Optimizing approaches and representations for predictive modeling of molecular mechanisms of action and binding affinities of bioactive molecules

Oršolić, Davor

Doctoral thesis / Disertacija

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Food Technology and Biotechnology / Sveučilište u Zagrebu, Prehrambeno-biotehnološki fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:159:223206

University of Zagreb

FACULTY OF FOOD TECHNOLOGY AND BIOTECHNOLOGY

Davor Oršolić

OPTIMIZING APPROACHES AND REPRESENTATIONS FOR PREDICTIVE MODELING OF MOLECULAR MECHANISMS OF ACTION AND BINDING AFFINITIES OF BIOACTIVE MOLECULES

DOCTORAL THESIS

Zagreb, 2024

University of Zagreb

FACULTY OF FOOD TECHNOLOGY AND BIOTECHNOLOGY

Davor Oršolić

OPTIMIZING APPROACHES AND REPRESENTATIONS FOR PREDICTIVE MODELING OF MOLECULAR MECHANISMS OF ACTION AND BINDING AFFINITIES OF BIOACTIVE MOLECULES

DOCTORAL THESIS

Supervisors: Tomislav Šmuc, PhD
Professor Antonio Starčević, PhD

Zagreb, 2024

Sveučilište u Zagrebu

PREHRAMBENO-BIOTEHNOLOŠKI FAKULTET

Davor Oršolić

OPTIMIRANJE METODA I REPREZENTACIJA ZA PREDIKTIVNO MODELIRANJE MEHANIZAMA DJELOVANJA I AFINITETA VEZANJA BIOLOŠKI AKTIVNIH MOLEKULA

DOKTORSKI RAD

Mentori: dr. sc. Tomislav Šmuc,
prof. dr. sc. Antonio Starčević

Zagreb, 2024.

The dissertation was made at the Ruđer Bošković Institute, Division of Electronics, Laboratory for Machine Learning and Knowledge Representation and supported in part by the Research Cooperability Program of the Croatian Science Foundation, funded by the European Union from the European Social Fund under the Operational Programme Efficient Human Resources 2014-2020, through the Grant 8525: Augmented Intelligence Workflows for Prediction, Discovery, and Understanding in Genomics and Pharmacogenomics and by the Croatian Government and the European Union under the European Regional Development Fund — the Competitiveness and Cohesion Operational Program, through the project Bioprospecting of the Adriatic Sea (KK.01.1.1.01.0002), granted to The Scientific Centre of Excellence for Marine Bioprospecting — BioProCro.

Supervisors: Tomislav Šmuc, PhD and Professor Antonio Starčević, PhD

The dissertation has: 95 pages

The dissertation number: 531.76:632.954:66.011(043.3)

This is an article-based doctoral thesis, known as Scandinavian model, which consists of already published scientific papers accompanied by a chapter with the critical review, which was written in accordance with Article 14 of the Doctoral Studies Regulations at the University of Zagreb (2016).

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Food Technology and Biotechnology
Postgraduate study in Biotechnology and Bioprocess Engineering

UDK: 531.76:632.954:66.011(043.3)
Scientific Area: Biotechnical Sciences
Scientific Field: Biotechnology

Optimizing approaches and representations for predictive modeling of molecular mechanisms of action and binding affinities of bioactive molecules

Davor Oršolić, mag.ing.biotechn.

Thesis performed in the Laboratory for Machine Learning and Knowledge Representations, Division of Electronic, Ruđer Bošković Institute

Supervisors: Tomislav Šmuc, PhD; Professor Antonio Starčević, PhD

Short abstract: This study investigates mechanisms of action in the context of direct physical interactions and in the context of biological readouts in the affected organisms. Two chemical spaces of interest include synthetic compounds with phytotoxic activity and protein kinase inhibitors, which regulate the functions of essential enzymes in cellular processes, in an effort to discover new bioactive molecules by delving into the immense chemical space. We tackle the difficulties associated with forecasting herbicidal activity and the interactions between compounds and protein kinases by employing several machine learning techniques. We evaluate and compare numerous algorithms, including the extreme gradient boosting algorithm (XGBoost) and graph convolutional networks (GCN). Additionally, the research paper presents a dynamic applicability domain (dAD) strategy, which improves the precision of predictions in QSAR modeling and offers an innovative framework for evaluating interactions within biomolecular complexes.

Number of pages: 95
Number of figures: 4
Number of tables: 3
Number of references: 52
Original in: English

Key words: dynamic applicability domain (dAD), conformal predictor, binding affinity, human kinome, herbicides, mechanism of action, confidence estimate, prediction region, xgboost, graph convolutional network (GCN).

