο αυτό, παρέχεται στο σύστημα ένα γνωστό σύνολο δεδομένων εκπαίδευσης (training set), στο οποίο τα δεδομένα είναι χαρακτηρισμένα με ετικέτες (labels). Με την χρήση αλγορίθμων το σύστημα προσπαθεί να κατατάξει τα δεδομένα εισόδου κάθε φορά στην αντίστοιχη επιθυμητή έξοδο.

- **Μη επιβλεπόμενη μάθηση:**

Στην συγκεκριμένη μάθηση, οι αλγόριθμοι του συστήματος προσπαθούν να εντοπίσουν ομοιότητες και διαφορές στα δεδομένα με στόχο την δημιουργία ομάδων (clusters) ανάλογα με τα κοινά στοιχεία που θα αναγνωρίσουν. Επομένως το σύστημα παίρνει αποφάσεις χωρίς να έχει εκπαιδευτεί από κάποιο σύνολο δεδομένων. Στην μη επιβλεπόμενη μάθηση δεν υπάρχει η δυνατότητα δημιουργίας ενός συνόλου δεδομένων εκπαίδευσης, οπότε το σύστημα πρέπει να ομαδοποιήσει τα δεδομένα βάσει κάποιων κοινών στοιχείων. Οι αλγόριθμοι που χρησιμοποιούνται συνήθως στην μη επιβλεπόμενη μάθηση είναι κατά βάση αλγόριθμοι ομαδοποίησης (clustering algorithms).

- **Ημι-επιβλεπόμενη μάθηση:**

Η ημι-επιβλεπόμενη μάθηση χρησιμοποιεί δεδομένα που έχουν χαρακτηριστεί με ετικέτες αλλά και δεδομένα που δεν έχουν χαρακτηριστεί με ετικέτες και γι' αυτό το λόγο μπορεί να θεωρηθεί ως ένας υβριδισμός μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης. Ο κύριος στόχος αυτής της μεθόδου είναι να παράξει όσο το δυνατόν καλύτερα αποτελέσματα πρόβλεψης, από την περίπτωση του να χρησιμοποιούνταν μόνο δεδομένα που είχαν χαρακτηριστεί με ετικέτες.

- **Ενισχυτική μάθηση:**

Αυτός ο τρόπος μάθησης αξιοποιεί την μέθοδο της ανταμοιβής και ποινής για να διατηρήσει την γνώση που λαμβάνει το σύστημα από το περιβάλλον του, αυξάνοντας κάθε φορά την ανταμοιβή. Ουσιαστικά, η ενισχυτική μάθηση χρησιμοποιεί μηχανισμούς ανάδρασης ώστε το σύστημα να αξιολογεί αυτόματα την επίδοσή του μέσα σε ένα συγκεκριμένο περιβάλλον, με απώτερο σκοπό να βελτιώσει την αποτελεσματικότητά του (Sarker, 2021).

1.9 Εικονική διαλογή (Virtual screening, VS)

Η εικονική διαλογή (VS) είναι μία από τις σημαντικότερες και ευρέως χρησιμοποιούμενες μεθόδους στην ανακάλυψη φαρμάκων που αναζητά μεγάλες χημικές βάσεις δεδομένων προκειμένου να εντοπιστούν εκείνες οι δομές που είναι πιο πιθανό να συνδεθούν με έναν στόχο φαρμάκου, συνήθως έναν υποδοχέα πρωτεΐνης ή ένα ένζυμο (Patrick Walters et al., 1998). Γενικά, οι τεχνικές εικονικού ελέγχου μπορούν να χωριστούν σε δύο τύπους: εικονική διαλογή με βάση τη δομή (structure-based VS, SBVS) και εικονική διαλογή με βάση τον προσδέτη (ligand-based VS, LBVS) (Schneider, 2010).

- **Εικονική διαλογή με βάση τη δομή του υποδοχέα (SBVS)**

Η μέθοδος εικονικής διαλογής που βασίζεται στη δομή περιλαμβάνει διαφορετικές τεχνικές υπολογισμού που λαμβάνουν υπόψη τη δομή του υποδοχέα που είναι ο μοριακός στόχος των ερευνώντων ενεργών προσδετών. Μερικές από αυτές τις τεχνικές περιλαμβάνουν μοριακή πρόσδεση, πρόβλεψη φαρμακοφόρου με βάση τη δομή και προσομοιώσεις μοριακής δυναμικής. Η μοριακή πρόσδεση είναι η πιο χρησιμοποιούμενη τεχνική που βασίζεται στη δομή, και εφαρμόζει μια λειτουργία βαθμολόγησης για την εκτίμηση της καταλληλότητας κάθε προσδέματος έναντι της θέσης δέσμευσης του μακρομοριακού υποδοχέα, βοηθώντας στην επιλογή των προσδετών με την υψηλότερη συγγένεια (Kooistra et al., 2016; Kroemer, 2007).

- **Εικονική διαλογή με βάση τον προσδέτη (LBVS)**

Μέσω της μεθόδου εικονικής διαλογής με βάση τον προσδέτη, μπορεί να δημιουργηθεί ένα μοντέλο του υποδοχέα το οποίο στηρίζεται σε ένα σύνολο δομικά διαφορετικών προσδετών οι οποίοι προσδένονται στον προς μελέτη υποδοχέα. Βέβαια, σκοπός για την υλοποίηση αυτού του μοντέλου είναι η χρήση των σωστών συλλογικών πληροφοριών που περιέχονται σε ένα τέτοιο σύνολο προσδετών (Santana et al., 2021). Όταν είναι γνωστή η τρισδιάστατη (τρειςδιάστατη) δομή ενός στόχου, η μέθοδος εικονικού ελέγχου με βάση τη δομή είναι πιο κατάλληλη για την εύρεση δομικά καινοτόμων μορίων. Όταν η τρισδιάστατη δομή είναι άγνωστη ή είναι δύσκολο να προβλεφθεί, η μέθοδος εικονικού ελέγχου με βάση τον προσδέτη είναι η προτιμώμενη μέθοδος για διαλογή και βελτιστοποίηση. Παρά τον αυξανόμενο αριθμό πειραματικά προσδιορισμένων τρισδιάστατων δομών φαρμάκων, η μέθοδος LBVS παραμένει σημαντική και χρησιμοποιείται ευρέως λόγω της γρήγορης ταχύτητάς της και του γεγονότος ότι τα κατοχυρωμένα με δίπλωμα ευρεσιτεχνίας ή τα αναφερόμενα ενεργά παραμένουν η κύρια πηγή έμπνευσης για την ανακάλυψη φαρμάκων (Schneider, 2010). Μια βασική αρχή της μεθόδου βασίζεται στην υπόθεση ότι παρόμοια μόρια είναι επιρρεπή να εμφανίζουν παρόμοιες βιολογικές δραστηριότητες (Jurafsky & Martin, 2020).

1.10 Χρήσιμα εργαλεία Χημειοπληροφορικής

1.10.1 Βάσεις Δεδομένων

Συνήθως, η διαθεσιμότητα δεδομένων θεωρείται το κλειδί για να κατασκευαστεί ένα μοντέλο μηχανικής μάθησης ή συστήματα που βασίζονται σε υπολογιστικά δεδομένα (Sarker et al. 2021). Έτσι, σχετικές τεχνολογίες προσανατολισμένες προς τα δεδομένα, όπως η μηχανική μάθηση, η τεχνητή νοημοσύνη, η προηγμένη ανάλυση κ.λπ. σχετίζονται με τη λήψη έξυπνων αποφάσεων που βασίζονται σε δεδομένα. Σήμερα, πολλοί ερευνητές αναλύοντας τα δεδομένα, χρησιμοποιούν τον όρο «επιστήμη δεδομένων» και περιγράφουν το διεπιστημονικό πεδίο συλλογής δεδομένων, προεπεξεργασίας, εξαγωγής συμπερασμάτων ή λήψης αποφάσεων. Για την κατανόηση και ανάλυση των πραγματικών φαινομένων που βασίζονται σε δεδομένα, χρησιμοποιούνται διάφορες επιστημονικές μέθοδοι, διαδικασίες και συστήματα, τα οποία είναι κοινώς γνωστά ως επιστημονικά δεδομένα. Σύμφωνα με τους Cao et al. «η επιστήμη των δεδομένων είναι ένα νέο διεπιστημονικό πεδίο που συνθέτει και βασίζεται στη στατιστική, στην πληροφορική, στους υπολογιστές, στην επικοινωνία, στην διαχείριση, και στην κοινωνιολογία με σκοπό τη μελέτη των δεδομένων και των περιβαλλόντων τους, την μετατροπή των δεδομένων σε ιδέες και

αποφάσεις ακολουθώντας μια σκέψη και μεθοδολογία δεδομένων για τη γνώση προς τη σοφία» (Cao et al. 2017). Βέβαια, ο όγκος της πληροφορίας σχετικά με δεδομένα χημικών ενώσεων είναι τεράστιος και αυξάνεται με την πάροδο των χρόνων, γεγονός που καθιστά την επεξεργασία τους ακατόρθωτη με συμβατικές μεθόδους. Η αποθήκευση και η αναζήτηση δεδομένων αναφορικά με χημικές ενώσεις αποτέλεσε ένα από τα πρώτα, αν όχι το πρώτο, ζητήματα της χημειοπληροφορικής. Η διαχείριση όλης αυτής της πολυδιάστατης πληροφορίας σχετικά με τις χημικές ενώσεις μπορούσε να επιτευχθεί μόνο με ηλεκτρονικά μέσα και συγκεκριμένα τις βάσεις δεδομένων. Μία βάση δεδομένων αποτελεί μια συλλογή από πληροφορίες ενός συγκεκριμένου θέματος (πχ. χημικές ενώσεις), οι οποίες συνήθως βρίσκονται σε πίνακες ή λίστες. Η ύπαρξη βάσεων δεδομένων βοηθάει στην καλύτερη οργάνωση της πληροφορίας, στην εύκολη πρόσβαση σε αυτήν και στην ανάκτηση της μέσα από απλές αναζητήσεις. Τα δεδομένα τοποθετούνται σε πίνακες οι οποίοι μπορούν να συνδέονται μεταξύ τους μέσα από κάποιο κοινό γνώρισμα, το οποίο και αναφέρεται ως πρωτεύον κλειδί. Μία βάση δεδομένων γενικά αποτελείται από πεδία (fields), εγγραφές (records), ερωτήματα (queries) και αναφορές (reports) (Karthikeyan & Vyas, 2014), τα οποία περιγράφονται επιγραμματικά παρακάτω):

- **Πεδία:** Κατά την δημιουργία πινάκων σε μια βάση δεδομένων οι πληροφορίες τοποθετούνται κάτω από συγκεκριμένα πεδία, τα οποία θα πρέπει να έχουν μοναδικό όνομα για να είναι πιο εύκολη η ανάκτηση των εγγραφών που περιέχουν. Τέτοια πεδία, για παράδειγμα, μπορούν να είναι τα ονόματα των στηλών σε έναν πίνακα.
- **Εγγραφές:** Οι εγγραφές είναι συγκεκριμένα χαρακτηριστικά ενός αντικείμενου στην βάση δεδομένων. Ως αντικείμενο μπορεί να θεωρηθεί και ένας πίνακας, και οι γραμμές που θα έχει αυτός ο πίνακας αποτελούν και τις εγγραφές του.
- **Ερωτήματα:** Το ερώτημα (query) αφορά στην ουσία την αναζήτηση που κάνει ο χρήστης για να ανακτήσει την πληροφορία που επιθυμεί, η οποία είναι αποθηκευμένη στην βάση δεδομένων. Ένα παράδειγμα είναι όταν ένας χρήστης επιθυμεί να μάθει πληροφορίες για μια χημική ένωση, από μία σχετική βάση δεδομένων, χρησιμοποιώντας το όνομα της ένωσης για την αναζήτηση.
- **Αναφορές:** Οι αναφορές είναι τα αποτελέσματα που ανακτώνται μετά από ένα ερώτημα, μια αναζήτηση στην βάση δεδομένων. Οι αναφορές μπορούν να προσαρμοστούν ανάλογα με τις ανάγκες του κάθε χρήστη, έτσι ώστε η πληροφορία που θα λάβει να είναι περισσότερο χρήσιμη και λειτουργική.

Πληροφορίες σχετικά με χημικές ουσίες μπορούν να βρεθούν σε βάσεις δεδομένων όπως η PubChem, η οποία περιλαμβάνει πληροφορίες για 100 εκατομμύρια ενώσεις, ή η Drugbank, η οποία περιλαμβάνει πληροφορίες για 10 χιλιάδες φάρμακα. Οι βάσεις δεδομένων που χρησιμοποιούνται για την ανακάλυψη φαρμάκων είναι αναπαραστάσεις κειμένου για χημικές ενώσεις και πρωτεΐνες. Ο Πίνακας 1.2 παραθέτει ορισμένες βάσεις δεδομένων που αποθηκεύουν διαφορετικούς τύπους βιοχημικών πληροφοριών (Öztürk et al., 2020).

Πίνακας 1.2. Πιο γνωστές βάσεις δεδομένων στην ανακάλυψη φαρμάκων (αναπροσαρμογή από πηγή: Öztürk et al., 2020 στις 16/01/2022)

Βάσεις Δεδομένων	Πηγή	Περιγραφή
UniProt (Arweiler, 2004)	https://www.uniprot.org/	Universal Protein Resource: αποθηκεύει πρωτεϊνικές αλληλουχίες και πληροφορίες των λειτουργιών τους
PDB (Berman et al., 2000)	https://www.rcsb.org/	Protein Data Bank: βάση δεδομένων που παρέχει πληροφορίες για την δομή περίπου 152 χιλ. μακρομοριακών δομών
PFam (Bateman et al., 2004)	https://pfam.xfam.org/	Protein Families Database: Μια πρωτεϊνική βάση δεδομένων που βασίζεται στην ευθυγράμμιση πολλαπλών ακολουθιών (MSA) και στα μοντέλα Markov (HMM)
PROSITE (Hulo et al., 2006)	https://prosite.expasy.org/	Μια βάση δεδομένων που περιέχει πρωτεϊνικούς δομές, μοτίβα, οικογένειες και λειτουργικά μέρη των πρωτεϊνών
PubChem (Y. Wang et al., 2009)	https://PubChem.ncbi.nlm.nih.gov/	Μία εκτεταμένη βάση δεδομένων για 96 εκατομμύρια ενώσεις και 265 εκατ. ουσίες. Η βάση δεδομένων PubChem λειτουργεί επίσης ως εργαλείο χημειοπληροφορικής παρέχοντας μια δυνατότητα που επιτρέπει τον υπολογισμό της 2D και 3D ομοιότητας των ενώσεων και εισάγει έναν 1D χημικό δείκτη
ChEMBL (Gaulton et al., 2012a)	https://www.ebi.ac.uk/chembl/	Μια ευρέως προσβάσιμη βάση δεδομένων που αποθηκεύει πληροφορίες σχετικά με πρωτεϊνικούς στόχους, χημικές ιδιότητες και βιοενεργότητες για 1,9 εκατ. Ενώσεις
DrugBank (Wishart et al., 2006)	https://www.drugbank.ca/	Μία διαδικτυακή βάση δεδομένων για χημικές, φαρμακολογικές και φαρμακευτικές πληροφορίες για 13 χιλ. φάρμακα και 5 χιλ. πρωτεΐνες (π.χ.

		στόχοι/ένζυμα φαρμάκων) που σχετίζονται με αυτά τα φάρμακα
BindingDB (Liu et al., 2007)	https://www.bindingdb.org/	Μια βάση δεδομένων αλληλεπιδράσεων μεταξύ πρωτεϊνών και μικρών μορίων που αποθηκεύει την συγγένεια αλληλεπίδρασής τους.
PDB-Bind (Wang et al. 2005)	www.pdbbind.org.cn/	Δημόσια βάση δεδομένων με δεδομένα για την συγγένεια δέσμευσης για σύμπλοκα πρωτεΐνης – προσδέτη
ZINC (Irwin & Shoichet 2005)	https://zinc.docking.org/	Μια βάση δεδομένων για πάνω από 230 εκατ. εμπορικά διαθέσιμες ενώσεις σε τρισδιάστατη μορφή

Η PubChem είναι μια ανοιχτή βάση δεδομένων Χημείας του Εθνικού Ινστιτούτου Υγείας των ΗΠΑ (National Institute of Health, USA). Η έννοια “ανοιχτή βάση δεδομένων” σημαίνει ότι το μπορούμε να εισάγουμε επιστημονικά δεδομένα στο PubChem και ότι αυτά μπορούν να τα χρησιμοποιήσουν άλλοι. Από την λειτουργία του το 2004, η PubChem έχει γίνει βασική πηγή χημικών πληροφοριών για επιστήμονες, φοιτητές και το ευρύ κοινό παρέχοντας έτσι δεδομένα σε πολλά εκατομμύρια χρήστες παγκοσμίως. Τα δεδομένα στη PubChem προέρχονται από ακαδημαϊκά εργαστήρια, κρατικές υπηρεσίες, εταιρείες προμήθειας χημικών και φαρμακευτικών προϊόντων, εκδότες περιοδικών και μεμονωμένους ερευνητές. Αυτή η βάση δεδομένων περιέχει κυρίως μικρά μόρια, αλλά και μεγαλύτερα μόρια όπως νουκλεοτίδια, υδατάνθρακες, λιπίδια, πεπτίδια και χημικά τροποποιημένα μακρομόρια όπως φαίνεται και στον Πίνακα 1.3. Συλλέγει πληροφορίες για χημικές δομές, χαρακτηριστικά τους, χημικές και φυσικές ιδιότητες, βιολογικές δραστηριότητες, διπλώματα ευρεσιτεχνίας, δεδομένα υγείας, ασφάλειας, τοξικότητας και πολλά άλλα (<https://pubchem.ncbi.nlm.nih.gov/>).

