



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

**Υπολογιστική Προσέγγιση για την Απόκριση Αντικαρκινικών
Φαρμάκων με Αλγόριθμους Συμβολικής Παλινδρόμησης**

Βασιλική Δ. Λιάλια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
Επιβλέπων
Μιχάλης Σαβελώνας

Λαμία, 2024



UNIVERSITY OF THESSALY

SCHOOL OF SCIENCE

INFORMATICS AND COMPUTATIONAL BIOMEDICINE

**Computational Approach for Anticancer Drug Response
with Symbolic Regression Algorithms**

Vasiliki D. Lialia

**Master thesis
Supervisor
Michalis Savelonas**

Lamia, 2024



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΠΛΗΡΟΦΟΡΙΚΗ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ
ΚΑΤΕΥΘΥΝΣΗ**

«ΥΠΟΛΟΓΙΣΤΙΚΗ ΙΑΤΡΙΚΗ ΚΑΙ ΒΙΟΛΟΓΙΑ»

**Υπολογιστική Προσέγγιση για την Απόκριση Αντικαρκινικών Φαρμάκων
με Αλγόριθμους Συμβολικής Παλινδρόμησης**

Βασιλική Δ. Λιάλια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Επιβλέπων
Μιχάλης Σαβελώνας**

Λαμία, 2024

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο Υπολογιστική Προσέγγιση για την Απόκριση Αντικαρκινικών Φαρμάκων με Αλγόριθμους Συμβολικής Παλινδρόμησης αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Η ΔΗΛΟΥΣΑ

Ημερομηνία

Υπογραφή

Υπολογιστική Προσέγγιση για την Απόκριση Αντικαρκινικών Φαρμάκων με Αλγόριθμους Συμβολικής Παλινδρόμησης

Βασιλική Δ. Λιάλια

Τριμελής Επιτροπή:

Μιχάλης Σαβελώνας, Επίκουρος Καθηγητής ΠΘ, επιβλέπων

Βασίλειος Δρακόπουλος, Αναπληρωτής Καθηγητής ΠΘ

Άγγελος Αβραμόπουλος, Αναπληρωτής Καθηγητής ΠΘ

ΠΕΡΙΛΗΨΗ

Η έρευνα του καρκίνου είναι σε εξέλιξη, καθώς αυτή η ασθένεια παραμένει μία από τις κύριες αιτίες θανάτου παγκοσμίως, με πολύπλοκους μηχανισμούς ανάπτυξης που διαφοροποιούνται μεταξύ των ασθενών. Η καλύτερη κατανόηση του καρκίνου μπορεί να οδηγήσει σε πιο αποτελεσματικές θεραπείες και ποσοστά επιβίωσης. Η φαρμακογονιδιωματική, η μελέτη του τρόπου με τον οποίο τα γονίδια επηρεάζουν την απόκριση ενός ατόμου στα φάρμακα, είναι μείζονος σημασίας στην καταπολέμηση του καρκίνου, διότι επιτρέπει την εξατομικευμένη ιατρική. Προσαρμόζοντας μία θεραπεία στο γενετικό προφίλ του εκάστοτε ασθενή, ενισχύεται η αποτελεσματικότητα και μειώνονται οι παρενέργειες των φαρμάκων.

Τα μοντέλα πρόβλεψης του ήμισυ της μέγιστης ανασταλτικής συγκέντρωσης (IC_{50}) αξιοποιούνται στην προσωποποιημένη ογκολογική θεραπεία. Δοθέντος της γονιδιακής έκφρασης ενός καρκίνου, αναπαριστούν την απαραίτητη συγκέντρωση ενός φαρμάκου που απαιτείται για την καταστολή του ήμισυ της λειτουργίας των καρκινικών κυττάρων. Ωστόσο, οι μέθοδοι πρόβλεψης που είναι διαθέσιμες στη βιβλιογραφία δεν είναι διαφανείς ως προς το αποτέλεσμα, κάτι που δυσχεραίνει τη διάδοση τους σε ιατρικά περιβάλλοντα. Η συμβολική παλινδρόμηση παράγει αναλυτικές εκφράσεις, διευκολύνοντας την ερμηνεία και την ανάλυση των φαινομένων. Στα πλαίσια της παρούσας εργασίας, έγινε εφαρμογή πέντε αλγορίθμων συμβολικής παλινδρόμησης για την κατασκευή 24 μοντέλων για 24 αντικαρκινικά φαρμάκων από δημόσια δεδομένα της Cancer Cell Line Encyclopedia. Για τα έξι φάρμακα δημιουργήθηκαν ικανοποιητικά μοντέλα πρόβλεψης με έξοδο τη μαθηματική έκφραση που αναπαριστά τη σχέση γονιδίων- IC_{50} . Αυτή η σχέση επιτρέπει στον ογκολόγο να κατανοήσει τους γενετικούς παράγοντες που επηρεάζουν την απόκριση σε φάρμακο και να προτείνει την κατάλληλη θεραπεία. Το καλύτερο μοντέλο πέτυχε συντελεστή προσδιορισμού $R^2 = 42\%$ και μέσο τετραγωνικό σφάλμα ($RMSE$) ίσο με 0,65 για το φάρμακο Panobinostat. Από τα 18989 γονίδια, η συνάρτηση πρόβλεψης περιέχει μόνο δύο γονίδια σε μία απλή γραμμική σχέση ως $\ln(IC_{50}) = 0.388 \cdot MYOF - 0.185 \cdot P2RX1 - 2.87$, όπου το πρώτο κωδικοποιεί την μυοφερλίνη ενώ το δεύτερο τον υποδοχέα P2X1, μέλος της οικογένειας των πουρινεργικών υποδοχέων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Φαρμακογονιδιωματική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: συμβολική παλινδρόμηση, πρόβλεψη απόκρισης φαρμάκων, φαρμακογονιδιωματική, βιοδείκτες

ABSTRACT

Cancer research is ongoing, as this disease remains one of the leading causes of death worldwide, with complex mechanisms that vary between individuals. Better understanding of cancer can lead in more effective treatments and survival rates. Pharmacogenomics, the study of how genes influence a person's response to drugs, is crucial in the fight against cancer, since it enables personalized medicine. By tailoring a treatment to the genetic profile of the cancer, the effectiveness is enhanced, and the drugs' side effects are reduced.

Half-maximal inhibitory concentration (IC_{50}) prediction models contribute to personalized oncology therapy. The necessary concentration of a drug needed to suppress half of the cancer cells' function can be predicted given the gene expression of the cancer. However, the existing prediction models are not transparent, which hinders their adoption in medical settings. Symbolic regression results in analytical expressions, facilitating interpretation and analysis. In this thesis, five symbolic regression algorithms were applied to build 24 models for 24 anticancer drugs, using public datasets from the Cancer Cell Line Encyclopedia. Sufficiently accurate models were created for six of the drugs. The resulting analytical expression enables the oncologist to understand the genetic factors that influence drug response and to recommend appropriate treatment. The best model has coefficient of determination $R^2 = 42\%$ and root means squared error $RMSE = 0.65$ for the drug Panobinostat. Out of 18989 genes, the prediction function contains only two genes in a simple linear relationship as $\ln(IC_{50}) = 0.388 \cdot MYOF - 0.185 \cdot P2RX1 - 2.87$, where the first encodes myoferlin, while the second encodes the P2X1 receptor, a member of the purinergic receptor family.

SUBJECT AREA: Pharmacogenomics

KEYWORDS: symbolic regression, drug response prediction, pharmacogenomics, biomarkers

Στους φίλους μου και στην οικογένειά μου

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τα μέλη της τριμελούς επιτροπής, τον Μιχάλη Σαβελώνα, Επίκουρο Καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, που μου παρείχε συνεχής καθοδήγηση καθ' όλη τη διάρκεια της συνεργασίας μας, τον Βασίλειο Δρακόπουλο, Αναπληρωτή Καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, για τις συμβουλές του, τον Άγγελο Αβραμόπουλο, Αναπληρωτή Καθηγητή του τμήματος Φυσικής του Πανεπιστημίου Θεσσαλίας.

Είμαι ευγνώμων στον Θεόδωρο Τζουραμάνη, Επίκουρο Καθηγητή, ως σύμβουλο μου στη διάρκεια των μεταπτυχιακών μου σπουδών.

Ευχαριστώ θερμά τους γονείς μου, την οικογένειά μου, και ιδιαίτερα τους φίλους και τις φίλες μου για τη συνεχή υποστήριξη, αγάπη, και υπομονή που δείχνουν στο πρόσωπό μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ.....	15
1. ΕΙΣΑΓΩΓΗ.....	16
2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	17
2.1 ΜΟΡΙΑΚΟ ΠΡΟΦΙΛ ΟΓΚΟΥ	17
2.2 ΒΙΟΔΕΙΚΤΕΣ.....	17
2.3 ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ.....	18
2.4 ΠΡΟΦΙΛ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ	19
2.5 ΔΙΑΛΟΓΗ ΥΨΗΛΗΣ ΑΠΟΔΟΣΗΣ	20
2.6 ΦΑΡΜΑΚΟΓΟΝΙΔΙΩΜΑΤΙΚΗ.....	21
2.7 ΉΜΙΣΥ ΤΗΣ ΜΕΓΙΣΤΗΣ ΑΝΑΣΤΑΛΤΙΚΗΣ ΣΥΓΚΕΝΤΡΩΣΗΣ - IC50.....	21
2.8 ΑΠΟΚΡΙΣΗ ΑΣΘΕΝΟΥΣ ΣΕ ΦΑΡΜΑΚΟ	21
2.9 ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	22
2.10 ΣΥΜΒΟΛΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....	22
2.11 ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ SR.....	23
2.11.1 Ορισμός κλάσης συνάρτησης F	23
2.11.2 Αναπαράσταση έκφρασης.....	24
2.12 PYSR.....	24
2.13 FEYN (QLATTICE)	25
2.14 FAST FUNCTION EXTRACTION (FFX)	25
2.15 INTERACTION-TRANSFORMATION EVOLUTIONARY ALGORITHM (ITEA)	25
2.16 EQUATION LEARNER (EQL)	26
3. ΜΕΘΟΔΟΛΟΓΙΑ	26
3.1 ΠΕΡΙΓΡΑΦΗ ΚΑΙ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	26
3.2 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ	29
4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΔΟΚΙΜΕΣ.....	30
5. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ	35
ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ.....	37
ΠΑΡΑΡΤΗΜΑ Ι – ΠΙΝΑΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ΜΟΝΤΕΛΩΝ ΚΑΙ ΠΑΡΑΓΟΜΕΝΕΣ ΕΞΙΣΩΣΕΙΣ.....	38
ΠΑΡΑΡΤΗΜΑ ΙΙ - ΔΙΑΓΡΑΜΜΑΤΑ ΔΙΑΣΠΟΡΑΣ ΠΡΟΒΛΕΠΟΜΕΝΗΣ ΕΝΑΝΤΙ ΠΡΑΓΜΑΤΙΚΗΣ ΤΙΜΗΣ LNIC50	53

ΑΝΑΦΟΡΕΣ	60
-----------------------	-----------

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1 – Ανάλυση των κυττάρων οδηγεί στο μοριακό προφίλ όγκου	17
Εικόνα 2 - Κυτταρικός θάνατος που προκαλείται από καταστολείς του EFGR.....	18
Εικόνα 3 - Ευθεία παλινδρόμησης για 50 τυχαία σημεία που προήλθαν από μία κανονική κατανομή. Οι παρατηρήσεις (ερυθρό χρώμα) αντιμετωπίζονται ως αποκλίσεις από την υποκείμενη σχέση (μπλε χρώμα) μεταξύ της εξαρτημένης με την ανεξάρτητη μεταβλητή.	22
Εικόνα 4 - Δέντρο έκφρασης για την $x^2 \cdot x + \text{tany}$	24
Εικόνα 5 – EQL αρχιτεκτονική για 3 επίπεδα και έναν νευρώνα ανά τύπο	26
Εικόνα 6 - Περιγραφή αρχείου δεδομένων .gcd (Data Formats - GeneSetEnrichmentAnalysisWiki, n.d.).....	27
Εικόνα 7 – QLattice (random_seed=157, n_epochs=18): Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	53
Εικόνα 8 – pySR (random_state=79, niterations=160, deterministic=True, procs=0, multithreading=None): Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	54
Εικόνα 9 – QLattice (random_seed=96, n_epochs=18) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	55
Εικόνα 10 – pySR (random_state=21, niterations=160, deterministic=True, procs=0, multithreading=None) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	56
Εικόνα 11 – FFX με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	57
Εικόνα 12 – EQL (num_layers=1, dim=train_X.shape[1]) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	58
Εικόνα 13 – ITEA (random_state=18, simplify_method='simplify_by_var') με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής lnIC50	59

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1 - Δείγμα συνόλου δεδομένων γονιδιακής έκφρασης σε καρκινικές κυτταρικές σειρές	28
Πίνακας 2 - Δείγμα συνόλου δεδομένων απόκρισης φαρμάκων σε κυτταρικές σειρές	28
Πίνακας 3 - Πλήθος δειγμάτων συνόλου εκπαίδευσης και ελέγχου ανά φάρμακο	29
Πίνακας 4 – Επτά σημαντικότερα γονίδια βάσει SelectKBest	31
Πίνακας 5 - Περιγραφικά στατιστικά του <i>lnIC50μMol</i>	32
Πίνακας 6 - Ικανοποιητικά μοντέλα ανά αλγόριθμο και φάρμακο.....	33
Πίνακας 7 - Συναρτήσεις ικανοποιητικών μοντέλων με μικρότερο πλήθος μεταβλητών	34
Πίνακας 8 - Ικανοποιητικά μοντέλα ανά αλγόριθμο και φάρμακο των Park et al. (2023).....	34
Πίνακας 9 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις Qlattice (random_seed=157, n_epochs=18).....	38
Πίνακας 10 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις PySR (random_state=79, niterations=160, deterministic=True, procs=0, multithreading=None)	40
Πίνακας 11 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με Qlattice (random_seed=96, n_epochs=18)	41
Πίνακας 12 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με PySR (random_state=21, niterations=160, deterministic=True, procs=0, multithreading=None)	42
Πίνακας 13 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με FFX	44
Πίνακας 14 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με EQL(num_layers=1, dim=train_X.shape[1])	50
Πίνακας 15 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με ITEA (random_state=18, simplify_method='simplify_by_var')	52

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών Πληροφορικής και Υπολογιστική Βιοϊατρική με ειδίκευση στην Υπολογιστική Ιατρική και Βιολογία των τμημάτων Πληροφορικής με Εφαρμογές στη Βιοϊατρική, και Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας υπό την επίβλεψη του Επίκουρου Καθηγητή Μιχάλη Σαβελώνα.

1. Εισαγωγή

Η ιατρική ακριβείας και οι εξατομικευμένες θεραπείες για τον καρκίνο συνιστούν μια κρίσιμη πρόκληση για την Ιατρική. Ο σκοπός της ογκολογίας ακριβείας, δηλαδή της ιατρικής ακριβείας στην περίπτωση του καρκίνου, είναι η βελτίωση των αποτελεσμάτων μιας θεραπείας, μέσω της ανάλυσης του προφίλ όγκων σε μοριακό επίπεδο, ώστε να σχεδιαστούν θεραπείες ειδικά προσαρμοσμένες στα χαρακτηριστικά της εκάστοτε μοριακής υποομάδας όγκων ή ακόμα και μεμονωμένων ασθενών. Η ογκολογία ακριβείας στηρίζεται λοιπόν στην αποτελεσματική μετάφραση των κλινικών δεδομένων μαζί με γονιδιακά δεδομένα σε προγνωστικούς βιοδείκτες.

Στα περισσότερα συστήματα ιατρικής περίθαλψης οι καρκίνοι ταξινομούνται κυρίως ανάλογα με τον τύπο του ιστού ή το μέρος του σώματος από το οποίο προέκυψαν: πνεύμονα, εγκέφαλο, μαστό, έντερο, πάγκρεας κ.λπ. Ωστόσο, οι ερευνητές έχουν εντοπίσει τα μοριακά αποτυπώματα του καρκίνου και τα χρησιμοποιούν για να χωρίσουν τις κάποτε ευρείες κατηγορίες σε πολύ πιο ακριβείς τύπους και υποτύπους. Έχουν ανακαλύψει επίσης ότι οι καρκίνοι που αναπτύσσονται σε εντελώς διαφορετικά μέρη του σώματος μπορεί συχνά, σε μοριακό επίπεδο, να έχουν πολλά κοινά. Από αυτή τη νέα προοπτική αναδύεται η ογκολογία ακριβείας, στην οποία οι γιατροί επιλέγουν θεραπείες με βάση την υπογραφή DNA (deoxyribonucleic acid - δεοξυριβονουκλεϊκό οξύ - αλληλουχίες νουκλεοτιδίων που ξεχωρίζουν έναν οργανισμό από τους υπόλοιπους) του όγκου ενός μεμονωμένου ασθενούς (*Precision Oncology*, 2023) (Phillippy et al., 2007).

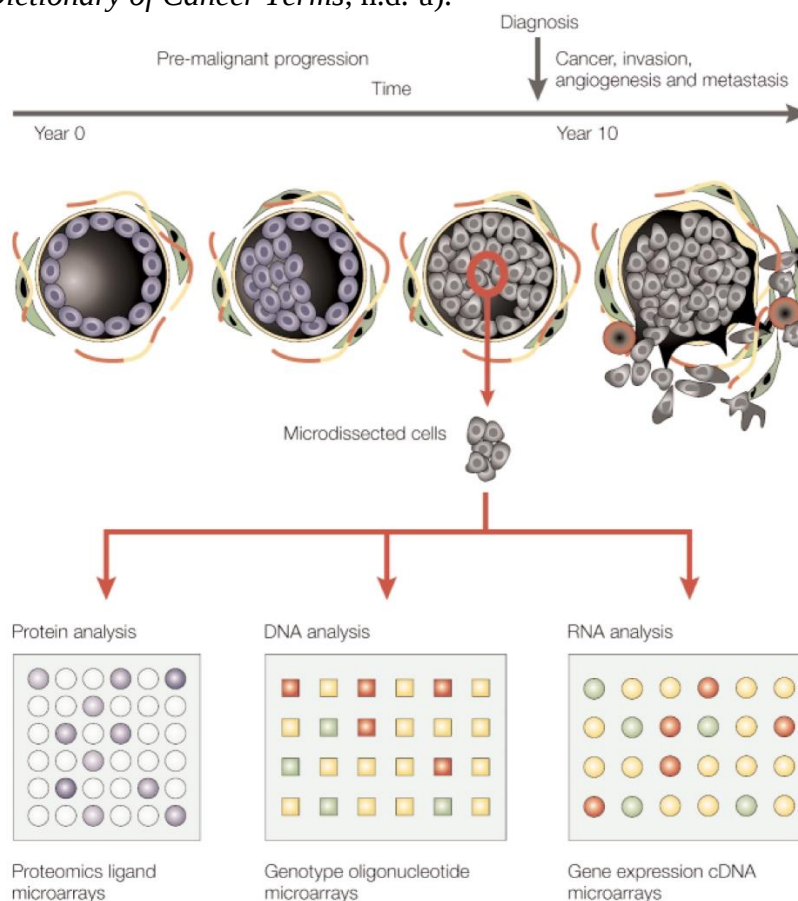
Εκτός από τον χαρακτηρισμό των όγκων σε μοριακό επίπεδο, μία άλλη αποστολή της ογκολογίας ακριβείας είναι να παράγει προφίλ απόκρισης φαρμάκων, τα οποία καλύπτουν ένα ευρύ φάσμα φαρμάκων και υποτύπων καρκίνου. Τα τελευταία χρόνια, έχουν δημοσιοποιηθεί αρκετά δεδομένα μεγάλης κλίμακας αξιολόγησης φαρμάκων από διάφορα ερευνητικά έργα, συμβάλλοντας στην προώθηση της ογκολογίας ακριβείας. Αυτά τα ερευνητικά έργα έχουν αξιολογήσει γνωστά και υποψήφια αντικαρκινικά φάρμακα έναντι καρκινικών κυτταρικών σειρών, οι οποίες έχουν χαρακτηριστεί σε μοριακό επίπεδο.

Τα παραπάνω δεδομένα έχουν ήδη επιτρέψει την αναγνώριση πιθανών βιοδεικτών απόκρισης φαρμάκων με χρήση υπολογιστικών μεθόδων, καθώς και την ανάπτυξη μοντέλων πρόβλεψης της απόκρισης φαρμάκων. Πράγματι, οι υπολογιστικές μέθοδοι διευκολύνουν την κατανόηση δεδομένων μεγάλου όγκου με στόχο την αξιολόγηση υποψήφιων φαρμάκων όμως καθώς η πειραματική αξιολόγηση όλων των πιθανών υποψήφιων φαρμάκων ή συνδυασμών φαρμάκων δεν είναι εφικτή, για πρακτικούς και οικονομικούς λόγους. Συνεπώς, η ανάπτυξη υπολογιστικών στρατηγικών για την πρόβλεψη απόκρισης φαρμάκων είναι απαραίτητη για να μειώσει το χώρο αναζήτησης και να καθοδηγήσει την διαδικασία ανακάλυψης, μειώνοντας τον απαιτούμενο πειραματικό κόπο μέσω διαλογής υψηλής απόδοσης (high throughput screening). Απαιτούνται, επίσης, υπολογιστικές μέθοδοι για να μεταφράσουν την αποκτωμένη γνώση από τις αναλύσεις βασισμένες σε κυτταρικές σειρές σε προφίλ απόκρισης φαρμάκων συγκεκριμένων ασθενών. Ήδη στην βιβλιογραφία υπάρχει μία πληθώρα υπολογιστικών μεθόδων για την πρόβλεψη απόκρισης φαρμάκων και για την ανακάλυψη βιοδεικτών απόκρισης φαρμάκων. Μεταξύ αυτών, έχουν αναφερθεί προσεγγίσεις μηχανικής μάθησης όπως είναι οι μηχανές ανυσμάτων στήριξης (support vector machines – SVM), μέθοδοι μάθησης Bayes πολλαπλών καθηκόντων και πυρήνων, (Bayesian multitask multiple kernel learning), τα τυχαία δάση (random forests), και τα μοντέλα νευρωνικών δικτύων (neural networks). Ωστόσο, υπάρχουν ακόμα αρκετά περιθώρια βελτίωσης για αυτά τα υπολογιστικά μοντέλα από άποψη απόδοσης, γενίκευσης, και ερμηνείας (Baptista et al., 2020).

2. Θεωρητικό υπόβαθρο

2.1 Μοριακό προφίλ όγκου

Το μοριακό προφίλ όγκου προέρχεται από μία εργαστηριακή τεχνική που εξετάζει δείγματα αίματος ή ιστού όγκου για να ανιχνεύσει συγκεκριμένα γονίδια, πρωτεΐνες ή μόρια. Στην Εικόνα 1 φαίνεται ότι ο μεταστατικός καρκίνος αναπτύσσεται μετά από μία παρατεταμένη προκακοήθη περίοδο. Ένας μικροτόμος επιτρέπει την εξέταση των μοριακών αλλαγών του φυσιολογικού ιστού έναντι του παθολογικού στον ίδιο ασθενή. Με ανάλυση DNA προδιορίζονται οι γενετικές αλλοιώσεις, όπως είναι μεταλλάξεις, ενώ η ανάλυση RNA και πρωτεϊνών δείχνουν αλλαγές στη μεταγραφή, στη πρωτεϊνοσύνθεση, και σε μετα-μεταγραφικές τροποποιήσεις των πρωτεϊνών (Liotta & Petricoin, 2000). Αν και δύο άνθρωποι μπορούν να έχουν τον ίδιο τύπο καρκίνου, οι καρκίνοι τους μπορεί να διαφέρουν σημαντικά σε μοριακό επίπεδο. Οι αναλύσεις που στοχεύουν στο μοριακό προφίλ προσδιορίζουν ορισμένες μοριακές ή γενετικές αλλαγές ενός όγκου, όπως είναι οι γονιδιακές μεταλλάξεις ή άλλες αλλαγές στο DNA του (Μοριακό Προφίλ | PrIme | Ιατρική Ακριβείας | Θεραπείες Καρκίνου, 2023). Έτσι οι ογκολόγοι μπορούν να καθορίσουν εάν υπάρχουν θεραπείες που σχετίζονται με οποιονδήποτε από τους εμφανιζόμενους βιοδείκτες και κατά προέκταση το μοριακό προφίλ μπορεί να χρησιμοποιηθεί για να βοηθήσει στον καθορισμό θεραπείας ή στην πρόβλεψη εάν ο καρκίνος θα επανέλθει ή θα εξαπλωθεί σε άλλα μέρη του σώματος (NCI Dictionary of Cancer Terms, n.d.-a).

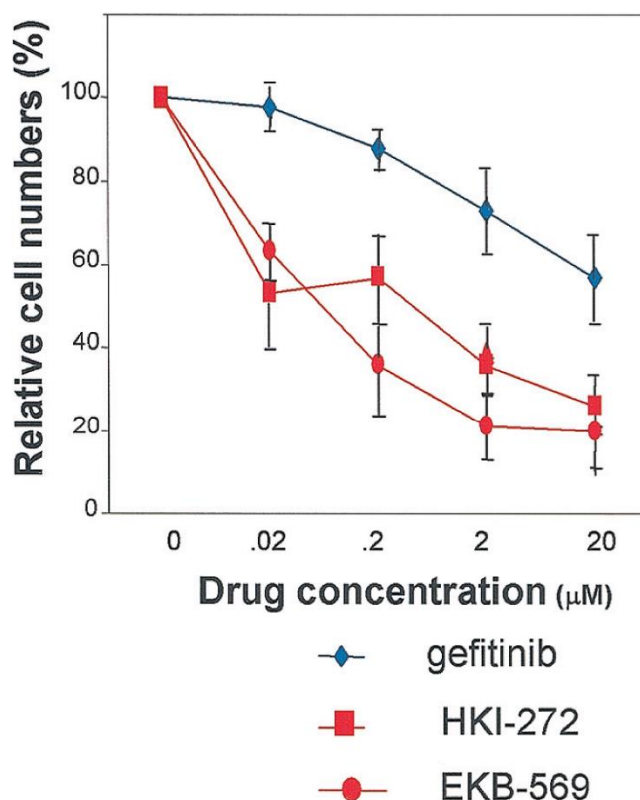


Εικόνα 1 – Ανάλυση των κυττάρων οδηγεί στο μοριακό προφίλ όγκου

2.2 Βιοδείκτες

Ένας βιοδείκτης είναι ένα βιολογικό μόριο που υπάρχει στο αίμα, στα σωματικά υγρά ή σε ιστούς και δείχνει αν μία βιολογική διαδικασία είναι φυσιολογική ή όχι, καθώς και αν υπάρχει ή όχι κάποια ασθένεια (NCI Dictionary of Cancer Terms, n.d.-b). Η λέξη βιοδείκτης είναι συντομογραφία του βιολογικού δείκτη (Biomarkers, n.d.). Μπορεί να γίνει εξέταση βιοδεικτών,

κατά την οποία το δείγμα αναλύεται για συγκεκριμένα γονίδια, πρωτεΐνες ή άλλα μόρια, για να εξεταστεί το ενδεχόμενο ασθένειας, ή αυξημένου κινδύνου για ασθένεια, όπως είναι ο καρκίνος (*NCI Dictionary of Cancer Terms*, n.d.-c). Ένα παράδειγμα είναι το γονίδιο EGFR. Ασθενείς με καρκίνο που έχουν γενετικές αλλοιώσεις σε αυτό το γονίδιο μπορούν να λάβουν στοχευμένη φαρμακευτική αγωγή, η οποία καταστέλλει το EFGR (*Biomarker Testing for Cancer Treatment*, 2021). Όπως φαίνεται στην Εικόνα 2 (Haber et al., 2005), η αύξηση συγκέντρωσης των καταστολέων του μεταλλαγμένου EFGR μειώνει το ποσοστιαίο πλήθος των παθολογικών κυττάρων. Συνεπώς, η ανίχνευση του συγκεκριμένου βιοδείκτη στο μοριακό προφίλ του όγκου οδηγεί σε στοχευμένη θεραπεία.