Date of the thesis defense: February 5, 2024

Reviewers:
1. Associate professor Jurica Žučko, PhD, Faculty of Food Technology and Biotechnology
2. Assistant professor Anita Horvatić, PhD, Faculty of Food Technology and Biotechnology
3. Assistant professor Krešimir Križanović, PhD, Faculty of Electrical Engineering and Computing
4. Senior research associate Bono Lučič, PhD, Ruđer Bošković Institute (substitute)

Thesis deposited in: Library of Faculty of Food Technology and Biotechnology, Kačićeva 23, and National and University Library, Hrvatske bratske zajednice 4, and University of Zagreb, Trg Republike Hrvatske 14.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prehrambeno-biotehnološki fakultet
Sveučilišni poslijediplomski studij Biotehnologija i bioprocesno inženjerstvo

UDK: 531.76:632.954:66.011(043.3)
Znanstveno područje: Biotehničke znanosti
Znanstveno polje: Biotehnologija

Optimiranje metoda i reprezentacija za prediktivno modeliranje mehanizama djelovanja i afiniteta vezanja biološki aktivnih molekula

Davor Oršolić, mag.ing.biotechn.

Rad je izrađen u Laboratoriju za strojno učenje i reprezentacije znanja, Zavod za elektroniku, Institut Ruđer Bošković

Mentori: dr.sc. Tomislav Šmuc; prof.dr.sc Antonio Starčević

Kratki sažetak: Ovo istraživanje analizira mehanizme djelovanja u okviru izravnih fizičkih interakcija te u kontekstu bioloških odgovora zahvaćenih organizama. Istražuju se dva područja kemijskih spojeva: sintetički spojevi s fitotoksičnim djelovanjem koji se koriste kao herbicidi i inhibitori proteinskih kinaza koji igraju ključnu ulogu u reguliranju staničnih procesa, s ciljem otkrivanja novih bioaktivnih molekula sistematiziranim pretraživanjem ogromnog kemijskog prostora. Prilikom modeliranja mehanizama djelovanja bioaktivnih molekula suočavamo se s izazovima predviđanja herbicidne aktivnosti i interakcija između spojeva i proteinskih kinaza korištenjem različitih metoda strojnog učenja. Procjenjujemo i uspoređujemo razne algoritme, uključujući algoritam ekstremnog gradijentnog pojačanja (XGBoost) i graf-konvolucijskih mreža (GCN). Nadalje, u radu se predstavlja strategija dinamičke domene primjenjivosti (dAD), koja povećava točnost predviđanja u QSAR modeliranju i nudi novi okvir za procjenu interakcija unutar biomolekularnih kompleksa.

Broj stranica: 95
Broj slika: 4
Broj tablica: 3
Broj literaturnih navoda: 52
Jezik izvornika: engleski

Ključne riječi: dinamička domena primjenjivosti (dAD), konformni prediktor, afinitet vezanja, ljudski kinom, herbicidi, mehanizam djelovanja, pouzdanost, predikcijski interval, graf konvolucijska mreža.

Datum obrane: 5. veljače 2024.

Stručno povjerenstvo za obranu:
1. izv.prof.dr.sc. Jurica Žučko, Prehrambeno-biotehnološki fakultet
2. doc.dr.sc. Anita Horvatić, Prehrambeno-biotehnološki fakultet
3. doc.dr.sc. Krešimir Križanović, Fakultet elektrotehnike i računarstva
4. dr.sc. Bono Lučič, viši znanstveni suradnik, Institut Ruđer Bošković (zamjena)

Rad je pohranjen u: Knjižnici Prehrambeno-biotehnološkog fakulteta u Zagrebu, Kačićeva 23 i Nacionalnoj i sveučilišnoj knjižnici u Zagrebu, Hrvatske bratske zajednice 4 i Sveučilištu u Zagrebu, Trg Republike Hrvatske 14.

Curriculum vitae of the Supervisors

Tomislav Šmuc, PhD is a Head of Laboratory for Machine Learning and Knowledge Representation at Ruđer Bošković Institute, Zagreb. His research interest is in development and use of machine learning and data mining methods for knowledge discovery in different domains of science. In this period, he has been leading or participating in a number of research projects financed by Croatian or European funding agencies. Tomislav Šmuc was mentor of 5 PhD students and a dozen of master's students at University of Zagreb. He was involved in organization of international conferences in the field such as ECML-PKDD and Discovery Science and a number of summer schools on data science and machine learning. He has published over 100 papers in journals and proceedings of international conferences and regularly serves as a reviewer for a number of established journals in the fields of computer science, machine learning and computational biology. He also regularly serves as an evaluator for several research funding agencies.