Πίνακας 1.3. Καταμέτρηση δεδομένων PubChem (αναπροσαρμογή από <https://pubchem.ncbi.nlm.nih.gov/> στις 03/01/2022)

Συλλογή δεδομένων	Καταμέτρηση	Περιγραφή
Ενώσεις	110.698.215	Μοναδικές χημικές δομές που εξάγονται από συνεισφερόμενα αρχεία ουσιών PubChem
Ουσίες	277.215.388	Πληροφορίες σχετικά με χημικές οντότητες που παρέχονται από τους συνεργάτες του PubChem
Βιολογικές Δοκιμασίες	1.391.569	Βιολογικά πειράματα που παρέχονται από τους συνεργάτες του PubChem
Βιολογικές	293.035.633	Σημεία δεδομένων βιολογικής δραστηριότητας που

Δραστικότητα		αναφέρονται στο PubChem BioAssays
Γονίδια	103.715	Στόχοι γονιδίων που δοκιμάστηκαν σε PubChem Βιολογικές Δοκιμασίες και εμπλέκονται σε PubChem μονοπάτια
Πρωτεΐνες	96.561	Πρωτεϊνικοί στόχοι που δοκιμάστηκαν σε PubChem Βιολογικές Δοκιμασίες και εμπλέκονται σε PubChem μονοπάτια
Ταξινόμηση	531.129	Οργανισμοί στόχοι που δοκιμάστηκαν σε PubChem Βιολογικές Δοκιμασίες και εμπλέκονται σε PubChem μονοπάτια
Μονοπάτια	238.609	Αλληλεπιδράσεις μεταξύ χημικών ουσιών, γονιδίων και πρωτεϊνών
Βιβλιογραφία	33.501.831	Επιστημονικές δημοσιεύσεις με συνδέσμους στο PubChem
Διπλώματα ευρεσιτεχνίας	28.231.359	Διπλώματα ευρεσιτεχνίας με συνδέσμους στο PubChem
Πηγές δεδομένων	836	Οργανισμοί που συνεισφέρουν δεδομένα στο PubChem

Μία πολύ εύχρηστη βάση δεδομένων για το επιστημονικό κοινό είναι η βάση δεδομένων ChEMBL. Η ChEMBL είναι μια βάση δεδομένων βιοδραστικών μορίων με ιδιότητες που μοιάζουν με φάρμακα. Είναι προσβάσιμη μέσω μιας απλής και εύκολης προς το χρήστη διεύθυνσης: <https://www.ebi.ac.uk/chembl/db>. Αυτή η διάδραση επιτρέπει στους χρήστες να αναζητούν ενώσεις, στόχους ή προσδιορισμούς ενδιαφέροντος με ποικίλους τρόπους. Για παράδειγμα, οι χρήστες που επιθυμούν να ανακτήσουν πιθανές ενώσεις για έναν στόχο ενδιαφέροντος μπορούν να πραγματοποιήσουν αναζήτηση χρησιμοποιώντας λέξεις-κλειδιά στη βάση δεδομένων βάζοντας ένα όνομα πρωτεΐνης, ένα συνώνυμο ή το αναγνωριστικό ChEMBL του στόχου που ενδιαφέρει τον χρήστη. Εναλλακτικά, οι στόχοι μπορούν να αναζητηθούν σύμφωνα με την οικογένεια πρωτεϊνών ή τον οργανισμό όπου ανήκει ο εκάστοτε στόχος. Τα δεδομένα στη βάση δεδομένων ChEMBL εξάγονται με μη αυτόματο τρόπο από το πλήρες κείμενο επιστημονικών δημοσιεύσεων με κριτές σε διάφορα περιοδικά, όπως τα Journal of Medicinal Chemistry, Bioorganic Medicinal Chemistry Letters και Journal of Natural Products. Από κάθε δημοσίευση, αφαιρούνται λεπτομέρειες για τις ενώσεις που δοκιμάστηκαν, τις δοκιμασίες που πραγματοποιήθηκαν και τυχόν πληροφορίες στόχου για αυτές τις δοκιμασίες. Πριν από τη φόρτωση στη βάση δεδομένων, οι δομές ελέγχονται για πιθανά προβλήματα και στη συνέχεια κανονικοποιούνται σύμφωνα με ένα σύνολο κανόνων, για να εξασφαλιστεί συνέπεια στην αναπαράσταση. Αυτό επιτρέπει στους χρήστες να προβάλλουν όλα τα δεδομένα που σχετίζονται με την ίδια μητρική ένωση. Εκτός από τα δεδομένα που προέρχονται από τη βιβλιογραφία, το ChEMBL περιέχει επίσης

δομές και σχολιασμούς για φάρμακα εγκεκριμένα από την Υπηρεσία Τροφίμων και Φαρμάκων (Food and Drug Administration, FDA) επιτρέποντας έτσι στους χρήστες των δεδομένων βιοδραστικότητας να αξιολογήσουν εάν μια ένωση ενδιαφέροντος είναι εγκεκριμένο φάρμακο. Δέχεται κατατεθειμένα αποτελέσματα από πολλά εργαστήρια και κέντρα διαλογής και περιέχει μεγάλη ποσότητα δεδομένων, κυρίως από πειράματα διαλογής υψηλής απόδοσης, που μετρούν την αναστολή ενός στόχου από μεγάλους αριθμούς ενώσεων, συχνά σε μία μόνο συγκέντρωση ένωσης (Gaulton et al., 2012). Σήμερα η ChEMBL περιέχει πληροφορίες βιοδραστικότητας που εξάγονται από >67.000 δημοσιεύσεις και διπλώματα ευρεσιτεχνίας, μαζί με κατατεθειμένα σύνολα δεδομένων και δεδομένα που ανταλλάσσονται με άλλες βάσεις δεδομένων, όπως η PubChem BioAssay και η BindingDB. Συνολικά, όπως φαίνεται και στην Εικόνα 1.9 περιέχει σχεδόν 15 εκατομμύρια μετρήσεις βιοδραστικότητας για περίπου 2 εκατομμύρια διακριτές ενώσεις. Οι αναλύσεις βασίζονται σε >1600 διακριτές κυτταρικές σειρές, 500 ιστούς/όργανα και 3600 (Mendez et al., 2019).



Εικόνα 1.9. Σύνολο δεδομένων της ChEMBL (<https://www.ebi.ac.uk/chembl/>, μεταμορφώθηκε στις 10/12/2021)

Οι πιο σημαντικοί τύποι δεδομένων στο ChEMBL αφορούν έγγραφα (από τα οποία εξάγονται τα δεδομένα), ενώσεις (ουσίες που έχουν ελεγχθεί για τη βιοδραστικότητά τους), αναλύσεις (μεμονωμένα πειράματα που έχουν διεξαχθεί για την αξιολόγηση της βιοδραστικότητας) και στόχους (οι πρωτεΐνες ή τα συστήματα που παρακολουθούνται με ανάλυση) (Gaulton et al., 2012).

1.10.2 Χρήσιμα Εργαλεία Λογισμικού στην Χημειοπληροφορική

- *Konstanz Information Miner (KNIME)*

Η KNIME είναι μια πλατφόρμα ανάλυσης ανοιχτού κώδικα, η οποία είναι το κορυφαίο εργαλείο για ευρεία επεξεργασία, ενσωμάτωση, ανάλυση και εξερεύνηση δεδομένων (Berthold et al., 2009). Επιτρέπει την οπτική δημιουργία ροών δεδομένων (pipelines), την επιλεκτική εκτέλεση καθορισμένων βημάτων ανάλυσης και την παρουσίαση των

αποτελεσμάτων μέσω διαδραστικής οπτικής σε μοντέλα και δεδομένα. Η πλατφόρμα KNIME προσφέρει διαισθητική χρήση και υψηλό επίπεδο επεκτασιμότητας, γεγονός που την καθιστά σήμερα τη δημοφιλέστερη πλατφόρμα για εφαρμογές χημειοπληροφορικής.

- *Chemistry Development Kit (CDK)*

Το CDK είναι ένα λογισμικό χημειοπληροφορικής ανοιχτού κώδικα και ανάπτυξης (Steinbeck et al., 2003). Οι κόμβοι του CDK παρέχουν χαρακτηριστικά σχετικά με το χειρισμό χημικών ενώσεων, όπως πολλές εφαρμογές μετατροπής αρχείων για μόρια, υπολογισμού και σχεδίασης δομών 2D και 3D, υπολογισμού ομάδων συμμετρίας, υπολογισμού δακτυλικών αποτυπωμάτων, σωστού χειρισμού ατόμων υδρογόνου και εκτιμήσεις μοριακής ιδιότητας.

- *Indigo*

Το Indigo είναι ένα εργαλείο λογισμικού στην οργανική χημεία (<http://lifescience.opensource.epam.com/indigo/>). Ο χειρισμός και η λειτουργικότητα των οργανικών δομών με τους κόμβους Indigo για το KNIME μπορούν να επιτευχθούν μέσω διάφορων διαδικασιών όπως οι μετατροπές τους σε ενώσεις Kekulé και αρωματικές ενώσεις, χειρισμό των ατόμων υδρογόνου, δημιουργία μοριακών ιδιοτήτων, σύγκριση δακτυλικών αποτυπωμάτων, αποσύνθεση της ομάδας R κ.α. Οι πρόσθετες λειτουργίες του συγκεκριμένου εργαλείου περιλαμβάνουν μετατροπές αρχείων μεταξύ μορφών SDF, SMILES και CML.

- *RDKit*

Το RDKit (<http://www.rdkit.org/>) παρέχει επίσης εφαρμογές χημειοπληροφορικής μέσω του KNIME, όπως είναι το φιλτράρισμα και η αναζήτηση χημικών υποομάδων, χημικές αντιδράσεις, η δημιουργία 2D και 3D δομών, μοριακά αποτυπώματα κ.α.

- *Online Chemical Modeling Environment (OCHEM)*

Το OCHEM είναι μια βάση δεδομένων που περιέχει χιλιάδες καταχωρήσεις και είναι φιλικό προς το χρήστη περιβάλλον εφαρμογής η οποία στοχεύει στην απλοποίηση των διαδικασιών για την εκτέλεση υπολογισμών QSAR (Sushko et al., 2011). Αυτό επιτυγχάνεται με τον συνδυασμό πειραματικών αποτελεσμάτων που λαμβάνονται από μια βάση δεδομένων και μια διαδικασία μοντελοποίησης.

- *Chemical Identifier Resolver (CIR)*

Οι κόμβοι Chemical Identifier Resolver (CIR, που αναπτύχθηκε από την ομάδα CADD στο Εθνικό Ινστιτούτο Καρκίνου) για το KNIME επιτρέπουν την αναγνώριση μιας χημικής δομής με την προϋπόθεση ότι είναι γνωστό ένα αναγνωριστικό (<http://cactus.nci.nih.gov/chemical/structure>) στην δομή που μας ενδιαφέρει. Το CIR

είναι μια λύση για διάφορα αναγνωριστικά δομής και μπορεί επίσης να μετατρέψει ένα συγκεκριμένο αναγνωριστικό δομής σε ένα άλλο.

- *Enalos KNIME nodes*

Οι κόμβοι Enalos KNIME (<https://tech.knime.Org/community/enalos-nodes>) αναπτύχθηκαν από τη NovaMechanics Ltd. Οι κόμβοι Enalos για την πλατφόρμα KNIME σχετίζονται με πολλές σημαντικές πτυχές σχετικά με την αξιοποίηση και την ανάλυση δεδομένων στον τομέα της χημειοπληροφορικής και της ναοπληροφορικής (Leoni et al., 2017).

Ενότητα 2. Εργαλεία Χημειοπληροφορικής Enalos

2.1 Εισαγωγή

Τα εργαλεία χημειοπληροφορικής Enalos έχουν αναπτυχθεί από την εταιρεία NovaMechanics Ltd η οποία είναι υπεύθυνη για την ανάπτυξη νέων αλγορίθμων και πλατφορμών για την επίλυση προβλημάτων χημειοπληροφορικής, βιοπληροφορικής, ναοπληροφορικής, μοντελοποίησης, προσομοίωσης, φαρμακευτικής και χημείας υλικών. Στόχος της είναι να μειωθεί ο κίνδυνος και πειραματικό κόστος με τη σύλληψη νέων ιδεών μοντελοποίησης και την επινόηση των προσομοιώσεων που απαιτούνται για τη δοκιμή τους. Τα *in silico* μοντέλα και τα εργαλεία προσομοίωσης της NovaMechanics είναι υπεύθυνα για την ερμηνεία ερευνητικών αποτελεσμάτων και το σχεδιασμό αποτελεσματικών πειραμάτων με σκοπό την επικύρωσή τους. Τα εργαλεία χημειοπληροφορικής Enalos περιλαμβάνουν :

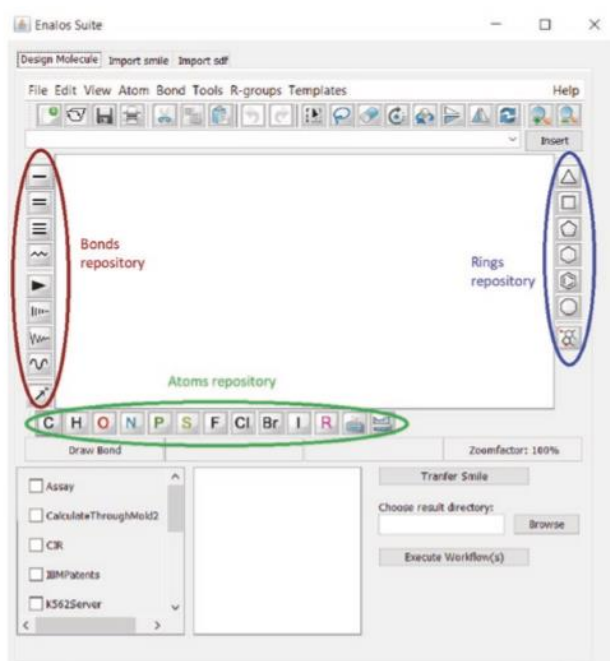
- **Πλατφόρμα Enalos Cloud:** περιλαμβάνει διαδικτυακές εφαρμογές που επιτρέπουν σε εργαστήρια και εταιρείες να ξεκινούν εύκολα εργασίες μοντελοποίησης, ανάλυσης δεδομένων και προσομοίωσης. Μέσω της πλατφόρμας Enalos Cloud οι *in silico* ροές εργασιών είναι διαθέσιμες στο διαδίκτυο μέσω διαδικτυακών υπηρεσιών και περιλαμβάνουν τα αντίστοιχα μοντέλα πρόβλεψης της NovaMechanics.
- **Enalos Suite:** είναι μια "offline" πλατφόρμα ανεξάρτητη της πλατφόρμας Enalos Cloud που έχει πρόσβαση σε πολλές βάσεις δεδομένων για εξόρυξη και χειρισμό δεδομένων.
- Οι **Κόμβοι Enalos+** για την πλατφόρμα ανάλυσης KNIME προσφέρουν ένα ευρύ φάσμα λειτουργιών χημειοπληροφορικής, βιοπληροφορικής και ναοπληροφορικής. Έχουν σχεδιαστεί για την εκτέλεση μοριακής

μοντελοποίησης και ο χρήστης έχει απευθείας πρόσβαση σε πολλαπλές βάσεις δεδομένων χημικών (<http://enalossuite.novamechanics.com/>).

2.2 Enalos Suite

Το Enalos Suite είναι ένα χρήσιμο εργαλείο στη διαδικασία ανακάλυψης φαρμάκων με τη βοήθεια υπολογιστή (*in silico*), καθώς παρέχει πολλές λειτουργίες, συμπεριλαμβανομένων του υπολογισμού μοριακών δεικτών, της εξόρυξης δεδομένων, και της αξιοποίησης βάσεων δεδομένων από δημοφιλείς χημικές βάσεις δεδομένων, (PubChem, UniChem, SureChEMBL, IBM διπλώματα ευρεσιτεχνίας κ.λ.π),. διευκολύνοντας την ανάκτηση δεδομένων για χιλιάδες χημικές ενώσεις. Παράλληλα, επιτρέπει την περιγραφή και την πρόβλεψη της δραστηριότητας/ιδιοτήτων για ενώσεις ενδιαφέροντος με ελάχιστα απαιτούμενα βήματα. Βασική υποδομή στην ανάπτυξη του Enalos Suite αποτελεί η πλατφόρμα KNIME (Konstanz Information Miner) (Varsou et al., 2018). Η επιφάνεια εργασίας του Enalos Suite αποτελείται από τρεις καρτέλες που επιτρέπουν την εισαγωγή χημικών δομών σε διαφορετικές μορφές:

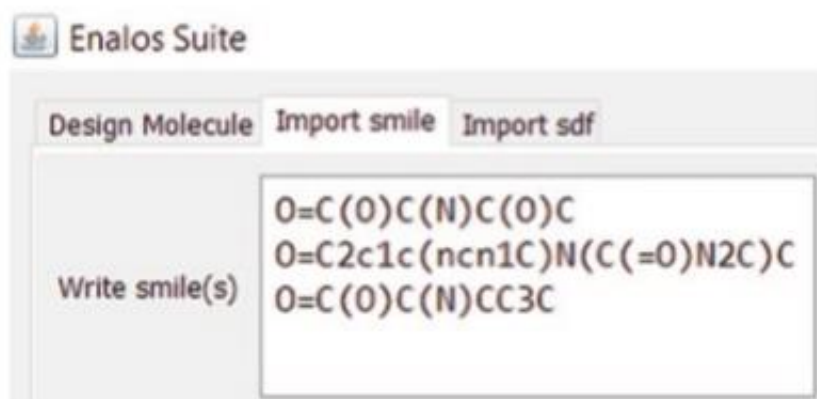
- Η κύρια καρτέλα (**Design Molecule**), όπου ο χρήστης μπορεί να σχεδιάσει και να υποβάλει προς ανάλυση αρκετές χημικές ενώσεις χρησιμοποιώντας ένα φιλικό εργαλείο σχεδίασης



Εικόνα 2.1. Καρτέλα “Design Molecule” Enalos Suite (Varsou et al., 2018)

Στην κύρια καρτέλα, η πλατφόρμα Enalos επιτρέπει στον χρήστη να σχεδιάσει εύκολα οποιαδήποτε χημική ένωση η οποία υποβάλλεται στη συνέχεια για περαιτέρω ανάλυση. Διάφορες δυνατότητες είναι διαθέσιμες σε διαφορετικά πάνελ, όπως η σχεδίαση διπλών και τριπλών δεσμών, η δημιουργία αλυσίδων και η δημιουργία stereo bonds. Διατίθενται επίσης για χρήση δακτύλιοι προπανίου, βουτανίου, πεντανίου, εξανίου, οκτανίου και βενζολίου, μαζί με μερικά πιο σύνθετα πρότυπα δομής όπως αλκαλοειδή, αμινοξέα, β-λακτάμες, υδατάνθρακες και στεροειδή. Ο χρήστης μπορεί επίσης να επιλέξει ανάμεσα σε διαφορετικά ετεροάτομα (N, P, S, F, Cl, Br, I) που συνήθως υπάρχουν σε οργανικά χημικά μόρια, να εισάγει ένα σύμβολο στοιχείου ή ομάδας μέσω πληκτρολογίου και να επιλέξει νέα σύμβολα σχεδίασης από τον περιοδικό πίνακα. Διατίθενται επίσης λειτουργίες, όπως η επιλογή, διαγραφή, περιστροφή, μετακίνηση και αφαίρεση ολόκληρων ή τμημάτων των σχεδιασμένων μορίων. Τέλος, υπάρχει η δυνατότητα ο χρήστης να ανοίξει, να αποθηκεύσει και να μετατρέψει αρχεία με ποικιλία χημικών μορφών, όπως π.χ. τα SMILES ή τα IUPAC Chemical Identifier.