Εικόνα 2 - Κυτταρικός θάνατος που προκαλείται από καταστολείς του EFGR

2.3 Κυτταρικές σειρές

Η κυτταρική καλλιέργεια είναι η διαδικασία ανάπτυξης των κυττάρων (βακτήρια, ζυμομύκητες, ζωικά κύτταρα) σε ένα ελεγχόμενο περιβάλλον, πέραν του φυσικού τους, όπως ένα εργαστήριο (*NCI Dictionary of Cancer Terms*, n.d.-d). Έτσι μπορεί να μελετηθεί η βιολογία, η φυσιολογία, και ο μεταβολισμός των κυττάρων, όπως και να διερευνηθούν αλληλεπιδράσεις μεταξύ φυσιολογικών και παθολογικών κυττάρων (Segeritz & Vallier, 2017). Η αρχική καλλιέργεια που προέρχεται απευθείας από δείγματα ιστού ονομάζεται πρωτογενής και μοιάζει πολύ με τον αρχικό ιστό. Μόλις όμως τα κύτταρα του όγκου πολλαπλασιαστούν σε βαθμό που καλύπτουν τη φιάλη καλλιέργειας, τότε πρέπει να μεταφερθούν σε νέα φιάλη. Σε αυτό το σημείο η ονοματολογία αλλάζει και από πρωτογενή κύτταρα όγκου δημιουργείται η λεγόμενη κυτταρική σειρά. Αξίζει να σημειωθεί ότι υπάρχουν δύο κατηγορίες κυτταρικών σειρών: οι «ορισμένου χρόνου» που έχουν προκαθορισμένη διάρκεια ζωής, και οι «συνεχείς» που μπορούν και πολλαπλασιάζονται επ' αόριστον (Meditz & Rinner, 2018). Κατά κανόνα, οι κυτταρικές σειρές καρκίνου ανήκουν στην δεύτερη κατηγορία.

Οι κυτταρικές σειρές έχουν πολλά πλεονεκτήματα στην μελέτη απόκρισης στα φάρμακα, καθώς σε πολλές απ' αυτές παρέχονται ομικά (omic) δεδομένα μέσω δημόσιων βάσεων δεδομένων. Αυτές οι πληροφορίες επιτρέπουν τη γενετική και γονιδιωματική χαρτογράφηση, με το κύριο

κόστος να σχετίζεται με το πως αποκρίνονται τα κύτταρα σε φάρμακα. Γενικά, οι κυτταρικές σειρές προσφέρουν καλά ελεγχόμενα περιβάλλοντα, διευκολύνοντας τη μέτρηση διαφόρων φαινοτύπων (όπως κυτταροτοξικότητα, ρυθμό ανάπτυξης, αλλαγές γονιδιακής έκφρασης) με λιγότερους συγχυτικούς παράγοντες (confounding factors) σε σύγκριση με τα κλινικά δεδομένα (Niu & Wang, 2015), καθώς οι κλινικές δοκιμές συνήθως δεν είναι εστιασμένες στην γενετική ανάλυση και οι πολύπλοκες θεραπείες εμποδίζουν τη δυνατότητα να διερευνηθεί η απόκριση σε μεμονωμένο φάρμακο από πλευράς γενετικής. Επίσης, οι κυτταρικές σειρές δεν έχουν ιδιαίτερους περιορισμούς στο πλήθος των δειγμάτων, σε αντίθεση με τα κλινικά δεδομένα, και η ανάλυσή τους μπορεί εύκολα να αυτοματοποιηθεί μέσω προσεγγίσεων ρομποτικής (Jack et al., 2014).

Αν και τα πλεονεκτήματα στην φαρμακογονιδιωματική είναι σαφή, υπάρχουν και περιορισμοί. Η αξιολόγηση του μικροπεριβάλλοντος και η επίδραση της φαρμακοκινητικής του φαρμάκου στην κλινική απόκριση δεν μπορεί να εκτιμηθεί σε μοντέλα κυτταρικών σειρών. Επιπλέον, το προφίλ γονιδιακής έκφρασης σε μοντέλα κυτταρικών σειρών διαφέρει από αυτό που παρατηρείται στους πρωτογενείς ιστούς, δηλαδή μπορεί να μην εκφράζεται κάποιο σημαντικό γονίδιο για την ενεργοποίηση ενός φαρμάκου. Τέλος, οι συνθήκες κυτταροκαλλιέργειας έχουν τη δυνατότητα να προκαλέσουν νέες μεταλλάξεις και να αλλάξουν τα χαρακτηριστικά της κυτταρικής σειράς (Niu & Wang, 2015).

2.4 Προφίλ γονιδιακής έκφρασης

Το κεντρικό δόγμα της βιολογίας περιγράφει τη μέθοδο με την οποία λαμβάνονται πληροφορίες από τα γονίδια και χρησιμοποιούνται για τη δημιουργία πρωτεϊνών. Η μεταγραφή του DNA παράγει RNA και στη συνέχεια η μετάφραση του RNA δημιουργεί πρωτεΐνες. Αυτή η διαδικασία είναι γνωστή ως έκφραση γονιδίων και όλες οι μορφές ζωής τη χρησιμοποιούν για να δημιουργήσουν τα δομικά στοιχεία της ζωής από γενετικές πληροφορίες. Ένα κύτταρο εκφράζει μόνο ένα υποσύνολο των γονιδίων που περιέχει ανά πάσα στιγμή, δηλαδή το κύτταρο μπορεί να ερμηνεύσει το γενετικό του κώδικα με διαφορετικούς τρόπους. Ο έλεγχος των γονιδίων που εκφράζονται επιτρέπει στο κύτταρο να ελέγχει το μέγεθος, το σχήμα, και τις λειτουργίες του. Οι τρόποι με τους οποίους τα κύτταρα ενός οργανισμού εκφράζουν τα γονίδια που περιέχουν επηρεάζουν τον φαινότυπο του οργανισμού.

Το προφίλ γονιδιακής έκφρασης μετρά ποια γονίδια εκφράζονται σε ένα κύτταρο ανά πάσα στιγμή. Αυτή η μέθοδος μπορεί να μετρήσει χιλιάδες γονίδια ταυτόχρονα, ενώ ορισμένα πειράματα μπορούν να μετρήσουν ολόκληρο το γονιδίωμα ταυτόχρονα. Το προφίλ γονιδιακής έκφρασης μετρά τα επίπεδα mRNA, εμφανίζοντας το μοτίβο των γονιδίων που εκφράζονται από ένα κύτταρο σε επίπεδο μεταγραφής. Αυτό συχνά σημαίνει μέτρηση των σχετικών ποσοτήτων mRNA σε δύο ή περισσότερες πειραματικές συνθήκες και στη συνέχεια την εκτίμηση των συνθηκών που οδήγησαν στην έκφραση συγκεκριμένων γονιδίων. Το προφίλ γονιδιακής έκφρασης χρησιμοποιείται από διάφορους ερευνητές βιοϊατρικής· από μοριακούς βιολόγους έως περιβαλλοντικούς τοξικολόγους. Αυτή η τεχνολογία μπορεί να παρέχει ακριβείς πληροφορίες για την έκφραση γονιδίων, προς αμέτρητους πειραματικούς στόχους. Χρησιμοποιούνται διαφορετικές τεχνικές για τον προσδιορισμό της γονιδιακής έκφρασης. Αυτές περιλαμβάνουν μικροσυστοιχίες DNA και τεχνολογίες προσδιορισμού αλληλουχίας. Η πρώτη μετρά τη δραστηριότητα συγκεκριμένων γονιδίων ενδιαφέροντος και η δεύτερη επιτρέπει στους ερευνητές να προσδιορίσουν όλα τα ενεργά γονίδια σε ένα κύτταρο.

Μόλις προσδιοριστεί η αλληλουχία ενός γονιδιώματος, ξέρουμε τι δυνατότητες έχει ένα κύτταρο - ποια χαρακτηριστικά και ποια λειτουργία μπορεί να έχει - με βάση τα γονίδια που περιέχει. Ωστόσο, η αλληλούχηση του γονιδιώματος δεν μας λέει ποια γονίδια εκφράζει ένα κύτταρο ή τις λειτουργίες ή τις διεργασίες που εκτελεί σε κάθε δεδομένη στιγμή. Για να τα προσδιορίσουμε, πρέπει να επεξεργαστούμε το προφίλ γονιδιακής έκφρασης του. Εάν ένα γονίδιο χρησιμοποιείται για την παραγωγή mRNA, θεωρείται «ενεργό», εάν δεν χρησιμοποιείται για την παραγωγή

mRNA, θεωρείται «απενεργοποιημένο». Ένα προφίλ γονιδιακής έκφρασης δείχνει πώς λειτουργεί ένα κύτταρο σε μια συγκεκριμένη χρονική στιγμή. Αυτό οφείλεται στο ότι η έκφραση του γονιδίου του κυττάρου επηρεάζεται από εξωτερικά και εσωτερικά ερεθίσματα, συμπεριλαμβανομένου του αν το κύτταρο διαιρείται, ποιοι παράγοντες υπάρχουν στο περιβάλλον του κυττάρου, τα σήματα που λαμβάνει από άλλα κύτταρα, ακόμη και την ώρα της ημέρας.

Το προφίλ γονιδιακής έκφρασης δίνει τη δυνατότητα να διερευνηθούν οι επιπτώσεις διαφορετικών συνθηκών στη γονιδιακή έκφραση αλλάζοντας το περιβάλλον στο οποίο εκτίθεται το κύτταρο και προσδιορίζοντας ποια γονίδια εκφράζονται. Εναλλακτικά, εάν είναι ήδη γνωστό ότι ένα γονίδιο εμπλέκεται σε μια συγκεκριμένη κυτταρική συμπεριφορά, το προφίλ γονιδιακής έκφρασης βοηθά να προσδιοριστεί εάν ένα κύτταρο εκτελεί αυτή τη λειτουργία. Για παράδειγμα, ορισμένα γονίδια είναι γνωστό ότι εμπλέκονται στην κυτταρική διαίρεση· συνεπώς εάν αυτά τα γονίδια είναι ενεργά σε ένα κύτταρο, αποκαλύπτει αν το κύτταρο υφίσταται διαίρεση ή αν είναι διαφοροποιημένο. Το προφίλ γονιδιακής έκφρασης χρησιμοποιείται συχνά στη δημιουργία υποθέσεων. Εάν πολύ λίγα είναι γνωστά για το πότε και γιατί θα εκφραστεί ένα γονίδιο, το προφίλ έκφρασης κάτω από διαφορετικές συνθήκες μπορεί να βοηθήσει στο σχεδιασμό μιας υπόθεσης για δοκιμή σε μελλοντικά πειράματα. Για παράδειγμα, εάν το γονίδιο A εκφράζεται μόνο όταν το κύτταρο εκτίθεται σε άλλα κύτταρα, αυτό το γονίδιο μπορεί να εμπλέκεται στη διακυτταρική επικοινωνία. Περαιτέρω πειράματα θα μπορούσαν να καθορίσουν εάν αυτό συμβαίνει.

Με το προφίλ γονιδίου μπορεί επίσης να διερευνηθεί η επίδραση μορίων που μοιάζουν με φάρμακα στην κυτταρική απόκριση. Μπορούν να αναγνωριστούν οι γονιδιακοί δείκτες του μεταβολισμού των φαρμάκων ή να προσδιοριστούν εάν τα κύτταρα όταν εκτίθενται στο φάρμακο εκφράζουν γονίδια που είναι γνωστό ότι σχετίζονται με απόκριση σε τοξικά περιβάλλοντα. Το γονιδιακό προφίλ μπορεί επίσης να χρησιμοποιηθεί ως διαγνωστικό εργαλείο. Εάν τα καρκινικά κύτταρα εκφράζουν υψηλότερα επίπεδα ορισμένων γονιδίων και αυτά τα γονίδια κωδικοποιούν έναν υποδοχέα πρωτεΐνης, αυτός ο υποδοχέας μπορεί να εμπλέκεται στον καρκίνο και η στόχευσή του με ένα φάρμακο μπορεί να θεραπεύσει την ασθένεια. Το προφίλ γονιδιακής έκφρασης μπορεί τότε να είναι ένα βασικό διαγνωστικό εργαλείο για άτομα με αυτόν τον τύπο καρκίνου (*Introduction to Gene Expression Profiling | Thermo Fisher Scientific - IE, n.d.*).

2.5 Διαλογή υψηλής απόδοσης

Η διαλογή υψηλής απόδοσης (high throughput screening - HTS) είναι μια διαδικασία ανακάλυψης φαρμάκων που επιτρέπει την αυτοματοποιημένη δοκιμή μεγάλου αριθμού χημικών ή/και βιολογικών ενώσεων για έναν συγκεκριμένο βιολογικό στόχο, για παράδειγμα μέσω binding assays. Μέθοδοι HTS χρησιμοποιούνται ευρέως στη φαρμακευτική βιομηχανία, αξιοποιώντας τη ρομποτική και τον αυτοματισμό για τη γρήγορη εξέταση της βιολογικής ή βιοχημικής δραστηριότητας μεγάλου αριθμού μορίων, συνήθως φαρμάκων. Επιταχύνουν την ανάλυση στόχων, καθώς οι βιβλιοθήκες χημικών ενώσεων μεγάλης κλίμακας μπορούν γρήγορα να ελεγχθούν με αποδοτικό τρόπο. Η HTS είναι ένα χρήσιμο εργαλείο για την αξιολόγηση, για παράδειγμα, φαρμακολογικών στόχων, αγωνιστών φαρμακολογικού προφίλ και ανταγωνιστών για υποδοχείς (όπως GPCRs) και ένζυμα. Ο πρωταρχικός στόχος της HTS είναι να εντοπίσει μέσω της εξέτασης μιας βιβλιοθήκης χημικών στοιχείων, υποψηφίους που επηρεάζουν τον στόχο με τον επιθυμητό τρόπο, τους λεγόμενους «hits» ή «leads». Αυτό συνήθως επιτυγχάνεται χρησιμοποιώντας συσκευές χειρισμού υγρών, ρομποτική, συσκευές ανάγνωσης δίσκων ως ανιχνευτές και εξειδικευμένο λογισμικό για τον έλεγχο οργάνων και την επεξεργασία δεδομένων.

Είναι σημαντικό ότι οι διαδικασίες HTS συνήθως δεν προσδιορίζουν φάρμακα, καθώς αρκετές ιδιότητες που είναι κρίσιμες για την ανάπτυξη ενός νέου φαρμάκου δεν μπορούν να αξιολογηθούν με έλεγχο μιας βιβλιοθήκης χημικών ενώσεων. Για παράδειγμα, η HTS δεν μπορεί να αξιολογήσει την τοξικότητα και τη βιοδιαθεσιμότητα. Ο πρωταρχικός ρόλος των χημικών αναλύσεων υψηλής απόδοσης είναι αντίθετα να εντοπίζουν «δυσνητικά leads» και να παρέχουν προτάσεις για τη

βελτιστοποιήσή τους. Τα αποτελέσματα των αναλύσεων HTS παρέχουν επομένως το σημείο εκκίνησης για περαιτέρω βήματα στον αγωγό ανακάλυψης φαρμάκων, όπως ο σχεδιασμός φαρμάκων, καθώς και για την κατανόηση της αλληλεπίδρασης ή του ρόλου μιας συγκεκριμένης βιοχημικής διαδικασίας. Κατά συνέπεια, η HTS θα πρέπει να θεωρείται ως μια γρήγορη σάρωση βιολογικών διεργασιών μέσω της οποίας ενώσεις με κακή ή καθόλου επίδραση μπορούν να αποκλειστούν γρήγορα από περαιτέρω ανάλυση (Pusterla, 2023).

2.6 Φαρμακογονιδιοματική

Υπάρχουν σημαντικές διαφοροποιήσεις σε κάθε ασθενή για το πώς ο οργανισμός του αποκρίνεται σε φάρμακα, με πολλούς να μην αποκρίνονται, να αποκρίνονται μερικώς ή να παρουσιάζουν ανεπιθύμητες παρενέργειες, ακόμα και όταν χορηγείται η συνιστώμενη δοσολογία. Η συγκέντρωση ενός φαρμάκου στο πλάσμα μπορεί να ποικίλλει περισσότερο από 600 φορές μεταξύ ατόμων του ίδιου βάρους στην ίδια δοσολογία λόγω γενετικών, φυσιολογικών, παθοφυσιολογικών ή περιβαλλοντικών παραγόντων. Οι γενετικές διαφορές επηρεάζουν ιδιαίτερα την απορρόφηση, την κατανομή, τον μεταβολισμό και τις αλληλεπιδράσεις των στόχων ενός φαρμάκου. Γενικά, οι γενετικοί παράγοντες συμβάλλουν στο 15%-30% των διαφορών στον μεταβολισμό και στην απόκριση του φαρμάκου, αλλά για ορισμένα φάρμακα μπορούν να ευθύνονται για έως και το 95% της διαφοροποίησης (Eichelbaum et al., 2006).

Ο στόχος της φαρμακογονιδιοματικής είναι να βρεθούν οι γενετικές παραλλαγές που επηρεάζουν τις αποκρίσεις των φαρμάκων στους ασθενείς. Έτσι θα μπορούν να αναπτυχθούν διαγνωστικές μέθοδοι και θεραπείες προσαρμοσμένες σε ασθενείς με ελαχιστοποίηση των ανεπιθύμητων παρενεργειών και μεγιστοποίηση των ωφέλιμων ενεργειών (Housman & Ledley, 1998).

2.7 Ήμισυ της μέγιστης ανασταλτικής συγκέντρωσης - IC50

Από τα πιο διαδεδομένα *in vitro* μέτρα που χρησιμοποιούνται για την αξιολόγηση της θεραπευτικής δυνατότητας φαρμακολογικών παραγόντων είναι το ήμισυ της μέγιστης ανασταλτικής συγκέντρωσης, γνωστή και ως IC50. Υποδεικνύει πόση ποσότητα από έναν συγκεκριμένο φαρμακολογικό παράγοντα απαιτείται για να ανασταλεί στο μισό μια δεδομένη βιολογική δραστηριότητα (Aykul & Martínez-Hackert, 2016). Οι τιμές της συνήθως εκφράζονται σε μοριακή συγκέντρωση (mol/L ή M).

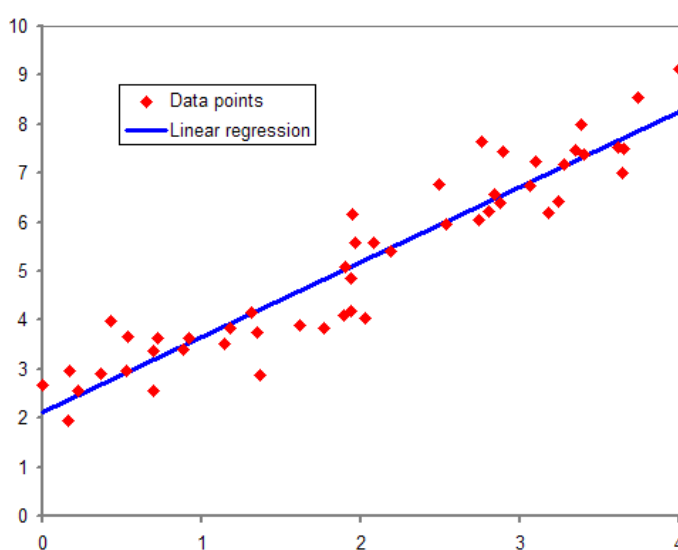
2.8 Απόκριση ασθενούς σε φάρμακο

Η πρόβλεψη της απόκρισης καρκινοπαθών σε φαρμακευτική αγωγή είναι υψίστης σημασίας για να έχουμε αποτελεσματική ιατρική. Συνήθως λαμβάνονται υπόψη τα αποτελέσματα από κλινικές μελέτες. Ωστόσο, συχνά αυτό είναι δύσκολο εξαιτίας ηθικών περιορισμών ή/και μεγάλου κόστους. Μια εναλλακτική προσέγγιση είναι η φαρμακογονιδιοματική σε κυτταρικές σειρές καρκίνου. Στο πλαίσιο αυτό έχουν κατασκευαστεί μοντέλα πρόβλεψης της απόκρισης του ασθενούς σε διάφορα φάρμακα, συγκρίνοντας τα γονιδιακά προφίλ μεταξύ ασθενών και της κυτταρικές σειρές που συσχετίζονται με το φάρμακο (Park et al., 2022). Με βάση αυτήν την προσέγγιση, χρησιμοποιούνται σύνολα δεδομένων φαρμακογονιδιοματικής, όπως είναι το Cancer Cell Line Encyclopedia (CCLE) και το Genomics of Drug Sensitivity in Cancer (GDSC), για την εκπαίδευση μοντέλων πρόβλεψης. Το γονιδιακό προφίλ μιας κυτταρικής σειράς καρκίνου χρησιμοποιείται ως δεδομένο εισόδου για την πρόβλεψη της απόκρισης φαρμάκου (Park et al., 2023), (Liu et al., 2024). Ωστόσο, οι περισσότερες έρευνες είναι επικεντρωμένες σε μοντέλα μηχανικής μάθησης που καταλήγουν στην πρόβλεψη της απόκρισης στη φαρμακευτική αγωγή με βάση τα γονιδιακά χαρακτηριστικά, χωρίς να αναδεικνύουν τις εξαρτήσεις των αποτελεσμάτων της ταξινόμησης με τα χαρακτηριστικά αυτά. Η εφαρμογή μοντέλων παλινδρόμησης (regression) επιτρέπει την ανάδειξη τέτοιων εξαρτήσεων. Ειδικότερα, η παρούσα διπλωματική εργασία διερευνά την εφαρμογή μοντέλων συμβολικής παλινδρόμησης (symbolic regression), που

επιτρέπει την ανακάλυψη βέλτιστων αναλυτικών εκφράσεων εισόδου-εξόδου με βάση τα δεδομένα, αντί της επιβολής κάποιας αυθαίρετης σχέσης εισόδου-εξόδου.

2.9 Ανάλυση παλινδρόμησης

Η ανάλυση παλινδρόμησης (regression analysis) είναι μια στατιστική μέθοδος που χρησιμοποιείται για τον προσδιορισμό της σχέσης μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών, που συχνά ονομάζονται δείκτες πρόβλεψης ή χαρακτηριστικά, και μιας εξαρτημένης μεταβλητής, η οποία μπορεί να ονομάζεται μεταβλητή απόκρισης. Στόχος της είναι να κατανοήσει πώς επηρεάζεται η εξαρτημένη μεταβλητή (συχνά συμβολίζεται ως y) καθώς μεταβάλλονται οι ανεξάρτητες μεταβλητές (συχνά συμβολίζονται ως $x_1, x_2, \dots, x_n$). Μέσω της ανάλυσης παλινδρόμησης, μπορεί κανείς να μοντελοποιήσει και να ποσοτικοποιήσει αυτή τη σχέση, επιτρέποντας προβλέψεις και συμπεράσματα για την εξαρτημένη μεταβλητή με βάση τις τιμές των ανεξάρτητων μεταβλητών.



Εικόνα 3 - Ευθεία παλινδρόμησης για 50 τυχαία σημεία που προήλθαν από μία κανονική κατανομή. Οι παρατηρήσεις (ερυθρό χρώμα) αντιμετωπίζονται ως αποκλίσεις από την υποκείμενη σχέση (μπλε χρώμα) μεταξύ της εξαρτημένης με την ανεξάρτητη μεταβλητή.

Αυτό γίνεται προσαρμόζοντας τα δεδομένα σε μία προκαθορισμένη μορφή συνάρτησης όπως είναι η $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \varepsilon$. Αυτή είναι η μορφή της πολλαπλής γραμμικής παλινδρόμησης, όπου η y εξαρτάται γραμμικά από τα ανεξάρτητα χαρακτηριστικά x_i με βάρη ή συντελεστές β_i , τον όρο β_0 που είναι το σημείο τομής της ευθείας με τον άξονα y , και τον όρο σφάλματος ε του μοντέλου. Μετέπειτα με προέκταση (extrapolation) ή παρεμβολή (interpolation) μπορεί να γίνει πρόβλεψη της y βάσει των x_i και των συντελεστών β_i . Η συνάρτηση που προκύπτει μπορεί να συνδράμει στην κατανόηση των σχέσεων μεταξύ των δεδομένων. Το διάγραμμα μίας απλουστευμένης περίπτωσης με $\beta_0 = 2$, $\beta_1 = 1.5$, λοιπά $\beta_i = 0$, και $\varepsilon = 0$ (για λόγους απλότητας) παρουσιάζεται στην Εικόνα 3. Ωστόσο, τα μοντέλα παλινδρόμησης μπορούν επίσης να λάβουν μη γραμμικές μορφές, επιτρέποντας την αποτύπωση πιο πολύπλοκων σχέσεων.

2.10 Συμβολική παλινδρόμηση

Ο περιορισμός των μοντέλων παλινδρόμησης είναι ότι συνήθως τα δεδομένα προσαρμόζονται σε μία προκαθορισμένη μορφή συνάρτησης, με σκοπό να βρεθούν οι άγνωστοι συντελεστές. Απεναντίας, η συμβολική παλινδρόμηση μπορεί να κατασκευάσει αναλυτικές εκφράσεις απευθείας από τα δεδομένα, μειώνοντας σημαντικά την ανάγκη πρότερης γνώσης του πεδίου εφαρμογής (Angelis et al., 2023). Παραδοσιακά, αντιμετωπίζεται με τεχνικές συνδυαστικής

βελτιστοποίησης (combinatorial optimization), όπως είναι οι γενετικοί αλγόριθμοι, κατά τους οποίους οι πιθανές λύσεις εξελίσσονται σε «γενιές». Ωστόσο, οι γενετικοί αλγόριθμοι είναι ευαίσθητοι στις αρχικές συνθήκες, μπορεί να παράγουν πολύπλοκες συναρτήσεις εξόδου που δυσχεραίνουν την κατανόηση των σχέσεων μεταξύ των δεδομένων, και έχουν υψηλές υπολογιστικές απαιτήσεις (Jin et al., 2019). Γι' αυτούς τους λόγους υπάρχουν πολλές εναλλακτικές μέθοδοι που θεραπεύουν αυτά τα προβλήματα. Ανάμεσά τους και ποικίλες υλοποιήσεις συμβολικής παλινδρόμησης (Arnaldo et al., 2014), (La Cava et al., 2016), (Virgolin et al., 2017), (William et al., 2018), (Virgolin et al., 2019), (Burlacu et al., 2020).

Το σημαντικότερο πλεονέκτημα που έχουν οι αλγόριθμοι συμβολικής παλινδρόμησης είναι η εν γένει δυνατότητα ερμηνείας του μοντέλου στη μορφή μίας συνάρτησης, σε αντίθεση με άλλα ερμηνεύσιμα μοντέλα που απλώς δίνουν μία ποσοτική τιμή για την σημαντικότητα των χαρακτηριστικών. Αν και τα νευρωνικά δίκτυα και άλλα μοντέλα μπορούν να οδηγήσουν σε αναλυτικές αναπαραστάσεις, η συμβολική παλινδρόμηση συνήθως στοχεύει στην εύρεση απλών συναρτήσεων, συχνά περιορισμένων σε πολυωνυμικές μορφές. Επομένως, μπορεί να χρησιμοποιηθεί για την ανακάλυψη ερμηνεύσιμων λειτουργικών σχέσεων από δεδομένα (Marcinkevičs & Vogt, 2023).