Antonio Starčević finished undergraduate studies in biotechnology at Faculty of Food Technology and Biotechnology, University of Zagreb and obtained his PhD degree in natural sciences (biology) at Kaiserslautern technical university. At the moment, prof. Antonio Starčević is a full professor at the Faculty of Food Technology and Biotechnology University of Zagreb, Laboratory for Bioinformatics, where he teaches courses in bioinformatics, proteomics and biotechnology for graduate and post-graduate students. Antonio Starčević has published 40 publications listed in the Web of Science Core Collection and has H-index of 14. Antonio Starčević was involved in several Croatian and international projects, both as participant and principal investigator. He has also filed two patent applications, one of them originating from his PhD, a WIPO (PCT) patent for the in silico method for the annotation of natural product gene-clusters and for the generation of novel biologically active chemical entities from DNA sequences – WO2009130520A1.

Životopis mentora

Tomislav Šmuc, dr. sc. voditelj je Laboratorija za strojno učenje i reprezentaciju znanja na Institutu Ruđer Bošković u Zagrebu. Njegovo područje istraživanja obuhvaća razvoj i primjenu metoda strojnog učenja i rudarenja podataka za otkrivanje znanja u različitim znanstvenim domenama. U tom razdoblju vodio je ili sudjelovao u nizu istraživačkih projekata financiranih od strane hrvatskih ili europskih agencija za financiranje. Tomislav Šmuc bio je mentor 5 doktorskih studenata i desetak magistarskih studenata na Sveučilištu u Zagrebu. Sudjelovao je u organizaciji međunarodnih konferencija u polju, poput *ECML-PKDD* i *Discovery Science* te niza ljetnih škola o podatkovnoj znanosti i strojnom učenju. Objavio je preko 100 radova u časopisima i zbornicima međunarodnih konferencija te redovito obavlja ulogu recenzenta za niz uglednih časopisa u područjima informatike, strojnog učenja i računalne biologije. Također redovito radi kao evaluator za nekoliko agencija za financiranje istraživanja.

Antonio Starčević diplomirao je na Sveučilištu u Zagrebu, Prehrambeno-biotehnološki fakultet. Doktorat je stekao na tehničkom sveučilištu u Kaiserslauternu (*njem. Technische Universität Kaiserslautern*), Njemačka. Trenutno, Antonio Starčević kao redoviti profesor na Prehrambeno-biotehnološkom fakultetu u Zagrebu, u Laboratoriju za Bioinformatiku drži kolegij iz bioinformatike, proteomike i biotehnologije na diplomskom i poslijediplomskom studiju. Dosada je publicirao 40 originalnih znanstvenih radova referenciranih u "Web of Science Core Collection" bazi podataka, gdje ima H-indeks 14. U ulozi suradnika i voditelja bio je uključen u niz nacionalnih i međunarodnih projekata, a sudjelovao je i u projektima transfera tehnologije kao što su "Proof of concept" - PoC projekti Ureda za transfer tehnologije Sveučilišta u Zagrebu i Tehnološki projekti Ministarstva Znanosti i Obrazovanja. Kao prvi autor, sudjelovao je u dvije patentne prijave, od kojih je jedna, proizišla iz rada na doktorskoj disertaciji prihvaćena kao WIPO (PCT) patent za "in silico metodu za anotaciju genskih nakupina sekundarnih metabolita i za generiranje novih biološki aktivnih kemijskih entiteta na temelju sekvencija DNA" – WO2009130520A1.

Acknowlegements

I would like to express my deepest gratitude to my thesis supervisors for their support, insightful feedback, and invaluable guidance throughout this research journey. Your expertise and mentorship have been pivotal in shaping both this work and my academic development.

I am also immensely thankful to other senior members of our laboratory at the Ruđer Bošković Institute - Višnja Stepanić, Bono Lučić, and Anja Barešić - for their constructive critiques and valuable suggestions that greatly enhanced the quality of this thesis and my research in general.

I would like to acknowledge my colleagues in the Laboratory for machine learning and knowledge representations for their collaboration, encouragement, and stimulating discussions that have contributed to my personal and academic growth and have made science fun and enjoyable to do.

A special thanks to my good friends Silvio and Anja for their endless love, moral support, patience, and amazing memories we made during the years we spent together in Zagreb. Living and studying in Zagreb would not be possible or that enjoyable without my roommate and good friend Valentino. Above all, I am profoundly grateful to my dearest of friends - Pegas - for providing unwavering emotional support. Despite the inherent stress and self-doubt that five years of PhD research can bring, one constant source of certainty and comfort was knowing that Pegas would always be there to greet me with hugs and affection upon my return home.

Also, my deep gratitude goes to my parents for making this whole journey possible and for believing in my success. This thesis is their work as much as it is mine.

Lastly, I would like to express my gratitude to everyone who was involved in this journey, directly or indirectly, for their invaluable contributions to my PhD experience.