- Η **καρτέλα εισαγωγής των SMILES** (Import Smile) η οποία δίνει τη δυνατότητα στον χρήστη να εισάγει και να υποβάλει χημικά είδη με την μορφή των SMILES (Εικόνα 2.2).



Εικόνα 2.2. Καρτέλα Εισαγωγής των SMILES του Enalos Suite (Varsou et al., 2018)

Εάν η γραμματοσειρά των SMILES είναι γνωστή, ο χρήστης μπορεί να την υποβάλει απευθείας στην Καρτέλα Εισαγωγής των SMILES. Όμως εάν η γραμματοσειρά των SMILES δεν είναι αρχικά γνωστή, δίνεται στον χρήστη η δυνατότητα να σχεδιάσει τη χημική δομή μέσω σχεδιαστικού εργαλείου και στη συνέχεια να μετατρέψει τη δομή σε μορφή των SMILES. Αυτό διευκολύνει τη δημιουργία πολλών δομών, επιτρέποντας την πραγματοποίηση πολλαπλών τροποποιήσεων με τη χρήση του

σχεδιαστικού εργαλείου στην κύρια καρτέλα και στη συνέχεια όλες οι δομές μπορούν να μεταφερθούν SMILES στην κατάλληλη καρτέλα επιτρέποντας την ανάλυση για ολόκληρο το σύνολο των παραγόμενων δομών.

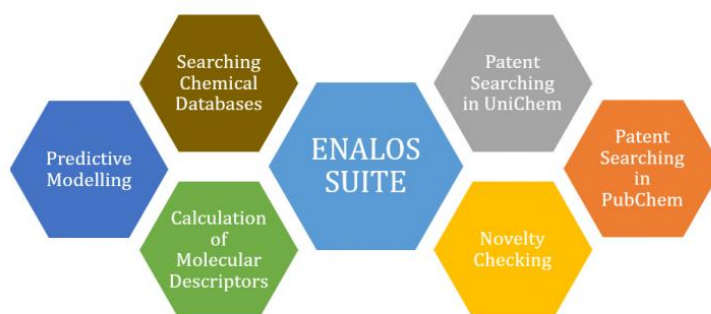
- Η **καρτέλα εισαγωγής sdf**, όπου ο χρήστης μπορεί να αναζητήσει το αρχείο .sdf της χημικής ένωσης (-ων) που τον ενδιαφέρει (Εικόνα 2.3)



Εικόνα 2.3. Καρτέλα «Εισαγωγής sdf» του Enalos Suite (Varsou et al., 2018)

Τα αρχεία .sdf περιέχουν αρχεία χημικής δομής, που χρησιμοποιούνται ως τυπική μορφή ανταλλαγής πληροφοριών χημικών δομών. Η δομή μιας ένωσης σε μορφή .sdf μπορεί να εξαχθεί από τη βάση δεδομένων PubChem ή άλλες βάσεις δεδομένων και να μεταφορτωθεί στο Enalos Suite μέσω του αντίστοιχου πεδίου (Varsou et al., 2018).

2.3 Λειτουργίες του Enalos Suite



Εικόνα 2.4. Οι λειτουργίες του Enalos Suite (<http://enalossuite.novamechanics.com/index.php/enalos-suite/>)

Οι λειτουργίες του Enalos Suite βασίζονται στις βάσεις δεδομένων και επιτρέπουν την πρόσβαση και την ανάκτηση δεδομένων από τα CIR, PubChem και UniChem.

2.3.1 Λειτουργία CIR

Η λειτουργία CIR επιτρέπει στο χρήστη να αποκτήσει άμεση πρόσβαση στο CIR (Chemical Identifier Resolver). Το CIR λειτουργεί ως αναλυτής για διαφορετικά αναγνωριστικά χημικής δομής και επιτρέπει τη μετατροπή ενός δεδομένου αναγνωριστικού δομής σε άλλη αναπαράσταση ή αναγνωριστικό δομής. Μέσω ενός μενού GUI μπορούν να επιλεγούν πολλές μορφές εξόδου. Τα διαθέσιμα αναγνωριστικά είναι τα εξής: Standard InChI, Standard InChIKey, InChIKey Simplified, SMILES, NCI/CADD FICTS Identifier, NCI/CADD FICuS Identifier, NCI/CADD Identifier, CACTVS HASHISY Hashcode, CACTVS HASHISY Hashcode, IUPACChemryS, Number CADD, CADD FICTS Identifier, Μοριακό βάρος, Χημικός τύπος, Αριθμός Δοτών Δεσμών Υδρογόνου, Αριθμός Αποδεκτών Δεσμών Υδρογόνου, Ο κανόνας των πέντε παραβιάσεων του Lipinski, ο αριθμός αποτελεσματικά περιστρεφόμενων δεσμών, λίστα χημικών ονομάτων για μια δομή και αρχείο SD μιας δομής.

2.3.2 Λειτουργίες του PubChem

Η βάση δεδομένων PubChem όπως αναφέρθηκε και παραπάνω είναι ένας δημόσιος χώρος αποθήκευσης χημικών πληροφοριών, συμπεριλαμβανομένων των μοριακών δομών και των βιοδραστικότητων τους που περιλαμβάνει σήμερα πληροφορίες για περισσότερες από 90 εκατομμύρια ενώσεις. Το PubChem προσφέρει επίσης ηλεκτρονικές υπηρεσίες που υποστηρίζουν την καινοτομία στα φάρμακα, όπως η τελειοποίηση και η ανάλυση των αποθηκευμένων ουσιών και αναλύσεων οι οποίες έχουν πολλές εφαρμογές στη χημική βιολογία, τη φαρμακευτική χημεία και στην πληροφορική. Επίσης, είναι ένας πόρος δευτερευόντων βάσεων δεδομένων και διαδικτυακών υπηρεσιών, συμβάλλοντας με αυτόν τον τρόπο στη διαδικασία ανακάλυψης φαρμάκων. Η κατηγορία Enalos Suite PubChem περιέχει πέντε λειτουργίες που παρέχουν άμεση πρόσβαση στη βάση δεδομένων PubChem για εξαγωγή χρήσιμων πληροφοριών:

1. Η λειτουργία PubChem επιτρέπει στο χρήστη να αναζητήσει στη βάση δεδομένων PubChem οποιαδήποτε ένωση θέλει και να εξάγει το PubChem CID, το όνομα IUPAC, το InChI, το InChI-Key, τον μοριακό της τύπο και το μοριακό της βάρος, το SMILES και την άμεση διεύθυνση URL του PubChem.
2. Η λειτουργία Assay δίνει στον χρήστη πρόσβαση στη βάση δεδομένων PubChem. Συγκεκριμένα, γνωρίζοντας τους κωδικούς αναγνώρισης που έχει μια ουσία ή μια ένωση [SID (SubstanceID) και CID (CompoundID) αντίστοιχα]

στο PubChem είναι δυνατό να ανακτηθούν οι αναλύσεις στις οποίες έχει δοκιμαστεί η ουσία ή ένωση αυτή.

3. Η λειτουργία Patent Coverage Information επιτρέπει στο χρήστη να εξαγάγει πληροφορίες σχετικά με την κάλυψη διπλωμάτων ευρεσιτεχνίας για χιλιάδες ενώσεις.
4. Μέσω της λειτουργίας Similar Compounds, ο χρήστης μπορεί να αναζητήσει σε ολόκληρη τη βάση δεδομένων PubChem παρόμοιες ενώσεις (χρησιμοποιώντας το Tanimoto Similarity) και να λάβει το PubChem CID, τον μοριακό τύπο, το βάρος και τον αριθμό των περιστρεφόμενων δεσμών (ο περιστρεφόμενος δεσμός είναι ο απλός δεσμός που βρίσκεται εκτός μιας δομής δακτυλίου χωρίς να συνδέεται σε ένα άτομο υδρογόνου και χωρίς να συνδέεται σε ένα τερματικό άτομο) των παρόμοιων ενώσεων με την ένωση που τον ενδιαφέρει.
5. Η λειτουργία Vendor δίνει τη δυνατότητα στον χρήστη να λάβει πληροφορίες σχετικά με την εμπορική διαθεσιμότητα χιλιάδων χημικών ενώσεων.

Έτσι, εξαλείφεται σημαντικά ο χρόνος που αφιερώνει ο κάθε ενδιαφερόμενος χρήστης στη μονότονη και κουραστική αναζήτηση πληροφοριών από διαφορετικούς πόρους, στη διαδικασία ανακάλυψης φαρμάκων.

2.3.3 Λειτουργία του UniChem

Το UniChem είναι μια απλή, μεγάλης κλίμακας βάση δεδομένων μεταξύ χημικών δομών και δεδομένων χημείας του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL)-Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (European Molecular Biology Laboratory, EMBL -European Bioinformatics Institute, EBI, Heidelberg, Germany). Η λειτουργία UniChem προσφέρει στον χρήστη άμεση πρόσβαση στις 27 διαθέσιμες βάσεις δεδομένων στο UniChem οι οποίες ομαδοποιούνται σε πέντε εύκολα αναγνωρίσιμες κατηγορίες. Η εξαγωγή δεδομένων από βάσεις δεδομένων UniChem, προσφέρει μεγάλη ευελιξία επειδή επιτρέπει την άμεση ανάλυση και τον χειρισμό τους καθώς και γρήγορη και αυτοματοποιημένη μοντελοποίηση. Επίσης, δύο σημαντικές λειτουργίες εύρεσης διπλωμάτων ευρεσιτεχνίας της UniChem στο Enalos Suite είναι τα Διπλώματα ευρεσιτεχνίας IBM και SureChEMBL.

- Διπλώματα ευρεσιτεχνίας της IBM

Αυτή η λειτουργία δίνει στον χρήστη πρόσβαση στη βάση δεδομένων διπλωμάτων ευρεσιτεχνίας της IBM μέσω του UniChem, η οποία περιέχει 31 βάσεις δεδομένων

χημικών (π.χ. ChEMBL, PubChem, DrugBank, KEGG, ZINC, BindinDB, eMolecules, πατέντες IBM, SureChEMBL κ.λπ.). Η λειτουργία διπλωμάτων ευρεσιτεχνίας IBM εξάγει τα InChIKeys των ουσιών που βρέθηκαν στη βάση δεδομένων ευρεσιτεχνιών της IBM και τα InChIKeys που δεν βρέθηκαν σε αυτήν τη βάση δεδομένων. Τα αποτελέσματα δίνονται σε δύο αρχεία .xls.

- Διπλώματα ευρεσιτεχνίας SureChEMBL

Αυτή η λειτουργία δίνει στον χρήστη πρόσβαση στη βάση δεδομένων SureChEMBL μέσω του UniChem. Η λειτουργία πατεντών SureChEMBL εξάγει τα InChIKeys των ενώσεων που βρέθηκαν στη βάση δεδομένων SureChEMBL και τα InChIKeys που δεν βρέθηκαν σε αυτήν τη βάση δεδομένων. Τα αποτελέσματα δίνονται σε δύο αρχεία .xls (<http://enalossuite.novamechanics.com/index.php/enalos-suite/>).

2.3.4 Έλεγχος Καινοτομίας

Για τον έλεγχο καινοτομίας νέων σχεδιασμένων χημικών ενώσεων, η NovaMechanics Ltd έχει αναπτύξει τη λειτουργία Novelty Checking Enalos Suite. Αυτή η λειτουργία δίνει στον χρήστη πρόσβαση στη βάση δεδομένων UniChem, η οποία περιέχει 31 βάσεις δεδομένων χημικών προκειμένου να ελέγξει εάν υπάρχει μια ένωση σε αυτές τις βάσεις δεδομένων. Η λειτουργία Έλεγχος καινοτομίας εξάγει τα InChIKeys των ενώσεων που βρέθηκαν και στις 31 βάσεις δεδομένων UniChem και τα InChIKeys που δεν βρέθηκαν σε αυτές τις βάσεις δεδομένων UniChem, που εμφανίζονται σε δύο αρχεία .xlsx.

2.3.5 Μοριακοί δείκτες του Enalos Suite

Μέσω του Enalos Suite υπάρχει και η δυνατότητα υπολογισμού μοριακών δεικτών για μία ή περισσότερες ενώσεις χρησιμοποιώντας το λογισμικό Mold2 το οποίο είναι ιδιαίτερα αξιόπιστο καθώς οι δείκτες Mold2 έχουν χρησιμοποιηθεί για την ανάπτυξη πλήρως αξιολογημένων μοντέλων. Ο υπολογισμός των μοριακών δεικτών χρησιμοποιώντας το λογισμικό Mold2 είναι διαθέσιμος στο Enalos Suite έγινε διαθέσιμος μέσω της λειτουργίας Calculate Through Mold2.

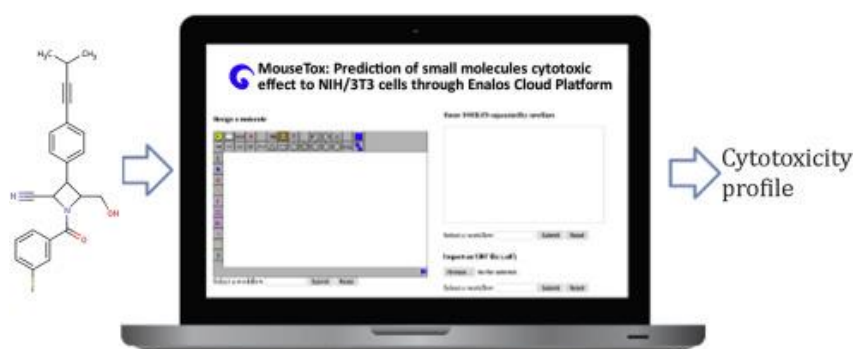
2.4 Μοντέλα πρόβλεψης Enalos Cloud & Enalos Suite

Οι έρευνες που γίνονται στο πλαίσιο της ανακάλυψης φαρμάκων επικεντρώνονται στην υποκατάσταση των πειραμάτων *in vivo* και *in vitro* με υπολογιστικές προσεγγίσεις *in silico*. Η έρευνα με τη βοήθεια υπολογιστή βοηθάει στον περιορισμό των δαπανηρών και εντατικών πειραματικών διαδικασιών και μπορεί επίσης να είναι

η απάντηση στο ερώτημα της ηθικής που προκύπτει από τη χρήση πειραματόζωων. Μία από τις πιο γνωστές *in silico* προσεγγίσεις είναι η ανάπτυξη μοντέλων πρόβλεψης που συνδέουν τα δομικά χαρακτηριστικά των μορίων με τη δραστικότητα ή το προφίλ της τοξικότητάς τους (μοντέλα ποσοτικών σχέσεων δομής δραστικότητας/τοξικότητας, QSAR/QSTR). Χρησιμοποιώντας αυτού του είδους μοντέλα, οι ερευνητές μπορούν να αναζητήσουν ισχυρά μοτίβα μεταξύ των πειραματικών δεδομένων και να χρησιμοποιήσουν αυτά τα πρότυπα για να κάνουν ακριβείς προβλέψεις σε μελλοντικά δεδομένα. Οι διαδικασίες μοντελοποίησης απαιτούν μια σειρά από σημαντικά βήματα, όπως η επιλογή μεταβλητών και η εσωτερική και εξωτερική αξιολόγηση, προκειμένου να δημιουργηθεί ένα ισχυρό μοντέλο που παράγει αξιόπιστες προβλέψεις. Όταν αναπτύσσεται ένα μοντέλο, είναι σημαντικό τα αποτελέσματά του να διαδίδονται στην επιστημονική κοινότητα, συμπεριλαμβανομένων των ερευνητών χωρίς υπολογιστικό υπόβαθρο, προκειμένου να χρησιμοποιηθούν σε πραγματικές εφαρμογές. Παρακάτω περιγράφονται τα αξιόπιστα μοντέλα πρόβλεψης που περιλαμβάνονται στο Enalos Suite καθώς και μοντέλα που είναι διαθέσιμα στο Enalos Cloud. (Varsou et al., 2018).

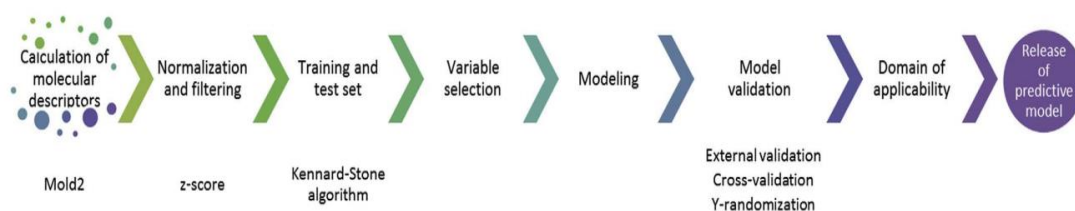
2.4.1 Μοντέλο κυτταροτοξικότητας MouseTox

Οι Varsou et al., 2017 ανέπτυξαν ένα μοντέλο πρόβλεψης της κυτταροτοξικότητας των χημικών ενώσεων το οποίο βασίστηκε αποκλειστικά σε *in silico* μοριακούς δείκτες οι οποίοι εξηγούν τα τοπολογικά, γεωμετρικά και δομικά χαρακτηριστικά αυτών. Το μοντέλο αυτό είναι το MouseTox το οποίο είναι ένα πλήρως επικυρωμένο μοντέλο QSTR και είναι διαθέσιμο διαδικτυακά μέσω της πλατφόρμας Enalos Cloud στην διεύθυνση <http://www.enaloscloud.novamechanics.com/EnalosWebApps/MouseTox/>.