2.11 Ορισμός προβλήματος SR

Δοθέντος ενός συνόλου δεδομένων $D = \{(x_i, y_i)\}_{i=1}^n$, όπου $x_i \in \mathbb{R}^d$ είναι το διάνυσμα εισόδου με πλήθος d χαρακτηριστικών και $y_i \in \mathbb{R}$ είναι η μονόμετρη έξοδος, αναζητείται μία συνάρτηση f που να μπορεί να προβλέψει την έξοδο y από την είσοδο x . Γι' αυτό ορίζεται μία κλάση συναρτήσεων F που περιλαμβάνει όλες τις πιθανές συναρτήσεις που αντιστοιχίζουν τις εισόδους στην έξοδο $f: \mathbb{R}^d \rightarrow \mathbb{R}$. Αυτές αναπαρίστανται ως συμβολικές εκφράσεις. Για την αξιολόγηση του πόσο καλά η συνάρτηση f προβλέπει την μεταβλητή έξοδος, χρησιμοποιείται μία συνάρτηση απώλειας $l(f) = \sum_{i=1}^n l(f(x_i), y_i)$. Για κάθε υπονήφια συνάρτηση f της κλάσης F , η συνάρτηση απώλειας υπολογίζει το άθροισμα της διαφοράς μεταξύ των προβλεπόμενων εξόδων $f(x_i)$ και των πραγματικών εξόδων y_i σε όλα τα δεδομένα. Μία συνήθης επιλογή για συνάρτηση απώλειας είναι το άθροισμα των τετραγώνων της διαφοράς μεταξύ προβλεπόμενων εξόδων και πραγματικών εξόδων ή $l(f) = \sum_i (y_i - f(x_i))^2$. Ωστόσο, πρέπει να επιλεγεί η βέλτιστη συνάρτηση f^* της κλάσης F που ελαχιστοποιεί την συνάρτηση απώλειας, $f^* = \arg_{f \in F} \min l(f)$ (Makke & Chawla, 2024).

Στον αντίποδα της τυπικής παλινδρόμησης, κατά την οποία η κλάση συναρτήσεων F είναι συνεχής, όπως είναι για παράδειγμα όλες οι γραμμικές συναρτήσεις στην γραμμική παλινδρόμηση, η συμβολική παλινδρόμηση χρησιμοποιεί ένα διακριτό σύνολο συναρτήσεων. Αυτό περιλαμβάνει διαφορετικές συμβολικές εκφράσεις όπως πολυώνυμα, λογαρίθμους, τριγωνομετρικές συναρτήσεις κ.λπ. Οι διαφορετικές προσεγγίσεις στο πρόβλημα της συμβολικής παλινδρόμησης εστιάζουν στην εξερεύνηση και στην αναπαράσταση του διακριτού συνόλου των πιθανών συναρτήσεων όπως και στην επιλογή της βέλτιστης από αυτές.

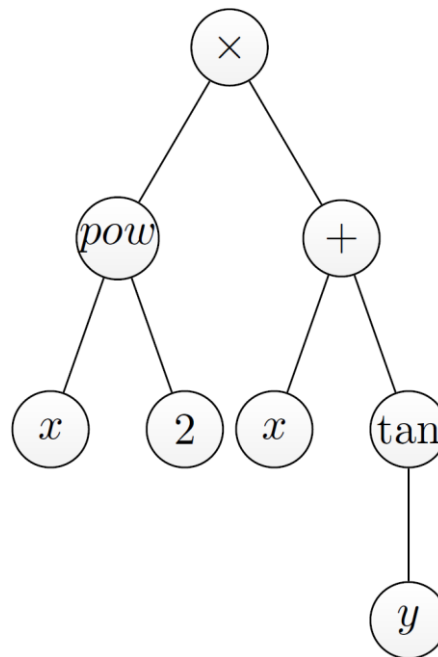
2.11.1 Ορισμός κλάσης συναρτήσεων F

Έστω P ένα σύνολο συναρτήσεων και μεταβλητών. Ο χώρος αναζήτησης της συμβολικής παλινδρόμησης F , περιλαμβάνει όλες τις συναρτήσεις που μπορούν να δημιουργηθούν συνδυάζοντας τα στοιχεία του P . Το σύνολο P μπορεί να περιλαμβάνει βασικές αλγεβρικές πράξεις· όπως πρόσθεση, αφαίρεση, πολλαπλασιασμός, και διαίρεση· υπερβατικές συναρτήσεις· όπως ημίτονο, συνημίτονο, λογάριθμο, και εκθετικά· σταθερούς αριθμούς· μεταβλητές του προβλήματος· όπως $x_1, x_2, \dots$ (Virgolin & Pissis, 2022).

Εφόσον έχει οριστεί το P , τότε καθορίζεται και η κλάση συνάρτησης F . Για παράδειγμα αν το $P = \{id(\cdot), add(\cdot, \cdot), sub(\cdot, \cdot), mul(\cdot, \cdot), \mathcal{R}\}$, όπου η συνάρτηση ταυτότητας $id(\cdot)$ περιλαμβάνει όλα τα πιθανά στοιχεία εισόδου $x_1, x_2, \dots$ και το $\mathcal{R}$ είναι ένα ειδικό στοιχείο του P που αναπαριστά μία κατανομή πιθανότητας από την οποία δειγματοληπτούνται σταθερές, τότε η F θα περιλαμβάνει όλα τα πιθανά πολυώνυμα οποιουδήποτε βαθμού σε όλες τις μεταβλητές $x_1, x_2, \dots$.

2.11.2 Αναπαράσταση έκφρασης

Οι συμβολικές εκφράσεις μπορούν να υλοποιηθούν ως δέντρο έκφρασης, μία δομή δεδομένων δέντρου που μπορεί να αναπαραστήσει μαθηματικές εκφράσεις. Κάθε κόμβος στο δέντρο αναπαριστά έναν τελεστή, μία συνάρτηση, μία μεταβλητή, ή μία σταθερά. Οι κόμβοι που αναπαριστούν τελεστές ή συναρτήσεις πρέπει να έχουν θυγατρικούς κόμβους που αντιστοιχούν σε κάθε απαιτούμενη παράμετρο, ενώ οι κόμβοι που περιέχουν μεταβλητές ή σταθερές είναι τερματικοί κόμβοι. Για παράδειγμα η έκφραση $x^2 \cdot (x + \tan(y))$ μπορεί να αναπαρασταθεί από το δέντρο της Εικόνα 4. Η απλότητα της έκφρασης μπορεί να αξιολογηθεί επιβάλλοντας ποινή στο bloatness, το μέγεθος του δέντρου (είτε ως πλήθος κόμβων είτε ως ύψος), λαμβάνοντας υπόψιν το πλήθος των μη γραμμικών συναρτήσεων που χρησιμοποιήθηκαν, την πολυπλοκότητα των σύνθετων συναρτήσεων, και το συνολικό πλήθος μεταβλητών (De Franca, 2018).



Εικόνα 4 - Δέντρο έκφρασης για την $x^2 \cdot (x + \tan(y))$

2.12 pySR

Ο pySR είναι ένας εξελικτικός αλγόριθμος συμβολικής παλινδρόμησης, ο οποίος διαθέτει πολλαπλούς πληθυσμούς που εξελίσσονται ασύγχρονα και ανεξάρτητα. Ο τρόπος που λειτουργεί είναι ο παρακάτω:

1. Αρχικοποίηση: Εκκίνηση με έναν πληθυσμό λύσεων (δηλαδή δέντρων έκφρασης), μία συνάρτηση καταλληλότητας (fitness function) και ένα σύνολο τελεστών μετάλλαξης. Η συνάρτηση καταλληλότητας αξιολογεί τα δέντρα έκφρασης και είναι προσαρμοσμένη στο πρόβλημα προς επίλυση.

2. Επιλογή: Παράλληλα, με κάθε διαδοχική γενιά, επιλέγεται ένα υποσύνολο του πληθυσμού που θα προχωρήσει στην επόμενη γενιά μέσω της διαδικασίας tournament selection.
3. Εξέλιξη: Γίνεται αντιγραφή του βέλτιστου δέντρου έκφρασης, και πάνω του εφαρμόζεται κάποιος τελεστής μετάλλαξης. Ύστερα αντικαθίσταται η γηραιότερη λύση με την μεταλλαγμένη βέλτιστη.

Εφαρμόζεται επίσης προσομοιωμένη ανόπτηση (simulated annealing) ώστε να απορριφθούν μεταλλάξεις βάσει κάποιας πιθανότητας που βασίζεται σε μία παράμετρο θερμοκρασίας, με σκοπό την εξερεύνηση διαφορετικών λύσεων, εστιάζοντας στα πιο υποσχόμενα δέντρα. Ωστόσο, αυτή η δυνατότητα δεν εφαρμόστηκε στην παρούσα εργασία. Ακόμη, στη διαδικασία tournament selection προστέθηκε ένα βήμα απλοποίησης των εξισώσεων που παράγονται μέσω μια σειράς αλγεβρικών ισοδυναμιών μαζί με εφαρμογή αλγορίθμου βελτιστοποίησης BFGS (Broyden-Fletcher-Goldfarb-Shanno), κάνοντας τα παραγόμενα δέντρα μετάλλαξης πιο κατανοητά στον άνθρωπο. Τέλος, το pySR υποστηρίζει την προσαρμοστική «φειδωλότητα» (adaptive parsimony). Στο pySR η απλότητα της έκφρασης εκφράζεται ως το πλήθος των κόμβων που περιλαμβάνονται στο δέντρο λύσης, ωστόσο η ποινή για το bloatness αντί να είναι σταθερή, είναι μία συνάρτηση της συχνότητας των εκφράσεων, αλλά και του πόσο πρόσφατες είναι. Έτσι προάγεται η αναζήτηση απλών και πολύπλοκων λύσεων (Cranmer, 2023).

2.13 Feyn (QLattice)

Το Feyn είναι ένα επιβλεπόμενο εργαλείο συμβολικής παλινδρόμησης, που λειτουργεί μέσω της μηχανής προσομοίωσης QLattice (Broløs et al., 2021). Ο ακριβής τρόπος που λειτουργεί δεν είναι δημοσίως γνωστός, καθώς είναι λογισμικό κλειστού κώδικα, ωστόσο είναι διαθέσιμη μία γενική περιγραφή του αλγορίθμου (*What Is the QLattice?* · Feyn Documentation, n.d.). Αρχικά, επιλέγεται μία ομάδα μοντέλων που εκπαιδεύονται μέσω οπισθοδιάδοσης. Έπειτα, τα μοντέλα αξιολογούνται και ύστερα από πολλαπλούς γύρους εκπαίδευσης κατατάσσονται από το καλύτερο στο χειρότερο βάσει συναρτήσεων σφάλματος. Εκείνα που θεωρούνται χειρότερα, απορρίπτονται. Η διαδικασία αυτή συνεχίζεται αναδρομικά, κρατώντας τα καλύτερα μοντέλα και προσθέτοντας νέα ομάδα μοντέλων, τα οποία εκπαιδεύονται ξανά, αξιολογούνται κ.ο.κ. Ως βιβλιοθήκη συναρτήσεων χρησιμοποιείται η Εξ. 1.

$$P = \{id(\cdot), add(\cdot, \cdot), sub(\cdot, \cdot), mul(\cdot, \cdot), tanh(\cdot), exp(\cdot), log(\cdot), inv(\cdot), \mathcal{R}\} \quad \text{Εξ. 1}$$

2.14 Fast Function Extraction (FFX)

Ο Fast Function Extraction δημοσιεύτηκε το 2011 (McConaghy, 2011) και έχει δοκιμαστεί σε πραγματικά δεδομένα με πλήθος χαρακτηριστικών από 13 έως περίπου 1500. Βασίζεται στην προσέγγιση pathwise regularized learning που διατηρεί μερικά μοντέλα από ένα μεγάλο σύνολο υποψήφιων μοντέλων. Το πρώτο βήμα του αλγορίθμου είναι να δημιουργήσει ένα μεγάλο σύνολο γραμμικών, αλλά και μη-γραμμικών, συναρτήσεων ως βάση για την επιλογή του μοντέλου πρόβλεψης. Το δεύτερο βήμα είναι προσδιορίσει τους συντελεστές αυτών των συναρτήσεων έναντι του χαρακτηριστικού πρόβλεψης, χρησιμοποιώντας pathwise regularized learning. Το τρίτο και τελευταίο βήμα είναι να εφαρμόσει μη-κυριαρχούμενο φίλτρο (non-dominated filter) στις συναρτήσεις βάσης βάσει του σφάλματος στην εκπαίδευση, δηλαδή ταξινομεί τα μοντέλα ώστε να βρεθεί αυτό που θεωρείται καλύτερο από άποψης πολυπλοκότητας και απόδοσης.

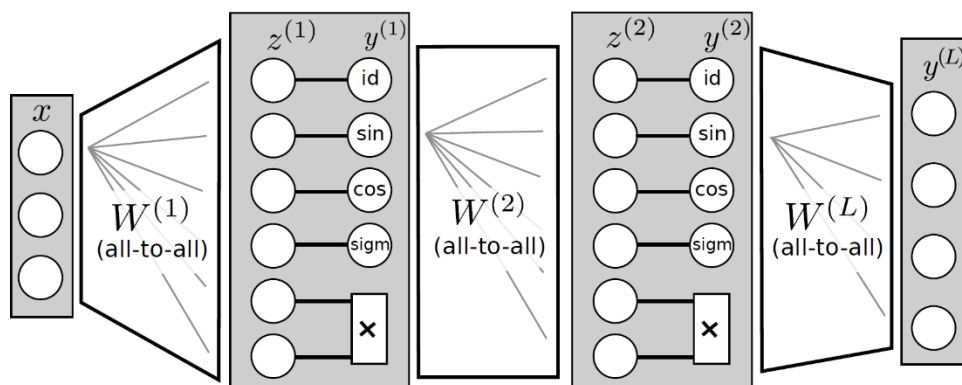
2.15 Interaction-Transformation Evolutionary Algorithm (ITEA)

Ο ITEA (Olivetti & Aldeia, 2019) είναι ένας αλγόριθμος συμβολικής παλινδρόμησης. Με έμπνευση την αναπαράσταση των πιθανών συμβολικών εκφράσεων ως δέντρου έκφρασης (expression tree), οι συγγραφείς εφαρμόζουν κάποιες βελτιώσεις. Συγκεκριμένα για να αντιμετωπίσει ο ITEA το πρόβλημα των πληθωρικών εκφράσεων που καθιστούν αδύνατη την

ερμηνεία στα δέντρα παράστασης, εφαρμόζουν μία ευριστική αναζήτηση στην αναπαράσταση Interaction-Transformation την οποία έχουν προτείνει στο παρελθόν (De Franca, 2018).

2.16 Equation Learner (EQL)

Ο σκοπός δημιουργίας του EQL είναι η δημιουργία ενός δικτύου που να επιτρέπει την εκμάθηση συμβολικών εκφράσεων αλλά παράλληλα να έχει τη δυνατότητα προέκτασης (extrapolation) σε απρόσμενα δεδομένα, για το σύνολο εκπαίδευσης, και μία στρατηγική για την επιλογή του κατάλληλου μοντέλου προέκτασης. Η αρχιτεκτονική του (Εικόνα 5) είναι ένα υπερσύνολο των κλασικών νευρωνικών δικτύων πρόσθιας τροφοδότησης με τη διαφορά ότι πέρα από τη σιγμοειδή συνάρτηση ως συνάρτηση ενεργοποίησης, χρησιμοποιεί επίσης τον πολλαπλασιασμό, το ημίτονο και το συνημίτονο ως μη-γραμμικά μέρη. Αυτές οι τριγωνομετρικές συναρτήσεις μπορούν να μοντελοποιήσουν τυχόν περιοδικότητα στα δεδομένα, ενώ ο πολλαπλασιασμός επιτρέπει στο δίκτυο να μάθει με μεγαλύτερη ακρίβεια φαινόμενα που είναι συχνά στον φυσικό κόσμο και έτσι προάγεται η εκμάθηση εξισώσεων ενός φυσικού συστήματος. Επίσης, με την προσθήκη της συνάρτησης ταυτότητας, κάθε νευρώνας έχει τη δυνατότητα να μην εφαρμόσει κάποια μη-γραμμικότητα στην είσοδο εάν δεν απαιτείται και έτσι να μπορεί δυνητικά να απλοποιηθεί η έξοδος. Το πιο απλό μοντέλο είναι εκείνο που εφαρμόζει τη συνάρτηση ταυτότητας τις περισσότερες φορές. Τα μοντέλα ταξινομούνται βάσει του σφάλματος στο σύνολο ελέγχου με την L2 νόρμα και προτιμάται εκείνο που έχει το μικρότερο σφάλμα αλλά και εκείνο που είναι αρκετά αραιό, δηλαδή εκείνο που έχει τους λιγότερους ενεργούς κρυμμένους νευρώνες.



Εικόνα 5 – EQL αρχιτεκτονική για 3 επίπεδα και έναν νευρώνα ανά τύπο

3. Μεθοδολογία

Στην παρούσα ενότητα παρουσιάζεται αναλυτικά ο τρόπος δημιουργίας των μοντέλων συμβολικής παλινδρόμησης.

3.1 Περιγραφή και προεπεξεργασία δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία αυτή είναι αυτά που χρησιμοποίησαν οι Park et al. (2023). Το σύνολο δεδομένων της γονιδιακής έκφρασης των κυτταρικών σειρών προήλθε από την Cancer Cell Line Encyclopedia (CCLE), όπως έχει δημοσιευτεί στις 17 Οκτωβρίου 2012: (https://data.broadinstitute.org/ccle_legacy_data/mRNA_expression/CCLE_Expression_Entrez_2012-09-29.gct) σε μορφή .gct. Ένα τέτοιο αρχείο περιλαμβάνει τις τιμές πλήθους γονιδίων και κυτταρικών σειρών, τις τιμές γονιδιακής έκφρασης, καθώς και την πληροφορία ονομάτων των κυτταρικών σειρών και των γονιδίων, όπως φαίνεται στην Εικόνα 6.

of samples

Third column onwards are sample names. These must be UNIQUE

Always "#1.2"

The # of rows (i.e probe sets)

Data starts on line 4

Column 1: Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE

Column 2: Row descriptions. Ignored by the program - can be dummy values (e.g. "na")

Each column contains expression values from 1 sample. Missing values are allowed (leave empty).

If editing, in Excel, make sure to save your data as "tab delimited text"

A	B	C	D	E	F	G
#1.2						
1000		130				
NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HSC	280.53	271.48	113.57	124.91	124.91
1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
117_at	X51757 /FEATURE=cds /DEFINITION=HSP70	51.27	61.12	24.1	41.44	43.56
121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
1255_g_at	L36861 /FEATURE=expanded_cds /DEFINITI	88.45	12.94	18.46	29.96	39
1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
1316_at	X55005 /FEATURE=mRNA /DEFINITION=HSC	106.87	45.11	30.05	46.65	36.5
1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTR	58.49	27.95	17.6	27.87	26.52
1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
1438_at	X75208 /FEATURE=cds /DEFINITION=HSPTR	80.87	9.77	15.33	11.18	44.59
1487_at	L38487 /FEATURE=mRNA /DEFINITION=HUI	64.26	80.61	102.9	59.77	105.72
1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
160020_at	Z48481 /FEATURE=cds /DEFINITION=HSM	411.94	171.16	130	234.76	266.96
1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
1773_at	L00635 /FEATURE= /DEFINITION=HUMFPT	62.82	45.96	41.15	23.1	28.41
177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	193.53

Εικόνα 6 - Περιγραφή αρχείου δεδομένων .gcd (Data Formats - GeneSetEnrichmentAnalysisWiki, n.d.)

Στην περίπτωση μας, τα δεδομένα γονιδιακής έκφρασης είναι ένας πίνακας δεκαδικών αριθμών μεγέθους 18989 γονιδίων $\times$ 1037 κυτταρικών σειρών που αποτυπώνει την έκφραση του εκάστοτε γονιδίου στις διαθέσιμες κυτταρικές σειρές. Ένα δείγμα του αρχείου με αυτές τις πληροφορίες φαίνεται στον Πίνακα 1. Η πληροφορία του ονόματος των γονιδίων εξάγεται από τη δεύτερη στήλη, Description, ωστόσο υπάρχουν 87 κενές τιμές ή $87/18988 = 0.45818\%$ ονομασίες γονιδίων που λείπουν. Επομένως, χρησιμοποιήθηκε η τιμή του probe set ID (μοναδικό αναγνωριστικό του συνόλου ανιχνευτών που αντιστοιχεί στο κάθε γονίδιο) ως υποκατάστατο της ονομασίας γονιδίου για να μπορεί να εκτελεστεί ορθά το μοντέλο. Επίσης, παρατηρήθηκε ότι το γονίδιο TTL εμφανίζεται δύο φορές με probe set ID 150465_at και probe set ID 646982_at, όπου στην πρώτη περίπτωση μετονομάστηκε σε *TTL_1* και στην δεύτερη σε *TT_2* αντίστοιχα, ώστε να μην υπάρχουν δύο ίδια ονόματα χαρακτηριστικών με διαφορετικές τιμές έκφρασης γονιδίων. Δεν έγινε περαιτέρω ανάλυση αυτών των μη αναμενόμενων εγγραφών, ούτε αποκλείστηκαν από την μελέτη, ώστε τα αποτελέσματα της παρούσας εργασίας να μπορούν να συγκριθούν με τα αποτελέσματα των Park et al. (2023). Επιπλέον, να σημειωθεί ότι τα συγκεκριμένα χαρακτηριστικά δεν εμφανίστηκαν στις συναρτήσεις που προέκυψαν στα παραγόμενα μοντέλα συμβολικής παλινδρόμησης. Τέλος, καθώς τα μοντέλα συμβολικής παλινδρόμησης που χρησιμοποιήθηκαν αξιοποιούν τη βιβλιοθήκη Python module SymPy για την μαθηματική αναπαράσταση των εξισώσεων, έγινε σχετική επεξεργασία των ονομάτων γονιδίων ώστε να εξαλειφθούν οι απαγορευμένοι για το SymPy χαρακτήρες. Ειδικότερα, έγινε αντικατάσταση του χαρακτήρα της παύλας, με δύο διαδοχικές κάτω παύλες, της τελείας με τρεις διαδοχικές κάτω παύλες, και του χαρακτήρα slash «/» με τέσσερις διαδοχικές κάτω παύλες.

Εφαρμόστηκε κανονικοποίηση Z-score Εξ. 2 των τιμών γονιδιακής έκφρασης χρησιμοποιώντας την βιβλιοθήκη scikit-learn της Python, ώστε να έχουν μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση.

$$z = \frac{x - \mu}{\sigma} \quad \text{Εξ. 2}$$

όπου z είναι η κανονικοποιημένη τιμή γονιδιακής έκφρασης, μ είναι η μέση τιμή του των τιμών γονιδιακής έκφρασης, σ είναι η τυπική απόκλιση των τιμών γονιδιακής έκφρασης και x είναι η εκάστοτε τιμή γονιδιακής έκφρασης.

Name	Description	LN18_ CENTRAL_ NERVOUS_ SYSTEM	...	CH157MN_ CENTRAL_ NERVOUS_ SYSTEM
100009676_at	LOC100009676	5.987545	...	5.720593
10000_at	AKT3	6.230233	...	5.757825
⋮	⋮	⋮	⋮	⋮
AFFX-TrpnX-M_at		3.418502	...	3.681598

Πίνακας 1 - Δείγμα συνόλου δεδομένων γονιδιακής έκφρασης σε καρκινικές κυτταρικές σειρές

Οι μετρήσεις της απόκρισης κυτταρικής σειράς σε φάρμακο είναι διαθέσιμες επίσης από την CCLE (δημοσιευμένες στις 24 Φεβρουαρίου 2015) και είναι διαθέσιμες εδώ: https://data.broadinstitute.org/ccle_legacy_data/pharmacological_profiling/CCLE_NP24.2009_Drug_data_2015.02.24.csv σε μορφή .csv. Περιλαμβάνουν, μεταξύ άλλων, το όνομα της κυτταρικής σειράς στην CCLE, το όνομα του φαρμάκου που εφαρμόστηκε, το όνομα του γονιδίου-στόχου που καταστέλλεται, και την τιμή IC(50) σε μM . Ένα δείγμα του αρχείου με αυτές τις πληροφορίες φαίνεται στον Πίνακα 2.

CCLE Cell Line Name	...	Compound	Target	...	IC50 (μM)	...
1321N1_CENTRAL_NERVOUS_SYSTEM	...	AEW541	IGF1R	...	8	...
22RV1_PROSTATE	...	AEW541	IGF1R	...	2.329924107	...
⋮	...	⋮	⋮	...	⋮	...
ZR7530_BREAST	...	Erlotinib	EGFR	...	8	...

Πίνακας 2 - Δείγμα συνόλου δεδομένων απόκρισης φαρμάκων σε κυτταρικές σειρές

Πιο αναλυτικά, τα δεδομένα αποτελούνται από 11670 κυτταρικές σειρές, από τις οποίες οι 504 είναι μοναδικές, και 24 αντικαρκινικά φάρμακα. Η τιμή των IC50(μM) κανονικοποιήθηκε σε $\ln(\text{IC50}(\mu\text{M}))$. Τα δεδομένα χωρίστηκαν σε 24 σύνολα, ένα για κάθε φάρμακο και στη συνέχεια σε σύνολο εκπαίδευσης και σύνολο ελέγχου, όμοια με τους Park et al. (2023) (Πίνακας 3). Επισημαίνεται ότι για κάθε φάρμακο, μέσω του εκάστοτε CCLE Cell Line Name έγινε αντιστοίχιση του IC50 με την συγκέντρωση των γονιδίων του συνόλου δεδομένων γονιδιακής έκφρασης. Έτσι, επιτεύχθηκε συσχέτιση της γονιδιακής έκφρασης των 18989 γονιδίων σε κάθε διαθέσιμη κυτταρική σειρά καρκίνου με το IC50 των 24 διαθέσιμων φαρμάκων και δημιουργήθηκαν 24 διαφορετικά μοντέλα, ένα για κάθε φάρμακο. Το σύνολο των μοναδικών CCLE Cell Line Name δεν είναι γνήσιο υποσύνολο των κυτταρικών σειρών του συνόλου δεδομένων γονιδιακής έκφρασης. Γι' αυτό για παράδειγμα ενώ υπάρχουν 500 κυτταρικές σειρές που ελέγχθηκε η απόκρισή τους στο φάρμακο Panobinostat, το πλήθος των δειγμάτων που χρησιμοποιήθηκαν στο μοντέλο είναι $390 + 97 = 487$, καθώς για τις υπόλοιπες 13 κυτταρικές σειρές δεν είχαμε δεδομένα γονιδιακής έκφρασης. Παρόμοιες αποκλίσεις υπάρχουν σε όλα τα φάρμακα.

Φάρμακο	Πλήθος δειγμάτων συνόλου εκπαίδευσης	Πλήθος δειγμάτων συνόλου ελέγχου
AEW541	392	98

Nilotinib	326	81
17-AAG	392	98
PHA-665752	392	98
Lapatinib	393	98
Nutlin-3	393	98
AZD0530	393	98
PF2341066	393	98
L-685458	383	95
ZD-6474	387	96
Panobinostat	390	97
Sorafenib	392	98
Irinotecan	245	61
Topotecan	393	98
LBW242	392	98
PD-0325901	393	98
PD-0332991	337	84
Paclitaxel	392	98
PLX4720	387	96
RAF265	358	89
TAE684	393	98
TKI258	392	98
Erlotinib	392	98
AZD6244	392	98

Πίνακας 3 - Πλήθος δειγμάτων συνόλου εκπαίδευσης και ελέγχου ανά φάρμακο

3.2 Μετρικές αξιολόγησης

Για την αξιολόγηση και σύγκριση των μοντέλων χρησιμοποιήθηκαν οι μετρικές ρίζας του μέσου τετραγωνικού σφάλματος (root mean squared error - RMSE) και συντελεστή προσδιορισμού R^2 . Ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της διακύμανσης του IC50 που μπορεί να εξηγήσει το μοντέλο συμβολικής παλινδρόμησης, και η ρίζα του μέσου τετραγωνικού σφάλματος δείχνει τη μέση απόσταση μεταξύ της πρόβλεψης IC50 μίας αγνώστου, για το σύνολο εκπαίδευσης, κυτταρικής σειράς έναντι της πειραματικής τιμής του IC50 για την ίδια κυτταρική σειρά. Οι μαθηματικοί τύποι υπολογισμού των μετρικών αυτών δίνονται στις Εξ. 3 και Εξ. 4.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad \text{Εξ. 3}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \text{ με } SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ και } SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Εξ. 4}$$

όπου n είναι το πλήθος των κυτταρικών σειρών του συνόλου ελέγχου, y_i είναι η παρατηρούμενη $\ln(IC_{50})$ τιμή, $\hat{y}_i$ είναι η προβλεπόμενη $\ln(IC_{50})$ τιμή, $\bar{y}$ είναι η μέση τιμή των παρατηρούμενων $\ln(IC_{50})$ τιμών.