Davor

Extended abstract

The vastness of the chemical space of compound scaffolds is humongous and it represents a large playground for potential lead drug discovery or repurposing. With the accumulation of experimental data over the years, together with the development of more complex statistical frameworks, screening of such elaborate chemical spaces is finally possible. There are several well-defined problem areas for drug screening efforts, the most popular being inhibition activity against a multitude of protein targets in human cells related to often occurring diseases. Some examples of highly targeted protein spaces include protein kinases, g-protein coupled receptors, and/or (non)selective serotonin re-uptake inhibitors. Mutation and dysregulation in any of the three mentioned protein groups can result in hereditary disorders, tumors, and mental disorders. Contrary to the available machine learning frameworks for prediction of direct physical interactions between compounds and protein targets, certain chemical activity predictions are not well-represented or defined in the literature, e.g., phytotoxic activity.

In this work, publicly available data is collected with regard to the experimentally measured binding affinities of diverse compounds against one of the most popular target protein families, protein kinases. This protein super-family is one of the most important enzyme groups responsible for the regulation of most of the important cellular processes, including cell metabolism, cell growth, and division. Protein kinases regulate biochemical cycles by transferring high energy phosphoryl group from adenosine-3-phosphate (ATP) to specific amino acid residues of the target protein substrates. All members of this enzyme family are characterized by the highly conserved protein kinase (PK) domain, but depending on the phosphorylation site and the activation mechanisms of individual members of this family, this superfamily can be divided into several kinase groups. Due the specific characteristics of this protein group and kinase inhibitors, it is important to investigate how each of these chemical or biological spaces impact models performance and how to achieve more optimal predictive performance.

On the other hand, we examine a different subspace of biological activity, focusing mostly on synthetic compounds with determined phytotoxic or herbicidal activity. We define this problem as a multiclass classification problem by using two predefined classification systems: main one, by the Herbicide Resistance Action Committee (HRAC), and the second one, by the Weed Science Society of America (WSSA). Considering that no defined machine learning

framework for modeling and prediction of herbicidal activity was publicly available, an effort was made to collect the representative data set and define the optimal computational approach to maximize the prediction accuracy for mechanism of action prediction. Considering that the classification of phytotoxic compounds was mostly performed by visual inspection of phenotypic changes in the affected weeds, there is a great need for an automated, systematic approach to this endeavor.

Due to the limited size of the collected data, consisting of molecules of known activity and grouped into known activity classes, we further tested several "shallow" learners. The panel of tested algorithms includes naive bayes (NB), support vector machines (SVM), extreme-gradient boosting approach (XGBoost) and random forest (RF). All the approaches mentioned were trained in a ten times repeated ten fold (10x10-fold) cross validation mode. A comparison of trained models over all hundred resamples was performed using a non-frequentist approach - Bayesian analysis. For the first time in herbicide activity modeling, we have implemented a computational framework from feature processing and selection to the training of several learners and, ultimately, a statistical comparison of their performance.

However, due to the sheer size of the publicly available experimental data for protein kinase inhibitors, modeling of physical interactions between small compound spaces and the human kinome has allowed for application of more complex modeling techniques. With this also came other challenges, such as defining and engineering the feature space for over 8000 compounds and learning representations for the nuanced protein kinase family.

Both of the aforementioned methods are founded on the QSAR (*Quantitative structure-activity relationship*) modeling principles. The definition of the applicability domain (AD) for a specified problem is one of the pillars of QSAR modeling. However, defining the boundaries of the chemical space within which the model can make accurate predictions is not simple and is dependent on the nature of the trained model. In the case of predicting general biological activity in the form of a phenotypic signal, as is the case with herbicidal activity, the applicability domain can be simply defined in two-dimensional space by considering the structural similarity of available molecules and a model output, such as the probability of belonging to a particular class. Predicting the physical interaction between any two entities, such as compounds and protein targets, adds complexity that cannot be accommodated by the conventional applicability domain.

In this instance, we intend to extend the standard applicability domain to include

information about both entities and generate a quantitative estimate of prediction confidence using the conformal prediction framework. Conformal predictors can reliably estimate a prediction region based on the computed nonconformity of test samples. The disadvantage of this method is that the nonconformity is defined in the label space of predefined calibration samples, resulting in estimates that work well in general but are not specific to any tested compound-target pair, thus failing for samples that are not already available in the training set. Combining concepts from both frameworks, we dynamically define similarity-based applicability domains or conformity regions for each new sample and then calculate nonconformity scores - we refer to this approach as the dynamic applicability domain (dAD).

The dAD approach was shown to produce tighter prediction regions when compared to the original conformal predictors algorithm. More importantly, complementary to the prediction regions, when it comes to realistic use-case scenarios (S2, S3), dAD achieves lower error rates for any confidence level. More importantly, merging the concept of applicability domain with a conformal predictor corrects for existing bottlenecks in the traditional applicability domain definition and allows for the evaluation of model behavior in an abstract interaction space between any number of interacting entities. This way, it is a valuable and informative approach for validation of data quality in subregions of interaction space specific to biomolecular complexes.