Εικόνα 2.5. Το μοντέλο του MouseTox στην πλατφόρμα Enalos Cloud (Varsou et al., 2017)

Αυτή η διαδικτυακή υπηρεσία προσφέρει στον χρήστη δωρεάν πρόσβαση στα αποτελέσματα του μοντέλου και επομένως μπορεί να λειτουργήσει ως εργαλείο πρόβλεψης τοξικότητας για την αξιολόγηση κινδύνου νέων ενώσεων, χωρίς ιδιαίτερες απαιτήσεις ή προηγούμενες δεξιότητες προγραμματισμού (Varsou et al., 2017). Το μοντέλο MouseTox δημιουργήθηκε με βάση μια ροή εργασίας η οποία αναπτύχθηκε στην πλατφόρμα ανάλυσης KNIME που περιλαμβάνει τους ιδιόκτητους κόμβους KNIME του Enalos+. Η ροή εργασίας περιλάμβανε όλα τα βήματα που απαιτούνται για την παροχή ενός πλήρως επικυρωμένου μοντέλου πρόβλεψης όπως φαίνεται στην Εικόνα 2.6. Το σύνολο δεδομένων που επιλέχθηκε για την ανάπτυξη του μοντέλου MouseTox αποτελείται από 5416 ενώσεις που δοκιμάστηκαν για κυτταροτοξικές επιδράσεις σε κύτταρα NIH/3T3 (κύτταρα εμβρυϊκών ινοβλαστών ποντικού), ως μέρος μιας μελέτης για τον εντοπισμό νέων φαρμάκων για τη θεραπεία της νόσου Chagas. Οι ενώσεις από το αρχικό σύνολο δεδομένων ταξινομήθηκαν ως «ενεργές» (3109 ενώσεις κυτταροτοξικές στην NIH/3T3 κυτταρική σειρά) ή «ανεργές» (2307 ενώσεις μη-κυτταροτοξικές στην NIH/3T3). Για κάθε ένωση του αρχικού συνόλου δεδομένων, υπολογίστηκαν 777 μοριακοί δείκτες με την χρήση του λογισμικού Mold2. Ορισμένοι δείκτες φιλτραρίστηκαν με αποτέλεσμα 424 να χρησιμοποιηθούν για την ανάπτυξη του μοντέλου QSTR. Στη συνέχεια, το αρχικό σύνολο δεδομένων (5416 ενώσεις) χωρίστηκε σε δύο ξεχωριστά και ταυτόχρονα αντιπροσωπευτικά υποσύνολα ενώσεων (με κυτταροτοξικές και μη-κυτταροτοξικές ενώσεις) χρησιμοποιώντας τους αλγόριθμους Kennard και Stones, το υποσύνολο με τα δεδομένα εκπαίδευσης (training compounds) και το υποσύνολο με τα δεδομένα ελέγχου (test compounds) σε αναλογία 75:25. Τα δεδομένα εκπαίδευσης είχαν αναλογία ενεργών/ανεργών ενώσεων ίση με 1,28, ενώ τα δεδομένα ελέγχου ίση με 1,58.



Εικόνα 2.6. Το συνολικό διάγραμμα ροής με τα πιο σημαντικά βήματα της ανάλυσης (Varsou et al., 2017)

Επίσης, προκειμένου να αφαιρεθούν δείκτες που δημιουργούν θόρυβο στην ανάλυση, η επιλογή μεταβλητής InfoGain (Information Gain) μαζί με τη μέθοδο του αξιολογητή Ranker εφαρμόστηκαν στο σύνολο των δεδομένων εκπαίδευσης,

προκειμένου να προσδιοριστούν 15 δείκτες ως οι πιο κρίσιμοι για την ανάπτυξη του μοντέλου (variable selection). Ιδιαίτερη προσοχή δόθηκε στην επικύρωση του προτεινόμενου μοντέλου, με τη χρήση διαφορετικών στρατηγικών για εξωτερική και εσωτερική επικύρωση ώστε να πληρούνται τα κριτήρια που προτείνει ο ΟΟΣΑ και ο τομέας εφαρμογής οριοθετήθηκε προκειμένου να καθοριστεί ο χώρος των αξιόπιστων προβλέψεων. Για την εσωτερική επικύρωση, η σταθερότητα του μοντέλου εξετάστηκε με την εκτέλεση δοκιμών διασταυρούμενης επικύρωσης (leave-k-out cross-validation tests). Επιπλέον, πραγματοποιήθηκε εξωτερική επικύρωση χρησιμοποιώντας τα δεδομένα ελέγχου. Στον Πίνακα 2.1 παρουσιάζονται τα σχετικά στατιστικά στοιχεία που αποδεικνύουν την ακρίβεια και την αξιοπιστία του αναπτυγμένου μοντέλου.

Πίνακας 2.1. Στατιστικά ακριβείας του μοντέλου πρόβλεψης (αναπροσαρμογή από πηγή: Varsou et al., 2018 στις 16/01/2022)

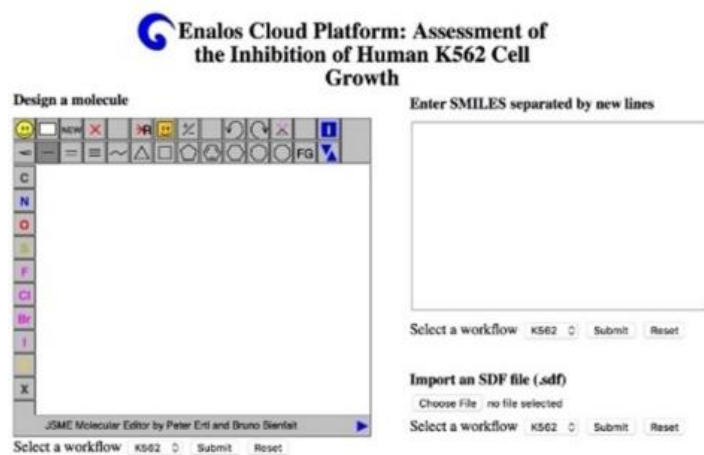
Ακρίβεια		Ευαισθησία		Ειδικότητα	
Δεδομένα εκπαίδευσης	Δεδομένα ελέγχου	Δεδομένα εκπαίδευσης	Δεδομένα ελέγχου	Δεδομένα εκπαίδευσης	Δεδομένα ελέγχου
0.915	0.832	0.928	0.866	0.898	0.779

Τέλος, η δοκιμή τυχαιοποίησης Y που αποτελεί ένα εργαλείο που χρησιμοποιείται για την επικύρωση των μοντέλων QSPR/QSAR, όταν εφαρμόστηκε στα δεδομένα απέδειξε την αξιοπιστία και τη στατιστική σημασία του προτεινόμενου μοντέλου. Είναι σημαντικό να αναφερθεί ότι ακόμα κι αν ένα μοντέλο είναι πλήρως επικυρωμένο και έχει υψηλή προγνωστική ισχύ, δεν μπορεί να προσφέρει αξιόπιστες προβλέψεις για όλες τις χημικές ενώσεις. Έτσι, είναι απαραίτητο να προσδιοριστεί εάν μια πρόβλεψη μπορεί να θεωρηθεί αξιόπιστη ή όχι με βάση το πεδίο εφαρμογής της. Το πεδίο εφαρμογής του μοντέλου MouseTox υπολογίστηκε χρησιμοποιώντας μετρήσεις ομοιότητας με βάση τον δείκτη της Ευκλείδειας απόστασης μεταξύ όλων των ενώσεων των δεδομένων εκπαίδευσης και ελέγχου. Η ενσωμάτωση του μοντέλου MouseTox στο Enalos Suite διευκόλυνε σημαντικά τον εικονικό έλεγχο για την αξιολόγηση της κυτταροτοξικότητας νέων ενώσεων. Με βάση το μοντέλο που αναπτύχθηκε και για να ληφθούν προβλέψεις κυτταροτοξικότητας για μια δεδομένη ένωση ενδιαφέροντος, ο χρήστης μπορεί να υποβάλει μία ή περισσότερες δομές είτε χρησιμοποιώντας την χημική δομή της ένωσης είτε υποβάλλοντας μια χημική μορφή SMILES ή ένα αρχείο .sdf. Έπειτα, έχει στη διάθεσή του το προφίλ τοξικότητας κάθε

ένωσης μαζί με μια ένδειξη της αξιοπιστίας του αποτελέσματος με βάση τα όρια του πεδίου εφαρμογής του μοντέλου (Varsou et al., 2018).

2.4.2 Μοντέλο αναστολής K562

Εκτός από το μοντέλο MouseTox, στο Enalos Suite ενσωματώνονται περισσότερα μοντέλα πρόβλεψης που καλύπτουν ένα ευρύ φάσμα βιολογικών δράσεων. Μεταξύ αυτών, περιλαμβάνεται ένα μοντέλο πρόβλεψης της αναστολής του πολλαπλασιασμού της κυτταρικής σειράς K562. Τα κύτταρα K562 αποτελούν την πρώτη ανθρώπινη κυτταρική σειρά μυελογενούς λευχαιμίας που έχει ανασταλεί η ανάπτυξή τους. Η αναστολή των κυττάρων K562 αποτελεί ενδιάμεση οδό για τον εντοπισμό φαρμακευτικών ενώσεων που επάγουν την παραγωγή της εμβρυϊκής αιμοσφαιρίνης F (HbF). Η αύξηση των επιπέδων της HbF θεωρείται ότι έχει μεγάλες δυνατότητες στη θεραπεία της β-θαλασσαιμίας και μπορεί να αντισταθμίσει τη μειωμένη παραγωγή της κύριας αιμοσφαιρίνης στους ενήλικες, της αιμοσφαιρίνης A (HbA). Έτσι, προτάθηκε ένα *in silico* και πλήρως επικυρωμένο μοντέλο για την πρόβλεψη της αναστολής των κυττάρων K562, που πιθανώς σχετίζεται με την επαγωγή της HbF. Αυτό το μοντέλο είναι διαθέσιμο στο διαδίκτυο μέσω της πλατφόρμας Enalos Cloud <http://enalos.insilicotox.com/K562> και βρίσκεται ενσωματωμένο επίσης στο Enalos Suite (Afantitis et al., 2018).



Εικόνα 2.7. Η πρόβλεψη της αναστολής της ανάπτυξης των ανθρώπινων κυττάρων K562 μέσω της πλατφόρμας Enalos Cloud (Afantitis et al., 2018)

Αρχικά σχεδιάστηκε μια ροή εργασιών KNIME που διευκολύνεται από τους κόμβους Enalos η οποία περιλαμβάνει τα εξής βήματα της διαδικασίας μοντελοποίησης: επιμέλεια και επεξεργασία δεδομένων, υπολογισμός των μοριακών δεικτών, επιλογή

των πιο σημαντικών μοριακών δεικτών (variable selection), ανάπτυξη και επικύρωση μοντέλου και προσδιορισμός του τομέα εφαρμογής του μοντέλου. Για την ανάπτυξη του μοντέλου, επιλέχθηκε ένα αρχικό σύνολο δεδομένων από 129 διαφορετικά μικρά μόρια- αναστολείς των κυττάρων K562 τα οποία αντλήθηκαν από τη βάση δεδομένων PubChem. Οι ενώσεις ταξινομήθηκαν ως «ενεργές» (67) και «ανεργές» (62) με βάση τη βιολογική τους δράση. Χωρίστηκαν σε δύο υποσύνολα δεδομένων: το υποσύνολο δεδομένων εκπαίδευσης και το υποσύνολο δεδομένων ελέγχου σε αναλογία 80:20 για να χρησιμοποιηθούν για την ανάπτυξη και την επικύρωση του μοντέλου, αντίστοιχα. Για κάθε ένωση, υπολογίστηκαν 777 μοριακοί δείκτες χρησιμοποιώντας το λογισμικό Mold2 και μετά από ένα βήμα φιλτραρίσματος, μερικοί από αυτούς αφαιρέθηκαν προκειμένου να παραμείνουν οι μοριακοί δείκτες που σχετίζονται καλύτερα με την ανασταλτική δραστηριότητα των ενώσεων για τα κύτταρα K562. Στη συνέχεια αναπτύχθηκαν τρία διαφορετικά μοντέλα τα οποία περιλαμβάνουν τρεις διαφορετικές μεθοδολογίες μοντελοποίησης (το kNN, random tree, and random forest) και δύο διαφορετικές τεχνικές επιλογής μεταβλητών (Gain Attribute evaluator & InfoGain Attribute Ratio Feature) τα οποία αργότερα συνδυάστηκαν και προέκυψε ένα "συνδυαστικό μοντέλο". Τα στατιστικά στοιχεία αξιοπιστίας τους παρουσιάζονται στον Πίνακα 2.2 και το συνδυαστικό μοντέλο παρουσιάζει την υψηλότερη ακρίβεια (84%).

Πίνακας 2.2. Αποτελέσματα επικύρωσης μοντέλου (Αναπροσαρμογή από πηγή: Varsou et al., 2018 στις 16/01/2022)

	Εξειδίκευση	Ευσαιθησία	Ακρίβεια (precision)	Ορθότητα (accuracy)
Μοντέλο I—random tree	0.6	0.733	0.733	0.68
Μοντέλο II—kNN	0.5	0.933	0.737	0.76
Μοντέλο III—random forest	0.7	0.733	0.786	0.72
Συνδυαστικό μοντέλο	0.7	0.933	0.824	0.84

Και σε αυτή την περίπτωση το πεδίο εφαρμογής του προτεινόμενου μοντέλου ορίστηκε με βάση τις Ευκλείδειες αποστάσεις. Αυτό το επικυρωμένο μοντέλο μπορεί να προσπελαστεί μέσω της πλατφόρμας Enalos Suite και θα μπορούσε εύκολα να χρησιμοποιηθεί για την ανακάλυψη φαρμάκων. Ο χρήστης μπορεί να υποβάλει μία ή περισσότερες δομές είτε χρησιμοποιώντας την χημική δομή της ένωσης είτε υποβάλλοντας μια χημική μορφή SMILES ή ένα αρχείο .sdf και να εξετάσει εάν αυτές οι ενώσεις παρουσιάζονται ως "ενεργές" ή "ανεργές" στην αναστολή ανάπτυξης των κυττάρων K562 (Afantitis et al., 2018).

2.4.3 Μοντέλο αναστολής του παράγοντα TNF (Tumor Necrosis Factor)

Το μοντέλο αναστολής του παράγοντα TNF βασίζεται σε ολοκληρωμένες υπολογιστικές και πειραματικές μεθόδους με σκοπό την ανακάλυψη νέων μικρών μορίων που είναι άμεσοι αναστολείς της λειτουργίας του TNF. Ο Παράγοντας Νέκρωσης Όγκων (Tumor Necrosis Factor) είναι μια προφλεγμονώδης κυτοκίνη η οποία θεωρείται βασικός μεσολαβητής της φυσιολογικής φλεγμονώδους απόκρισης ενάντια στα μικρόβια και τη βλάβη των ιστών. Επίσης, εμπλέκεται σε μια σειρά παθολογικών καταστάσεων όπως η ρευματοειδής αρθρίτιδα (RA), η ψωριασική αρθρίτιδα, η νόσος του Crohn και η σκλήρυνση κατά πλάκας. Για τον έλεγχο των δυσμενών λειτουργιών του TNF, οι μέχρι τώρα προσπάθειες έχουν επικεντρωθεί στην παρεμπόδιση της δέσμευσης του TNF στους υποδοχείς του οδηγώντας στην κινητοποίηση του ενδιαφέροντος της βιομηχανίας για παραγωγή anti-TNF ενώσεων. Το μοντέλο σχεδιάστηκε με βάση την πλατφόρμα KNIME που διευκολύνεται από τους κόμβους Enalos. Δημιουργήθηκε συνδυάζοντας μοντελοποίηση βασισμένη στην δομή του TNF και την δομή του υποδοχέα του χρησιμοποιώντας το μεγαλύτερο διαθέσιμο σύνολο γνωστών αναστολέων TNF (2481 μικρά μόρια) από τη βάση δεδομένων PubChem. 1149 που εμφάνιζαν αναστολή πάνω από το 50% χαρακτηρίστηκαν ως ενεργές και οι υπόλοιπες 1332 χαρακτηρίστηκαν ως ανενεργές. Για κάθε ένωση, υπολογίστηκαν 777 μοριακοί δείκτες χρησιμοποιώντας το λογισμικό Mold2 οι οποίοι φιλτραρίστηκαν και παρέμειναν 616 για τη δημιουργία του μοντέλου. Για λόγους επικύρωσης, το αρχικό σύνολο δεδομένων χωρίστηκε επίσης σε υποσύνολα εκπαίδευσης και υποσύνολα δοκιμών. Κατασκευάστηκαν τρία διαφορετικά μοντέλα από τα οποία προέκυψε ένα συνδυαστικό μοντέλο το οποίο είχε την καλύτερη απόδοση από όλα τα άλλα με βάση τις μετρήσεις επικύρωσης δηλαδή την ειδικότητα, την ευαισθησία και την ακρίβεια. Το πεδίο εφαρμογής του μοντέλου MouseTox υπολογίστηκε χρησιμοποιώντας μετρήσεις ομοιότητας με βάση τον δείκτη της Ευκλείδειας απόστασης μεταξύ όλων των ενώσεων των δεδομένων εκπαίδευσης και ελέγχου. Με βάση αυτά τα μικρά μόρια δημιουργήθηκε μια λίστα με εννέα μικρά μόρια ως υποψήφια για άμεση αναστολή της λειτουργίας του TNF. Η in vitro αξιολόγηση αυτών των ενώσεων οδήγησε στην επιλογή δύο μικρών μορίων που δρουν ως ισχυροί άμεσοι αναστολείς της λειτουργίας του TNF, με τιμές IC₅₀ (IC₅₀: η συγκέντρωση μιας ανασταλτικής ουσίας που απαιτείται για την αναστολή, in vitro, μιας βιολογικής διεργασίας ή βιολογικού συστατικού κατά 50%) συγκρίσιμες με αυτές ενός γνωστού άμεσου αναστολέα (SPD304), ενώ παράλληλα παρουσιάζουν σημαντικά μικρή τοξικότητα. Αυτές οι ενώσεις, συγκεκριμένα οι T8 και T23, οι οποίες

βιοσυγκέντρωση αναφέρεται στη διαδικασία πρόσληψης και συσσώρευσης χημικών ουσιών σε ζωντανούς οργανισμούς.

MSlogBP: Το μοντέλο προβλέπει το σημείο βρασμού χρησιμοποιώντας την μέθοδο της ομαδοποίησης ουσιών και την σύγκρισή τους (read across predicting modelling). Το μοντέλο είναι βασισμένο σε >4000 ενώσεις από τη βάση δεδομένων EPA.

MSlogP: Το μοντέλο MSlogP προβλέπει τον συντελεστή κατανομής οκτανόλης/νερού χρησιμοποιώντας την ανάγνωση (read across). Το μοντέλο είναι βασισμένο σε >14K ενώσεις από τη βάση δεδομένων PHYSPROP.

MSlogS: Το μοντέλο MSlogS προβλέπει τη διαλυτότητα στο νερό χρησιμοποιώντας την μέθοδο της ομαδοποίησης ουσιών και την σύγκρισή τους. Το μοντέλο είναι βασισμένο σε ενώσεις από τη βάση δεδομένων PHYSPROP.