4. Αποτελέσματα και δοκιμές

Η γλώσσα Python χρησιμοποιήθηκε για την υλοποίηση της εργασίας μαζί με τις βιβλιοθήκες `pysr`, `feyn` (QLattice), `FFX`, `ITEA`, `EQL`. Για την εκτέλεση των αλγορίθμων δεν χρησιμοποιήθηκε εξειδικευμένη (dedicated) κάρτα γραφικών, παρόλο που στην εκπαίδευση συμπεριλήφθηκαν όλα τα γονίδια. Τα αποτελέσματα του PySR και του feyn (QLattice) παρατίθενται στο ΠΑΡΑΡΤΗΜΑ I – Πίνακες αξιολόγησης μοντέλων και παραγόμενες εξισώσεις. Απαιτήθηκε 1 ώρα και 20 λεπτά για την εκπαίδευση του QLattice συνολικά, 33 λεπτά για την εκπαίδευση του PySR, ενώ ως διαθέσιμοι τελεστές χρησιμοποιήθηκαν οι προεπιλεγμένοι και στις δύο υλοποιήσεις. Τα καλύτερα αποτελέσματα επιτυγχάνονται στην περίπτωση του φαρμάκου Panobinostat, στην οποία το PySR και ο feyn συμφωνούν ότι το γονίδιο MYOF έχει την μεγαλύτερη επίδραση στην πρόβλεψη του IC50.

Αντιθέτως, ο FFX δεν μπόρεσε να δώσει αποτελέσματα καθώς κατά τη διαδικασία της εκπαίδευσης, καθώς σε δεδομένα εισόδου 390×18988 απαίτησε δέσμευση μνήμης 454 GiB για έναν πίνακα 246844×246844 που δημιουργεί εσωτερικά, κάτι το οποίο δεν μπορούσε να εκπληρωθεί από το υπολογιστικό σύστημα που ήταν διαθέσιμο κατά τη διάρκεια της εργασίας. Ο δε ITEA, χρειάστηκε 13.5 ώρες για εκπαίδευση σε ένα από τα 24 φάρμακα (panobinostat) με παράλογα αποτελέσματα μετρικών $RMSE = 1.359056e + 12$ και $R^2 = -2.520523e + 24$ και μία εξίσωση που περιλάμβανε περισσότερες από 151 χιλιάδες λέξεις. Ο EQL με 4 επίπεδα κρυφών νευρώνων, αν και στο Panobinostat δημιούργησε μοντέλο με $RMSE = 0.72263$, $R^2 = 0.287395$ μετά από 5 μέρες εκτέλεσης δεν κατάφερε να εξάγει τη σχετική εξίσωση που ερμηνεύει το αποτέλεσμα. Είναι σαφές, ότι για τους FFX, ITEA, EQL, απαιτείται μείωση διάστασης προτού εκτελεστούν για να μπορεί να εκτιμηθεί η ερμηνεία των μοντέλων. Δηλαδή, για δεδομένους υπολογιστικούς πόρους, πρέπει να γίνει μείωση διάστασης, ώστε τα μοντέλα να συγκλίνουν σε αποτέλεσμα με αποδεκτό μέγεθος συναρτήσεων και τιμή μετρικών. Έτσι, στα δεδομένα εισόδου εφαρμόστηκε η SelectKBest για την επιλογή 7 γονιδίων (Πίνακας 4). Επιλέχθηκε το 7, καθώς είναι το μέγιστο πλήθος γονιδίων από τις παραγόμενες εξισώσεις του PySR και του feyn (Πίνακας 9 και Πίνακας 10). Για λόγους πληρότητας συμπεριλήφθηκε ο PySR και ο feyn σε αυτή τη διαδικασία.

Φάρμακο	Γονίδια
AEW541	CUX2, AMPD1, MT2A, GPRC5D, CORO7, CD28, TP53INP1
Nilotinib	CISH, C15orf26, TRDC, PRG2, SAMSN1, ANKRD55, APOL4
17-AAG	NQO1, DNAJC18, GNAO1, CTDSP1, GNPAT1, TXN, SLK
PHA-665752	PPP1R27, ELANE, KCNE1L, TLR8, MS4A4A, PRTN3, ECRP
Lapatinib	ERBB2, TBC1D30, GRB7, LAD1, EPN3, CATSPERB, PGAP3
Nutlin-3	IGSF6, ARL11, PPP1R27, C5orf20, TLR8, MS4A4A, LPPR3
AZD0530	SPINT2, EVPL, JUP, ADAP2, RREB1, EPX, NAPS
PF2341066	SH2D3C, IL2RG, AMOTL2, PTPN7, SLA, SPN, CBFA2T3
L-685458	IKZF1, TAGAP, ITGAL, PIK3CG, LAX1, SELPLG, ARHGAP9
ZD-6474	CHN1, CAPN12, CTSH, SLC39A2, ACE2, SYTL1, TNK1
Panobinostat	PRSS23, MYOF, CYR61, GPX8, TNFRSF12A, AMOTL2, GNG12
Sorafenib	CSF3R, GAPT, SIGLECP3, P2RX1, CCL23, STAR, TYROBP
Irinotecan	CPSF6, NRF1, RFXAP, S100A13, ARHGAP19, SLFN11, CD63
Topotecan	HMGB2, PPIC, RFXAP, PRPF38A, ARHGAP19, SLFN11, CD63
LBW242	LOC285205, KLK3, KCNF1, IPO5, RNF186, SRD5A2, CABP1
PD-0325901	LOC100127888, GJB1, FMN1, RXRG, CAPN3, TRIM63, SPRYD5
PD-0332991	LOC100506972, RBM33, KCNA3, P2RX1, CECR1, RNASE6, TRAF3IP3
Paclitaxel	ANP32B, CNTRL, ABCB1, RPL4, SSRP1, C4BPB, ANP32A
PLX4720	DCT, MLANA, SLC45A2, ROPN1, RXRG, CAPN3, SPRYD5

RAF265	SH2B3, C9orf7, RILPL2, AK2, PSMD9, SUGP1, C7orf70
TAE684	PUM2, WWTR1, MT2A, MYB, PCF11, CASP2, P4HA2
TKI258	C1orf228, ITGAL, ORC1, EIF3L, DPEP2, LST1, KDM2B
Erlotinib	CDH3, EVPL, TMPRSS11E, IRF6, KRT16, GPR87, PRSS8
AZD6244	CPN1, MLANA, GJB1, LYZ, RXRG, CAPN3, SPRYD5

Πίνακας 4 – Επτά σημαντικότερα γονίδια βάσει SelectKBest

Με 7 χαρακτηριστικά εισόδου, όλοι οι αλγόριθμοι έφεραν αποτελέσματα που παρουσιάζονται στο ΠΑΡΑΡΤΗΜΑ I – Πίνακες αξιολόγησης μοντέλων και παραγόμενες εξισώσεις στους πίνακες Πίνακας 11 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με Qlattice (random_seed=96, n_epochs=18) έως Πίνακας 15 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με ITEA (random_state=18, simplify_method='simplify_by_var'). Εφαρμόστηκε περαιτέρω επεξεργασία για την σωστή αναπαράστασή των εξισώσεων σε μορφή LaTeX ώστε να μπορεί να αποτυπωθεί στην παρούσα εργασία.

Ένα ικανοποιητικό μοντέλο προσδιορίζεται από τις τιμές του R^2 και του RMSE. Τιμές R^2 κοντά στο 0 σημαίνουν ότι το μοντέλο δεν εξηγεί τη διασπορά στα δεδομένα και μπορεί να κάνει χειρότερες προβλέψεις συγκριτικά με ένα μοντέλο που πάντα προβλέπει τη μέση τιμή των IC50 του συνόλου εκπαίδευσης. Τιμές R^2 από 0.3 έως 0.6 είναι ικανοποιητικές, υπό την έννοια ότι εξηγείται σε έναν βαθμό η διασπορά των IC50, αλλά ένα μεγάλο μέρος της δεν μπορεί να προβλεφθεί. Για εύρος R^2 από 0.6 έως 0.8, το μοντέλο εξηγεί ένα σημαντικό κομμάτι της διασποράς και το μοντέλο θεωρείται καλό, ενώ από 0.8 και πάνω το μοντέλο είναι εξαιρετικό καθώς εξηγεί σχεδόν όλη τη διασπορά της μεταβλητής απόκρισης. Στα πλαίσια της εργασίας, τα πλήρη αποτελέσματα παρουσιάζονται στο ΠΑΡΑΡΤΗΜΑ I – Πίνακες αξιολόγησης μοντέλων και παραγόμενες εξισώσεις. Στην παρούσα ενότητα παρουσιάζονται τα μοντέλα που είναι ικανοποιητικά κατά R^2 , δηλαδή μοντέλα με $R^2 \geq 0.3$.

Η αξιολόγηση των τιμών του RMSE απαιτεί γνώση των περιγραφικών στατιστικών του IC50. Αν για παράδειγμα το IC50 ποικίλει από 0 έως 100, τότε ένα $RMSE = 0.7$ είναι αμελητέο, ενώ αν το IC50 ποικίλει από 0 έως 1, τότε το $RMSE = 0.7$ είναι σημαντικό. Για κάθε φάρμακο τα περιγραφικά στατιστικά δίνονται στον Πίνακα 5.

Φάρμακο	Μέση Τιμή	Τυπική Απόκλιση	Ελάχιστο	Μέγιστο	Διάμεσος	Q1	Q3
AEW541	1.551	0.822	-1.990	2.079	2.029	1.270	2.079
Nilotinib	1.793	0.980	-5.825	2.079	2.079	1.994	2.079
17-AAG	-1.684	1.504	-5.974	2.079	-1.813	-2.703	-0.946
PHA-665752	1.976	0.492	-3.472	2.079	2.079	2.079	2.079
Lapatinib	1.750	0.786	-2.716	2.079	2.079	1.955	2.079
Nutlin-3	2.022	0.398	-5.112	2.079	2.079	2.079	2.079
AZD0530	1.683	0.817	-3.829	2.079	2.079	1.637	2.079
PF2341066	1.711	0.791	-3.778	2.079	2.079	1.638	2.079
L-685458	1.885	0.651	-2.471	2.079	2.079	2.079	2.079
ZD-6474	1.628	0.786	-3.222	2.079	2.079	1.351	2.079
Panobinostat	-2.871	0.870	-5.991	0.131	-2.773	-3.336	-2.270
Sorafenib	1.883	0.660	-4.914	2.079	2.079	2.053	2.079
Irinotecan	-1.201	1.321	-5.888	2.079	-1.263	-2.085	-0.288
Topotecan	-0.737	1.573	-4.748	2.079	-0.762	-1.917	0.415
LBW242	1.886	0.859	-5.614	2.079	2.079	2.079	2.079
PD-0325901	0.545	2.205	-5.991	2.079	2.079	-0.782	2.079
PD-0332991	1.860	0.591	-1.362	2.079	2.079	2.079	2.079

Paclitaxel	-4.183	2.063	-5.991	2.079	-4.662	-5.638	-3.799
PLX4720	1.897	0.630	-3.394	2.079	2.079	2.079	2.079
RAF265	0.879	1.170	-2.490	2.079	1.108	0.102	2.079
TAE684	1.099	1.106	-4.643	2.079	1.361	0.565	2.079
TKI258	1.591	0.755	-3.571	2.079	2.008	1.316	2.079
Erlotinib	1.842	0.702	-3.246	2.079	2.079	2.079	2.079
AZD6244	1.335	1.482	-5.974	2.079	2.079	1.620	2.079

Πίνακας 5 - Περιγραφικά στατιστικά του $\ln(IC_{50}(\mu Mol))$

Από τα περιγραφικά στατιστικά, η τυπική απόκλιση θα χρησιμοποιηθεί ως βάση για την αξιολόγηση του RMSE. Συνεπώς, RMSE που είναι ίσο ή μεγαλύτερο της τυπικής απόκλισης της εξαρτημένης μεταβλητής χαρακτηρίζει ένα μη ικανοποιητικό μοντέλο. Τιμές RMSE που κυμαίνονται από το μισό της τυπικής απόκλισης έως την τυπική απόκλιση χαρακτηρίζονται ικανοποιητικές, ενώ καλές θεωρούνται αν είναι μικρότερες από το μισό της τυπικής απόκλισης. Τέλος, το μοντέλο θεωρείται άριστο εάν έχει τιμές RMSE κοντά στο μηδέν, παραδείγματος χάρη αν κυμαίνονται από το μηδέν έως το ένα τέταρτο της τυπικής απόκλισης.

Ακολουθώντας τα παραπάνω, τα μοντέλα που είναι ικανοποιητικά βάσει του R^2 και του RMSE παρουσιάζονται στον Πίνακα 6. Παρατηρούμε ότι από τα 24 φάρμακα, μόνο στα 6 έχουμε ικανοποιητικά μοντέλα. Όλοι οι αλγόριθμοι, εκτός του FFX έχουν παρουσιάσει τουλάχιστον ένα ικανοποιητικό μοντέλο. Το πιο ικανοποιητικό από τα μοντέλα παράγεται από τον αλγόριθμο QLattice, μετά από εφαρμογή του SelectKBest για το φάρμακο Panobinostat. Το Panobistat είναι το μόνο φάρμακο για το οποίο κατασκευάστηκαν ικανοποιητικά μοντέλα για τέσσερις από τους πέντε αλγορίθμους συμβολικής παλινδρόμησης που εξετάστηκαν.

Αλγόριθμος	Φάρμακο	RMSE	R^2
pySR	17-AAG	1.129959	0.306472
pySR Select7Best	17-AAG	1.129967	0.306461
QLattice Select7Best	Panobinostat	0.645778	0.430909
QLattice	Panobinostat	0.649938	0.423552
ITEA Select7Best	Panobinostat	0.656784	0.411344
EQL Select7Best	Panobinostat	0.658837	0.407659
pySR	Panobinostat	0.683738	0.362038
pySR Select7Best	Panobinostat	0.696533	0.337937
EQL Select7Best	Irinotecan	0.984114	0.378784
QLattice Select7Best	Irinotecan	1.002543	0.355300
EQL Select7Best	Topotecan	1.231213	0.339449
pySR Select7Best	PD-0332991	0.480822	0.393369
QLattice Select7Best	PD-0332991	0.482168	0.389968
QLattice	PD-0332991	0.501483	0.340113

EQL Select7Best	AZD6244	1.173328	0.337040
--------------------	---------	----------	----------

Πίνακας 6 - Ικανοποιητικά μοντέλα ανά αλγόριθμο και φάρμακο

Για κάθε φάρμακο για το οποίο κατασκευάστηκε ένα ικανοποιητικό μοντέλο, οι πιο απλές συναρτήσεις είναι αυτές με τις λιγότερες μεταβλητές και δίνονται στον Πίνακα 7. Είναι αξιοσημείωτο ότι ο EQL για τα φάρμακα Topotecan και AZD6244 ήταν ο μόνος αλγόριθμος που κατασκεύασε ικανοποιητικό μοντέλο, αν και οι συναρτήσεις που κατασκεύασε είναι ιδιαίτερα μακροσκελείς και δυσνόητες. Αυτό συμβαίνει γιατί δεν γίνεται καλή απλοποίηση της έκφρασης, μιας και μπορούν να γίνουν πράξεις και να μειωθεί το μέγεθος της συνάρτησης, όπως επίσης διότι ο EQL πάντα χρησιμοποιεί όλα τα γονίδια εισόδου πολλαπλές φορές. Δρα δηλαδή όπως μία γραμμική παλινδρόμηση που σε είσοδο 19000 γονιδίων θα επιστρέψει εξίσωση με 19000 γονίδια, αν και ο EQL θα εμφανίσει 5 φορές το κάθε γονίδιο εισόδου· δύο φορές ως γραμμικό συνδυασμό, μία φορά ως γραμμικό συνδυασμό μέσα σε ημίτονο, μία φορά ως γραμμικό συνδυασμό μέσα σε συνημίτονο, και μία τελευταία φορά πάλι ως γραμμικό συνδυασμό στον εκθέτη του παρονομαστή της σιγμοειδής συνάρτησης. Συνεπώς, ο συγκεκριμένος αλγόριθμος χρήζει βελτίωσης ως προς την περιεκτικότητα των παραγόμενων εκφράσεων, καθώς οι παραγόμενες εκφράσεις έχουν 5 φορές μεγαλύτερο πλήθος μεταβλητών από τις μεταβλητές εισόδου.

Η πιο απλή συνάρτηση παράγεται από το QLattice για το Panobinostat. Πρόκειται για μία απλή γραμμική παλινδρόμηση με δύο μεταβλητές/γονίδια και έναν σταθερό όρο. Ο QLattice παράγει επίσης την πιο απλή συνάρτηση για τα Irinotecan και PD-0332991 με μόνο τρία γονίδια. Γι' αυτά, οι σχέσεις δεν είναι ακριβώς γραμμικές, καθώς περιλαμβάνουν επίσης το υπερβολικό ημίτονο. Τέλος, στο 17-AAG, μόνο το pySR παράγει ένα ικανοποιητικό μοντέλο το οποίο έχει τέσσερα γονίδια. Ο αριθμητής είναι ένας γραμμικός συνδυασμός των γονιδίων CTDSP1, GNAO1, και GNPAT1, και ο παρονομαστής περιλαμβάνει το γονίδιο NQO1 και έναν σταθερό όρο. Συνολικά, το πιο ικανοποιητικό μοντέλο είναι για το Panobinostat από το QLattice με είσοδο όλο το πλήθος των γονιδίων.

Αλγόριθμος	Φάρμακο	Εξίσωση
pySR Select7Best	17-AAG	$\ln(IC_{50}) = 0.40098 + \frac{-0.76709 \cdot CTDSP1 + GNAO1 - GNPAT1 - 6.54116}{3.3877 - NQO1}$
QLattice	Panobinostat	$\ln(IC_{50}) = 0.388 \cdot MYOF - 0.185 \cdot P2RX1 - 2.87$
QLattice Select7Best	Irinotecan	$\ln(IC_{50}) = -0.376 \cdot ARHGAP19 - 1.05 \cdot \tanh(-0.853 \cdot CD63 + 0.797 \cdot SLFN11 + 0.406) - 1.03$
EQL Select7Best	Topotecan	$\ln(IC_{50}) = -0.0697943017519389 \cdot ARHGAP19 + \dots - 0.288084218918485 \cdot SLFN11 - (0.0706525564654238 \cdot ARHGAP19 - \dots + 0.0581333661221653 \cdot SLFN11 + 0.101253865800475) \cdot (0.249702796339989 \cdot ARHGAP19 - \dots + 0.247055694460869 \cdot SLFN11 + 0.3181092441082) - 0.504336476325989 \cdot \sin(-0.0623004212975502 \cdot ARHGAP19 - \dots + 0.859442055225372 \cdot SLFN11 + 0.210722014307976) - 0.325596511363983 \cdot \cos(0.0630510672926903 \cdot ARHGAP19 + \dots - 0.407091408967972 \cdot SLFN11 + 0.321089565753937) - 0.29024053112699 \cdot \frac{0.361397713422775}{0.809222815838597 \cdot e^{-0.0308107901364565 \cdot ARHGAP19 + \dots - 0.73715478181839 \cdot SLFN11} + 1}$
QLattice Select7Best	PD-0332991	$\ln(IC_{50}) = 1.06 - 0.906 \cdot \tanh(0.464 \cdot LOC100506972 + (0.086 - 0.719 \cdot P2RX1) \cdot (0.341 - 0.901 \cdot CECR1) - 2.0)$

EQL Select7Best	AZD6244	$\ln(IC_{50})$ $= 0.0214658983103668 \cdot CAPN3 + \dots - 0.0866502631953963 \cdot SPRYD5$ $- (0.0182234146602267 \cdot CAPN3 + \dots - 0.00598157806769873 \cdot SPRYD5$ $- 0.213824012033811)$ $\cdot (0.169173166155815 \cdot CAPN3 + \dots + 0.0553271435201168 \cdot SPRYD5$ $+ 0.0492627397179604) - 0.279703289270401$ $\cdot \sin(0.0824903771281242 \cdot CAPN3 + \dots + 0.0331891141831875 \cdot SPRYD5$ $- 0.492052465677261) + 0.491845160722733$ $\cdot \cos(-0.0379341877996922 \cdot CAPN3 - \dots + 0.132581442594528 \cdot SPRYD5$ $+ 0.161847099661827) + 0.648938834251565$ 0.421451419591904 $+ \frac{0.55348243783415 \cdot e^{-0.189851924777031 \cdot CAPN3 - \dots + 0.0669686123728752 \cdot SPRYD5} + 1}{1}$
--------------------	---------	--

Πίνακας 7 - Συναρτήσεις ικανοποιητικών μοντέλων με μικρότερο πλήθος μεταβλητών

Στη βιβλιογραφία, μία παρόμοια μελέτη των Park et al. (2023), στα ίδια δεδομένα, διερεύνησε την εφαρμογή αλγορίθμων μηχανικής μάθησης, και κατέληξε σε 7 μοντέλα από τα 24 φάρμακα (Πίνακας 8). Για τα Panobinostat, Irinotecan, Topotecan, PD-0332991, AZD6244 οι διαφορές στο R^2 μεταξύ των μοντέλων της παρούσας εργασίας και των Park et al. (2023) κυμαίνονται από 0.6% στο Topotecan έως 5% στο Panobinostat, ενώ οι διαφορές του RMSE ποικίλει από 0.0055 στο Topotecan έως 0.02702 στο Panobinostat. Παρότι για τα Lapatinib, PD-0325901 οι αλγόριθμοι συμβολικής παλινδρόμησης που δοκιμάσαμε δεν έφεραν αποτελέσματα, κατασκευάσαμε ένα ικανοποιητικό μοντέλο για το 17-AAG. Επίσης, οι συναρτήσεις που υπολογίστηκαν για τα 17-AAG, Panobinostat, Irinotecan, PD-0332991 αποτελούνται από το πολύ τέσσερα γονίδια και έτσι προσφέρουν έναν απλό, αναλυτικό και ερμηνεύσιμο τρόπο αξιολόγησης του μοντέλου.

Αλγόριθμος	Φάρμακο	RMSE	R^2
XGBoost	Lapatinib	0.634628	0.412699
Ridge	Panobinostat	0.622922	0.470478
Random Forest	Irinotecan	0.994741	0.365294
Random Forest	Topotecan	1.225712	0.345339
Ridge	PD-0325901	1.631838	0.407561
XGBoost	PD-0332991	0.501174	0.340928
Ridge	AZD6244	1.136689	0.377797

Πίνακας 8 - Ικανοποιητικά μοντέλα ανά αλγόριθμο και φάρμακο των Park et al. (2023)

5. Συμπεράσματα και μελλοντικές προσεγγίσεις

Η παρούσα εργασία επισήμανε την ανάγκη εφαρμογής της συμβολικής παλινδρόμησης στον ιατροφαρμακευτικό τομέα, καθώς μπορεί να δημιουργήσει ικανοποιητικά μοντέλα που περιγράφονται από μία περιεκτική μαθηματική έκφραση. Έγινε σύγκριση των αποτελεσμάτων στις διαφορετικές υλοποιήσεις συμβολικής παλινδρόμησης και παρουσιάστηκαν οι αδυναμίες και οι δυνατότητες της κάθε υλοποίησης.

Μετά τις δοκιμές που εκτελέστηκαν έχουμε τα παρακάτω συμπεράσματα:

- Ο QLattice δημιούργησε το καλύτερο μοντέλο πρόβλεψης IC50 στο Panobinostat με χρήση μόνο 2 από τα 19898 γονίδια σε μία γραμμική σχέση με RMSE ίσο με 0.649938 και R^2 ίσο με 0.423552.
- Κατασκευάστηκαν ικανοποιητικά μοντέλα για 6 φάρμακα από τα 24.
- Ο FFX είχε τα χειρότερα αποτελέσματα, καθώς δεν κατάφερε να κατασκευάσει ούτε ένα ικανοποιητικό μοντέλο.
- Ο EQL χρησιμοποιεί σταθερά το πενταπλάσιο πλήθος μεταβλητών από τις μεταβλητές εισόδου, οπότε δυσχεραίνεται η ερμηνεία των αποτελεσμάτων.
- Ο ITEA κατάφερε να δημιουργήσει ένα ικανοποιητικό μοντέλο, που όμως περιλάμβανε 16 μεταβλητές ενώ η είσοδος είχε μόνο 7.
- Ο pySR είχε παρόμοιο πλήθος ικανοποιητικών μοντέλων με το QLattice με μικρότερο πλήθος μεταβλητών από αυτές που δόθηκαν ως είσοδο.
- Συγκριτικά με δημοσίευση σε όμοια δεδομένα των Park et al. (2023), οι υλοποιήσεις μας προσφέρουν παρόμοιες τιμές μετρικών αξιολόγησης, με το πλεονέκτημα ότι οι συναρτήσεις για τα 17-AAG, Panobinostat, Irinotecan, PD-0332991 αποτελούνται από το πολύ τέσσερα γονίδια και έτσι προσφέρουν ένα απλό, μαθηματικό ερμηνεύσιμο τρόπο αξιολόγησης του μοντέλου.