Keywords: dynamic applicability domain (dAD), conformal predictor, binding affinity, human kinome, herbicides, mechanism of action, confidence estimate, prediction region, xgboost, graph convolutional network (GCN).

Prošireni sažetak

Veličina prostora potencijalnih kemijskih struktura je ogromna te omogućava pretraživanje i testiranje novih potencijalnih terapeutika ili prenamjenu već postojećih u svrhu ciljanja drugih proteina. Kroz vrijeme, sve veće nakupljanje eksperimentalnih podataka i razvoja naprednih statističkih pristupa omogućilo je učinkovito ciljano pretraživanje kemijskog prostora. Postoji nekoliko dobro definiranih problematičnih područja gdje se automatizirano pretraživanje novih terapeutika pokazalo učinkovitim, a najpopularnija je inhibicija aktivnosti mnoštva ciljanih proteina u ljudskim stanicama povezanih s učestalnim bolestima. Među proteinske skupine od velikog interesa spadaju proteinske kinaze, g-protein spregnuti receptori i/ili (ne)selektivni inhibitori ponovne pohrane serotonina. Mutacija i disregulacija u bilo kojoj od tri navedene skupine proteina može rezultirati nasljednim poremećajima, tumorima i mentalnim poremećajima. Suprotno dostupnim okvirima strojnog učenja za predviđanje izravnih fizičkih interakcija između spojeva i proteina od interesa, određena predviđanja kemijske aktivnosti nisu dobro predstavljena ili definirana u literaturi, npr. herbicidno djelovanje.

U ovom radu prikupljena je većina javno dostupnih podataka s eksperimentalno izmjerenim afinitetima vezanja različitih spojeva protiv jedne od najpopularnijih proteinskih porodica od interesa, proteinskih kinaza. Ova super-porodica proteina jedna je od najvažnijih enzimskih skupina odgovornih za regulaciju većine važnih staničnih procesa, uključujući regulaciju staničnog metabolizma, rasta i diobe stanica. Kinaze reguliraju biokemijske cikluse prijenosom fosforilnih skupina visoke energije s molekule adenozin-3-fosfata (ATP) na specifične aminokiselinske bočne lance ciljnih proteinskih supstrata. Svi članovi ove obitelji enzima karakterizirani su visoko očuvanom proteinskom kinaznom (PK) domenom, ali ovisno o mjestu fosforilacije i mehanizmima aktivacije, članovi ove porodice mogu se podijeliti u nekoliko kinaznih skupina. S obzirom na specifičnost proteinske porodice kinaza, kao i kinaznih inhibitora, vrlo je važno analizirati utjecaj svakog pojedinačnog kemijskog, odnosno biološkog prostora, na izvedbu i učinkovitost samog modela, kao i način za postizanje optimalnijeg riješenja.

S druge strane, osim prostora proteinskih kinaznih inhibitora, ispitujemo i drugačiji potprostor biološke aktivnosti, fokusirajući se uglavnom na sintetičke primjere molekula s izmjerenom fitotoksičnom aktivnošću. Budući da ova specifična aktivnost, u smislu fizičke

interakcije između spojeva i ciljanih proteina, obično nije dobro dokumentirana za ovaj specifični zadatak - ovaj problem definiramo kao problem klasifikacije s više oznaka uzimajući unaprijed definirane sustave klasifikacije od strane Odbora za otpornost na herbicide (*engl. Herbicide Resistance Action Committee, HRAC*) i Američkog društva za znanost o korovima (*engl. Weed Science Society of America, WSSA*). Zbog nedostatka javno dostupnih definiranih okvira strojnog učenja za modeliranje i predviđanje učinkovitosti herbicida tijekom provedenog istraživanja, nastojimo sakupiti reprezentativan skup podataka i uspostaviti optimalan računalni pristup radi povećanja točnosti predviđanja mehanizma djelovanja (MoA). Imajući u vidu da se klasifikacija fitotoksičnih spojeva obično vrši vizualnom inspekcijom promjene fenotipa biljaka nakon izlaganja, postoji izražena potreba za automatizacijom ovog pristupa.

Zbog ograničene veličine prikupljenih podataka koji se sastoje od molekularnih struktura poznate aktivnosti i označenih MoA skupinom, dodatno testiramo nekoliko "plitkih" modela strojnog učenja. Panel testiranih algoritama uključuje Naive Bayes (NB), stroj potpornih vektora (*engl. support vector machine, SVM*), pristup ekstremnog pojačanja gradijenta (*engl. extreme gradient boosting, XGBoost*) i nasumične šume (*engl. random forest, RF*). Svi spomenuti pristupi naučeni su u deset puta ponovljenom desetostrukom (10x10-strukom) načinu unakrsne validacije. Usporedba treniranih modela na svih stotinu ponovnih uzoraka provedena je nefrekvencijskim pristupom - Bayesovom analizom. Po prvi put za modeliranje aktivnosti herbicida, implementirali smo računalni okvir od obrade značajki i odabira, do učenja nekoliko modela, i konačno, statističke usporedbe njihove izvedbe.