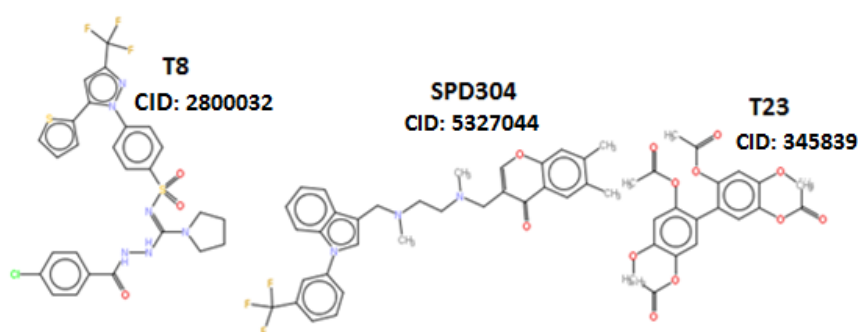
MSlogVP: Το μοντέλο προβλέπει την πίεση ατμών χρησιμοποιώντας την μέθοδο της ομαδοποίησης ουσιών και την σύγκρισή τους (read across predicting modelling). Το μοντέλο βασίζεται σε >2000 ενώσεις από τη βάση δεδομένων EPA.

Mutagenicity: Το μοντέλο προβλέπει, το δυναμικό μεταλλαξιογένεσης Ames, δηλαδή την ικανότητα μιας χημικής ουσίας να προκαλεί σημειακές μεταλλάξεις στο DNA του βακτηρίου *Salmonella typhimurium*. Το μοντέλο βασίζεται σε >4000 ενώσεις από το Σύνολο Δεδομένων Benchmark (TU Berlin).

Ενότητα 3. Μελέτη περίπτωσης - Μεθοδολογία

3.1 Εισαγωγή

Για την συγκεκριμένη μελέτη χρησιμοποιήθηκαν τρεις αναστολείς του παράγοντα TNF (T8, SPD304 και T23). Ο SPD304 αποτελεί γνωστό αναστολέα και οι T8 και T23 έχουν προταθεί πρόσφατα ως αναστολείς (Melagraki et al., 2017).



Εικόνα 3.1. Χημικές δομές των ενώσεων SPD304 και T8, T23 (Ανάκτηση δομών μέσω του Enalos Suite)

Με βάση αυτές τις δομές, πραγματοποιήθηκε αναζήτηση νέων ενώσεων με στόχο την ισχυρή αναστολή της λειτουργίας του παράγοντα νέκρωσης όγκου (TNF). Αρχικά, όλες οι ενέργειες ανάκτησης δεδομένων για τις τρεις αυτές ενώσεις πραγματοποιήθηκαν στο Enalos Suite ενώ έπειτα αναζητήθηκαν όλες οι παρόμοιες ενώσεις με ομοιότητα Tanimoto $\geq 85\%$. Για αυτές τις δομές αξιολογήθηκε *in silico* η αναστολή του TNF και η τοξικότητα τους μέσω των διαθέσιμων διαδικτυακών εργαλείων Enalos Cloud και MouseTox. Τέλος μέσω του Enalos Suite αξιολογήθηκε η εμπορική διαθεσιμότητα των πιο υποσχόμενων ενώσεων καθώς και αν υπάρχει κάλυψή τους με διπλώματα ευρεσιτεχνίας. Στα ερχόμενα κεφάλαια, παρατίθενται βασικές πληροφορίες για τον παράγοντα TNF και των γνωστών αναστολέων του και στη συνέχεια θα αναλυθούν τα αποτελέσματα της συγκεκριμένης μελέτης για εύρεση νέων αναστολέων.

3.2 Παράγοντας νέκρωσης όγκων (Tumor Necrosis Factor, TNF)

Ο Παράγοντας Νέκρωσης Όγκων (Tumor Necrosis Factor) είναι μια προφλεγμονώδης κυτοκίνη η οποία διαδραματίζει κεντρικό ρόλο στην φλεγμονή, την ανάπτυξη του ανοσοποιητικού συστήματος, την απόπτωση και το μεταβολισμό των λιπιδίων. Επίσης, παίζει βασικό ρόλο στον κυτταρικό πολλαπλασιασμό, τον κυτταρικό μεταβολισμό, τη διαφοροποίηση των κυττάρων και την απόπτωση. Η υπερπαραγωγή ή η απορυθμισμένη έκλυση TNF σχετίζεται με πολλές παθολογικές καταστάσεις, όπως η ρευματοειδής αρθρίτιδα (RA), η ψωριασική αρθρίτιδα, η νόσος του Crohn και η σκλήρυνση κατά πλάκας. Ο TNF παράγεται κυρίως από λεμφοκύτταρα και ενεργοποιημένα μονοκύτταρα/μακροφάγα και σε μικρότερο βαθμό από άλλα κύτταρα (Olesen et al., 2016). Παράγεται ως διαμεμβρανική πρωτεΐνη (tmTNF), η οποία διασπάται πρωτεολυτικά από το ένζυμο μετατροπής του παράγοντα νέκρωσης όγκου (TACE) στη διαλυτή του μορφή (sTNF). Οι sTNF και tmTNF συνδέονται με δύο διαφορετικούς υποδοχείς, τον TNFR1 (τύπος 1 υποδοχέα TNF) και τον TNFR2 (υποδοχέας TNF τύπου 2). Κατά την σύνδεση του TNF, οι υποδοχείς σχηματίζουν τριμερή και αυτό έχει ως αποτέλεσμα την ενεργοποίηση πολλών μονοπατιών που περιλαμβάνουν ενεργοποιημένες από μιτογόνα πρωτεϊνικές κινάσες και τον πυρηνικό παράγοντα-κΒ (NF-κΒ) παράγοντας την έκφραση πολλαπλών προφλεγμονωδών και αντι-αποπτωτικών γονιδίων (Chu, 2013). Έχει αποδειχθεί ότι ο tmTNF και ο sTNF διαφέρουν ως προς τις φυσιολογικές τους λειτουργίες και οι αναστολές που τα στοχεύουν μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα. Για τον έλεγχο των δυσμενών λειτουργιών του TNF, οι μέχρι τώρα προσπάθειες έχουν επικεντρωθεί στην παρεμπόδιση της δέσμευσης του

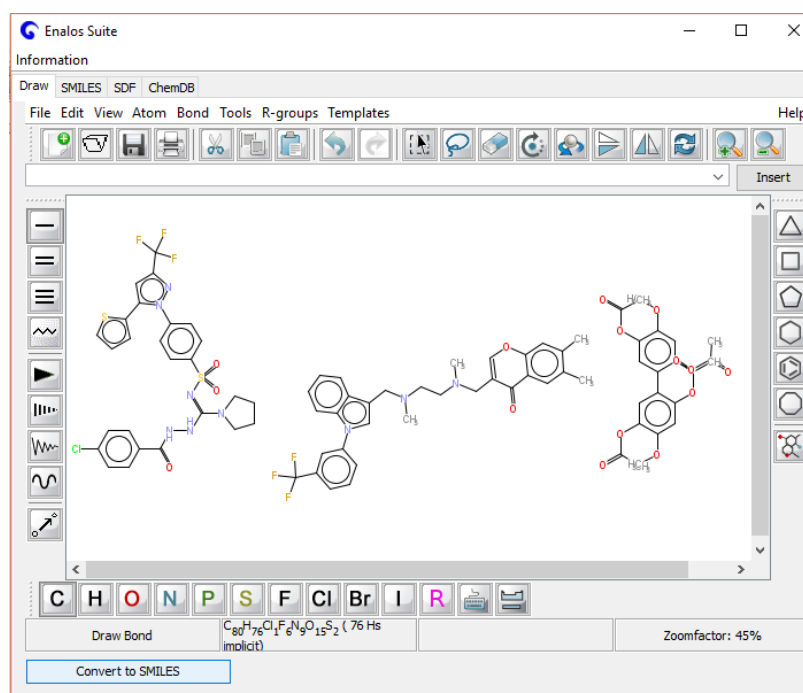
TNF στους υποδοχείς του οδηγώντας στην κινητοποίηση του ενδιαφέροντος της βιομηχανίας για παραγωγή anti-TNF ενώσεων (Melagraki et al., 2017).

3.3 Αναστολείς TNF

Μέχρι στιγμής, τέσσερα συνθετικά αντισώματα έχουν εγκριθεί από τον FDA ως αποτελεσματικοί αναστολείς του TNF και αυτοί είναι το infliximab, το adalimumab, το certolizumab pegol, το golimumab καθώς και το Fc- γ 75 etanercept. Ωστόσο, η χρήση τους δεν περιόρισε τις συνεχώς αυξανόμενες ερευνητικές προσπάθειες για την ανάπτυξη νέων αντι-TNF φαρμάκων, κυρίως λόγω των ανεπιθύμητων παρενέργειών τους (π.χ. υψηλός κίνδυνος ηπατίτιδας B και φυματίωσης, αλλά και λόγω της ανεπαρκούς κλινικής ανταπόκρισης, της ανάγκης για ενδοφλέβια χορήγηση τους και το υψηλό κόστος (Melagraki et al., 2017).¹

3.4 Ανάκτηση Δεδομένων & έλεγχος ομοιότητας

Αρχικά μέσω του Enalos Suite έγινε ο σχεδιασμός των τριών ενώσεων με αποτέλεσμα να γίνει ανάκτηση των SMILES των τριών ενώσεων, όπως φαίνεται στην Εικόνα 3.2.



Εικόνα 3.2. Σχεδιασμός των τριών αναστολέων στο Enalos Suite

Έπειτα, μέσω της λειτουργίας PubChem στο Enalos Suite πραγματοποιήθηκε αναζήτηση στη βάση δεδομένων PubChem για τις τρεις αυτές ενώσεις και ανακτήθηκαν με ένα αίτημα οι ακόλουθες πληροφορίες: το PubChem CID (Compound ID), το όνομα IUPAC, τα InChI & InChIkey, ο μοριακός τύπος, το μοριακό βάρος, τα SMILES και το URL στο PubChem. Για την ανάκτηση των πληροφοριών χρησιμοποιήθηκαν τα SMILES των ενώσεων.

Πίνακας 3.1. Αποτελέσματα συνάρτησης PubChem Information του Enalos Suite για τους τρεις αναστολείς

Input CID	InChI	InChIKey	Molecular Formula	Canonical SMILES	PubChem URL
2800032	3137(H3234)	CRVWIGCXKLMHBL-UHFFFAOYSA-N	C ₂₆ H ₂₂ O ₃ N ₃ S ₂	C1CCN(C1)C(=NS(=O)(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	https://pubchem.ncbi.nlm.nih.gov/compound/2800032
5327044	1S/C32H32F3N3O2/c1-21-14-28-30(15-22(21)2)40-20-24(31(28)39)18-37(4)13-12-36(3)17-23-19-38(29-11-6-5-10-27(23)29)26-9-7-8-25(16-26)32(3334)35/h5-1114-1619-20H12-1317-18H21-4H3	JZNXLPJRFECJ-UHFFFAOYSA-N	C ₃₂ H ₃₂ F ₃ N ₃ O ₂	CC1=CC2=C(C=C1C)OC=C(C2=O)CN(C)CCN(C)C3=CN(C4=CC=CC=C4)C5=CC=CC(=C5)C(F)(F)F	https://pubchem.ncbi.nlm.nih.gov/compound/5327044
345839	InChI=1S/C2	QBGAHSSWDUWJM	C ₂₂ H ₂₂ O ₁₀	CC(=O)OC1=CC(=C(C=C1)C2=CC(=C(C=C2)OC(=O)C)OC(=O)C)OC(=O)C)OC	https://pubchem.ncbi.nlm.nih.gov/compound/345839

2H22	X-			39
O10/c	UHFFFAO			
1-	YSA-N			
11(23)				
29-17-				
9-				
19(27-				
5)21(3				
1-				
13(3)2				
5)7-				
15(17)				
16-8-				
22(32-				
14(4)2				
6)20(2				
8-				
6)10-				
18(16)				
30-				
12(2)2				
4/h7-				
10H1-				
6H3				

Αφού ανακτήθηκαν οι πληροφορίες από το PubChem σε αρχείο xlxs στη συνέχεια μέσω της λειτουργίας Similarity για τις τρεις υπό μελέτη ενώσεις, πραγματοποιήθηκε αναζήτηση σε ολόκληρη τη βάση δεδομένων PubChem για παρόμοιες ενώσεις (Tanimoto Similarity $\geq 85\%$) και λήφθηκαν οι ακόλουθες πληροφορίες με ένα αίτημα: PubChem CID (Compound ID), μοριακός τύπος, μοριακό βάρος και αριθμός περιστρεφόμενων δεσμών.

	A	B	C	D	E
1	Input_SMILES	CID	MolecularFormula	MolecularWeight	RotatableBondC
2	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	131801155	"C27H25F3N6O3S2"	602.7	8
3	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	145976275	"C27H25F3N6O3S2"	602.7	8
4	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583147	"C26H23ClF3N7O3S2"	638.1	8
5	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	152004422	"C21H16ClN5O3S2"	486.0	6
6	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	5712878	"C26H22ClF3N6O3S2"	623.1	8
7	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800029	"C26H22ClF3N6O2S2"	607.1	8
8	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800031	"C25H22ClF3N6O4S3"	659.1	9
9	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800032	"C26H22ClF3N6O3S2"	623.1	8
10	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800034	"C27H23F4N7O4S2"	649.6	8
11	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800037	"C26H23ClF3N7O3S2"	638.1	8
12	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	2800040	"C28H26ClF3N6O2S2"	635.1	8
13	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583143	"C26H22ClF3N6O2S2"	607.1	8
14	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583144	"C25H22ClF3N6O4S3"	659.1	9
15	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583145	"C26H22ClF3N6O3S2"	623.1	8
16	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583146	"C27H23F4N7O4S2"	649.6	8
17	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	9583149	"C28H26ClF3N6O2S2"	635.1	8
18	C1CCN(C1)C(=NS(=O))(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl	18832609	"C22H14ClF3N4O2S"	490.9	4

Εικόνα 3.3. Εξαγόμενο αρχείο των παρόμοιων ενώσεων μέσω της συνάρτησης Similar Compounds του Enalos Suite

3.5 Υπολογισμός αναστολής & τοξικότητας

Για την εκτίμηση της αναστολής της παραγωγής του παράγοντα νέκρωσης όγκου καθώς και της τοξικότητας που εμφανίζουν οι ενώσεις που αναφέρθηκαν παραπάνω χρησιμοποιήθηκαν δύο μοντέλα του Enalos Cloud: το μοντέλο Enalos TNF και το μοντέλο MouseTox τα οποία είναι προσβάσιμα διαδικτυακά στις αντίστοιχες διευθύνσεις: <http://enalos.insilicotox.com/TNFPubChem/>,

<http://enalos.insilicotox.com/MouseTox/>. Μέσω του Enalos Suite χρησιμοποιώντας τους CID κωδικούς, ανακτήθηκαν τα SMILES για όλες τις παρόμοιες ενώσεις των τριών αναστολέων προκειμένου να χρησιμοποιηθούν ως πληροφορία εισόδου στις πλατφόρμες Enalos TNF και MouseTox.

3.6 Εμπορική διαθεσιμότητα & κάλυψη διπλωμάτων ευρεσιτεχνίας

Μετά την ανάκτηση των δεδομένων για τις παρόμοιες ενώσεις των τριών υπό μελέτη αναστολέων και την διαλογή τους με βάση τα μοντέλα MouseTox και TNF του Enalos Cloud, πραγματοποιήθηκε αξιολόγηση των πλέον υποσχόμενων ενώσεων, δηλαδή αυτών για τις οποίες υπήρχε αξιόπιστη πρόβλεψη και εντός του χώρου εφαρμογής του μοντέλου, ότι είναι δραστικές και μη τοξικές. Στην συγκεκριμένη περίπτωση, οι πιο πολλά υποσχόμενες ενώσεις είναι αυτές των αναστολέων SPD304 (Tanimoto \geq 85%) και T23 (Tanimoto \geq 90%) για τις οποίες ελέγχθηκε η εμπορική διαθεσιμότητα και η κάλυψη διπλωμάτων ευρεσιτεχνίας. Η διαδικασία αυτή πραγματοποιήθηκε μέσω του μοντέλου PubChem Vendor και PubChem Patent του Enalos Suite αντίστοιχα. Μέσω της συνάρτησης Vendor εξάγονται πληροφορίες από το PubChem σχετικά με το PubChem SID, το Compound URL, τον Προμηθευτή και την Κατηγορία του Προμηθευτή. Τα δεδομένα για την εμπορική διαθεσιμότητα λήφθηκαν σε αρχεία xlsx μέσω των κωδικών CID των ενώσεων που αποτελούσαν πληροφορία εισόδου στο Enalos Suite.

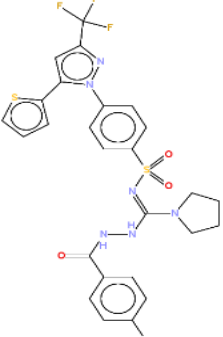
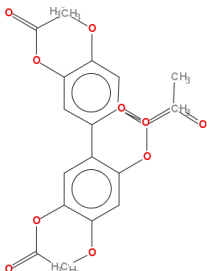
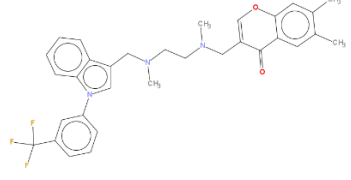
Ενότητα 4. Αποτελέσματα

4.1 Αποτελέσματα ελέγχου ομοιότητας μέσω του συντελεστή Tanimoto

Για Tanimoto Similarity \geq 85% προέκυψαν 19 παρόμοιες ενώσεις για τον αναστολέα T8, 740 παρόμοιες ενώσεις για τον αναστολέα SPD304 και 11503 παρόμοιες ενώσεις για τον αναστολέα T23. Επειδή ο αναστολέας T23 εμφάνισε ομοιότητα για

πολλές ενώσεις σε τιμές Tanimoto $\geq 85\%$, πραγματοποιήθηκε ξανά αναζήτηση παρόμοιων ενώσεων σε τιμές Tanimoto $\geq 90\%$ και τα αποτελέσματα περιορίστηκαν σε 1462 παρόμοιες ενώσεις με τις οποίες και συνεχίστηκε η μελέτη.

Πίνακας 4.1. Αριθμός παρόμοιων ενώσεων από την συνάρτηση PubChem Similarity του Enalos Suite, με Tanimoto $\geq 85\%$ για τα T8 και SPD304 και Tanimoto $\geq 90\%$ για το T23.