Μετά την εφαρμογή των αλγορίθμων συμβολικής παλινδρόμησης σε δεδομένα γονιδιακής έκφρασης, γίνεται εμφανές ότι οι αλγόριθμοι έχουν περιθώρια βελτίωσης, ιδιαίτερα ως προς την απλοποίηση της καλύτερης μαθηματικής έκφρασης. Επίσης, θα είναι χρήσιμο να διερευνηθούν εφαρμογές σε διαφορετικά σύνολα δεδομένων αλλά και να γίνει σύγκριση μεγαλύτερης κλίμακας με άλλες προσεγγίσεις ανάλυσης παλινδρόμησης.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
adaptive parsimony	προσαρμοστική φειδωλότητα
backpropagation	οπισθοδιάδοση
Bayesian multitask multiple kernel learning	μάθηση Bayes πολλαπλών καθηκόντων και πυρήνων
biomarker	βιοδείκτης
cell line	κυτταρική σειρά
combinatorial optimization	συνδυαστική βελτιστοποίηση
confounding factor	συγχυτικός παράγοντας
drug response	απόκριση ασθενούς σε φάρμακο
expression tree	δέντρο έκφρασης
extrapolation	προέκταση
feedforward neural network	νευρωνικό δίκτυο πρόσθιας τροφοδότησης
fitness function	συνάρτηση καταλληλότητας
gene expression profile	προφίλ γονιδιακής έκφρασης
half maximal inhibitory concentration	ήμισυ της μέγιστης ανασταλτικής συγκέντρωσης
high-throughput screening	διαλογή υψηλής απόδοσης
interpolation	παρεμβολή
loss function	συνάρτηση απώλειας
neural network	νευρωνικό δίκτυο
non-dominated filter	μη-κυριαρχούμενο φίλτρο
omic	ομικά
pharmacogenomics	φαρμακογονιδιωματική
random forest	τυχαίο δάσος
regression	παλινδρόμηση
root mean squared error	ρίζα του μέσου τετραγωνικού σφάλματος
simulated annealing	προσομοιωμένη ανόπτηση
support vector machine	μηχανή ανυσμάτων στήριξης
symbolic regression	συμβολική παλινδρόμηση
test dataset	σύνολο δεδομένων ελέγχου
training dataset	σύνολο δεδομένων εκπαίδευσης
tumor molecular profile	μοριακό προφίλ (κατατομή) όγκου
adaptive parsimony	προσαρμοστική φειδωλότητα

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

BFGS	Broyden–Fletcher–Goldfarb–Shanno
CCLE	Cancer Cell Line Encyclopedia
DNA	deoxyribonucleic acid
EGFR	epidermal growth factor receptor
EQL	Equation Learner
FFX	Fast Function Extraction
GDSC	Genomics of Drug Sensitivity in Cancer
GPCR	G protein-coupled receptor
HTS	high-throughput screening
IC50	Half maximal inhibitory concentration
ITEA	Interaction-Transformation Evolutionary Algorithm
mRNA	messenger ribonucleic acid
pySR	python Symbolic Regression
RF	random forest
RMSE	root mean squared error
RNA	ribonucleic acid
SVM	support vector machines

ΠΑΡΑΡΤΗΜΑ Ι – Πίνακες αξιολόγησης μοντέλων και παραγόμενες εξισώσεις

Φάρμακο	RMSE	R^2	Εξίσωση
AEW541	0.773314	0.057962	$\ln(IC_{50}) = 1.69 - \frac{0.059}{(1 - 0.11 \cdot TAL_2)^2 \cdot (0.246 \cdot PIGM - 1)^2 \cdot (-0.271 \cdot PIK3C2A - 1)^2}$
Nilotinib	0.691892	-0.59013	$\ln(IC_{50}) = 1.96 - 12.4 \cdot e^{-0.482 \cdot (-0.152 \cdot ATL2 - 1)^2 - 8.36 \cdot (0.394 \cdot ABL1 + (0.441567 - 0.267483 \cdot MAP3K7) \cdot (0.574 \cdot CRYBB2P1 - 1.71) - 1)^2}$
17-AAG	1.354661	0.003218	$\ln(IC_{50}) = 0.329 \cdot RBFOX1 - 0.368 \cdot TXN - 1.71$
PHA-665752	0.848821	-0.02321	$\ln(IC_{50}) = (-6.851 \cdot RCOR1 - 4.41626) \cdot e^{-5.22 \cdot (0.275 \cdot TMEM167A - 1)^2 - 4.63 \cdot (0.519 \cdot TMEM214 + 1)^2} + 2.03$
Lapatinib	0.872193	-0.1093	$\ln(IC_{50}) = 1.89 - 5.46 \cdot e^{-4.61 \cdot (KLK15 + 0.295)^2 \cdot (-0.0708 \cdot TRAV131 - 1)^2 - 4.68 \cdot (0.129 \cdot NOP14 - 1)^2 \cdot (-0.0156 \cdot CRYL1 - 0.196 \cdot PGAP3 + 1)^2}$
Nutlin-3	0.802268	-0.01633	$\ln(IC_{50}) = 2.05 - 2.56 \cdot e^{-2.32 \cdot (1 - 0.596 \cdot LOC100505487)^2 - 14.5 \cdot (-0.361 \cdot B3GALNT2 + e^{0.776 \cdot KIAA1671 + 0.409})^2}$
AZD0530	0.768161	-0.1767	$\ln(IC_{50}) = -0.161 \cdot SLC48A1 + 2.0 + \frac{0.358}{0.243 \cdot MAS1 - 0.978}$
PF2341066	0.839005	-0.60263	$\ln(IC_{50}) = 1.74 - 9.74 \cdot e^{-19.9 \cdot (-0.196 \cdot CELSR3 - 0.218 \cdot SH2D2A - 0.193 \cdot ZNF671 + 1)^2 - 2.0 \cdot \log(0.00871 - 0.705 \cdot EHF)^2}$
L-685458	0.822723	-0.2099	$\ln(IC_{50}) = 1.99 - 5.41 \cdot e^{-0.418 \cdot (-0.64 \cdot PGLYRP3 - 1)^2 - 2.0 \cdot \left((1.57 - 0.863 \cdot RCS1) \cdot (3.11 - 0.46 \cdot LOC338651) + \frac{0.0514 \cdot ACACB - 1.92}{-0.0744 \cdot PDE2A - 0.871} \right)^2}$
ZD-6474	0.814524	-0.03697	$\ln(IC_{50}) = 0.143 \cdot ATG12 - 3.99 \cdot e^{-2.26 \cdot (1 - 0.129 \cdot MUC17)^2 - 2.5 \cdot (1 - 0.723 \cdot SHANK2AS3)^2} + 1.75$
Panobinostat	0.649938	0.423552	$\ln(IC_{50}) = 0.388 \cdot MYOF - 0.185 \cdot P2RX1 - 2.87$
Sorafenib	1.111895	-1.47899	$\ln(IC_{50}) = 1.96 - 8.01 \cdot e^{-5.89 \cdot (1 - 0.248 \cdot FRY)^2 - 21.6 \cdot (1 - 0.106 \cdot ZNF717)^2 \cdot ((0.17 - AP1AR)^2 + (0.44688 - 0.030723 \cdot TNFRSF19) \cdot (0.379 \cdot PLAC8L1 - 1.09))^2}$
Irinotecan	1.213581	0.055309	$\ln(IC_{50}) = -1.86 \cdot \tanh(0.331 \cdot NCAPH + 0.205 \cdot SMC1B - 0.274 \cdot TRNP1 - 0.318) - 1.7$
Topotecan	1.387619	0.160964	$\ln(IC_{50}) = -0.482 \cdot ANP32B - 0.185 \cdot BCAT1 + 0.335 \cdot CRIM1 - 0.133 \cdot KIAA1430 - 0.713$
LBW242	0.947192	-11.636969	$\ln(IC_{50}) = 1.92 - 9.26 \cdot e^{-33.8 \cdot (0.266 \cdot SH3GLB2 + 1)^2 \cdot (0.258 \cdot PACSIN1 + 0.808 \cdot e^{-2.63 \cdot (0.35 \cdot C10orf113 + 1)^2 - 0.718 \cdot (0.906 \cdot CENPF - 1)^2 - 1})^2}$
PD-0325901	1.910107	0.188282	$\ln(IC_{50}) = -0.526 \cdot ATP8B4 + 0.407 \cdot CCDC99 - 0.825 \cdot FAM167B + 0.558 \cdot FTSJD2 + 0.571$
PD-0332991	0.501483	0.340113	$\ln(IC_{50}) = 1.13 \cdot \tanh(0.508 \cdot FDX1L - 0.404 \cdot MTDH + 0.452 \cdot TMEM141 - 0.382 \cdot ZDHHC18 + 2.1) + 0.798$
Paclitaxel	2.36568	-0.07819	$\ln(IC_{50}) = 0.54 \cdot CHST9 - 0.381 \cdot MLLT10 + 0.314 \cdot SAMD4A - 4.11$
PLX4720	0.664106	0.019037	$\ln(IC_{50}) = 1.92 \cdot e^{-0.853 \cdot (-MLANA - 0.202)^2 - 2.0 \cdot e^{-7.92 \cdot (1 - 0.554 \cdot KNCN)^2 - 9.75 \cdot (-0.446 \cdot TMTC3 - 1)^2}} + 0.0933$
RAF265	1.300989	-0.04895	$\ln(IC_{50}) = 1.8 - \frac{1.53}{-0.245 \cdot CBL + 0.579 \cdot LOC553103 + 2.05}$
TAE684	1.28028	0.124929	$\ln(IC_{50}) = -0.226 \cdot ALS2CR8 - 0.161 \cdot KCNE3 - 0.154 \cdot SPAST + 1.1$
TKI258	0.672026	0.068885	$\ln(IC_{50}) = -0.134 \cdot XPNPEP3 - (-0.107338 \cdot HM13 + 0.0560798 \cdot STAR + 0.058548) \cdot e^{0.299 \cdot C9orf47} + 1.65$
Erlotinib	0.619705	-0.25418	$\ln(IC_{50}) = 1.94 - 6.46 \cdot e^{-4.64 \cdot (1 - 0.25 \cdot TMPRSS11E)^2 - 4.16 \cdot (0.548 \cdot SMC1A + 1)^2 \cdot (0.951 \cdot APLF - 0.693 \cdot GATC - 0.118 \cdot GGTLC1 - 1)^2}$
AZD6244	1.355102	0.115714	$\ln(IC_{50}) = -0.288 \cdot MS4A3 - 1.45 \cdot \tanh(0.702 \cdot EDNRB + 0.52 \cdot RIPK3 - 1.15) + 0.296$

Πίνακας 9 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις Qlattice (random_seed=157, n_epochs=18)

Φάρμακο	RMSE	R ²	Εξίσωση
AEW541	0.799565	-0.007079	$\ln(IC_{50}) = 0.12693 \cdot CNN3 - 0.12693 \cdot PHOX2A + 0.12693 \cdot PIM2 \cdot (-ART5 + BMPR2 - 2.30755030459664 \cdot CD99L2) + 0.12693 \cdot RAB9A + 1.5975$
Nilotinib	0.569765	-0.078323	$\ln(IC_{50}) = 1.8398 - \frac{APOL4 \cdot TRAF3IP2^{AS1} \cdot (GOPC + NFIX) + \frac{CHST5}{1.9388 - EDAR}}{FEZF2 + 3.5257}$
17-AAG	1.129959	0.306472	$\ln(IC_{50}) = 0.22446 \cdot NQO1 \cdot (-HPS4 \cdot (ALDH1A3 - DPH2 - LOC100129960) + LAT2 - 3.7185) - 1.6204$
PHA-665752	0.855018	-0.038207	$\ln(IC_{50}) = -0.056109 \cdot GMFG - 0.056109 \cdot TLR8 \cdot (2.0154 \cdot LOC100628307 + 0.1176973446) + 2.0142 - \frac{0.00267134949 \cdot LOC100506922}{FLJ10661}$
Lapatinib	0.749557	0.180722	$\ln(IC_{50}) = 0.15634 \cdot CATSPERB \cdot (1.370131915 \cdot LOC100506233 - 1.370131915 \cdot ZBTB33 - 1.30067992822865) - 0.15634 \cdot STARD3 - 0.15634 \cdot ZBTB33 + 1.78868594$
Nutlin-3	0.798482	-0.006759	$\ln(IC_{50}) = -0.033263 \cdot FAIM3 \cdot SLC22A16 \cdot SRSF3 \cdot (LOC100505622 - LOC100507675 + SRSF9) + 2.04968674405$
AZD0530	0.725501	-0.049631	$\ln(IC_{50}) = \left(0.074377 - \frac{0.01794438}{KRT23}\right) \cdot (-HSH2D + HUS1 - RREB1 + THAP3 + 0.17784) + 1.6847$
PF2341066	0.689583	-0.082621	$\ln(IC_{50}) = -MET \cdot (0.16694 \cdot MET + 0.16694 \cdot ST^{AS17} \cdot TMC07 + 0.16694 \cdot ST^{AS17} - 0.109412476) + 1.87524$
L-685458	0.718793	0.076473	$\ln(IC_{50}) = -0.20538 \cdot SMAD4 \cdot (0.76117 \cdot CD27 \cdot (-ADRB3 + MOB2) + MPO + PIK3CG + 0.41043) + 2.005$
ZD-6474	0.798729	0.002857	$\ln(IC_{50}) = -0.12004 \cdot ARHGEF5 - 0.12004 \cdot RAB37 + 0.12004 \cdot TFE3 + 1.6334 - \frac{0.00055275 \cdot (ARHGEF5 - PHLDA3 + 0.12541)}{ENPP2}$
Panobinostat	0.683738	0.362038	$\ln(IC_{50}) = 0.106367111028246 \cdot ARSJ + 0.106367111028246 \cdot FND C1 + 0.310742363506417 \cdot MYOF - 0.106367111028246 \cdot POC1B - 0.106367111028246 \cdot SPN - 2.89432815325813$
Sorafenib	0.723446	-0.049445	$\ln(IC_{50}) = -0.16808 \cdot ZBTB48 \cdot (PER4 \cdot PSMG4 + RPL37A \cdot SPATA9 \cdot (-CD22 + UCN) + 0.46066) + 1.914$
Irinotecan	1.194676	0.084512	$\ln(IC_{50}) = -0.22404964 \cdot C4orf46 + 0.22404964 \cdot CCND1 - 0.22404964 \cdot CUL2 - 0.22404964 \cdot HIST1H2AL + 0.22404964 \cdot (FBXO5 + 1.134936) \cdot (LOC100506421 - 1.413026) - 0.813254442263512$
Topotecan	1.473423	0.053993	$\ln(IC_{50}) = -0.332006210281035 \cdot ANP32B - 0.421766840752852 \cdot FBL - 0.691937899403854 \cdot SLFN11 - 0.836571132184772$
LBW242	0.329094	-0.525482	$\ln(IC_{50}) = KLK3 \cdot (-0.31294 \cdot CES1P1 + 0.31294 \cdot (C2CD4C - 1.1071) \cdot (MGC12916 - 0.28643)) \cdot (LRRC16B + 0.35003) + 1.8782$
PD-0325901	1.874987	0.217856	$\ln(IC_{50}) = -0.405545016 \cdot CYGB + 0.405545016 \cdot HIVEP1 - 0.769812676296957 \cdot LYZ - 0.405545016 \cdot MLKL - 0.769812676296957 \cdot PLP1 + 0.55226714733864$
PD-0332991	0.522380	0.283974	$\ln(IC_{50}) = -0.10048 \cdot GPR84 + RBBP5 \cdot (-0.171370060660595 \cdot CECR1 + 0.171370060660595 \cdot LOC100506165 - 0.11046) + 1.87498$
Paclitaxel	2.246881	0.027382	$\ln(IC_{50}) = 0.5652 \cdot ABCB1 - 0.5652 \cdot ANP32B + 0.5652 \cdot C4BPB + 0.481629528 \cdot PRDM4 \cdot (TAF9 - 0.39294) - 4.314930864$
PLX4720	0.609275	0.174333	$\ln(IC_{50}) = RXRG \cdot (0.16132 \cdot ACTR1A + 0.16132 \cdot C1orf52 + 0.16132 \cdot C22orf45 \cdot LOC100288432 \cdot MS4A4A - 0.16132 \cdot LOC100288432 - 0.291118072) + 1.9445$

RAF265	1.276681	-0.010118	$\ln(IC_{50}) = 0.22402 \cdot ARHGAP19 \cdot RGS6 - 0.22402 \cdot ERAP1 + 0.22402 \cdot LOC100216545 - 0.22402 \cdot LST1 - 0.22402 \cdot PSMD9 + 0.22402 \cdot RGS6 + 0.91904$
TAE684	1.275089	0.132011	$\ln(IC_{50}) = -0.19601 \cdot ANP32B + 0.19601 \cdot BCAR3 + 0.19601 \cdot BEST3 \cdot (-GUSB + LOC100653004 + 0.43608) - 0.19601 \cdot MAB21L1 - 0.19601 \cdot SUDS3 + 1.116316152$
TKI258	0.664631	0.089265	$\ln(IC_{50}) = -CCDC28B \cdot CSF3R \cdot MIR600HG \cdot (0.373871272897035 \cdot RBM4B + 0.1853126208) - 0.25313 \cdot KDM2B + 1.5952$
Erlotinib	0.592065	-0.144796	$\ln(IC_{50}) = \frac{CDH3}{CDH3 - 0.69216 \cdot HMX1 + 0.69216 \cdot LOC1720 - 4.8337} + 1.87793$
AZD6244	1.275287	0.216814	$\ln(IC_{50}) = 0.20939 \cdot C10orf90 \cdot (-2.045 \cdot MGC24103 + MSH3) - 0.20939 \cdot LYZ \cdot (-LEPREL4 - LYG1 + 2.8011) + 1.5053$

Πίνακας 10 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις PySR (random_state=79, niterations=160, deterministic=True, procs=0, multithreading=None)

Φάρμακο	RMSE	R ²	Εξίσωση
AEW541	0.936056	-0.380258	$\ln(IC_{50}) = -0.169 \cdot CUX2 - 1.72 \cdot \tanh(0.135 \cdot GPRC5D - (1.48 - 0.368 \cdot CORO7) \cdot (0.636 \cdot MT2A + 3.51) + 0.701) - 0.106$
Nilotinib	0.769837	-0.968591	$\ln(IC_{50}) = 1.85 - 9.07 \cdot e^{-1.09 \cdot (0.214 \cdot APOL4 - 1)^2 - 14.7 \cdot (-0.155 \cdot APOL4 + (0.130626 \cdot C15orf26 - 0.77121) \cdot (0.336 \cdot CISH - 1.05) + 1)^2}$
17-AAG	1.225206	0.184625	$\ln(IC_{50}) = -0.238 \cdot GNPNTAT1 - 0.432 \cdot NQO1 - (0.2366 \cdot DNAJC18 - 0.6565) \cdot (-0.353 \cdot CTDSP1 + 0.547 \cdot GNAO1 - 0.821) - 1.19$
PHA-665752	0.864336	-0.060959	$\ln(IC_{50}) = (0.11448 - 0.34128 \cdot KCNE1L) \cdot (-0.231 \cdot PPP1R27 + 0.273 \cdot TLR8 + 0.0656) + 2.0$
Lapatinib	0.736538	0.208934	$\ln(IC_{50}) = -0.0906 \cdot LAD1 - 2.98 \cdot \tanh(0.319 \cdot CATSPERB + 0.41 \cdot ERBB2 + 0.953 \cdot LAD1 - 4.29) - 1.14$
Nutlin-3	0.799315	-0.008861	$\ln(IC_{50}) = 1.27 \cdot \tanh((1.95 - 0.25 \cdot ARL11) \cdot (2.53 - 0.202 \cdot MS4A4A)) + 0.771$
AZD0530	1.175980	-1.757785	$\ln(IC_{50}) = 0.979 \cdot \log(-0.325 \cdot ADAP2 - 0.182 \cdot EPX - 0.26 \cdot EVPL - 0.0802 \cdot NAPS A + 1.92) + 1.08$
PF2341066	0.695621	-0.101661	$\ln(IC_{50}) = -0.179 \cdot CBFA2T3 - 5.13 \cdot e^{-3.28 \cdot (1 - 0.196 \cdot SH2D3C)^2 - 6.3 \cdot (1 - 0.346 \cdot SLA)^2} + 1.74$
L-685458	0.694958	0.136707	$\ln(IC_{50}) = 5.39 \cdot \tanh(e^{-0.0182 \cdot (0.215 - ITGAL)^2 \cdot (0.913 \cdot LAX1 - 1)^2 - 0.0666 \cdot (1 - 0.348 \cdot LAX1)^2 \cdot (SELPLG - 0.261)^2}) - 2.03$
ZD-6474	0.887094	-0.229979	$\ln(IC_{50}) = 1.74 - 4.48 \cdot e^{-1.98 \cdot (1 - 0.56 \cdot SYTL1)^2 - 2.0 \cdot (0.276 \cdot CAPN12 + (1.1 - 0.288 \cdot ACE2) \cdot (0.201 \cdot CHN1 + 0.317 \cdot SLC39A2 - 1.11) - 0.584)^2}$
Panobinostat	0.645778	0.430909	$\ln(IC_{50}) = -0.797 \cdot (0.156 \cdot AMOTL2 + \tanh(0.969 \cdot GNG12 + 1.27 \cdot MYOF - 1.61) - 0.045)^2 - 2.36$
Sorafenib	0.956483	-0.834435	$\ln(IC_{50}) = 1.97 - 11.3 \cdot e^{-2.05 \cdot (1 - 0.126 \cdot TYROBP)^2 \cdot (0.928 \cdot TYROBP - 1)^2 - 3.64 \cdot (0.258 \cdot P2RX1 - 1)^2}$
Irinotecan	1.002543	0.355300	$\ln(IC_{50}) = -0.376 \cdot ARHGAP19 - 1.05 \cdot \tanh(-0.853 \cdot CD63 + 0.797 \cdot SLFN11 + 0.406) - 1.03$
Topotecan	1.279186	0.286971	$\ln(IC_{50}) = 0.333 \cdot PPIC - 1.13 \cdot \tanh(-0.772 \cdot CD63 + 0.421 \cdot PRPF38A + 0.963 \cdot SLFN11 + 0.088) - 0.813$
LBW242	0.601180	-4.090677	$\ln(IC_{50}) = -3.83 \cdot \tanh(0.269 \cdot KCNF1 + (0.669 - 0.366 \cdot LOC285205) \cdot (0.551 \cdot SRD5A2 - 0.853) - 1.78) - 1.84$
PD-0325901	1.956103	0.148718	$\ln(IC_{50}) = 1.22 - 5.35 \cdot \tanh((-0.422 \cdot LOC100127888 - 0.351) \cdot (-0.391 \cdot FMN1 - 0.229 \cdot GJB1 - 0.33))$
PD-0332991	0.482168	0.389968	$\ln(IC_{50}) = 1.06 - 0.906 \cdot \tanh(0.464 \cdot LOC100506972 + (0.086 - 0.719 \cdot P2RX1) \cdot (0.341 - 0.901 \cdot CECR1) - 2.0)$
Paclitaxel	2.124337	0.130582	$\ln(IC_{50}) = 0.511 \cdot ABCB1 + 2.98 \cdot \tanh(0.479 \cdot C4BPB - 1.97) - 3.36 - \frac{2.98}{-0.505 \cdot CNTRL - 0.195 \cdot SSRP1 - 1.83}$

PLX4720	0.687895	-0.052502	$\ln(IC_{50}) = -0.272 \cdot SLC45A2 - 3.56 \cdot e^{-2.33 \cdot (0.186 \cdot CAPN3 - 1)^2 - 2.64 \cdot (0.437 \cdot MLANA - 1)^2} + 1.95$
RAF265	1.323388	-0.085379	$\ln(IC_{50}) = (0.63504 \cdot C7orf70 - 0.7434 \cdot SH2B3 - 1.512) \cdot (0.128 \cdot PSMD9 + 0.163 \cdot SUGP1 + 0.315) + 1.37$
TAE684	1.270992	0.137580	$\ln(IC_{50}) = -0.197 \cdot PCF11 - 0.131 \cdot (PUM2 + 0.701)^2 + 1.31$
TKI258	0.628885	0.184594	$\ln(IC_{50}) = 1.81 - 1.45 \cdot e^{(0.407 \cdot DPEP2 - 1.76) \cdot (-0.544 \cdot EIF3L - 0.339 \cdot ORC1 + 1.65)}$
Erlotinib	0.633320	-0.309891	$\ln(IC_{50}) = -0.201 \cdot KRT16 + (-1.8981 \cdot CDH3 - 0.34029) \cdot e^{-5.63 \cdot (EVPL - 0.503)^2 \cdot (0.5 \cdot KRT16 - 1)^2 - 5.61 \cdot (0.243 \cdot TMPRSS11E - 1)^2} + 1.92$
AZD6244	1.233041	0.267844	$\ln(IC_{50}) = -0.692 \cdot LYZ + (0.6314 - 1.1808 \cdot RXRG) \cdot (0.173 \cdot CPN1 + 0.0894) + 1.42$

Πίνακας 11 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με Qlattice (random_seed=96, n_epochs=18)

Φάρμακο	RMSE	R ²	Εξίσωση
AEW541	0.818831	-0.056198	$\ln(IC_{50}) = 0.040497096 \cdot CORO7^2 \cdot MT2A - 0.12984 \cdot CUX2 - 0.12984 \cdot GPRC5D + 0.12984 \cdot MT2A + 1.56503$
Nilotinib	0.549355	-0.002451	$\ln(IC_{50}) = -0.053676 \cdot APOL4 \cdot CISH \cdot \left(TRDC - \left(-0.23737 + \frac{0.25269}{TRDC} \right) \cdot (ANKRD55 - C15orf26) \right) + 1.8427$
17-AAG	1.129967	0.306461	$\ln(IC_{50}) = 0.40098 + \frac{-0.76709 \cdot CTDSP1 + GNAO1 - GNPAT1 - 6.54116}{3.3877 - NQO1}$
PHA-665752	2.486527	-7.780500	$\ln(IC_{50}) = - \frac{0.022643 \cdot MS4A4A}{-ECRP + TLR8} + 2.0012$ $\frac{2 \cdot ELANE + 0.020246 \cdot MS4A4A}{+ 0.0052783}$
Lapatinib	0.734118	0.214124	$\ln(IC_{50}) = LAD1 \cdot (-0.0615668628 \cdot GRB7 - 0.037354 \cdot TBC1D30) \cdot (CATSPERB + LAD1 + TBC1D30 - 0.89692) + 1.8285$
Nutlin-3	0.798431	-0.006630	$\ln(IC_{50}) = -0.0109257433804596 \cdot MS4A4A \cdot PPP1R27 - 0.018976 \cdot (ARL11 - 1.8836) \cdot (ARL11 - LPPR3) + 2.0583$
AZD0530	0.705923	0.006255	$\ln(IC_{50}) = -0.172041746288402 \cdot ADAP2 \cdot (NAPSA + SPINT2) - 0.172041746288402 \cdot ADAP2 - 0.172041746288402 \cdot EPX - 0.172041746288402 \cdot EVPL - 0.172041746288402 \cdot RREB1 + 1.7207$
PF2341066	0.632117	0.090301	$\ln(IC_{50}) = 0.1143 \cdot AMOTL2 - 0.1143 \cdot CBFA2T3 + 1.734961619 - \frac{0.062470665 \cdot SH2D3C^2 \cdot SLA}{CBFA2T3}$
L-685458	0.696231	0.133541	$\ln(IC_{50}) = 0.096966 \cdot LAX1 - 0.096966 \cdot SELPLG \cdot (ARHGAP9 + 0.38003) + 2.0364 - \frac{0.096966 \cdot LAX1^2 \cdot TAGAP}{IKZF1}$
ZD-6474	0.807793	-0.019903	$\ln(IC_{50}) = -0.147572046548468 \cdot CAPN12 - 0.147572046548468 \cdot SLC39A2 - 0.147572046548468 \cdot SYTL1 \cdot (ACE2 \cdot (SYTL1 + 0.43925) + 0.66305) + 1.6636$
Panobinostat	0.696533	0.337937	$\ln(IC_{50}) = 0.1309 \cdot CYR61 + 0.1309 \cdot GPX8 \cdot (-AMOTL2 + MYOF + 1.246) + 0.1309 \cdot MYOF + 0.1309 \cdot PRSS23 - 2.85876519$
Sorafenib	0.725327	-0.054909	$\ln(IC_{50}) = \frac{GAPT}{TYROBP - 8.97146334319008 + \frac{TYROBP}{CSF3R^2}} + 1.937648$