Obje navedene metode temelje se na principima kvantitativnog modeliranja odnosa između strukture i aktivnosti (*engl. quantitative structure-activity relationship, QSAR*). Definicija domene primjenjivosti za određeni problem jedan je od temelja QSAR-a. Međutim, definiranje granica kemijskog prostora unutar kojeg model može napraviti točna predviđanja nije jednostavno i ovisi o prirodi naučenog modela. U slučaju predviđanja opće biološke aktivnosti u obliku fenotipskog signala, kao što je slučaj s herbicidnom aktivnošću, domena primjenjivosti može se jednostavno definirati u dvodimenzionalnom prostoru uzimajući u obzir strukturnu sličnost dostupnih molekula i modelnog produkta kao npr. vjerojatnost pripadnosti određenoj klasi. Predviđanje fizičke interakcije između bilo koja dva entiteta, kao što su spojevi i proteinski ciljevi, dodaje složenost koja se ne može prilagoditi konvencionalnoj domeni primjenjivosti.

U ovom slučaju, namjeravamo proširiti standardnu domenu primjenjivosti kako bismo uključili oba entiteta i generirali kvantitativnu procjenu pouzdanosti predviđanja korištenjem okvira predviđanja nesukladnosti primjera (*engl. conformal predictors*). Navedenim postpukom može se pouzdano procijeniti područje predviđanja na temelju izračunate nesukladnosti ispitnih uzoraka. Nedostatak ove metode je taj što je nesukladnost definirana u prostoru oznaka unaprijed definiranih kalibracijskih uzoraka, što rezultira procjenama koje općenito dobro funkcioniraju, ali nisu specifične ni za jedan testirani par kemijskog spoja i proteina, stoga nisu uspješne za uzorke koji su malo izvan distribucije podataka u skupu za učenje. Kombinirajući koncepte iz oba okvira, dinamički definiramo domene primjenjivosti temeljene na sličnosti, što nazivamo regijama sukladnosti za svaki novi uzorak, a zatim izračunavamo rezultate nesukladnosti - ovaj pristup nazivamo dinamičkom domenom primjenjivosti (*engl. dynamic applicability domain, dAD*).

Pokazalo se da dAD pristup proizvodi strože intervale predviđanja u usporedbi s izvornim algoritmom konformnih prediktora. Još važnije, komplementarno regijama predviđanja, dAD postiže niže stope pogreške za bilo koju razinu pouzdanosti. Što je posebno važno za teže scenarije testiranja, kao što su scenariji otkrivanja (S2) i prenamjene (S3) biološki aktivnih spojeva.

Ključne riječi: dinamička domena primjenjivosti, konformni prediktor, afinitet vezanja, ljudski kinom, herbicidi, mehanizam djelovanja, pouzdanost, predikcijski interval, graf konvolucijska mreža.

Table of Contents

1. General introduction

1.1 Motivation and related work

Over the past three decades, significant efforts have been made in pharmaceutical sciences and other fields to model and predict the behavior and activity of examined compounds. Automated drug activity and toxicity screening across a panel of cell lines, or more specifically, protein targets, lowers the overall costs of selecting and testing vast spaces of promising molecules, allowing safer compounds into clinical trials, and assisting researchers in selecting compounds targeting specific diseases. This process necessitates advancements on several fronts, particularly given the ever-expanding space of experimentally measured bioactivities, as well as the rapid expansion of computational frameworks without a clear guide for their application in the domain of biological response prediction. Since mode of action modeling is one of the focal points of this study, in this work we recognize two very different bioactivity profiling problems.

The first problem is related to herbicides, which are small molecular compounds that play an essential role in modern agriculture and land management practices by allowing farmers and landowners to control weeds and other unwanted plants without resorting to more labor-intensive practices (Bloch et al., 2021). The advantages of utilizing intelligent herbicide application include enhanced agricultural productivity and decreased labor expenses, which are important considering increasing world food consumption.