Ένωση	SMILES	Παρόμοιες ενώσεις
 <p>T8</p>	<chem>C1CCN(C1)C(=NS(=O)(=O)C2=CC=C(C=C2)N3C(=CC(=N3)C(F)(F)F)C4=CC=CS4)NNC(=O)C5=CC=C(C=C5)Cl</chem>	19
 <p>T23</p>	<chem>C(C=C1C2=CC(=C(C=C2)OC(=O)C)OC(=O)C)OC(=O)C)OC(=O)C)OC</chem>	1462
 <p>SPD304</p>	<chem>CC1=CC2=C(C=C1C)OC=C(C2=O)CN(C)CCN(C)CC3=CN(C4=CC=CC=C4)C5=CC=CC(=C5)C(F)(F)F</chem>	740

4.2 Αποτελέσματα αναστολής & τοξικότητας

Από τις 19 παρόμοιες ενώσεις του αναστολέα T8 καμία δεν εμφάνισε δραστηριότητα ενάντια στον TNF παράγοντα ενώ η πρόβλεψη με βάση το μοντέλο MouseTox υπέδειξε 2 μη κυτταροτοξικές ενώσεις με αξιοπιστία στο μοντέλο (CID: 131801155, 145976275). Ο αναστολέας T23 εμφάνισε 671 παρόμοιες ενώσεις οι οποίες παρουσιάζουν αναστολή έναντι του παράγοντα TNF και 791 μη δραστικές. Από τις 671 ενώσεις όμως οι 129 ενώσεις ξεχώρισαν ως προς την αξιοπιστία τους με βάση το πεδίο εφαρμογής του μοντέλου TNF, δηλαδή με βάση την κατηγορία των

δεδομένων που έχουν χρησιμοποιηθεί για την δημιουργία του μοντέλου. Σύμφωνα με το μοντέλο MouseTox από τις 1462 ενώσεις οι 1386 δεν εμφανίζουν κυτταροτοξικότητα σε σχέση με τις υπόλοιπες 76 ενώσεις ενώ παράλληλα τα αποτελέσματα αυτά έδειξαν ότι είναι αναξιόπιστα με βάση τον τομέα εφαρμογής του μοντέλου πρόβλεψης. Σύμφωνα και με τα δύο μοντέλα οι ενώσεις που είναι δραστικές έναντι του TNF και δεν εμφανίζουν κυτταροτοξικότητα είναι 129, όμως δεν είναι αξιόπιστες ως προς το μοντέλο MouseTox. Για τις 740 παρόμοιες ενώσεις του αναστολέα SPD304 βρέθηκαν 466 δραστικές ενώσεις και 274 μη δραστικές έναντι στον TNF παράγοντα. Ενώ σύμφωνα με το μοντέλο MouseTox 274 ενώσεις έδειξαν μη κυτταροτοξικό χαρακτήρα και 466 ήταν κυτταροτοξικές. Συνδυαστικά, υπολογίστηκαν σύμφωνα και με τα δύο μοντέλα ότι οι περισσότερα υποσχόμενες ενώσεις είναι 50. Συγκεκριμένα, επιλέχθηκαν οι ενώσεις οι οποίες όπως φαίνεται εμφανίζουν δραστηριότητα ενάντια στον TNF παράγοντα ενώ παράλληλα δεν είναι κυτταροτοξικές σύμφωνα με το μοντέλο MouseTox. Παράλληλα, είναι σημαντικό να αναφερθεί ότι οι 50 αυτές ενώσεις ξεχώρισαν και λόγω της αξιοπιστίας τους με βάση τον τομέα εφαρμογής των δύο μοντέλων.

Πίνακας 4.2. Αποτελέσματα αναστολής του TNF και τοξικότητας των παρόμοιων ενώσεων για τους τρεις αναστολείς βασισμένα στις πλατφόρμες του Enalos Cloud (<http://www.enaloscloud.novamechanics.com/EnalosWebApps/TNF/>, <http://www.enaloscloud.novamechanics.com/EnalosWebApps/MouseTox/>)

TNF inhibitors	Similar compounds	Enalos TNF Extraction Platform			MouseTox	
		Active	Inactive	Non-cytotoxic & active compounds (reliable predictions)	Active	Inactive
T8 (CID:2800032)	19	19	0	0	6	13
Spd304 (CID: 5327044)	740	274	466	50	466	274
T23 (CID: 343859)	1462	791	671	0	76	1386

4.3 Αποτελέσματα εμπορικής διαθεσιμότητας και διπλωμάτων ευρεσιτεχνίας

Στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα της εμπορικής διαθεσιμότητας για τις 129 παρόμοιες ενώσεις για τον αναστολέα T23 και για τις 50 παρόμοιες ενώσεις του αναστολέα SPD304 και τα δεδομένα για την κάλυψη διπλωμάτων ευρεσιτεχνίας.

Πίνακας 4.3. Αποτελέσματα της εμπορικής διαθεσιμότητας των περισσότερα υποσχόμενων ενώσεων για τον αναστολέα T23. Δεδομένα τα οποία λήφθηκαν από την συνάρτηση PubChem Vendor του Enalos Suite

Input CID	SID	URL	Supplier	Category
38347541	69924576	http://zinc.docking.org/substances/ZINC31156147/	"ZINC"	"Chemical Vendors"
38347547	69924581	http://zinc.docking.org/substances/ZINC31156152/	"ZINC"	"Chemical Vendors"
6540177	114617008		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
38347535	69924571	http://zinc.docking.org/substances/ZINC31156142/	"ZINC"	"Chemical Vendors"
6540166	114616986		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
5386548	113950084		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
5386548	117554487		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
5386547	113950081		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
353128	104515038		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
353128	117575444		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
182878	104438173		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
182877	104438170		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
9932416	252949743	https://mcule.com/MCULE-2181940319/	"Mcule"	"Chemical Vendors"
9932416	290041632	http://online.aurorafinechemicals.com/info?ID=K06.124.181	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - - "Legacy Depositors"
9932416	366185067	http://www.ambinter.com/reference/10549645	"Ambinter"	"Chemical Vendors"
9932416	436967076	https://chem-space.com/CSSS00159279374	"Chem-Space.com Database"	"Chemical Vendors"

1325613	3527584	http://www.hit2lead.com/search.asp?db=SC&ids=6945902	"ChemBridge"	"Chemical Vendors"
1325613	88478953	https://www.molport.com/shop/molecule-link/MolPort-000-830-921	"MolPort"	"Chemical Vendors"
1325613	110286567		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
1325613	164244563	http://www.chembase.cn/molecule-188653.html	"Chembase.cn"	"Chemical Vendors" - - "Legacy Depositors"
1325613	168681108	https://mculc.com/MCULE-2816191565/	"Mculc"	"Chemical Vendors"
1325613	218947944	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=AKOS016390207	"AKos Consulting & Solutions"	"Chemical Vendors"
1325613	256051829	http://zinc.docking.org/substances/ZINC1156130/	"ZINC"	"Chemical Vendors"
1325613	290256235	http://online.aurorafinechemicals.com/info?ID=K00.218.770	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - - "Legacy Depositors"
1325613	309816621	http://online.aurorafinechemicals.com/info?ID=A17.588.904	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - - "Legacy Depositors"
1325613	313011178	http://www.request.vitasmlab.biz/index.php?option=com_search_stk&Itemid=22&stk=STL458595&utm_source=pubchem&utm_medium=p_search_link&utm_campaign=pubchem_search&utm_content=pubchem_slink	"Vitas-M Laboratory"	"Chemical Vendors"
1325613	329472270	http://chemistryondemand.com:8080/eShop/search_results.jsp?s_type=txt&idnumber=Y040-1174	"ChemDiv"	"Chemical Vendors"
1325613	346734962	https://labnetwork.com/frontend-app/p/#!/moleculdetails/LN00348053	"LabNetwork -- a WuXi AppTec Company"	"Chemical Vendors"
1325613	369038642	http://www.ambinter.com/reference/1847666	"Ambinter"	"Chemical Vendors"
1325613	384125759		"Innovapharm"	"Chemical Vendors"
1325613	385623048		"Eximed Laboratory"	"Chemical Vendors"
1325613	437722184	https://chem-space.com/CSSS00160297750	"Chem-Space.com Database"	"Chemical Vendors"
1325597	3527549	http://www.hit2lead.com/search.asp?db=SC&ids=6945813	"ChemBridge"	"Chemical Vendors"
1325597	89271818	https://www.molport.com/shop/molecule-link/MolPort-002-228-886	"MolPort"	"Chemical Vendors"
1325597	110286519		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"

1325597	164245235	http://www.chembase.cn/molecule-189325.html	"Chembase.cn"	"Chemical Vendors" - - "Legacy Depositors"
1325597	168681178	https://mcule.com/MCULE-9210674861/	"Mcule"	"Chemical Vendors"
1325597	218921397	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=A KOS016363659	"AKos Consulting & Solutions"	"Chemical Vendors"
1325597	256051820	http://zinc.docking.org/substances/ZINC1156114/	"ZINC"	"Chemical Vendors"
1325597	290256220	http://online.aurorafinechemicals.com/info?ID=K00.219.256	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - - "Legacy Depositors"
1325597	309816597	http://online.aurorafinechemicals.com/info?ID=A17.588.880	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - - "Legacy Depositors"
1325597	313010076	http://www.request.vitasmlab.biz/index.php?option=com_search_stk&Itemid=22&stk=STL457493&utm_source=pubchem&utm_medium=p_search_link&utm_campaign=pubchem_search&utm_content=pubchem_slink	"Vitas-M Laboratory"	"Chemical Vendors"
1325597	329471700	http://chemistryondemand.com:8080/eShop/search_results.jsp?stype=txt&idnumber=Y040-0587	"ChemDiv"	"Chemical Vendors"
1325597	346757508	https://labnetwork.com/frontend-app/p/#!/moleculdetails/LN00370655	"LabNetwork -- a WuXi AppTec Company"	"Chemical Vendors"
1325597	368995197	http://www.ambinter.com/reference/18476015	"Ambinter"	"Chemical Vendors"
1325597	384125763		"Innovapharm"	"Chemical Vendors"
1325597	385621796		"Eximed Laboratory"	"Chemical Vendors"
1325597	437722061	https://chem-space.com/CSSS00160297617	"Chem-Space.com Database"	"Chemical Vendors"
356770	104524638		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
1325363	3526900	http://www.hit2lead.com/search.asp?db=SC&ids=6944083	"ChemBridge"	"Chemical Vendors"
1325363	89271594	https://www.molport.com/shop/molecule-link/MolPort-002-228-662	"MolPort"	"Chemical Vendors"
1325363	110285869		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
1325363	168682947	https://mcule.com/MCULE-6746930101/	"Mcule"	"Chemical Vendors"
1325363	218950049	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=A KOS016392312	"AKos Consulting & Solutions"	"Chemical Vendors"

1325363	256051679	http://zinc.docking.org/substances/ZINC_1155843/	"ZINC"	"Chemical Vendors"
1325363	290255936	http://online.aurorafinechemicals.com/info?ID=K01.844.729	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - "Legacy Depositors"
1325363	309816119	http://online.aurorafinechemicals.com/info?ID=A17.588.402	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - "Legacy Depositors"
1325363	313011145	http://www.request.vitasmlab.com/index.php?option=com_search_stk&Itemid=22&stk=STL458562&utm_source=pubchem&utm_medium=p_search_link&utm_campaign=pubchem_search&utm_content=pubchem_slink	"Vitas-M Laboratory"	"Chemical Vendors"
1325363	329472238	http://chemistryondemand.com:8080/eshop/search_results.jsp?s_type=txt&idnumber=Y040-1141	"ChemDiv"	"Chemical Vendors"
1325363	368407434	http://www.ambinter.com/reference/17975173	"Ambinter"	"Chemical Vendors"
1325363	384125876		"Innovapharm"	"Chemical Vendors"
1325363	385623008		"Eximed Laboratory"	"Chemical Vendors"
1325363	437721889	"https://chem-space.com/CSSS00160297425"	"Chem-Space.com Database"	"Chemical Vendors"
1940249	3526705	http://www.hit2lead.com/search.asp?db=SC&ids=6943514	"ChemBridge"	"Chemical Vendors"
1940249	87490668	http://www.request.vitasmlab.biz/index.php?option=com_search_stk&Itemid=22&stk=STK829417&utm_source=pubchem&utm_medium=p_search_link&utm_campaign=pubchem_search&utm_content=pubchem_slink	"Vitas-M Laboratory"	"Chemical Vendors"
1940249	89271522	https://www.molport.com/shop/molecule-link/MolPort-002-228-590	"MolPort"	"Chemical Vendors"
1940249	110752524		"ABI Chem"	"Chemical Vendors" - "Legacy Depositors"
1940249	131896525	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=AKOS005613898	"AKos Consulting & Solutions"	"Chemical Vendors"
1940249	169649288	https://mcule.com/MCULE-9665433560/	"Mcule"	"Chemical Vendors"
1940249	256459713	http://zinc.docking.org/substances/ZINC_2325358/	"ZINC"	"Chemical Vendors"
1940249	290255849	http://online.aurorafinechemicals.com/info?ID=K00.312.399	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - "Legacy Depositors"
1940249	309815980	http://online.aurorafinechemicals.com/info?ID=A17.588.263	"Aurora Fine Chemicals LLC"	"Chemical Vendors" - "Legacy Depositors"

				Depositors"
1940249	329477771	http://chemistryondemand.com:8080/eShop/search_results.jsp?s_type=txt&idnumber=Y040-7813	"ChemDiv"	"Chemical Vendors"
1940249	347079642	https://labnetwork.com/frontend-app/p/#!/moleculedetails/LN00730704	"LabNetwork -- a WuXi AppTec Company"	"Chemical Vendors"
1940249	369028875	http://www.ambinter.com/reference/18477847	"Ambinter"	"Chemical Vendors"
1940249	384125764		"Innovapharm"	"Chemical Vendors"
1940249	385624872		"Eximed Laboratory"	"Chemical Vendors"
1940249	437721784	https://chem-space.com/CSSS00160297309	"Chem-Space.com Database"	"Chemical Vendors"
5373894	114177312		"ABI Chem"	"Chemical Vendors" - - "Legacy Depositors"
38347552	69924585	http://zinc.docking.org/substances/ZINC31156156/	"ZINC"	"Chemical Vendors"
92150327	256223786	http://zinc.docking.org/substances/ZINC1607539/	"ZINC"	"Chemical Vendors"
92278114	257833461	http://zinc.docking.org/substances/ZINC5758747/	"ZINC"	"Chemical Vendors"
158529935	Not found	Not found	Not found	Not found
159868909	Not found	Not found	Not found	Not found

Πίνακας 4.4. Αποτελέσματα της εμπορικής διαθεσιμότητας των περισσότερα υποσχόμενων ενώσεων για τον αναστολέα SPD304. Δεδομένα τα οποία λήφθηκαν από την συνάρτηση PubChem Vendor του Enalos Suite.

Input_CI	D	SID	URL	Supplier	Category
57507911	263942239		http://zinc.docking.org/substances/ZINC64859686/	«ZINC»	«Chemical Vendors»
57507911	381992477		https://www.aablocks.com/prod/1345966-76-8	«AA BLOCKS»	«Chemical Vendors»
57507911	442034803		https://www.thebiotek.com/product/others/bt-508856	«THE BioTek»	«Chemical Vendors»
57507911	444131059		https://www.a2bchem.com/1345966-76-8.html	«A2B Chem»	«Chemical Vendors»
57507911	444687307		https://www.smolecule.com/products/s1781036	«Smolecule»	«Chemical Vendors»
57507911	445756753		https://www.benchchem.com/product/b583761	«BenchChem»	«Chemical Vendors»
45224579	92616629		https://www.molport.com/shop/molecule-link/MolPort-005-108-335	«MolPort»	«Chemical Vendors»
45224579	116223985		http://www.hit2lead.com/search.asp?db=SC&ids=59961424	«ChemBridge»	«Chemical Vendors»

45224579	16785120 0	https://mcule.com/MCULE-6711673891/	«Mcule»	«Chemical Vendors»
45224579	27769912 6	http://online.aurorafinechemicals.com/info?ID=K08.084.946	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»
45224579	36674310 5	http://www.ambinter.com/reference/11090043	«Ambinter»	«Chemical Vendors»
45224579	43749548 6	https://chem-space.com/CSSS00160028751	«Chem-Space.com Database»	«Chemical Vendors»
62505419	14793171 3	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=AKOS011795294	«AKos Consulting & Solutions»	«Chemical Vendors»
62505419	29454941 2	http://online.aurorafinechemicals.com/info?ID=A06.282.549	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»
62505420	14793171 4	http://akoscompounds.de/catalogue/akosamplesretrieval.php?IDNUMBERS=AKOS011795295	«Akos Consulting & Solutions»	«Chemical Vendors»
62505420	29454941 3	http://online.aurorafinechemicals.com/info?ID=A06.282.550	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»
82039854	29253852 1	http://online.aurorafinechemicals.com/info?ID=A01.339.429	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»
82039854	36926867 7	http://www.ambinter.com/reference/17137054	«Ambinter»	«Chemical Vendors»
11019326 2	28901493 3	http://online.aurorafinechemicals.com/info?ID=K16.455.631	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»
11654708 0	30882012 5	http://online.aurorafinechemicals.com/info?ID=A15.839.386	«Aurora Fine Chemicals LLC»	«Chemical Vendors» - - «Legacy Depositors»

				Depositors »
12620472 9	33255399 9	http://zinc.docking.org/substances/ZINC409198503/	«ZINC»	«Chemical Vendors»
15590388 6	44117851 7	http://www.bldpharm.com/products/P001281867.html	«BLD Pharm»	«Chemical Vendors»
15590536 5	44119991 3	http://www.bldpharm.com/products/P001281870.html	«BLD Pharm»	"Chemical Vendors"

Πίνακας 4.5. Αποτελέσματα πληροφοριών κάλυψης διπλωμάτων ευρεσιτεχνίας τα οποία λήφθηκαν μέσω της συνάρτησης PubChem Patent Enalos Suite για τον αναστολέα T23. Στον πίνακα απεικονίζονται μόνο τα CID των ενώσεων που δεν εμφάνισαν κάλυψη διπλώματος ευρεσιτεχνίας