Irinotecan	1.047721	0.295885	$\ln(IC_{50}) = -0.28966 \cdot ARHGAP19 + 0.28966 \cdot CD63 - \frac{0.0541685043072869 \cdot CD63}{RFXAP} - 0.57932 \cdot SLFN11 - 1.2391$
Topotecan	1.473479	0.053920	$\ln(IC_{50}) = -0.241196333815726 \cdot HMGB2 + 0.22155 \cdot PPIC - 0.241196333815726 \cdot PRPF38A - 0.22155 \cdot RFXAP - 0.597249338815726 \cdot SLFN11 - 0.750699469368066$
LBW242	0.427961	-1.579734	$\ln(IC_{50}) = 0.062723 \cdot IPO5 - 0.062723 \cdot KCNF1 - 0.062723 \cdot LOC285205 \cdot (LOC285205 - SRD5A2 \cdot (KCNF1 + LOC285205)) - 0.062723 \cdot SRD5A2 + 1.9235$
PD-0325901	1.941828	0.161098	$\ln(IC_{50}) = \frac{CAPN3 + FMN1 + GJB1 + LOC100127888 + RXRG - 0.75212}{-0.46727549 \cdot CAPN3 - 2.21475022}$
PD-0332991	0.480822	0.393369	$\ln(IC_{50}) = -0.107 \cdot LOC100506972 - 0.107 \cdot R33 \cdot (0.24282 \cdot P2RX1 \cdot (CECR1 + 0.0623) + RNASE6 + 1.0007) + 1.9301$
Paclitaxel	2.250285	0.024433	$\ln(IC_{50}) = 0.64141 \cdot ABCB1 - 0.35859 \cdot ANP32A - 0.35859 \cdot ANP32B + 0.35859 \cdot C4BPB - 0.35859 \cdot SSRP1 - 4.210463862$
PLX4720	0.645599	0.072950	$\ln(IC_{50}) = \frac{-0.07561 \cdot SPRYD5 - 0.146018032}{0.07561 + \frac{0.34677}{ROPN1} + \frac{0.40141}{MLANA}} + 1.9778$
RAF265	1.309336	-0.062453	$\ln(IC_{50}) = 0.19682 \cdot C7orf70 + 0.19682 \cdot C9orf7 - 0.19682 \cdot PSMD9 - 0.1412636186 \cdot RILPL2 - 0.19682 \cdot SUGP1 + 0.91137501 + \frac{0.0033654}{RILPL2}$
TAE684	1.274869	0.132311	$\ln(IC_{50}) = -0.17343 \cdot CASP2 - 0.17343 \cdot MYB \cdot (-CASP2 + PUM2 + 0.19857) - 0.17343 \cdot PCF11 - 0.17343 \cdot PUM2 + 1.1119$
TKI258	0.676957	0.055170	$\ln(IC_{50}) = -0.090919 \cdot EIF3L - 0.090919 \cdot KDM2B - 0.181838 \cdot ORC1 + 0.090919 \cdot (C1orf228 - ORC1) \cdot (-DPEP2 + ORC1) + 1.6818$
Erlotinib	0.622251	-0.264504	$\ln(IC_{50}) = 1.9177 + \frac{CDH3 \cdot TMPRSS11E + IRF6 \cdot KRT16}{-KRT16 + \frac{-EVPL - 5.99212}{CDH3}}$
AZD6244	1.226761	0.275283	$\ln(IC_{50}) = 0.63947 \cdot CPN1 \cdot (0.036752064 - 0.2528 \cdot RXRG) - 0.63947 \cdot LYZ - 0.161658016 \cdot SPRYD5 + 1.444051154$

Πίνακας 12 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με PySR (random_state=21, niterations=160, deterministic=True, procs=0, multithreading=None)

Φάρμακο	RMSE	R^2	Εξίσωση
AEW541	1.505904	-2.572325	$\ln(IC_{50}) = -0.0699 \cdot AMPD1^2 + 0.0194 \cdot AMPD1 - 0.0392 \cdot CD28^2 + 0.0266 \cdot CD28 + \dots - 0.353 \cdot \max(0, TP53INP1 - 0.618) - 0.354 \cdot \max(0, TP53INP1 + 0.321) + 12.2$
Nilotinib	0.902371	-1.704756	$\ln(IC_{50}) = 0.0453 \cdot APOL4^2 + 0.42 \cdot APOL4 + 0.0225 \cdot C15orf26^2 - 0.0216 \cdot C15orf26 + \dots - 0.0246 \cdot \max(0, TRDC - 2.11) - 0.334 \cdot \max(0, TRDC - 1.04) + 4.63$

17-AAG	3.844898	-7.029871	$\ln(IC_{50}) = 0.976 \cdot CTDSP1^2 - 0.621 \cdot CTDSP1 - 1.67 \cdot DNAJC18^2 + 0.446 \cdot DNAJC18 - \dots + 0.579 \cdot \max(0, TXN + 1.67) - 0.222 \cdot \max(0, TXN + 2.46) - 4.09$
PHA-665752	0.850269	-0.026704	$\ln(IC_{50}) = -0.0568 \cdot KCNE1L + 0.028 \cdot PPP1R27 + 0.0224 \cdot PRTN3 - 0.0739 \cdot TLR8 + 0.0236 \cdot \max(0, 0.651 - PPP1R27) + 0.0478 \cdot \max(0, 1.45 - ECRP) + 0.0561 \cdot \max(0, 1.57 - PRTN3) - 0.0796 \cdot \max(0, 1.68 - ELANE) + 0.00473 \cdot \max(0, 2.08 - MS4A4A) - 0.0969 \cdot \max(0, ELANE - 6.14) - 0.0462 \cdot \max(0, MS4A4A - 11.5) - 0.0631 \cdot \max(0, MS4A4A - 9.65) + 0.763 \cdot \max(0, PPP1R27 - 8.3) - 0.255 \cdot \max(0, TLR8 - 10.7) - 0.14 \cdot \max(0, TLR8 - 8.83) - 0.269 \cdot \max(0, TLR8 - 6.97) + 1.99$
Lapatinib	15.542975	-351.282121	$\ln(IC_{50}) = 0.2 \cdot CATSPERB^2 - 0.0899 \cdot CATSPERB - 0.16 \cdot EPN3^2 - 0.215 \cdot EPN3 \cdot EPN3 + \dots + 0.355 \cdot \max(0, TBC1D30 - 1.83) - 0.459 \cdot \max(0, TBC1D30 - 1.18) - 0.949 \cdot \max(0, TBC1D30 - 0.526) + 3.81$
Nutlin-3	1.248088	-1.459721	$\ln(IC_{50}) = -0.00899 \cdot C5orf20^2 - 0.0114 \cdot C5orf20 + 0.0277 \cdot IGSF6^2 + 0.0428 \cdot IGSF6 + \dots + 0.312 \cdot \max(0, TLR8 - 9.69) - 0.192 \cdot \max(0, TLR8 - 5.64) - 0.0582 \cdot \max(0, TLR8 - 3.61) + 2.46$
AZD0530	1.988474	-6.884976	$\ln(IC_{50}) = -0.112 \cdot ADAP2^2 + 0.0225 \cdot ADAP2 - 0.0199 \cdot EPX^2 - 0.0588 \cdot EPX + \dots + 0.27 \cdot \max(0, SPINT2 + 0.0689) - 0.0663 \cdot \max(0, SPINT2 + 0.484) + 1.95$
PF2341066	1.808543	-6.446648	$\ln(IC_{50}) = -1.86 \cdot AMOTL2^2 - 0.0676 \cdot AMOTL2 - 0.408 \cdot CBFA2T3^2 - 0.601 \cdot CBFA2T3 + \dots - 2.27 \cdot \max(0, SPN - 1.66) + 0.95 \cdot \max(0, SPN - 1.01) - 0.813 \cdot \max(0, SPN - 0.358) + 3.7$
L-685458	1.311052	-2.072423	$\ln(IC_{50}) = 0.554 \cdot ARHGAP9^2 + 0.0841 \cdot ARHGAP9 + 0.102 \cdot IKZF1^2 - 0.00557 \cdot IKZF1 - \dots - 0.436 \cdot \max(0, TAGAP - 1.22) + 0.107 \cdot \max(0, TAGAP - 0.44) + 3.23$
ZD-6474	5.104643	-39.727636	$\ln(IC_{50}) = -0.0583 \cdot ACE2^2 - 0.0898 \cdot ACE2 - 0.878 \cdot CAPN12^2 - 0.329 \cdot CAPN12 - \dots + 0.0228 \cdot \max(0, TNK1 + 0.514) - 0.0225 \cdot \max(0, TNK1 + 1.14) + 2.53$
Panobinostat	0.942686	-0.212689	$\ln(IC_{50}) = -1.69 \cdot AMOTL2^2 + 0.205 \cdot AMOTL2 + 0.0669 \cdot CYR61^2 + 0.561 \cdot CYR61 + \dots + 0.725 \cdot \max(0, TNFRSF12A + 1.27) + 0.0986 \cdot \max(0, TNFRSF12A + 1.79) - 1.77$
Sorafenib	17.049092	-581.841198	$\ln(IC_{50}) = 0.13 \cdot CCL23^2 + 0.0469 \cdot CCL23 + 0.081 \cdot CSF3R^2 + 0.00353 \cdot CSF3R - \dots + 0.447 \cdot \max(0, TYROBP - 3.36) - 0.551 \cdot \max(0, TYROBP - 2.24) - 0.0216 \cdot \max(0, TYROBP - 1.11) + 2.71$
Irinotecan	20.272523	-262.613519	$\ln(IC_{50}) = -1.68 \cdot ARHGAP19^2 + 0.119 \cdot ARHGAP19 + 0.37 \cdot CD63^2 - 1.1 \cdot CD63 - \dots - 0.191 \cdot \max(0, SLFN11 - 0.308) - 0.884 \cdot \max(0, SLFN11 + 0.124) + 0.601 \cdot \max(0, SLFN11 + 0.556) - 8.31$
Topotecan	2.570464	-1.879141	$\ln(IC_{50}) = -2.12 \cdot ARHGAP19^2 + 0.0581 \cdot ARHGAP19 - 1.36 \cdot HMGB2^2 - 0.702 \cdot HMGB2 + \dots + 4.26 \cdot \max(0, SLFN11 - 0.79) + 0.917 \cdot \max(0, SLFN11 - 0.349) - 5.65 \cdot \max(0, SLFN11 + 0.0917) - 0.0644$
LBW242	0.586174	-3.839726	$\ln(IC_{50}) = -0.275 \cdot CABP1 + 0.198 \cdot IPO5 + 0.0774 \cdot KCNF1 + 0.0254 \cdot KLK3 - \dots - 0.237 \cdot \max(0, RNF186 - 3.09) + 0.217 \cdot \max(0, RNF186 - 1.26) - 1.13 \cdot \max(0, SRD5A2 - 5.25) - 0.0286 \cdot \max(0, SRD5A2 - 1.77) + 2.7$
PD-0325901	67.727762	-1019.524443	$\ln(IC_{50}) = 0.618 \cdot CAPN3^2 - 0.0403 \cdot CAPN3 - 1.64 \cdot FMN1^2 - 0.149 \cdot FMN1 + 1.63 \cdot GJB1^2 - \dots - 0.195 \cdot \max(0, TRIM63 - 3.39) - 1.72 \cdot \max(0, TRIM63 - 2.5) - 1.01 \cdot \max(0, TRIM63 - 1.61) + 0.926 \cdot \max(0, TRIM63 - 0.715) - 11.0$
PD-0332991	4.102676	-43.166290	$\ln(IC_{50}) = -0.0351 \cdot CECR1^2 + 0.0282 \cdot CECR1 + 0.268 \cdot KCNA3^2 + 0.397 \cdot KCNA3 + \dots - 1.12 \cdot \max(0, TRAF3IP3 - 0.856) + 0.6 \cdot \max(0, TRAF3IP3 - 0.161) - 4.94$
Paclitaxel	47.135268	-427.029726	$\ln(IC_{50}) = 3.22 \cdot ABCB1^2 - 2.98 \cdot ABCB1 + 0.603 \cdot ANP32A^2 - 0.335 \cdot ANP32A + \dots - 5.63 \cdot \max(0, SSRP1 - 1.26) - 2.18 \cdot \max(0, SSRP1 - 0.254) + 6.36$

PLX4720	1.257106	-2.514974	$\ln(IC_{50}) = -0.369 \cdot CAPN3^2 - 0.179 \cdot CAPN3 - 0.0233 \cdot DCT^2 - 0.139 \cdot DCT + \dots - 0.187 \cdot \max(0, SPRYD5 - 2.9) + 0.287 \cdot \max(0, SPRYD5 - 2.07) - 0.346 \cdot \max(0, SPRYD5 - 1.24) + 0.257 \cdot \max(0, SPRYD5 - 0.418) + 4.46$
RAF265	2.219663	-2.053383	$\ln(IC_{50}) = -0.599 \cdot C7orf70^2 - 0.0364 \cdot C7orf70 + 1.18 \cdot C9orf7^2 - 0.287 \cdot C9orf7 + \dots + 2.76 \cdot \max(0, SUGP1 - 0.993) + 0.378 \cdot \max(0, SUGP1 - 0.132) + 0.116 \cdot \max(0, SUGP1 + 1.59) + 2.47$
TAE684	10.815195	-61.445613	$\ln(IC_{50}) = -1.09 \cdot MYB^2 - 0.485 \cdot MYB + 1.03 \cdot P4HA2^2 - 0.576 \cdot P4HA2 - 1.21 \cdot PCF11^2 + \dots + 0.587 \cdot \max(0, WWTR1 - 0.253) + 0.0465 \cdot \max(0, WWTR1 + 0.297) + 0.0788 \cdot \max(0, WWTR1 + 0.847) + 0.512$
TKI258	1.318395	-2.583627	$\ln(IC_{50}) = 0.0661 \cdot C1orf228^2 + 0.0623 \cdot C1orf228 + 0.0398 \cdot DPEP2^2 - 0.221 \cdot DPEP2 - \dots - 0.231 \cdot \max(0, ORC1 - 1.61) - 0.278 \cdot \max(0, ORC1 - 0.826) - 2.05 \cdot \max(0, ORC1 - 0.0392) + 0.169 \cdot \max(0, ORC1 + 0.747) + 7.08$
Erlotinib	2.383947	-17.560195	$\ln(IC_{50}) = -0.459 \cdot CDH3^2 - 0.13 \cdot CDH3 + 0.11 \cdot EVPL^2 - 0.235 \cdot EVPL - \dots - 1.96 \cdot \max(0, TMPRSS11E - 2.16) + 0.225 \cdot \max(0, TMPRSS11E - 1.35) + 0.193 \cdot \max(0, TMPRSS11E - 0.541) + 0.922$
AZD6244	1.505904	-2.572325	$\ln(IC_{50}) = -0.102 \cdot CAPN3^2 + 0.312 \cdot CAPN3 - 0.207 \cdot CPN1^2 + 0.433 \cdot CPN1 + \dots + 0.82 \cdot \max(0, SPRYD5 - 2.86) - 1.41 \cdot \max(0, SPRYD5 - 2.03) - 1.48 \cdot \max(0, SPRYD5 - 1.2) + 0.748 \cdot \max(0, SPRYD5 - 0.372) - 14.9$

Πίνακας 13 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με FFX

Φάρμακο	RMSE	R^2	Εξίσωση
AEW541	0.854828	-0.151102	$\ln(IC_{50}) = -0.0383928353869147 \cdot AMPD1 - \dots - 0.061298849925163 \cdot TP53INP1$ $+ (0.0406491302838408 \cdot AMPD1 - \dots + 0.057693906708619 \cdot TP53INP1 + 0.38856512835859)$ $\cdot (0.0832347646355629 \cdot AMPD1 - \dots + 0.0875449925661087 \cdot TP53INP1 + 0.556357562541962)$ $- 0.400789529085159$ $\cdot \sin(0.469616830348969 \cdot AMPD1 - \dots + 0.27860414981842 \cdot TP53INP1 - 0.589265942573547)$ $+ 0.257741808891296$ $\cdot \cos(0.0401993915438652 \cdot AMPD1 - \dots + 0.128800898790359 \cdot TP53INP1 - 0.153957486152649)$ $+ 0.543495496394018 + \frac{0.432567745447159}{0.677046815502507 \cdot e^{0.206169873476028 \cdot AMPD1 - \dots + 0.126029178500175 \cdot TP53INP1} + 1}$

Nilotinib	0.556036	-0.026982	$\ln(IC_{50}) = -0.0319117343617088 \cdot ANKRD55 - \dots + 0.00520352140093833 \cdot SAMS1 - 0.015192404104956 \cdot TRDC$ $+ (0.00209123337317985 \cdot ANKRD55 - \dots - 0.0984643943771245 \cdot TRDC + 0.324727493246115)$ $\cdot (0.219764187932014 \cdot ANKRD55 + \dots + 0.262310028076172 \cdot TRDC + 0.571885049343109)$ $- 0.467307597398758$ $\cdot \sin(0.0740100964903831 \cdot ANKRD55 + \dots + 0.0185082200914621 \cdot TRDC - 0.637901842594147)$ $+ 0.477982550859451$ $\cdot \cos(0.106039308011532 \cdot ANKRD55 + \dots + 0.140751332044601 \cdot TRDC - 0.185045421123505)$ $+ 0.594428417634882 + \frac{0.541182518005371}{0.591984500487514 \cdot e^{0.171615570783615 \cdot ANKRD55 + \dots - 0.025081392377615 \cdot TRDC} + 1}$
17-AAG	1.242864	0.160953	$\ln(IC_{50}) = -0.150781242496116 \cdot CTDSP1 + \dots - 0.103132204866739 \cdot TXN$ $- (0.0142308474674287 \cdot CTDSP1 + \dots - 0.0890640958636626 \cdot TXN + 0.203529993282761)$ $\cdot (0.0447552539408207 \cdot CTDSP1 + \dots - 0.195620238780975 \cdot TXN + 0.47743684053421)$ $- 0.609167695045471$ $\cdot \sin(0.119848698377609 \cdot CTDSP1 - \dots + 0.0981440618634224 \cdot TXN + 0.510238111019135)$ $- 0.397218853235245$ $\cdot \cos(0.373847603797913 \cdot CTDSP1 + \dots - 0.260353326797485 \cdot TXN + 0.0732168331742287)$ $- 0.683351831242192 - \frac{0.357717573642731}{0.731537422735441 \cdot e^{-0.0371087044477463 \cdot CTDSP1 - \dots - 0.0213779080659151 \cdot TXN} + 1}$
PHA-665752	0.863945	-0.059997	$\ln(IC_{50}) = 0.0598787710161619 \cdot ECRP + \dots + 0.00637984558144689 \cdot TLR8$ $- (-0.02448002807796 \cdot ECRP + \dots + 0.0948596373200417 \cdot TLR8 + 0.722687423229218)$ $\cdot (0.037885162494232 \cdot ECRP + \dots + 0.0229271970404557 \cdot TLR8 - 0.361518123853227)$ $- 0.475895553827286$ $\cdot \sin(0.286750435829163 \cdot ECRP + \dots + 0.171277612447739 \cdot TLR8 - 0.671014726161957)$ $+ 0.461808681488037$ $\cdot \cos(0.0466577894985676 \cdot ECRP + \dots + 0.174131885170937 \cdot TLR8 - 0.31396272778511)$ $+ 0.740185265317727 + \frac{0.434418737888336}{0.571319861236457 \cdot e^{0.109882883727551 \cdot ECRP - \dots - 0.00331234629265964 \cdot TLR8} + 1}$
Lapatinib	0.714931	0.254666	$\ln(IC_{50}) = -0.0292555132043648 \cdot CATSPERB - \dots - 0.101627628309706 \cdot TBC1D30$ $- (-0.0194453615695238 \cdot CATSPERB + \dots + 0.0661927312612534 \cdot TBC1D30$ $+ 0.520062804222107)$ $\cdot (0.040958632450632 \cdot CATSPERB - \dots + 0.0214055661624899 \cdot TBC1D30 - 0.144282159820506)$ $+ 0.596807956695557$ $\cdot \sin(0.0208250097930431 \cdot CATSPERB + \dots - 0.0295650120824575 \cdot TBC1D30$ $+ 0.625838160514832) + 0.515604019165039$ $\cdot \cos(0.137000113725662 \cdot CATSPERB - \dots + 0.405521214008331 \cdot TBC1D30 - 0.54210352897644)$ $+ 0.767084299103168$ $+ \frac{0.429107546806335}{0.483072413001164 \cdot e^{0.0614253878593445 \cdot CATSPERB - \dots + 0.182096615433693 \cdot TBC1D30} + 1}$

Nutlin-3	0.805334	-0.024112	$\ln(IC_{50}) = -0.0377277993414173 \cdot ARL11 + \dots + 0.0310926381512893 \cdot TLR8$ $+ (0.0113907398014576 \cdot ARL11 - \dots - 0.00103500602793606 \cdot TLR8 + 0.0178996293773964)$ $\cdot (0.220390841364861 \cdot ARL11 - \dots - 0.0795974582433701 \cdot TLR8 + 0.314442306756973)$ $- 0.65543794631958$ $\cdot \sin(-0.0223407316952944 \cdot ARL11 + \dots + 0.102065183222294 \cdot TLR8 - 0.660381734371185)$ $+ 0.424889802932739$ $\cdot \cos(-0.312603980302811 \cdot ARL11 + \dots - 0.0193206630647182 \cdot TLR8 + 0.162396773695946)$ $+ 0.945840080344993 + \frac{0.436692535877228}{0.546141893581773 \cdot e^{-0.00803623721003532 \cdot ARL11 - \dots - 0.145447611808777 \cdot TLR8} + 1}$
AZD0530	0.714115	-0.016942	$\ln(IC_{50}) = -0.082415354934863 \cdot ADAP2 - \dots + 0.0342642245791769 \cdot SPINT2$ $- (0.01266246962911 \cdot ADAP2 + \dots - 0.0787348374589349 \cdot SPINT2 + 0.0897721950041975)$ $\cdot (0.0810511857271194 \cdot ADAP2 + \dots + 0.465775698423386 \cdot SPINT2 - 0.357956141233444)$ $- 0.520180642604828$ $\cdot \sin(0.00440635019913316 \cdot ADAP2 + \dots - 0.060454998165369 \cdot SPINT2 - 0.671943426132202)$ $+ 0.447325438261032$ $\cdot \cos(0.477528661489487 \cdot ADAP2 + \dots + 0.251232802867889 \cdot SPINT2 - 0.246834069490433)$ $+ 0.696970272266356 + \frac{0.39019700884819}{0.585625884991283 \cdot e^{0.0919095575809479 \cdot ADAP2 + \dots - 0.066566102206707 \cdot SPINT2} + 1}$
PF2341066	0.624906	0.110938	$\ln(IC_{50}) = -0.0444323121143182 \cdot AMOTL2 - \dots - 0.0471422507450805 \cdot SPN$ $- (-0.0610114724127717 \cdot AMOTL2 + \dots + 0.0238917505348257 \cdot SPN - 0.0340740624315905)$ $\cdot (0.31614550948143 \cdot AMOTL2 - \dots - 0.0335843190550804 \cdot SPN + 0.21216008067131)$ $- 0.358036905527115$ $\cdot \sin(-0.749536633491516 \cdot AMOTL2 + \dots + 0.432523727416992 \cdot SPN - 0.22176867723465)$ $+ 0.486603796482086$ $\cdot \cos(0.149700492620468 \cdot AMOTL2 + \dots + 0.194354817271233 \cdot SPN - 0.222095742821693)$ $+ 0.660354949776691 + \frac{0.465581864118576}{0.608050385968191 \cdot e^{0.20104618370533 \cdot AMOTL2 + \dots + 0.132784321904182 \cdot SPN} + 1}$
L-685458	0.743476	0.011958	$\ln(IC_{50}) = -0.0538146678058915 \cdot ARHGAP9 - \dots - 0.0430575346166231 \cdot TAGAP$ $+ (0.0628494670019463 \cdot ARHGAP9 + \dots - 0.00220302582318567 \cdot TAGAP + 0.204319343448418)$ $\cdot (0.114459335803986 \cdot ARHGAP9 + \dots + 0.0601550079882145 \cdot TAGAP + 0.593214809894562)$ $- 0.438441902399063$ $\cdot \sin(0.336620151996613 \cdot ARHGAP9 + \dots + 0.327439814805984 \cdot TAGAP - 0.596537590026855)$ $+ 0.463247448205948$ $\cdot \cos(0.176415756344795 \cdot ARHGAP9 + \dots - 0.161491677165031 \cdot TAGAP + 0.335541009902954)$ $+ 0.645312467158028 + \frac{0.463121801614761}{0.492122011861413 \cdot e^{0.251779168844223 \cdot ARHGAP9 + \dots + 0.133995458483696 \cdot TAGAP} + 1}$

ZD-6474	0.804794	-0.012343	$\ln(IC_{50}) = -0.052793476251022 \cdot ACE2 - \dots + 0.00596927749273357 \cdot TNK1$ $+ (0.043827761448325 \cdot ACE2 - \dots + 0.0590089335979014 \cdot TNK1 - 0.0667445464487403)$ $\cdot (0.233449548482895 \cdot ACE2 + \dots + 0.216857492923737 \cdot TNK1 - 0.365786582231522)$ $- 0.530112683773041$ $\cdot \sin(0.52008044719696 \cdot ACE2 + \dots - 0.0718662738800049 \cdot TNK1 - 0.626292884349823)$ $+ 0.395022809505463$ $\cdot \cos(0.0913972482085228 \cdot ACE2 + \dots + 0.140794903039932 \cdot TNK1 - 0.0480803772807121)$ $+ 0.59966475228029 + \frac{0.453720599412918}{0.586774230756211 \cdot e^{0.255279839038849 \cdot ACE2 + \dots + 0.0125558590516448 \cdot TNK1} + 1}$
Panobinostat	0.658837	0.407659	$\ln(IC_{50}) = 0.0205204685409279 \cdot AMOTL2 + \dots - 0.0374497347953719 \cdot TNFRSF12A$ $- (0.0395812227666211 \cdot AMOTL2 + \dots + 0.0188104552922627 \cdot TNFRSF12A$ $- 0.0433104314063675)$ $\cdot (0.237883910536766 \cdot AMOTL2 + \dots + 0.0678532049059868 \cdot TNFRSF12A - 0.189845368266106)$ $- 0.829602003097534$ $\cdot \sin(0.160284131765366 \cdot AMOTL2 - \dots + 0.137320697307587 \cdot TNFRSF12A$ $+ 0.975382626056671) - 0.525240778923035$ $\cdot \cos(0.144496604800224 \cdot AMOTL2 + \dots + 0.22505259513855 \cdot TNFRSF12A + 0.336577862501144)$ $- 1.29884045500772$ $- \frac{0.526631414890289}{0.451112372998829 \cdot e^{-0.0759551152586937 \cdot AMOTL2 + \dots - 0.149743661284447 \cdot TNFRSF12A} + 1}$
Sorafenib	0.744153	-0.110381	$\ln(IC_{50}) = -0.0306560053509757 \cdot CCL23 - \dots + 0.0138678610759185 \cdot TYROBP$ $- (0.0115822867394314 \cdot CCL23 + \dots + 0.00899492200759813 \cdot TYROBP - 0.0324487234187321)$ $\cdot (0.0637812092900276 \cdot CCL23 + \dots + 0.216057434678078 \cdot TYROBP + 0.138913288712502)$ $- 0.448133289813995$ $\cdot \sin(0.120003394782543 \cdot CCL23 + \dots + 0.187845230102539 \cdot TYROBP - 0.617390215396881)$ $+ 0.469020158052444$ $\cdot \cos(0.016497015953064 \cdot CCL23 + \dots + 0.0946232080459595 \cdot TYROBP - 0.277789860963821)$ $+ 0.839736136653908 + \frac{0.484889149665833}{0.512554122827166 \cdot e^{0.0110034076496959 \cdot CCL23 + \dots - 0.0447750948369503 \cdot TYROBP} + 1}$