In recent years, there has been a rise in awareness regarding the dangers of herbicide overuse. Excessive use of herbicides may result in the emergence of herbicide resistance, whereby specific weeds or plants acquire resistance to the chemical agents employed for their management. The aforementioned phenomenon has the potential to result in increased application of herbicides, thereby endangering detrimental ecological niches. According to a review disseminated in the journal Pest Management Science, the issue of herbicide resistance

is progressively escalating on a global scale (Beckie et al., 2021). As of 2021, over 500 instances of herbicide-resistant weeds have been documented in 70 countries, with the most common emerging resistance including resistance to acetolactate synthase inhibitors, photosystem-II inhibitors, enolpyruvylshikimate phosphate synthase inhibitors (glyphosate), and acetyl-CoA carboxylase inhibitors (Beckie et al., 2021). To address the problem of herbicide resistance, researchers are developing new herbicides and exploring alternative weed management strategies, such as crop rotation and integrated pest management (IPM) approaches (Powles and Yu, 2010). Herbicides have a crucial function in contemporary agriculture and land management methodologies; however, their excessive usage may result in the emergence of herbicide resistance, thereby causing detrimental ecological and financial consequences. To mitigate this problem, farmers and land managers should adopt sustainable weed management strategies and use herbicides judiciously (Beckie et al., 2021; Powles and Yu, 2010).

To avoid repeated use of chemicals with similar activity, rotational programs heavily rely on classification of herbicides into distinct classes based on their mechanism of action, thus slowing down the emergence of resistance. But it is also worth mentioning that due to the inconsistent classification of herbicides into different classes and having different subclasses with some compounds with exact known protein targets and others only with known location of biological activity due to phenotypic observation, both mode of action (MoA) and site of action (SoA) are used interchangeably. With that said, this is the domain that would most benefit from an optimized machine learning framework, given that the rotational programs today rely on manually curated data and there had been no prior efforts to optimize a systematic phytotoxicity screening pipeline.

The second problem is related to regulation of protein kinase activity. Protein kinase inhibitors constitute a significant and interesting chemical space. Unlike herbicides, kinase inhibitors are considerably better documented in the scientific literature, with thousands of bioactivities measured across the entire human kinome. Protein kinases are enzymes that play a vital role in cell signaling and have been linked to a variety of illnesses, including cancer, inflammation, and metabolic disorders (Roskoski Jr, 2015). Kinase inhibitors specifically target and inhibit protein kinases, affecting cellular signaling pathways. As of 2022, there were 72 small molecule protein kinase inhibitors FDA-approved for clinical use in the treatment of cancer, inflammation, and other malignancies as potent therapeutic agents in the treatment of

numerous diseases Roskoski Jr (2022); Cohen et al. (2021); Taylor et al. (2017); Garg et al. (2017). Cancer is a disease characterized by uncontrolled cell proliferation and expansion. Protein kinases are engaged in multiple signaling pathways that regulate cell growth and survival, and aberrant regulation of these pathways can result in cancer. Thus, kinase inhibitors have emerged as a promising class of anti-cancer drugs that target specific kinases implicated in the growth and survival of cancer cells. Imatinib for chronic myelogenous leukemia and non-small cell lung cancer, vemurafenib for melanoma, and lapatinib for breast cancer are examples of kinase inhibitors approved for clinical usage in the treatment of different forms of cancer (Roskoski Jr, 2015; Cohen et al., 2021).

Inflammation, a natural reaction to a tissue injury or infection that involves multiple signaling pathways, is another extremely complex illness. Similar as it was with cancer, protein kinases are essential for controlling the inflammatory response, and kinase activity dysregulation can result in persistent inflammation and tissue damage (Castelo-Soccio et al., 2023; Cohen et al., 2021; Ferguson and Gray, 2018). Kinase inhibitors approved for inflammation treatment include tofacitinib, a JAK inhibitor authorized for rheumatoid arthritis, which has been developed for clinical use in the treatment of various disorders (Cohen et al., 2021; Ferguson and Gray, 2018). In addition, since dysregulation of metabolic pathways characterizes even metabolic diseases such as diabetes and obesity, several kinase inhibitors have been developed for their treatment, including sotagliflozin, an FDA-approved dual SGLT1 and SGLT2 inhibitor for the treatment of type 1 diabetes (Cefalo et al., 2019; Garg et al., 2017). Similar to the previous chemical group, any chemical group employed to treat diseases in humans will surely have several off-target effects that result in various outcomes from the primary effect. The difference, however, comes in the fact that in herbicidal activity, the end user typically does not need to know all potential targets of the utilized chemical nor the precise mode of action (MoA), which is frequently defined as a site of action (SoA) based on the organ with the observed phenotype. When it comes to human application, these factors are more important.

In these situations, machine learning methods can be used to test the bioactivity and off-target effects of a large number of compounds across the human kinome quickly and efficiently. Generally, the benchmark datasets used by baseline machine learning approaches found in the literature impose restrictions (Davis et al., 2011; Metz et al., 2011). Computational models constructed from available benchmark datasets are frequently constrained by their size and lack

of sample diversity, resulting in overconfident results on stratified test sets (Pahikkala et al., 2015; Cichońska et al., 2021). The aforementioned behavior is reflected in the lack of practical application of the majority of cutting-edge approaches proposed in recent years. This calls for careful problem construction with a focus on the data and representation of compound and protein target spaces in the training set Cichonska et al. (2017); Cichońska et al. (2021).