Input_ CID of compounds			Patent ID
13577736	1325363	101023179	Not found
46879082	1940249	101086582	Not found
38347541	5373894	101462861	Not found
38347547	10525267	101607119	Not found
6540177	10621129	101636943	Not found
46865220	10623672	101639956	Not found
46228397	10645555	101686717	Not found
44207634	10709042	101686718	Not found
38347535	10715172	101686719	Not found
6540166	10813673	101686720	Not found
5386548	10813975	101686779	Not found
5386547	10835060	101686780	Not found
353128	11122570	101760061	Not found
182878	11143124	101915259	Not found
182877	11339082	101915260	Not found
142737661	11797631	101996127	Not found
141049106	12063600	101996128	Not found
140297603	15483076	101996304	Not found
140218383	21160693	102093386	Not found
140172294	21676235	102232463	Not found
139815734	23259404	102311138	Not found
135012610	24754564	132521421	Not found
142737661	25074382	132534870	Not found
129846666	38347552	132556726	Not found
86061031	44474760	155519283	Not found
46887466	50909264	155524404	Not found
46887465	71733613	101996128	Not found
9932416	92150327	101996304	Not found
1325613	92278114	102093386	Not found
1325597	101008002	102232463	Not found
356770	101023178	102311138	Not found
132521421	132556726	155524404	Not found
132534870	155519283	158529935	Not found

	159868909	Not found
--	-----------	-----------

Πίνακας 4.6. Αποτελέσματα πληροφοριών κάλυψης διπλωμάτων ευρεσιτεχνίας τα οποία λήφθηκαν μέσω της συνάρτησης PubChem Patent Enalos Suite για τις παρόμοιες ενώσεις του αναστολέα SPD304. Στον πίνακα απεικονίζονται μόνο τα CID των ενώσεων που δεν εμφάνισαν κάλυψη διπλώματος ευρεσιτεχνίας

Input_CID	Patent ID
57507910	Not found
134133236	Not found
144788739	Not found
43161959	Not found
45224579	Not found
53236306	Not found
53236308	Not found
53236312	Not found
53236551	Not found
62505419	Not found
62505420	Not found
82039854	Not found
110193262	Not found
116547080	Not found
126204729	Not found
155903886	Not found
155905365	Not found

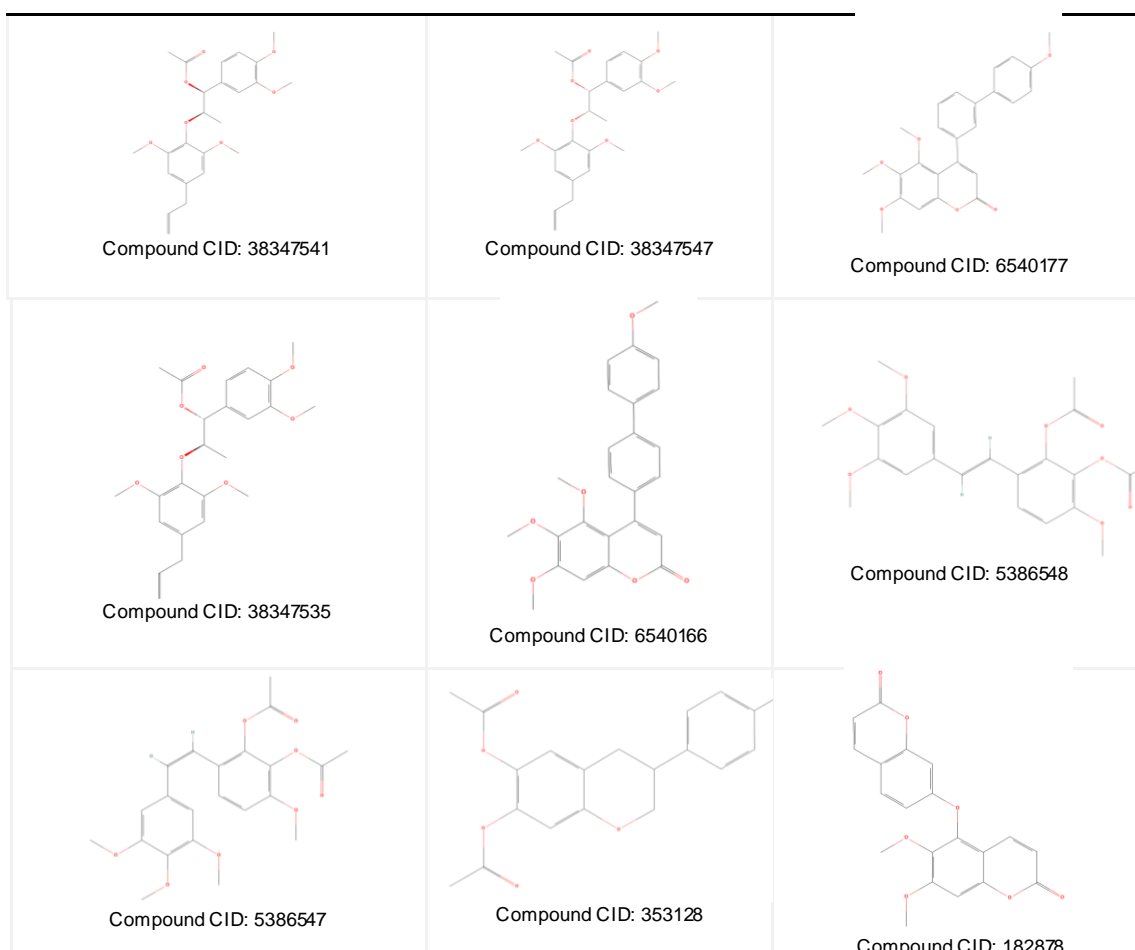
4.4 Σχολιασμός Αποτελεσμάτων

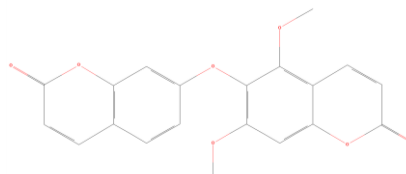
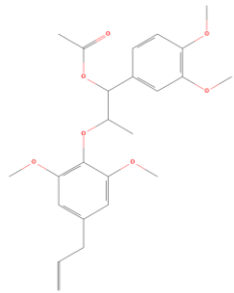
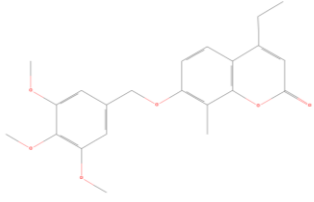
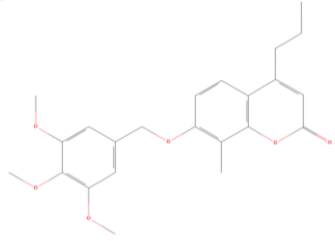
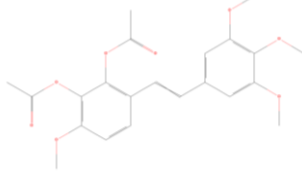
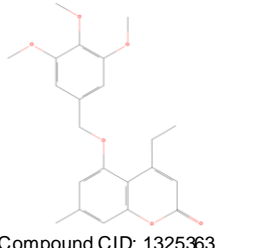
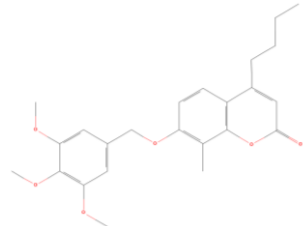
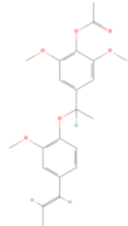
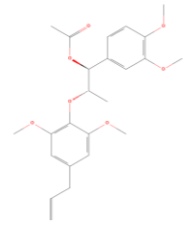
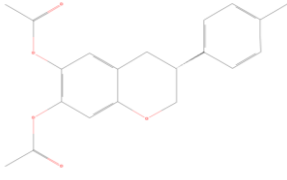
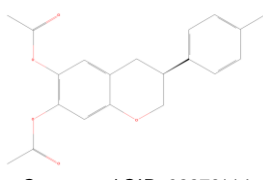
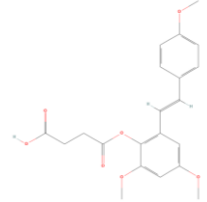
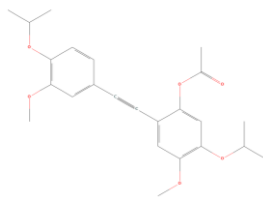
Στη συγκεκριμένη εργασία πραγματοποιήθηκε με την χρήση της συνάρτησης PubChem Similarity της πλατφόρμας Enalos Suite αναζήτηση όλων των ενώσεων που παρουσιάζουν ομοιότητα με τους αναστολείς T23, T8 και SPD304. Κατά την διαδικασία αναζήτησης αυτών των δεδομένων σκοπός ήταν να βρεθούν οι πιο συγγενικές ενώσεις με τους τρεις αυτούς αναστολείς που να εμφανίζουν αναστολή στον παράγοντα TNF και να μην είναι κυτταροτοξικές σε κύτταρα NIH/3T3. Η αναζήτηση βασίστηκε στον συντελεστή Tanimoto, τον πιο δημοφιλή δείκτη ομοιότητας για τη σύγκριση χημικών δομών. Με βάση τα αποτελέσματα τα οποία αναφέρθηκαν στην παράγραφο 4.1, ο αναστολέας T23 παρουσίασε τον μεγαλύτερο αριθμό παρόμοιων ενώσεων (1462 ενώσεις) για τιμές Tanimoto $\geq 90\%$ σε σχέση με τους άλλους δύο αναστολείς οι οποίοι για τιμές Tanimoto $\geq 80\%$ εμφάνισαν 19 ενώσεις ο T8 και 740 ενώσεις ο SPD304. Η αναζήτηση των παρόμοιων ενώσεων για τον αναστολέα T23 βασίστηκε σε τιμές Tanimoto $\geq 90\%$ καθώς σε μικρότερες τιμές ο συγκεκριμένος αναστολέας εμφάνιζε πολύ μεγάλο αριθμό παρόμοιων ενώσεων,

γεγονός που δυσκόλευε την επεξεργασία των δεδομένων λόγω του μεγάλου όγκου τους. Αφού συγκεντρώθηκαν όλες οι παρόμοιες ενώσεις και για τους τρεις αναστολείς πραγματοποιήθηκε λήψη των πληροφοριών τους μέσω της συνάρτησης PubChem Information για να ανακτηθούν τα SMILES τους. Τα SMILES χρησιμοποιήθηκαν ως πληροφορία εισόδου στα μοντέλα πρόβλεψης με τα οποία συνεχίστηκε η μελέτη. Μέσω του μοντέλου πρόβλεψης TNF ελέγχθηκε η ικανότητά τους να αναστέλλουν τον TNF (<http://www.enaloscloud.novamechanics.com/EnalosWebApps/TNF/>) ενώ μέσω του μοντέλου MouseTox (<http://www.enaloscloud.novamechanics.com/EnalosWebApps/MouseTox/>) ελέγχθηκε η κυτταροτοξικότητά τους σε κύτταρα NIH/3T3. Τα συγκεκριμένα μοντέλα όπως αναφέρθηκε στην Ενότητα 2 βασίζονται σε *in silico* προσεγγίσεις που συνδέουν τα δομικά χαρακτηριστικά των μορίων με τη δραστηριότητα ή το προφίλ της τοξικότητάς τους. Είναι αξιοσημείωτο να αναφερθεί ότι για την δημιουργία των συγκεκριμένων μοντέλων έχουν χρησιμοποιηθεί δεδομένα τα οποία αναφέρονται αποκλειστικά σε ενώσεις που αφορούν την αναστολή του παράγοντα TNF και την κυτταροτοξικότητα σε κύτταρα NIH/3T3. Επομένως, οι προβλέψεις για όλες τις παρόμοιες ενώσεις υπολογίστηκαν με βάση τα συγκεκριμένα πεδία εφαρμογής των δύο μοντέλων. Στην συγκεκριμένη μελέτη βάσει του πεδίου εφαρμογής του μοντέλου TNF ο αναστολέας T8 δεν εμφάνισε καμία παρόμοια ένωση που να αναστέλλει τον παράγοντα TNF και για αυτό δεν μελετήθηκε μετέπειτα και η κυτταροτοξικότητά τους μέσω του μοντέλου MouseTox. Ο αναστολέας T23 παρουσίασε 129 παρόμοιες ενώσεις που αναστέλλουν τον παράγοντα TNF με αξιόπιστες προβλέψεις εντός του πεδίου εφαρμογής του μοντέλου TNF οι οποίες παράλληλα εμφάνισαν κυτταροτοξικότητα στα κύτταρα NIH/3T3 χωρίς βέβαια η πρόβλεψη αυτή να είναι αξιόπιστη βάσει του μοντέλου (εκτός των ορίων του πεδίου εφαρμογής). Τέλος ο αναστολέας SPD304 παρουσίασε 50 παρόμοιες μη κυτταροτοξικές ενώσεις οι οποίες αναστέλλουν τον παράγοντα TNF (με αξιοπιστία και στα δύο μοντέλα). Σε αντιστοιχία με όσα αναφέρθηκαν παραπάνω οι παρόμοιες ενώσεις του SPD304 σύμφωνα με τα δύο μοντέλα πρόβλεψης είναι πολλά υποσχόμενες. Σχετικά με τις παρόμοιες ενώσεις του αναστολέα T23 θεωρείται αξιοσημείωτο ότι και οι 129 αναστέλλουν τον TNF παράγοντα και παράλληλα δεν εμφανίζουν κυτταροτοξικότητα σε κύτταρα NIH/3T3. Βέβαια, χρήζει περαιτέρω μελέτη το γεγονός ότι η πρόβλεψη κυτταροτοξικότητας τους δεν ήταν αξιόπιστη με βάση πεδίο εφαρμογής του μοντέλου MouseTox, γεγονός που μπορεί να οδηγήσει την επιστημονική κοινότητα σε περαιτέρω μελέτη αυτών των δομών με στόχο την βελτίωσή τους. Η μελέτη συνεχίστηκε με την αξιολόγηση της εμπορικής διαθεσιμότητας των 129 παρόμοιων ενώσεων για τον αναστολέα T23 και

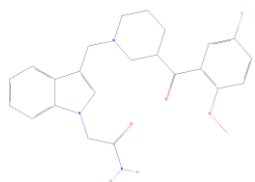
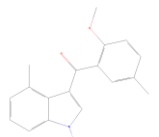
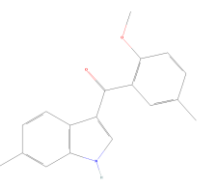
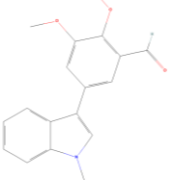
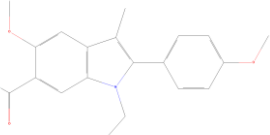

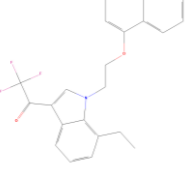
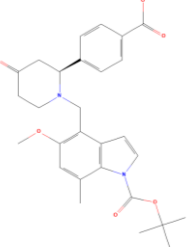
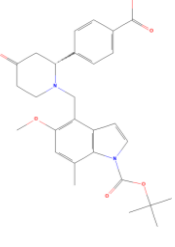
των 50 παρόμοιων ενώσεων για τον αναστολέα SPD304. Ο έλεγχος της εμπορικής διαθεσιμότητας αυτών των ενώσεων είναι ένα σημαντικό σημείο καθώς επιτρέπει την αγορά τους από το εμπόριο και την περαιτέρω δοκιμή τους σε *in vitro* πειράματα. Η διαδικασία ελέγχου πραγματοποιήθηκε μέσω της συνάρτησης PubChem Vendor στο Enalos Suite. Όπως φαίνεται στους Πίνακες 4.3 και 4.4 δεν παρουσιάζουν όλες οι ενώσεις εμπορική διαθεσιμότητα. Συγκεκριμένα, για τον T23 μόνο 22 από τις 219 ενώσεις είναι εμπορικά διαθέσιμες ενώ για τον SPD304 μόλις 9 ενώσεις. Παράλληλα, κάποιες από αυτές είναι διαθέσιμες σε περισσότερες από μια βάσεις δεδομένων. Οι 22 αυτές ενώσεις όπως φαίνεται και στους πίνακες της προηγούμενης ενότητας δεν εμφανίζουν κάλυψη διπλώματος ευρεσιτεχνίας το οποίο επιτρέπει την ελεύθερη έρευνα πάνω σε αυτές τις δομές και την περαιτέρω βελτιστοποίησή τους στοχεύοντας τελικά στην ανάπτυξη νέων θεραπειών για μια σειρά φλεγμονωδών και αυτοάνοσων ασθενειών. Αναλυτικά οι δομές τους παρουσιάζονται στον Πίνακα 4.7 και στον Πίνακα 4.8.