Irinotecan	0.984114	0.378784	$\ln(IC_{50}) = -0.0934792938326003 \cdot ARHGAP19 + \dots - 0.197385836676284 \cdot SLFN11$ $- (0.0297665635289412 \cdot ARHGAP19 - \dots + 0.0411551478661254 \cdot SLFN11$ $+ 0.0439996465122938)$ $\cdot (0.0688021332025528 \cdot ARHGAP19 - \dots + 0.211780309677124 \cdot SLFN11 + 0.219069436192513)$ $- 0.518098115921021$ $\cdot \sin(0.134301796555519 \cdot ARHGAP19 - \dots + 0.61832857131958 \cdot SLFN11 + 0.515163779258728)$ $- 0.426208704710007$ $\cdot \cos(0.326393097639084 \cdot ARHGAP19 - \dots + 0.291065096855164 \cdot SLFN11$ $+ 0.000919792335480452) - 0.430479822300562$ 0.386193096637726 $- \frac{0.599069488940032 \cdot e^{-0.148753881454468 \cdot ARHGAP19 + \dots - 0.544402778148651 \cdot SLFN11} + 1}{1}$
Topotecan	1.231213	0.339449	$\ln(IC_{50}) = -0.0697943017519389 \cdot ARHGAP19 + \dots - 0.288084218918485 \cdot SLFN11$ $- (0.0706525564654238 \cdot ARHGAP19 - \dots + 0.0581333661221653 \cdot SLFN11 + 0.101253865800475)$ $\cdot (0.249702796339989 \cdot ARHGAP19 - \dots + 0.247055694460869 \cdot SLFN11 + 0.3181092441082)$ $- 0.504336476325989$ $\cdot \sin(-0.0623004212975502 \cdot ARHGAP19 - \dots + 0.859442055225372 \cdot SLFN11$ $+ 0.210722014307976) - 0.325596511363983$ $\cdot \cos(0.0630510672926903 \cdot ARHGAP19 + \dots - 0.407091408967972 \cdot SLFN11$ $+ 0.321089565753937) - 0.29024053112699$ 0.361397713422775 $- \frac{0.809222815838597 \cdot e^{-0.0308107901364565 \cdot ARHGAP19 + \dots - 0.73715478181839 \cdot SLFN11} + 1}{1}$
LBW242	0.403192	-1.289763	$\ln(IC_{50}) = -0.0501647978233066 \cdot CABP1 + \dots - 0.0788877423805268 \cdot SRD5A2$ $+ (-0.087960459291935 \cdot CABP1 + \dots - 0.041129183024168 \cdot SRD5A2 - 0.748088359832764)$ $\cdot (0.0189255973495419 \cdot CABP1 + \dots + 0.0343751649217976 \cdot SRD5A2 - 0.435326585689019)$ $+ 0.57026219367981$ $\cdot \sin(0.173572242259979 \cdot CABP1 + \dots + 0.094735361635685 \cdot SRD5A2 + 0.700207889080048)$ $+ 0.464535713195801$ $\cdot \cos(0.0314651429653168 \cdot CABP1 + \dots - 0.313488364219666 \cdot SRD5A2 - 0.346247732639313)$ 0.289761245250702 $+ 0.686430382949666 + \frac{0.796456106323834 \cdot e^{-0.214637473225594 \cdot CABP1 - \dots - 0.0176049172878265 \cdot SRD5A2} + 1}{1}$
PD-0325901	1.953712	0.150798	$\ln(IC_{50}) = -0.145863176719117 \cdot CAPN3 - \dots - 0.0894271521798138 \cdot TRIM63$ $- (-0.160134553909302 \cdot CAPN3 + \dots - 0.111392185091972 \cdot TRIM63 + 0.336367696523666)$ $\cdot (-0.0618529688804154 \cdot CAPN3 + \dots - 0.0312187800803802 \cdot TRIM63 + 0.11802668165072)$ $- 0.708837866783142$ $\cdot \sin(0.397627949714661 \cdot CAPN3 + \dots - 0.203472033143044 \cdot TRIM63 - 0.0758784934878349)$ $+ 0.382890522480011$ $\cdot \cos(-0.121331073343754 \cdot CAPN3 + \dots + 0.126467749476433 \cdot TRIM63 + 0.250133275985718)$ 0.0370023213326931 $+ 0.169043741652591 + \frac{1.13554053803539 \cdot e^{0.0176316257566214 \cdot CAPN3 + \dots + 0.0953004881739616 \cdot TRIM63} + 1}{1}$

PD-0332991	0.567796	0.154057	$\ln(IC_{50}) = -0.0225250208035108 \cdot CECR1 + \dots + 0.0436100482427322 \cdot TRAF3IP3$ $+ (0.017326571233862 \cdot CECR1 + \dots + 0.040408504627275 \cdot TRAF3IP3 - 0.151174204326321)$ $\cdot (0.0982911065220833 \cdot CECR1 + \dots + 0.128034338355064 \cdot TRAF3IP3 - 0.525022625923157)$ $- 0.199250459671021$ $\cdot \sin(0.417340159416199 \cdot CECR1 + \dots + 0.583121240139008 \cdot TRAF3IP3 - 0.547408521175385)$ $+ 0.464864164590836$ $\cdot \cos(0.258850127458572 \cdot CECR1 + \dots + 0.0879783406853676 \cdot TRAF3IP3 - 0.339129358530045)$ $+ 0.946239587342205 + \frac{0.490981089533046 \cdot e^{-0.23110681772321 \cdot CECR1 - \dots - 0.226600348949432 \cdot TRAF3IP3} + 1}{0.490981089533046}$
Paclitaxel	2.230213	0.041759	$\ln(IC_{50}) = 0.365879432227764 \cdot ABCB1 - \dots - 0.256836630029504 \cdot SSRP1$ $- (-0.171991392970085 \cdot ABCB1 + \dots - 0.0214373767375946 \cdot SSRP1 + 0.964126169681549)$ $\cdot (-0.0669506217324286 \cdot ABCB1 + \dots - 0.000625907616553817 \cdot SSRP1 + 0.345100699552681)$ $+ 1.08793020248413$ $\cdot \sin(0.027042668312788 \cdot ABCB1 - \dots + 0.143155157566071 \cdot SSRP1 - 1.15887522697449)$ $- 0.796992361545563$ $\cdot \cos(-0.24960295855999 \cdot ABCB1 + \dots + 0.25225031375885 \cdot SSRP1 + 0.21514467895031)$ $- 1.81112782229575 - \frac{0.756839871406555}{0.346337110833773 \cdot e^{0.204377219080925 \cdot ABCB1 - \dots - 0.188603058457375 \cdot SSRP1} + 1}$
PLX4720	0.642340	0.082284	$\ln(IC_{50}) = -0.0958520269661198 \cdot CAPN3 - \dots - 0.0204736780536424 \cdot SPRYD5$ $+ (0.0142737704450767 \cdot CAPN3 + \dots + 0.0186419964808168 \cdot SPRYD5 - 0.100968766030747)$ $\cdot (0.050442755223206 \cdot CAPN3 + \dots + 0.0991767272353172 \cdot SPRYD5 - 0.462305724620819)$ $- 0.325947165489197$ $\cdot \sin(0.260130912065506 \cdot CAPN3 + \dots + 0.472223937511444 \cdot SPRYD5 - 0.713068068027496)$ $+ 0.508555948734283$ $\cdot \cos(0.1228301897645 \cdot CAPN3 + \dots + 0.108603447675705 \cdot SPRYD5 - 0.174087330698967)$ $+ 0.756429424602329 + \frac{0.462642043828964}{0.548513634728603 \cdot e^{-0.233648672699928 \cdot CAPN3 - \dots - 0.30969774723053 \cdot SPRYD5} + 1}$
RAF265	1.366854	-0.157848	$\ln(IC_{50}) = -0.0597548900446192 \cdot AK2 + \dots - 0.191235837937725 \cdot SUGP1$ $- (-0.0191202068625795 \cdot AK2 + \dots + 0.166101862048933 \cdot SUGP1 + 0.112322959909292)$ $\cdot (-0.00498305587098002 \cdot AK2 - \dots - 0.222151100635529 \cdot SUGP1 - 0.224360212683678)$ $+ 0.345675259828568$ $\cdot \sin(0.0591106340289116 \cdot AK2 + \dots - 0.17129972577095 \cdot SUGP1 + 0.444293469190598)$ $+ 0.351380407810211$ $\cdot \cos(0.262828886508942 \cdot AK2 - \dots + 0.20024736225605 \cdot SUGP1 - 0.068990133702755)$ $+ 0.414443596432022 + \frac{0.232036009430885}{0.851611267530992 \cdot e^{-0.0665842369198799 \cdot AK2 - \dots + 0.3575539290905 \cdot SUGP1} + 1}$

TAE684	1.265584	0.144903	$\ln(IC_{50}) = -0.0546438845446524 \cdot CASP2 + \dots - 0.0111384151929627 \cdot WWTR1$ $+ (-0.109137900173664 \cdot CASP2 + \dots + 0.237138226628304 \cdot WWTR1 + 0.168925598263741)$ $\cdot (-0.017016986365488 \cdot CASP2 + \dots + 0.0365758791947943 \cdot WWTR1 + 0.0306272943941081)$ $+ 0.378867834806442$ $\cdot \sin(-0.264357060194016 \cdot CASP2 + \dots + 0.108326271176338 \cdot WWTR1 + 0.594939291477203)$ $+ 0.353124260902405$ $\cdot \cos(0.020853353664279 \cdot CASP2 + \dots - 0.0475377067923546 \cdot WWTR1 + 0.110969625413418)$ $+ 0.403527105321193 + \frac{0.625578876598796 \cdot e^{0.134557247161865 \cdot CASP2 + \dots + 0.0599746964871883 \cdot WWTR1} + 1}{0.323940187692642}$
TKI258	0.622303	0.201573	$\ln(IC_{50}) = -0.0352686947350676 \cdot C1orf228 - \dots - 0.00751141480438677 \cdot ORC1$ $+ (0.0332375105189746 \cdot C1orf228 - \dots + 0.0863799576077539 \cdot ORC1 - 0.194472720262461)$ $\cdot (0.201920345425606 \cdot C1orf228 + \dots + 0.185305565595627 \cdot ORC1 - 0.680621206760406)$ $+ 0.5070680975914$ $\cdot \sin(-0.0411964394152164 \cdot C1orf228 + \dots + 0.102314665913582 \cdot ORC1 + 0.684940695762634)$ $+ 0.393782198429108$ $\cdot \cos(0.0722227096557617 \cdot C1orf228 + \dots - 0.529388844966888 \cdot ORC1 - 0.309654861688614)$ $+ 0.581004472944173$ $+ \frac{0.395039439201355}{0.545176196761727 \cdot e^{-0.0768337175250053 \cdot C1orf228 + \dots + 0.0212045609951019 \cdot ORC1} + 1}$
Erlotinib	0.600369	-0.177133	$\ln(IC_{50}) = -0.100148882694197 \cdot CDH3 - \dots - 0.0638108927619307 \cdot TMPRSS11E$ $+ (0.0279966714172708 \cdot CDH3 + \dots - 0.0505616835773104 \cdot TMPRSS11E + 0.0142830717640676)$ $\cdot (0.202526807785034 \cdot CDH3 + \dots - 0.167943552136421 \cdot TMPRSS11E + 0.0834684520959854)$ $- 0.640294373035431$ $\cdot \sin(0.0189005807042122 \cdot CDH3 + \dots + 0.131301447749138 \cdot TMPRSS11E - 0.821081757545471)$ $+ 0.485994011163712$ $\cdot \cos(0.465230852365494 \cdot CDH3 + \dots + 0.10785748809576 \cdot TMPRSS11E - 0.356593459844589)$ $+ 0.824254312185225$ $+ \frac{0.422121495008469}{0.515362024083309 \cdot e^{0.0796997919678688 \cdot CDH3 + \dots + 0.0207446403801441 \cdot TMPRSS11E} + 1}$
AZD6244	1.173328	0.337040	$\ln(IC_{50}) = 0.0214658983103668 \cdot CAPN3 + \dots - 0.0866502631953963 \cdot SPRYD5$ $- (0.0182234146602267 \cdot CAPN3 + \dots - 0.00598157806769873 \cdot SPRYD5 - 0.213824012033811)$ $\cdot (0.169173166155815 \cdot CAPN3 + \dots + 0.0553271435201168 \cdot SPRYD5 + 0.0492627397179604)$ $- 0.279703289270401$ $\cdot \sin(0.0824903771281242 \cdot CAPN3 + \dots + 0.0331891141831875 \cdot SPRYD5 - 0.492052465677261)$ $+ 0.491845160722733$ $\cdot \cos(-0.0379341877996922 \cdot CAPN3 - \dots + 0.132581442594528 \cdot SPRYD5 + 0.161847099661827)$ $+ 0.648938834251565 + \frac{0.421451419591904}{0.55348243783415 \cdot e^{-0.189851924777031 \cdot CAPN3 - \dots + 0.0669686123728752 \cdot SPRYD5} + 1}$

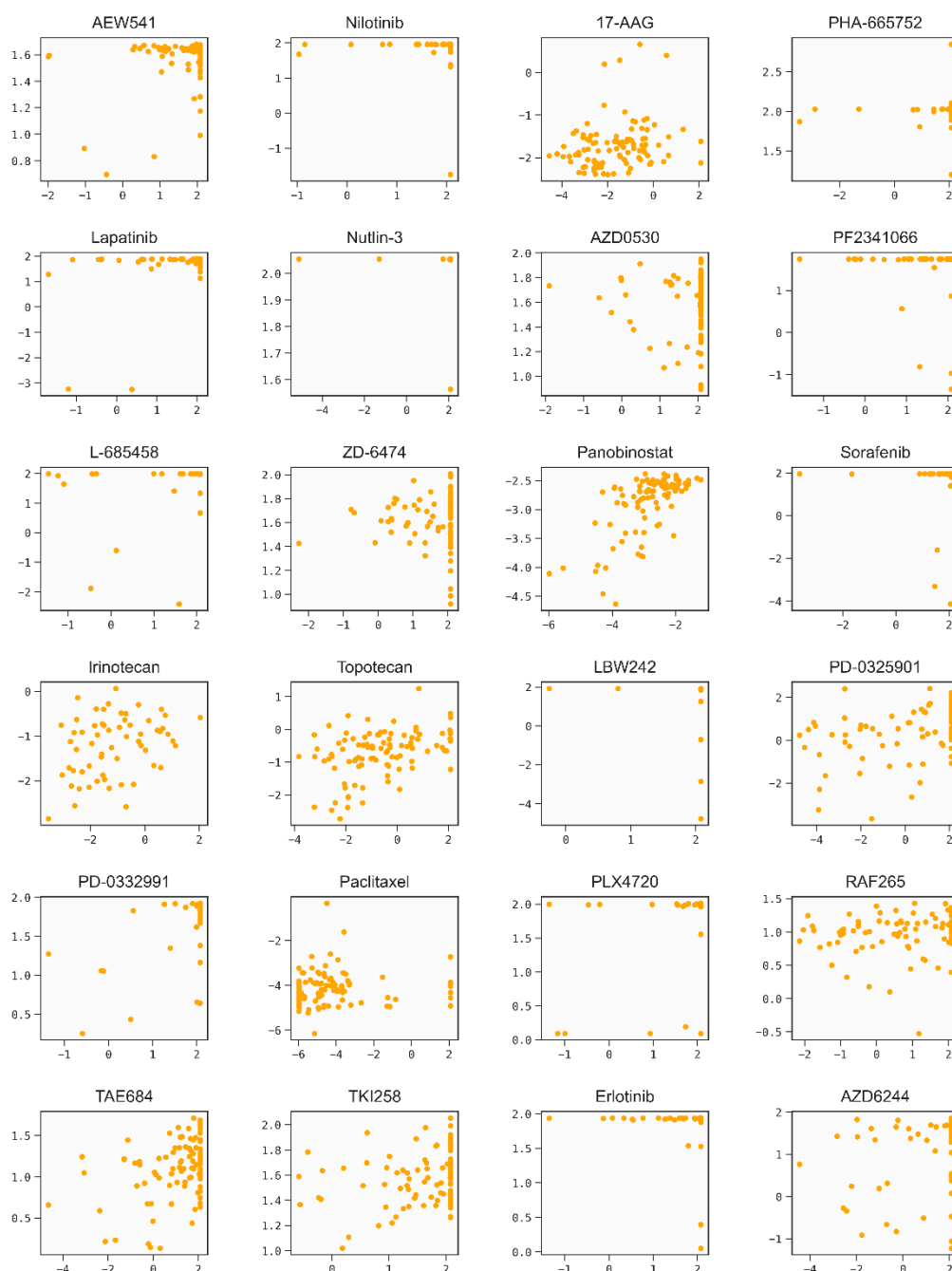
Πίνακας 14 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με EQL(num_layers=1, dim=train_X.shape[1])

Φάρμακο	RMSE	R ²	Εξίσωση
AEW541	9.501497	-141.213363	$\ln(IC_{50}) = 1.595$
Nilotinib	0.550551	-0.006823	$\ln(IC_{50}) = 1.861$
17-AAG	3.800998	-6.847554	$\ln(IC_{50}) = -1.729$
PHA-665752	0.850427	-0.027087	$\ln(IC_{50}) = 2.015$
Lapatinib	1.344981	-1.637878	$\ln(IC_{50}) = -\frac{0.005 \cdot CATSPERB^2 \cdot EPN3^2 \cdot ERBB2^2 \cdot LAD1}{0.004 \cdot CATSPERB \cdot EPN3^2 \cdot PGAP3^2} - \frac{0.001 \cdot CATSPERB^2 \cdot EPN3^2 \cdot PGAP3^2}{ERBB2 \cdot TBC1D30^2}$ $- \frac{LAD1 \cdot TBC1D30}{0.001 \cdot CATSPERB \cdot EPN3^2 \cdot GRB7^2 \cdot LAD1 \cdot PGAP3^2 \cdot TBC1D30} + \frac{ERBB2^2}{TBC1D30^2 + 1.855} - 0.011 \cdot GRB7^2 \cdot LAD1^2$
Nutlin-3	0.800490	-0.011830	$\ln(IC_{50}) = 2.047$
AZD0530	0.710304	-0.006118	$\ln(IC_{50}) = 0.015 \cdot ADAP2^2 \cdot EPX^2 \cdot JUP \cdot NAPSA - 0.055 \cdot ADAP2^2 \cdot EPX^2 \cdot SPINT2 - 0.077 \cdot ADAP2^2 \cdot NAPSA$ $- 0.036 \cdot EPX^2 \cdot JUP + 0.092 \cdot EPX \cdot EVPL \cdot NAPSA + 1.704$
PF2341066	1.969331	-7.829593	$\ln(IC_{50}) = -0.003 \cdot AMOTL2^2 \cdot CBFA2T3^2 \cdot IL2RG \cdot PTPN7 + 0.011 \cdot AMOTL2^2 \cdot IL2RG^2 \cdot SPN^2$ $- \frac{0.02 \cdot AMOTL2^2 \cdot IL2RG \cdot SH2D3C \cdot SLA \cdot SPN}{CBFA2T3 \cdot PTPN7} + 0.073 \cdot AMOTL2 \cdot IL2RG^2 \cdot SPN + 1.786$ $+ \frac{0.002 \cdot IL2RG \cdot SPN^2}{CBFA2T3^2 \cdot PTPN7}$
L-685458	1.847614	-5.101887	$\ln(IC_{50}) = 1.999$
ZD-6474	38.917001	-2366.212092	$\ln(IC_{50}) = 1.661$
Panobinostat	0.656784	0.411344	$\ln(IC_{50}) = -0.017 \cdot AMOTL2^2 \cdot CYR61^2 \cdot PRSS23 \cdot TNFRSF12A + \frac{0.122 \cdot AMOTL2^2 \cdot PRSS23 \cdot TNFRSF12A}{CYR61}$ $- 0.097 \cdot AMOTL2 \cdot CYR61 \cdot PRSS23^2 \cdot TNFRSF12A + 0.167 \cdot CYR61^2 \cdot PRSS23 + 0.12 \cdot CYR61 \cdot PRSS23^2 - 2.762$
Sorafenib	0.853887	-0.462004	$\ln(IC_{50}) = 1.92$
Irinotecan	79.06265	-4008.543109	$\ln(IC_{50}) = \frac{0.001 \cdot ARHGAP19^2 \cdot CD63^2 \cdot CPSF6 \cdot NRF1^2 \cdot RFXAP \cdot SLFN11^2}{S100A13}$ $+ \frac{0.001 \cdot ARHGAP19 \cdot CD63 \cdot S100A13^2 \cdot SLFN11^2}{CPSF6 \cdot RFXAP^2} - 1.131$ $+ \frac{0.001 \cdot CD63^2 \cdot CPSF6^2 \cdot NRF1^2 \cdot RFXAP^2 \cdot S100A13}{ARHGAP19 \cdot SLFN11^2}$
Topotecan	1.305802	0.256989	$\ln(IC_{50}) = 0.005 \cdot HMGB2 \cdot PPIC^2 \cdot SLFN11^2 + 0.031 \cdot HMGB2 \cdot SLFN11 - 0.522 \cdot HMGB2 - 0.074 \cdot PPIC^2$ $\cdot SLFN11^2 - 0.599 \cdot SLFN11 - 0.675$

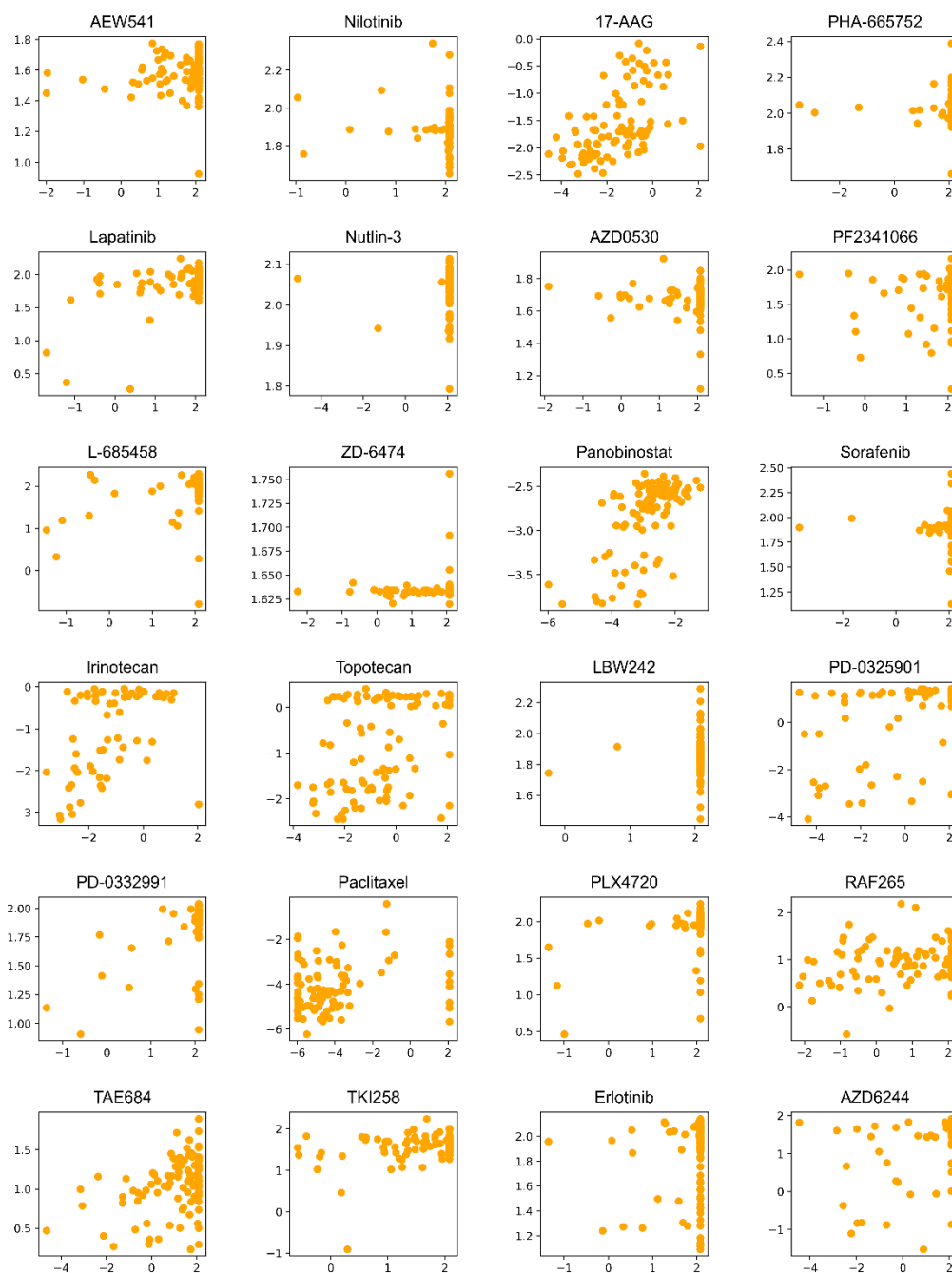
LBW242	0.351456	-0.739839	$\ln(IC_{50}) = \frac{0.01 \cdot CABP1 \cdot IPO5 \cdot KLK3^2 \cdot RNF186^2 \cdot SRD5A2^2}{KCNF1} + 1.892$ $+ \frac{0.02 \cdot IPO5^2 \cdot KCNF1^2 \cdot KLK3 \cdot LOC285205^2 \cdot SRD5A2^2}{CABP1}$
PD-0325901	24.456600	-132.070613	$\ln(IC_{50}) = -\frac{0.007 \cdot CAPN3^2 \cdot FMN1 \cdot GJB1 \cdot RXRG \cdot TRIM63}{SPRYD5^2} - \frac{0.002 \cdot CAPN3^2 \cdot FMN1 \cdot GJB1 \cdot RXRG^2}{LOC100127888 \cdot TRIM63^2}$ $+ \frac{0.004 \cdot CAPN3^2 \cdot GJB1 \cdot RXRG^2 \cdot TRIM63^2}{0.018 \cdot CAPN3^2 \cdot RXRG \cdot TRIM63^2} - \frac{FMN1 \cdot SPRYD5^2}{FMN1^2}$ $- \frac{0.003 \cdot FMN1^2 \cdot GJB1^2 \cdot LOC100127888^2}{TRIM63^2} + 0.795$
PD-0332991	0.763010	-0.527629	$\ln(IC_{50}) = -0.004 \cdot CECR1^2 \cdot RB33^2 \cdot TRAF3IP3 - 0.001 \cdot CECR1 \cdot RB33^2 \cdot RNASE6^2 \cdot TRAF3IP3$ $- \frac{0.001 \cdot KCNA3^2 \cdot P2RX1^2}{LOC100506972 \cdot RB33 \cdot TRAF3IP3} + 1.929 - \frac{0.006 \cdot LOC100506972 \cdot RB33^2 \cdot TRAF3IP3^2}{KCNA3}$ $- \frac{0.002 \cdot LOC100506972^2 \cdot RB33^2 \cdot RNASE6 \cdot TRAF3IP3^2}{CECR1 \cdot KCNA3}$
Paclitaxel	2.281237	-0.002589	$\ln(IC_{50}) = -4.257$
PLX4720	175.29670	-68346.961875	$\ln(IC_{50}) = -\frac{0.034 \cdot CAPN3^2 \cdot RXRG}{SLC45A2} + \frac{0.018 \cdot CAPN3 \cdot RXRG^2}{SLC45A2} + \frac{0.006 \cdot CAPN3 \cdot MLANA \cdot ROPN1 \cdot RXRG^2}{DCT \cdot SLC45A2}$ $- \frac{0.001 \cdot DCT^2 \cdot SLC45A2^2 \cdot SPRYD5^2}{MLANA \cdot RXRG^2} + 1.997 - \frac{0.027 \cdot MLANA \cdot ROPN1 \cdot RXRG^2}{DCT}$
RAF265	1.304569	-0.054730	$\ln(IC_{50}) = 0.936$
TAE684	5.307634	-14.039563	$\ln(IC_{50}) = 1.168$
TKI258	0.667249	0.082072	$\ln(IC_{50}) = 1.636$
Erlotinib	0.576919	-0.086975	$\ln(IC_{50}) = \frac{0.001 \cdot CDH3^2 \cdot GPR87 \cdot TMPRSS11E^2}{KRT16 \cdot PRSS8^2} - \frac{0.001 \cdot EVPL \cdot GPR87 \cdot PRSS8^2 \cdot TMPRSS11E}{IRF6^2 \cdot KRT16}$ $- \frac{0.001 \cdot GPR87^2 \cdot KRT16^2 \cdot PRSS8^2 \cdot TMPRSS11E^2}{IRF6^2} + 1.898$
AZD6244	6.891855	-21.872891	$\ln(IC_{50}) = 1.476$

Πίνακας 15 – Αξιολόγηση μοντέλων και παραγόμενες εξισώσεις SelectKBest με ITEA (random_state=18, simplify_method='simplify_by_var')

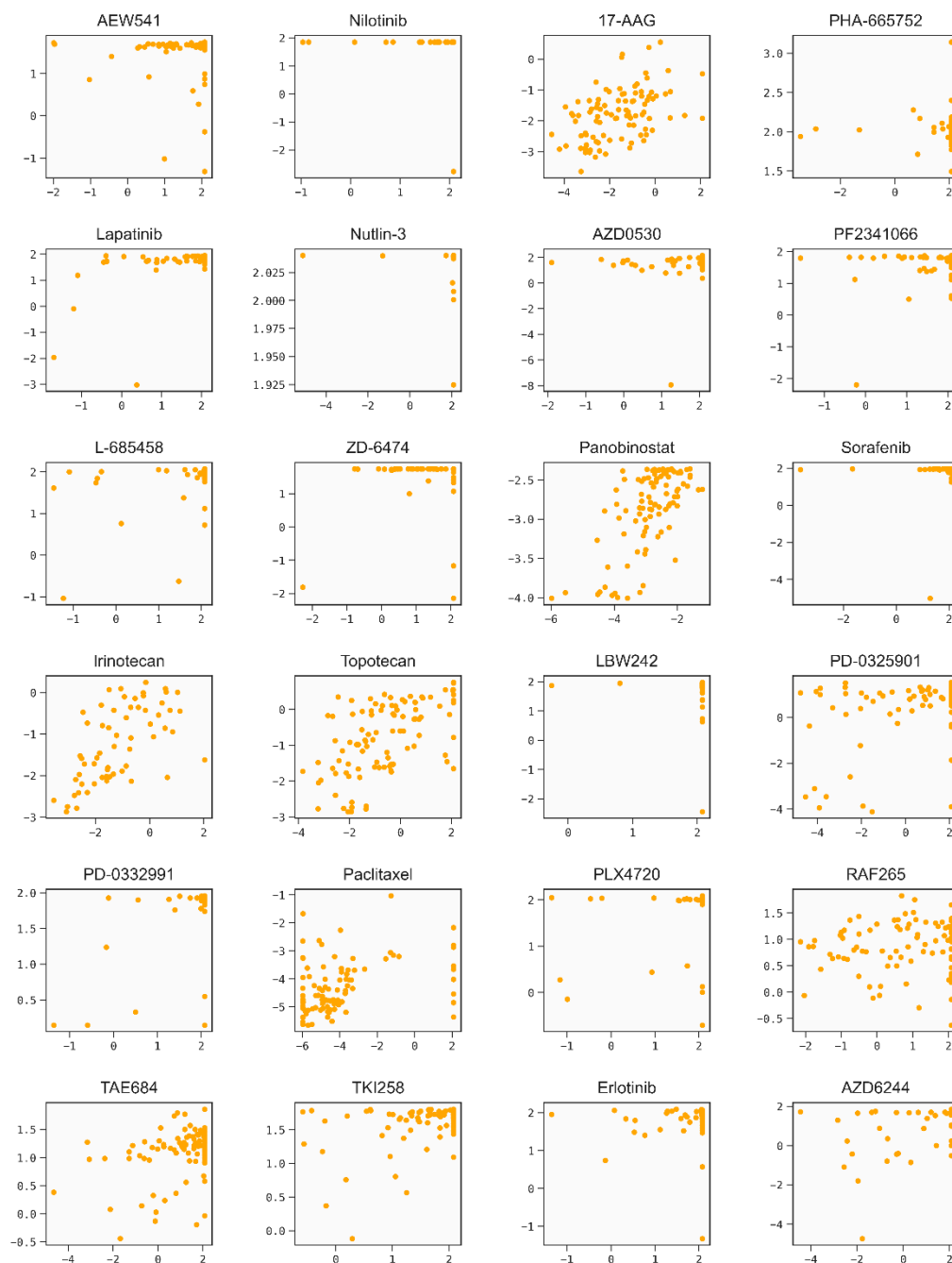
ΠΑΡΑΡΤΗΜΑ II - Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



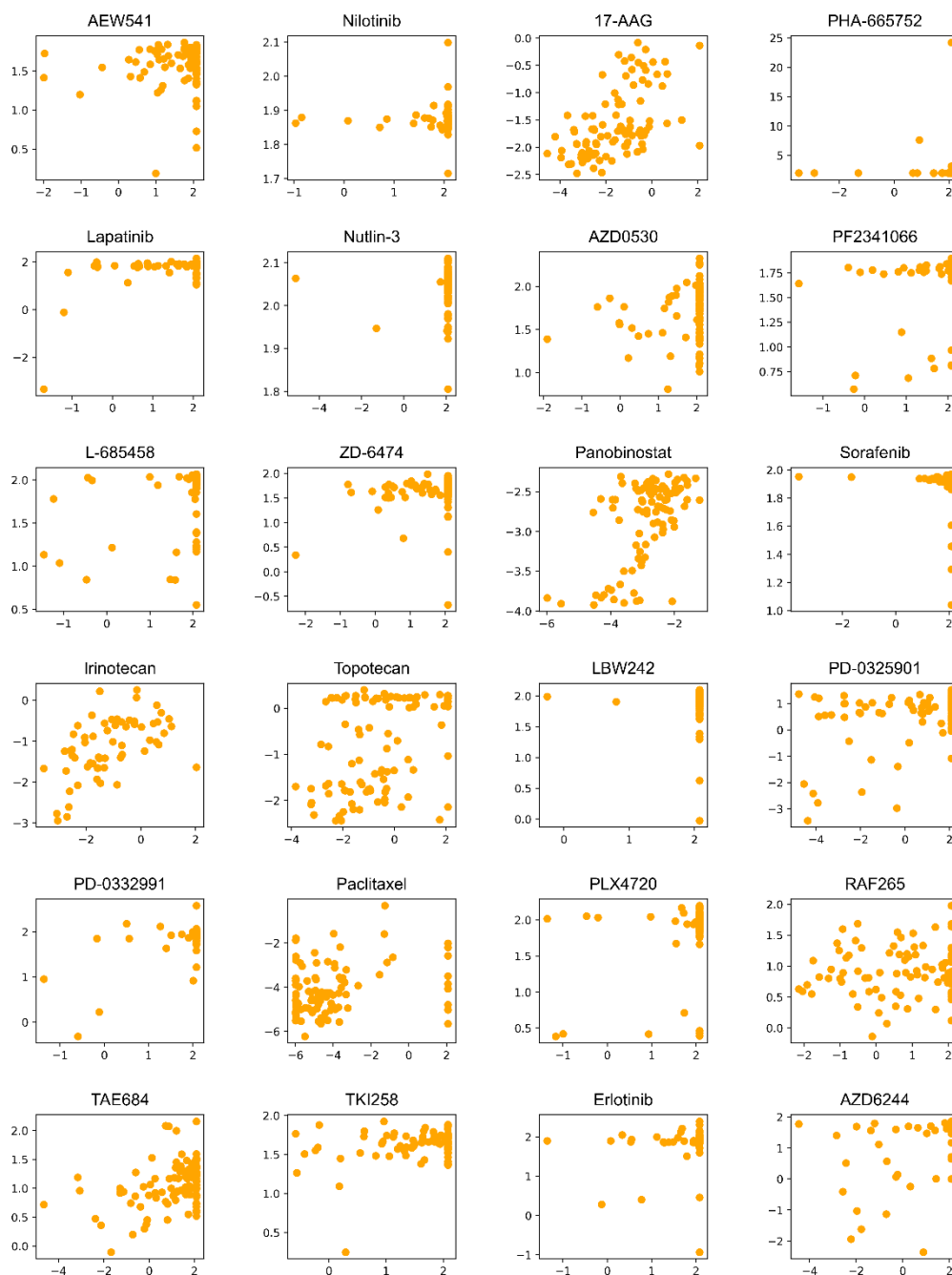
Εικόνα 7 – QLattice (random_seed=157, n_epochs=18): Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



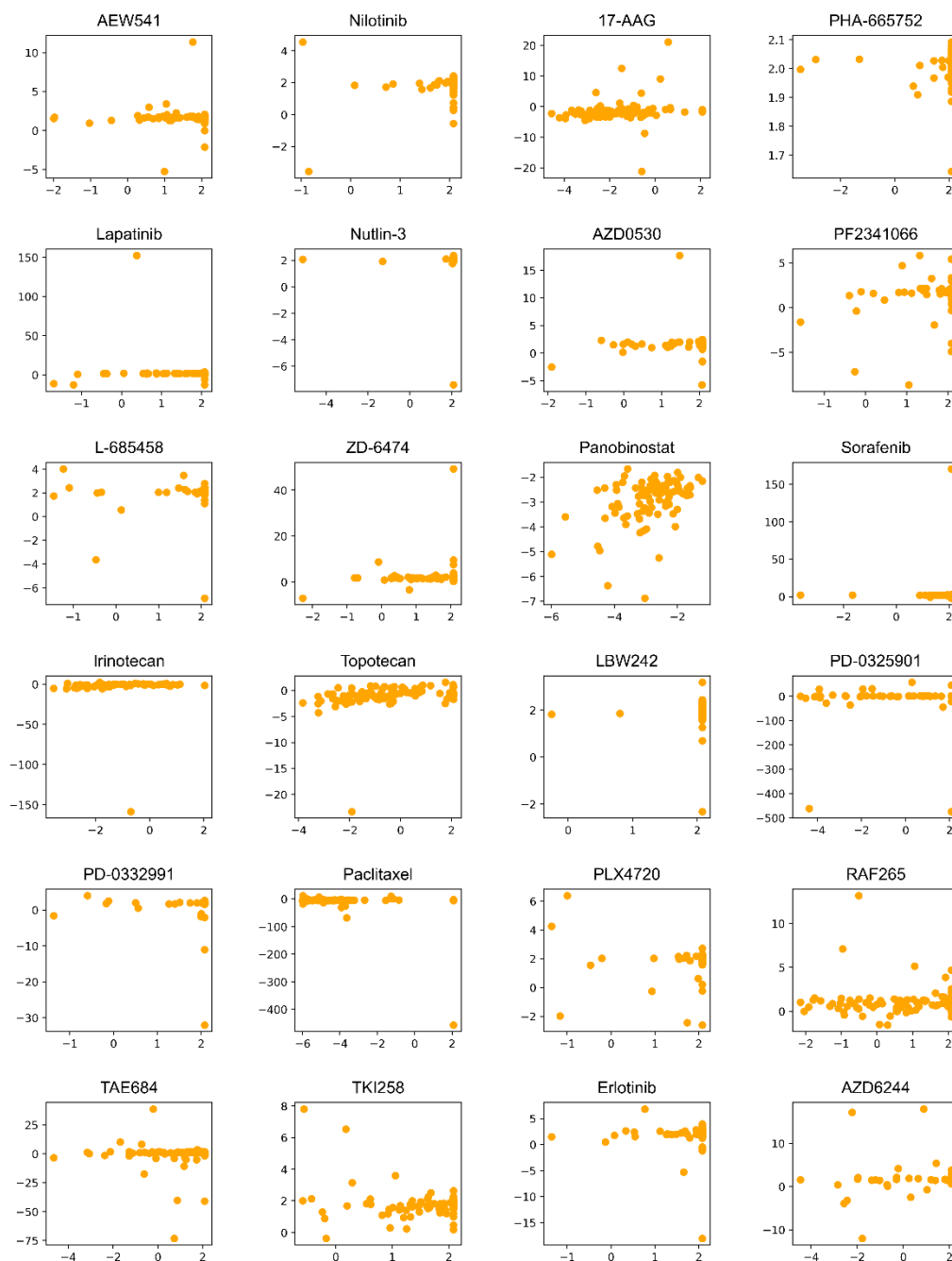
Εικόνα 8 – pySR (random_state=79, niterations=160, deterministic=True, procs=0, multithreading=None): Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



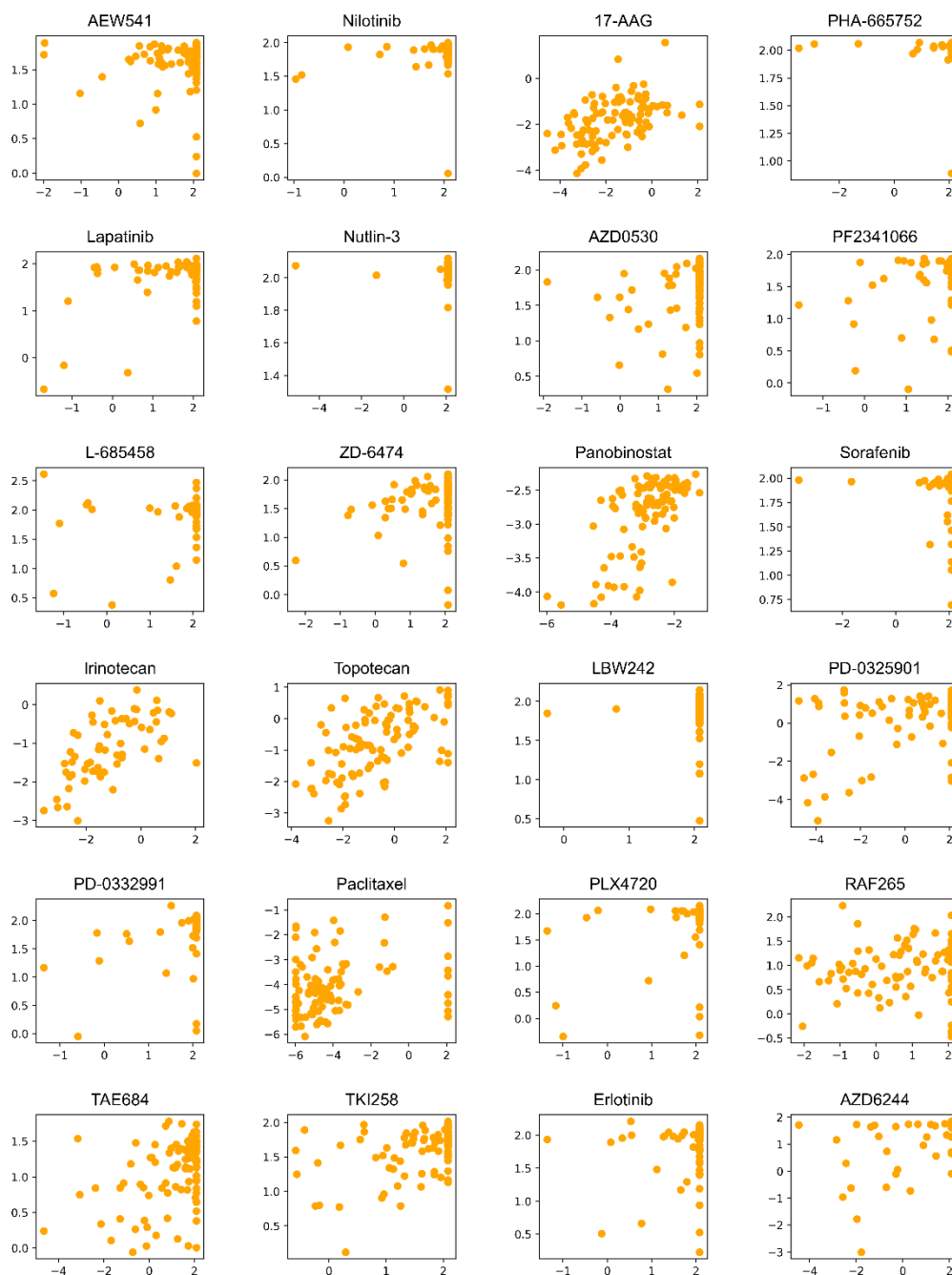
Εικόνα 9 – QLattice (random_seed=96, n_epochs=18) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



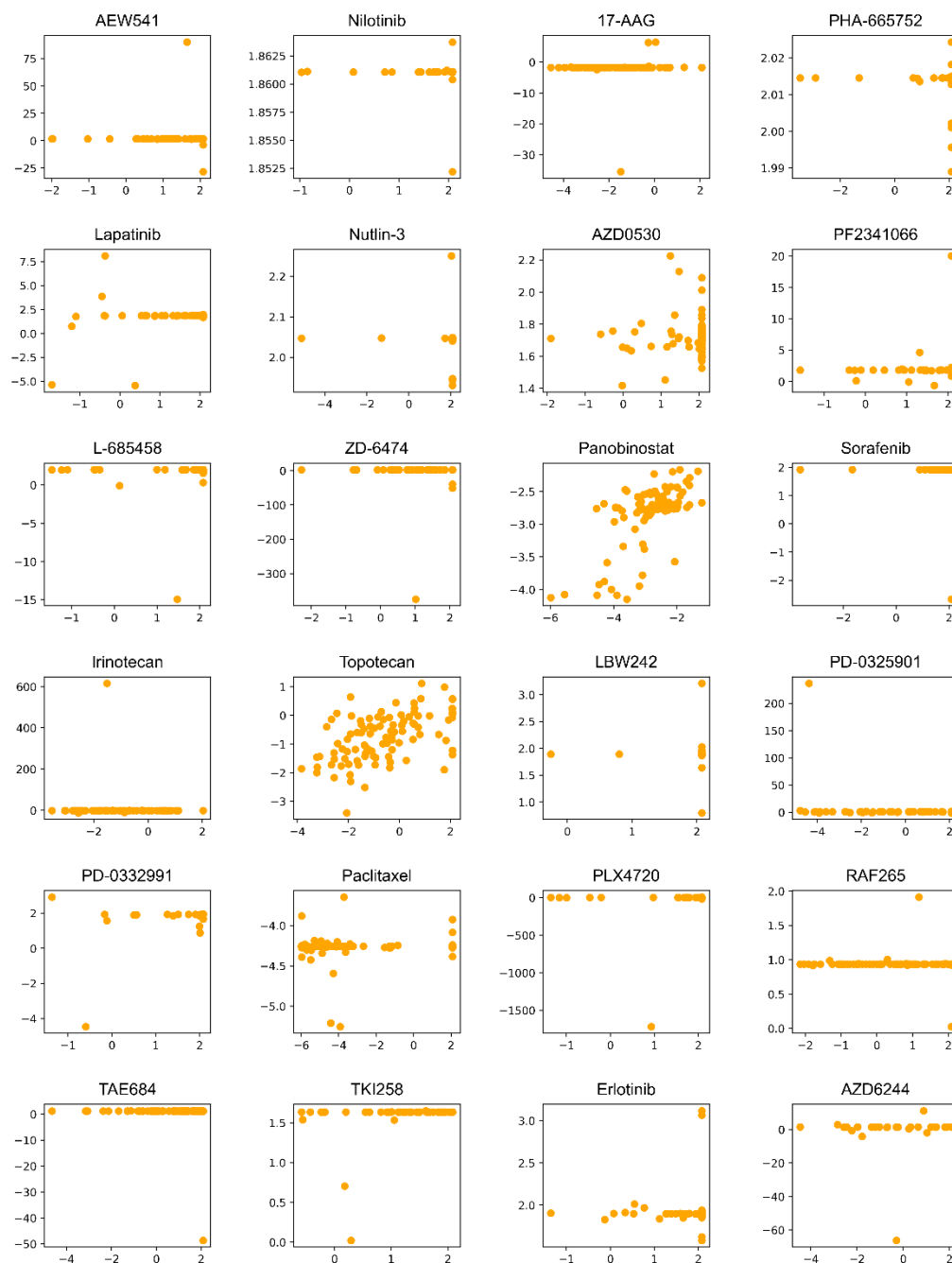
Εικόνα 10 – pySR (random_state=21, niterations=160, deterministic=True, procs=0, multithreading=None) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



Εικόνα 11 – FFX με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



Εικόνα 12 – EQL (num_layers=1, dim=train_X.shape[1]) με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$



Εικόνα 13 – ITEA (random_state=18, simplify_method='simplify_by_var') με SelectKBest: Διαγράμματα διασποράς προβλεπόμενης έναντι πραγματικής τιμής $\ln(IC_{50})$

ΑΝΑΦΟΡΕΣ

- [1] Angelis, D., Sofos, F., & Karakasidis, T. E. (2023). Artificial intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives. *Archives of Computational Methods in Engineering*, 30(6), 3845–3865. <https://doi.org/10.1007/s11831-023-09922-z>
- [2] Arnaldo, I., Krawiec, K., & O'Reilly, U. (2014). *Multiple regression genetic programming*. <https://doi.org/10.1145/2576768.2598291>
- [3] Aykul, S., & Martínez-Hackert, E. (2016). Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical Biochemistry*, 508, 97–103. <https://doi.org/10.1016/j.ab.2016.06.025>
- [4] Baptista, D., Ferreira, P. G., & Rocha, M. (2020). Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 22(1), 360–379. <https://doi.org/10.1093/bib/bbz171>
- [5] *Biomarker testing for cancer treatment*. (2021, December 14). Cancer.gov. <https://www.cancer.gov/about-cancer/treatment/types/biomarker-testing-cancer-treatment>
- [6] *Biomarkers*. (n.d.). National Institute of Environmental Health Sciences. <https://www.niehs.nih.gov/health/topics/science/biomarkers>
- [7] Broløs, K. R., Machado, M. V., Cave, C., Kasak, J., Stentoft-Hansen, V., Batanero, V. G., Jelen, T., & Wilstrup, C. (2021, April 12). *An approach to Symbolic Regression using Feyn*. arXiv.org. <https://arxiv.org/abs/2104.05417>
- [8] Burlacu, B., Kronberger, G., & Kommenda, M. (2020). *Operon C++*. <https://doi.org/10.1145/3377929.3398099>
- [9] Cranmer, M. (2023). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2305.01582>
- [10] *Data formats - GeneSetEnrichmentAnalysisWiki*. (n.d.). https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats
- [11] De França, F. O. (2018). A greedy search tree heuristic for symbolic regression. *Information Sciences*, 442–443, 18–32. <https://doi.org/10.1016/j.ins.2018.02.040>
- [12] Eichelbaum, M., Ingelman-Sundberg, M., & Evans, W. E. (2006). Pharmacogenomics and individualized drug therapy. *Annual Review of Medicine*, 57(1), 119–137. <https://doi.org/10.1146/annurev.med.56.082103.104724>
- [13] Haber, D. A., Bell, D. W., Sordella, R., Kwak, E. L., Godin-Heymann, N., Sharma, S. V., Lynch, T. J., & Settleman, J. (2005). Molecular targeted therapy of lung cancer: EGFR mutations and response to EGFR inhibitors. *Cold Spring Harbor Symposia on Quantitative Biology/Cold Spring Harbor Symposia on Quantitative Biology*, 70(0), 419–426. <https://doi.org/10.1101/sqb.2005.70.043>
- [14] *Introduction to Gene Expression Profiling | Thermo Fisher Scientific - IE*. (n.d.). <https://www.thermofisher.com/gr/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/gene-expression-analysis-real-time-pcr-information/introduction-gene-expression-profiling.html>
- [15] Jack, J., Rotroff, D. M., & Motsinger-Reif, A. (2014). Lymphoblastoid Cell Lines Models of Drug Response: Successes and Lessons from this Pharmacogenomic Model. *Current Molecular Medicine*, 14(7), 833–840. <https://doi.org/10.2174/1566524014666140811113946>
- [16] Jin, Y., Fu, W., Kang, J., Guo, J., & Guo, J. (2019). Bayesian Symbolic regression. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1910.08892>
- [17] La Cava, W., Spector, L., & Danai, K. (2016). *Epsilon-Lexicase Selection for Regression*. <https://doi.org/10.1145/2908812.2908898>
- [18] Liotta, L. A., & Petricoin, E. F. (2000). Molecular profiling of human cancer. *Nature Reviews. Genetics*, 1(1), 48–56. <https://doi.org/10.1038/35049567>

- [19] Liu, H., Peng, W., Dai, W., Lin, J., Fu, X., Liu, L., Liu, L., & Yu, N. (2024). Improving anti-cancer drug response prediction using multi-task learning on graph convolutional networks. *Methods*, 222, 41–50. <https://doi.org/10.1016/j.ymeth.2023.11.018>
- [20] Makke, N., & Chawla, S. (2024). Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1). <https://doi.org/10.1007/s10462-023-10622-0>
- [21] Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 13(3). <https://doi.org/10.1002/widm.1493>
- [22] McConaghy, T. (2011). FFX: Fast, Scalable, Deterministic Symbolic Regression Technology. In *Genetic and evolutionary computation* (pp. 235–260). https://doi.org/10.1007/978-1-4614-1770-5_13
- [23] Meditz, K., & Rinner, B. (2018). Establishment of tumor cell lines: from primary tumor cells to a tumor cell line. In *Learning materials in biosciences* (pp. 61–73). https://doi.org/10.1007/978-3-319-74854-2_4
- [24] *NCI Dictionary of Cancer Terms*. (n.d.-a). National Cancer Institute. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-profiling>
- [25] *NCI Dictionary of Cancer Terms*. (n.d.-b). Cancer.gov. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker>
- [26] *NCI Dictionary of Cancer Terms*. (n.d.-c). Cancer.gov. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker-testing>
- [27] *NCI Dictionary of Cancer Terms*. (n.d.-d). Cancer.gov. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cell-culture>
- [28] Niu, N., & Wang, L. (2015). In vitro human cell line models to predict clinical response to anticancer drugs. *Pharmacogenomics*, 16(3), 273–285. <https://doi.org/10.2217/pgs.14.170>
- [29] Olivetti, D. F. F., & Aldeia, G. S. I. (2019). Interaction-Transformation Evolutionary Algorithm for Symbolic regression. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1902.03983>
- [30] Park, A., Joo, M., Kim, K., Son, W., Lim, G., Lee, J., Kim, J. H., Lee, D. H., & Nam, S. (2022). A comprehensive evaluation of regression-based drug responsiveness prediction models, using cell viability inhibitory concentrations (IC50 values). *Bioinformatics*, 38(10), 2810–2817. <https://doi.org/10.1093/bioinformatics/btac177>
- [31] Park, A., Lee, Y., & Nam, S. (2023). A performance evaluation of drug response prediction models for individual drugs. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-39179-2>
- [32] Phillippy, A. M., Mason, J. A., Ayanbule, K., Sommer, D., Taviani, E., Huq, A., Colwell, R. R., Knight, I. T., & Salzberg, S. L. (2007). Comprehensive DNA signature discovery and validation. *PLOS Computational Biology/PLoS Computational Biology*, 3(5), e98. <https://doi.org/10.1371/journal.pcbi.0030098>
- [33] *Precision oncology*. (2023, November 16). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/what-we-do/nih-turning-discovery-into-health/promise-precision-medicine/precision-oncology>
- [34] Pusterla, T. (2023, May 4). *High-throughput screening (HTS)*. BMG LABTECH GmbH. <https://www.bmglabtech.com/en/blog/high-throughput-screening/>
- [35] Segeritz, C., & Vallier, L. (2017). Cell culture. In *Elsevier eBooks* (pp. 151–172). <https://doi.org/10.1016/b978-0-12-803077-6.00009-6>
- [36] Virgolin, M., Alderliesten, T., & Bosman, P. a. N. (2019). Linear scaling with and within semantic backpropagation-based genetic programming for symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*. <https://doi.org/10.1145/3321707.3321758>
- [37] Virgolin, M., Alderliesten, T., Witteveen, C., & Bosman, P. a. N. (2017). Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block

- learning. *Proceedings of the Genetic and Evolutionary Computation Conference*. <https://doi.org/10.1145/3071178.3071287>
- [38] Virgolin, M., & Pissis, S. P. (2022). Symbolic Regression is NP-hard. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2207.01018>
- [39] *What is pharmacogenomics?: MedlinePlus Genetics*. (n.d.). <https://medlineplus.gov/genetics/understanding/genomicresearch/pharmacogenomics/>
- [40] *What is the QLattice? · Feyn Documentation*. (n.d.). https://docs.abzu.ai/docs/guides/getting_started/qlattice
- [41] William, L. C., Singh, T. R., Taggart, J., Suri, S., & Moore, J. H. (2018). Learning concise representations for regression by evolving networks of trees. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1807.00981>
- [42] *Μοριακό προφίλ | PrIME | Ιατρική Ακριβείας | Θεραπείες Καρκίνου*. (2023, September 6). Prime DX. <https://myprimetest.com/moriako-profil-iatriki-akrivias/?lang=el>