Πίνακας 4.7. Δομές των 22 πιο πολλά υποσχόμενων ενώσεων του αναστολέα T23 που είναι εμπορικά διαθέσιμες και δεν παρουσιάζουν κάλυψη διπλώματος ευρεσιτεχνίας (Δομές που λήφθηκαν από Enalos Suite)



 <p>Compound CID: 182877</p>	 <p>Compound CID: 9932416</p>	 <p>Compound CID: 1325613</p>
 <p>Compound CID: 1325597</p>	 <p>Compound CID: 356770</p>	 <p>Compound CID: 1325363</p>
 <p>Compound CID: 1940249</p>	 <p>Compound CID: 5373894</p>	 <p>Compound CID: 38347552</p>
 <p>Compound CID: 92150327</p>	 <p>Compound CID: 92278114</p>	 <p>Compound CID: 158529935</p>
 <p>Compound CID: 159868909</p>		

Πίνακας 4.8. Δομές των 9 πιο πολλά υποσχόμενων ενώσεων του αναστολέα SPD304 που είναι εμπορικά διαθέσιμες και δεν παρουσιάζουν κάλυψη διπλώματος ευρεσιτεχνίας (Δομές που λήφθηκαν από Enalos Suite)

 <p>CID Compound: 45224579</p>	 <p>CID Compound: 62505419</p>	 <p>CID Compound: 62505420</p>
 <p>CID Compound: 82039854</p>	 <p>Compound: 110193262</p>	 <p>CID Compound: 116547080</p>
 <p>CID Compound: 126204729</p>	 <p>CID Compound: 155903886</p>	 <p>CID Compound: 155905365</p>

Συμπεράσματα

Σκοπός της παρούσας εργασίας ήταν να μελετηθεί η χρησιμότητα των εργαλείων χρήση των εργαλείων χημειοπληροφορικής Enalos στην ανακάλυψη νέων βιοδραστικών ενώσεων με ικανοποιητικό προφίλ τοξικότητας και ιδιοτήτων. Αρχικά αναφέρθηκαν βασικές έννοιες του κλάδου της χημειοπληροφορικής όπως είναι η μοριακή απεικόνιση, η μοριακή ομοιότητα και οι μοριακοί δείκτες. Αναλύθηκαν ορολογίες όπως η εικονική διαλογή και η μηχανική μάθηση που έχουν συνεισφέρει σημαντικά στην ανακάλυψη φαρμάκων με την βοήθεια *in silico* προσεγγίσεων. Επίσης, αξιοσημείωτα θεωρούνται οι βάσεις χημικών δεδομένων καθώς και διάφορα εργαλεία λογισμικού στην χημειοπληροφορική μέσω των οποίων έχει καταστεί πιο γρήγορη και πιο εύκολη η επεξεργασία του μεγάλου όγκου δεδομένων. Σύμφωνα με τα όσα αναφέρθηκαν είναι αντιληπτό ότι ο τομέας της χημειοπληροφορικής έχει επηρεάσει σημαντικά τον τομέα ανάπτυξης φαρμάκων. Τα εργαλεία

χημειοπληροφορικής Enalos της εταιρείας NovaMechanics Ltd βασίζονται στην ανάπτυξη νέων αλγορίθμων και πλατφορμών για την επίλυση προβλημάτων χημειοπληροφορικής, μοντελοποίησης και προσομοίωσης. Μέσω των λειτουργιών του Enalos Suite και του Enalos Cloud ερμηνεύονται και εφαρμόζονται επικυρωμένα μοντέλα πρόβλεψης σε ανάλυση μεγάλων δεδομένων με την βοήθεια του ηλεκτρονικού υπολογιστή. Κάθε ένα από αυτά τα μοντέλα είναι πλήρως επικυρωμένα σύμφωνα με τις αρχές του ΟΟΣΑ και δίνουν την δυνατότητα στον χρήστη να αναζητήσει πληροφορίες σε μεγάλες βάσεις δεδομένων όπως η PubChem για κάθε επιθυμητή δομή ή ομάδα δομών σε σύντομο χρονικό διάστημα. Το μοντέλο κυτταροτοξικότητας MouseTox, το μοντέλο αναστολής των K562 κυττάρων και το μοντέλο αναστολής του παράγοντα TNF αποτελούν πιο σημαντικά αξιόπιστα μοντέλα τα οποία είναι διαθέσιμα στο Enalos Cloud. Για να κατανοηθεί με βέλτιστο τρόπο η χρήση των βάσεων δεδομένων με την βοήθεια των εργαλείων χημειοπληροφορικής Enalos, στην παρούσα διπλωματική αναλύθηκαν ενδεικτικά τρεις γνωστοί αναστολείς TNF (T8, T23 & SPD304), σύμφωνα με πρόσφατη βιβλιογραφία. Όλη η διαδικασία πραγματοποιήθηκε μέσω της πλατφόρμας του Enalos Suite καθώς και σε μοντέλα του Enalos Cloud. Για τις τρεις αυτές δομές βρέθηκαν όλες οι παρόμοιες ενώσεις με βάση τον συντελεστή ομοιότητας Tanimoto και έπειτα όλες οι δομές ελέγχθηκαν για την δυνατότητα αναστολής του παράγοντα TNF καθώς και την κυτταροτοξικότητά τους μέσω του μοντέλου αναστολής του παράγοντα TNF και του μοντέλου κυτταροτοξικότητας MouseTox αντίστοιχα. Οι πλέον υποσχόμενες συγγενείς ενώσεις, δηλαδή αυτές των αναστολέων T23 και SPD304, ελέγχθηκαν περαιτέρω για την εμπορική τους διαθεσιμότητα και την κάλυψη διπλωμάτων ευρεσιτεχνίας μέσω των εργαλείων του Enalos Suite. Τα κριτήρια πληρούσαν 22 ενώσεις για τον αναστολέα T23 και 9 ενώσεις για τον αναστολέα SPD304 οι οποίες είναι και οι πλέον ενδιαφέρουσες για περαιτέρω μελέτες με στόχο την θεραπεία ασθενειών που σχετίζονται με τον παράγοντα TNF.

Αναφορές-Πηγές

- Afantitis, A., Leonis, G., Gambari, R., & Melagraki, G. (2018). Consensus Predictive Model for Human K562 Cell Growth Inhibition through Enalos Cloud Platform. *ChemMedChem*, 13(6), 555–563. <https://doi.org/10.1002/cmdc.201700675>
- Apweiler, R. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(90001), 115D – 119. <https://doi.org/10.1093/nar/gkh131>
- Bajusz, D., Rácz, A., & Héberger, K. (2017a). Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In *Comprehensive Medicinal Chemistry III* (Vols. 3–8, pp. 329–378). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-409547-2.12345-5>
- Bajusz, D., Rácz, A., & Héberger, K. (2017b). Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In *Comprehensive Medicinal Chemistry III* (Vols. 3–8, pp. 329–378). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-409547-2.12345-5>
- Bajusz, D., Rácz, A., & Héberger, K. (2017c). Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In *Comprehensive Medicinal Chemistry III* (Vols. 3–8, pp. 329–378). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-409547-2.12345-5>
- Barnard, J. M., & Downs, G. M. (1997). *Chemical Fragment Generation and Clustering Software*. *Journal of Chemical Information and Computer Sciences*, 37(1), 141–142. doi:10.1021/ci960090k
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32(DATABASE ISS.). <https://doi.org/10.1093/nar/gkh121>
- Bender Andreas, & Brown Nathan. (2018). Special Issue: C heminformatics in Drug Discovery. *Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim*, 13, 467–469. <https://doi.org/10.1002/cmdc.v13.6.issue-toc>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... Wiswedel, B. (2009). *KNIME - the Konstanz information miner. ACM SIGKDD Explorations Newsletter*, 11(1), 26. doi:10.1145/1656274.1656280
- Cao, Longbing (2017). *Data Science. ACM Computing Surveys*, 50(3), 1–42. doi:10.1145/3076253
- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015a). Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C), 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>
- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015b). Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C), 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>
- Chu W. M. (2013). Tumor necrosis factor. *Cancer letters*, 328(2), 222–225. <https://doi.org/10.1016/j.canlet.2012.10.014>
- Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2009). How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, 20(3–4), 241–266. <https://doi.org/10.1080/10629360902949567>
- Demchuk, E., Ruiz, P., Chou, S., & Fowler, B. A. (2011). SAR/QSAR methods in public health practice. In *Toxicology and Applied Pharmacology* (Vol. 254, Issue 2, pp. 192–197). <https://doi.org/10.1016/j.taap.2010.10.017>
- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>
- Flower, D. R. (1998). *On the Properties of Bit String-Based Measures of Chemical Similarity. Journal of Chemical Information and Computer Sciences*, 38(3), 379–386. doi:10.1021/ci970437z
- Gasteiger, J. (2016). Cheminformatics: Achievements and challenges, a personal view. In *Molecules* (Vol. 21, Issue 2). MDPI AG. <https://doi.org/10.3390/molecules21020151>
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012a). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1). <https://doi.org/10.1093/nar/gkr777>
- Grisoni, F., Consonni, V., & Todeschini, R. (2018). Impact of Molecular Descriptors on Computational Models. In *Methods in Molecular Biology* (Vol. 1825, pp. 171–209). Humana Press Inc. https://doi.org/10.1007/978-1-4939-8639-2_5
- HANSCH, C., MALONEY, P., FUJITA, T. *et al.* Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194**, 178–180 (1962). <https://doi.org/10.1038/194178b0>
- Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1). <https://doi.org/10.1186/s13321-015-0068-4>
- Henkel, T., Brunne, R. M., Müller, H., & Reichel, F. (1999). *Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. Angewandte Chemie International Edition*, 38(5), 643–647. doi:10.1002/(sici)1521-3773(19990301)38:5<643::aid-anie643>3.0.co;2-g

- Holliday. (2002). *Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings*. *Combinatorial Chemistry & High Throughput Screening*, 5(2). doi:10.2174/1386207024607338
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Pagni, M., & Sigrist, C. J. A. (2006). The PROSITE database. *Nucleic Acids Research*, 34(Database issue). <https://doi.org/10.1093/nar/gki063>
- Irwin, J. J., & Shoichet, B. K. (2005). ZINC--a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1), 177–182. <https://doi.org/10.1021/ci049714+>
- Jaworska, J. S., Comber, M., Auer, C., & van Leeuwen, C. J. (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environmental Health Perspectives*, 111(10), 1358–1360. <https://doi.org/10.1289/ehp.5757>
- Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of Machine Learning Tools in Healthcare Decision Making. In *Journal of Healthcare Engineering* (Vol. 2021). Hindawi Limited. <https://doi.org/10.1155/2021/6679512>
- Johann Gasteiger, & Thomas Engel. (2003). *Cheminformatics—A Textbook.* Wiley-VCH Verlag GmbH & Co. KGaA.
- Kooistra, A. J., Vischer, H. F., McNaught-Flores, D., Leurs, R., de Esch, I. J. P., & de Graaf, C. (2016). Function-specific virtual screening for GPCR ligands using a combined scoring method. *Scientific Reports*, 6. <https://doi.org/10.1038/srep28288>
- Kroemer R. T. (2007). Structure-based drug design: docking and scoring. *Current protein & peptide science*, 8(4), 312–328. <https://doi.org/10.2174/138920307781369382>
- Kumar, A., & Zhang, K. (2018). Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in chemistry*, 6, 315. <https://doi.org/10.3389/fchem.2018.00315>
- Kurdi, M. Z. (2016). *Natural Language Processing and Computational Linguistics* 1. doi:10.1002/9781119145554
- Leoni, G., Melagraki, G., & Afantitis, A. (2017). Open source cheminformatics software including KNIME analytics. In *Handbook of Computational Chemistry* (pp. 2201–2230). Springer International Publishing. https://doi.org/10.1007/978-3-319-27282-5_57
- Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(SUPPL. 1). <https://doi.org/10.1093/nar/gkl999>
- Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018a). Machine learning in cheminformatics and drug discovery. In *Drug Discovery Today* (Vol. 23, Issue 8, pp. 1538–1546). Elsevier Ltd. <https://doi.org/10.1016/j.drudis.2018.05.010>
- Maggiara, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. In *Journal of Medicinal Chemistry* (Vol. 57, Issue 8, pp. 3186–3204). American Chemical Society. <https://doi.org/10.1021/im401411z>
- Melagraki, G., Ntougkos, E., Rinotas, V., Papaneophytou, C., Leonis, G., Mavromoustakos, T., Kontopidis, G., Douni, E., Afantitis, A., & Kollias, G. (2017a). Cheminformatics-aided discovery of small-molecule Protein-Protein Interaction (PPI) dual inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF- κ B Ligand (RANKL). *PLoS Computational Biology*, 13(4). <https://doi.org/10.1371/journal.pcbi.1005372>
- Melagraki, G., Ntougkos, E., Rinotas, V., Papaneophytou, C., Leonis, G., Mavromoustakos, T., Kontopidis, G., Douni, E., Afantitis, A., & Kollias, G. (2017b). Cheminformatics-aided discovery of small-molecule Protein-Protein Interaction (PPI) dual inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF- κ B Ligand (RANKL). *PLoS Computational Biology*, 13(4). <https://doi.org/10.1371/journal.pcbi.1005372>
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., de Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., ... Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1), D930–D940. <https://doi.org/10.1093/nar/gky1075>
- Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2016). Machine learning: Algorithms and applications. In *Machine Learning: Algorithms and Applications*. CRC Press. <https://doi.org/10.1201/9781315371658>
- Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. In *Expert Opinion on Drug Discovery* (Vol. 11, Issue 2, pp. 137–148). Taylor and Francis Ltd. <https://doi.org/10.1517/17460441.2016.1117070>
- Nikolova, N., & Jaworska, J. (2004). Approaches to Measure Chemical Similarity - A Review. In *QSAR and Combinatorial Science* (Vol. 22, Issues 9–10, pp. 1006–1026). Wiley-VCH Verlag. <https://doi.org/10.1002/qsar.200330831>
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An Open chemical toolbox. *Journal of Cheminformatics*, 3(10). <https://doi.org/10.1186/1758-2946-3-33>
- Olesen, C. M., Coskun, M., Peyrin-Biroulet, L., & Nielsen, O. H. (2016). Mechanisms behind efficacy of tumor necrosis factor inhibitors in inflammatory bowel diseases. In *Pharmacology and Therapeutics* (Vol. 159, pp. 110–119). Elsevier Inc. <https://doi.org/10.1016/j.pharmthera.2016.01.001>

- Öztürk, H., Özgür, A., Schwaller, P., Laino, T., & Ozkirimli, E. (2020). Exploring chemical space using natural language processing methodologies for drug discovery. In *Drug Discovery Today* (Vol. 25, Issue 4, pp. 689–705). Elsevier Ltd. <https://doi.org/10.1016/j.drudis.2020.01.020>
- Ruiz, P., Begluzzi, G., Tincher, T., Wheeler, J., & Mumtaz, M. (2012). Prediction of acute mammalian toxicity using QSAR methods: A case study of sulfur mustard and its breakdown products. *Molecules*, 17(8), 8982–9001. <https://doi.org/10.3390/molecules17088982>
- Saldívar-González, F. I., Huerta-García, C. S., & Medina-Franco, J. L. (2020a). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12(1). <https://doi.org/10.1186/s13321-020-00466-z>
- Saldívar-González, F. I., Huerta-García, C. S., & Medina-Franco, J. L. (2020b). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12(1). <https://doi.org/10.1186/s13321-020-00466-z>
- Santana, K., do Nascimento, L. D., Lima e Lima, A., Damasceno, V., Nahum, C., Braga, R. C., & Lameira, J. (2021). Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. In *Frontiers in Chemistry* (Vol. 9). Frontiers Media S.A. <https://doi.org/10.3389/fchem.2021.662688>
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). <https://doi.org/10.1007/s42979-021-00592-x>
- Sarker, I. H., Hoque, M. M., Uddin, M. K., & Alsanoosy, T. (2021). Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions. *Mobile Networks and Applications*, 26(1), 285–303. <https://doi.org/10.1007/s11036-020-01650-z>
- Schneider, G. (2010). Virtual screening: An endless staircase? In *Nature Reviews Drug Discovery* (Vol. 9, Issue 4, pp. 273–276). <https://doi.org/10.1038/nrd3139>
- Schneider, M., Pons, J. L., Labesse, G., & Bourguet, W. (2019). *In silico* predictions of endocrine disruptors properties. In *Endocrinology (United States)* (Vol. 160, Issue 11, pp. 2709–2716). Endocrine Society. <https://doi.org/10.1210/en.2019-00382>
- Segler, M., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS central science*, 4(1), 120–131. <https://doi.org/10.1021/acscentsci.7b00512>
- Shin, W. H., Zhu, X., Bures, M. G., & Kihara, D. (2015). Three-dimensional compound comparison methods and their application in drug discovery. *Molecules*, 20(7), 12841–12862. <https://doi.org/10.3390/molecules200712841>
- Stahura, F. L., Godden, J. W., Xue, L., & Bajorath, J. (2000). Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *Journal of Chemical Information and Computer Sciences*, 40(5), 1245–1252. <https://doi.org/10.1021/ci0003303>
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493–500. <https://doi.org/10.1021/ci025584y>
- Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. v., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., ... Tetko, I. v. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design*, 25(6), 533–554. <https://doi.org/10.1007/s10822-011-9440-2>
- Tovar, A., Eckert, H., & Bajorath, J. (2007). Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem*, 2(2), 208–217. <https://doi.org/10.1002/cmdc.200600225>
- Tropsha, A., Gramatica, P., & Gombar, V. (2003). *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*. *QSAR & Combinatorial Science*, 22(1), 69–77. doi:10.1002/qsar.200390007
- Varnek, A., & Baskin, I. (2012). Machine learning methods for property prediction in chemoinformatics: Quo Vadis? In *Journal of Chemical Information and Modeling* (Vol. 52, Issue 6, pp. 1413–1437). <https://doi.org/10.1021/ci200409x>
- Varsou, D. D., Melagraki, G., Sarimveis, H., & Afantitis, A. (2017). MouseTox: An online toxicity assessment tool for small molecules through Enalos Cloud platform. *Food and Chemical Toxicology*, 110, 83–93. <https://doi.org/10.1016/j.fct.2017.09.058>
- Varsou, D. D., Nikolakopoulos, S., Tsoumanis, A., Melagraki, G., & Afantitis, A. (2018a). Enalos suite: New cheminformatics platform for drug discovery and computational toxicology. In *Methods in Molecular Biology* (Vol. 1800, pp. 287–311). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7899-1_14
- Vogt, M., & Bajorath, J. (2020). ccbmlib - a Python package for modeling Tanimoto similarity value distributions. *F1000Research*, 9, Chem Inf Sci-100. <https://doi.org/10.12688/f1000research.22292.2>
- Walters, W.P., Stahl, M.T., & Murcko, M.A. (1998). Virtual screening : an overview. *Drug Discovery Today*, 3, 160-178.
- Wang, R., Fang, X., Lu, Y., Yang, C. Y., & Wang, S. (2005). The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 48(12), 4111–4119. <https://doi.org/10.1021/jm048957q>

- Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. In *Nucleic Acids Research* (Vol. 37, Issue SUPPL. 2). <https://doi.org/10.1093/nar/gkp456>
- Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. In *Drug Discovery Today* (Vol. 11, Issues 23–24, pp. 1046–1053). <https://doi.org/10.1016/j.drudis.2006.10.005>
- Willett, P. (2009). *Similarity methods in chemoinformatics*. , 43(1), 1–117. doi:10.1002/aris.2009.1440430108
- William Edward Kearns, P. (2008). ENVIRONMENT DIRECTORATE JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY Developments in OECD Delegations on the Safety Ass. <https://doi.org/10.13140/RG.2.2.19606.37446>
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*, 34(Database issue). <https://doi.org/10.1093/nar/gki067>

Διαδικτυακοί Τόποι

ChEMBL, <https://www.ebi.ac.uk/chembl/>
Consensus Predictive Model for the prediction of Human K562 Cell Growth Inhibition through Enalos Cloud Platform, <http://enalos.insilicotox.com/K562>
Enalos Platform by NovaMechanics, <https://www.insilicotox.com/index.php/products/predictive-models-web-services/tnf/>
Enalos TNF Extraction Platform, <http://www.enaloscloud.novamechanics.com/EnalosWebApps/TNF/>
Enalos Nodes for KNIME, <https://tech.knime.Org/community/enalos-nodes>
Indigo Toolkit, <http://lifescience.opensource.epam.com/indigo/>
GSI Technology, <https://www.qsitechnology.com>
MouseTox: Prediction of small molecules cytotoxic effect to NIH/3T3 cells through Enalos Cloud, <http://www.enaloscloud.novamechanics.com/EnalosWebApps/MouseTox/>
Nova Mechamnic Ltd, <https://novamechanics.com/>
OECD, <http://www.oecd.org>
PubChem, <https://pubchem.ncbi.nlm.nih.gov/>
Wikipedia, <https://en.wikipedia.org/wiki/>
Wiktionary, the free dictionary, <https://en.wiktionary.org>