Both of the problems mentioned, herbicidal and kinase inhibitory activity, require similar solutions, including smarter dataset construction in an effort to mimic the real use-case scenarios, feature optimization to correctly reconstruct the chemical spaces, and rigorous evaluation of trained models by careful inspection and applicability domain definition.

1.2 Objectives

Publicly available data is often miscellaneous and requires a systematic approach to be organized in a meaningful way. This is especially true for data meant to be used for building computational models. For this purpose, the datasets used need to reflect the problem that wants to be solved. Hence, for the large chemical space used in this study, it is necessary to define clear boundaries for samples representing the problem of interest. Accordingly, one of the objectives of this study is to define a machine learning framework for prediction of MoA, encompassing every step from feature selection and hyperparameter tuning to defining the limits of the applicability domain for such an approach.

MoA prediction, or specifically the binding affinity of kinase inhibitors, is a layered problem that evokes rigorous optimization of predictive models through a combination of molecular and protein representations. This is shown to be a challenge, especially for the protein target space, where the protein kinase superfamily is defined by the evolutionary conserved protein kinase domain. The structural characteristics of the members of the protein kinase superfamily are responsible for the low selectivity of FDA-approved kinase inhibitors.

Taking into account many available computational models already available in the literature with high accuracy scores on benchmark datasets, the focus of this study shifts more to the evaluation of trained models and their interpretability. Both of these measures are detrimental to the real use-case of trained models and thorough screening for potential new leads.

To summarize, the main objective is to define a straightforward framework, beginning with data processing and ending with the evaluation of trained models on real-world data. We believe that such a pipeline is necessary, particularly for life science research, with the objective of direct application of the aforementioned framework as opposed to building each model from scratch.

2. Theoretical overview

2.1 Bioactivity of small compounds

Small compounds or small molecules are organic compounds with a low molecular weight that play a central role in numerous biological processes (Cichońska et al., 2021). For the purpose of this thesis small compounds were defined by the molecular weights lower or equal to 900 Da (Cichońska et al., 2021; Oršolić and Šmuc, 2023). Their unique physicochemical properties enable them to interact with specific biomolecular targets, thereby making them important tools in understanding and modulating cellular pathways. The effects of small compound bioactivities on biological systems are diverse, ranging from enzyme inhibition and receptor activation to gene expression regulation (Stockwell, 2004).

Bioactive compounds bind to particular targets, such as proteins, nucleic acids, and lipids, in order to exert their bioactivity, which is usually facilitated by van der Waals forces, hydrophobic contacts and hydrogen bonds. Furthermore, the binding affinity and specificity are determined by the structural compatibility of the small compound and its target. For this reason, understanding the principles of molecular recognition is essential for designing small-molecule compounds with minimal off-target effects, thus ensuring higher selectivity (Liao, 2007; Davis et al., 2011; Kairys et al., 2019).

Small bioactive compounds frequently function as potent enzyme inhibitors, modulating essential enzyme activities in the cells. Enzyme inhibition can disrupt enzyme-substrate interactions or prevent catalysis through reversible or irreversible interactions between compounds and target proteins. These inhibitors can target essential enzymes in disease pathways, providing a foundation for drug development in a variety of therapeutic fields, including cancer, infectious diseases, and metabolic disorders (Ferguson and Gray, 2018).

In addition, they can act as agonists or antagonists of cell surface receptors or intracellular receptors, altering the signaling pathways of cells. Agonists stimulate receptor activity,

resulting in signaling cascades, whereas antagonists inhibit receptor activation, thereby inhibiting signal transduction (Liao, 2007; Ferguson and Gray, 2018). These bioactive compounds are extensively used as pharmaceutical agents to regulate cellular responses and treat a variety of diseases, such as neurological and cardiovascular disorders. Moreover, some compounds can alter patterns of gene expression by targeting epigenetic enzymes that modify DNA and histones, thereby influencing chromatin structure and gene accessibility (Altucci and Rots, 2016). By modulating epigenetic modifications, small molecules can influence cell differentiation, reprogram cell fate, and potentially reverse disease-associated aberrant gene expression. This epigenetic targeting paves the way for novel therapeutic approaches, especially in the context of cancer and other epigenetically driven diseases (Roskoski Jr, 2015; Zhang et al., 2023; Altucci and Rots, 2016).

Small molecules are indispensable resource in drug discovery, and not only for the treatment of human disease but also for their application in agriculture as pesticides (Bloch et al., 2021; Zhang et al., 2023). They allow scientists to investigate biological pathways and determine the functions of specific targets in cellular processes. Utilizing small compound libraries in high-throughput screening permits the identification of novel drug candidates and the investigation of new targets implicated in the biology of disease Davis et al. (2011); Cichońska et al. (2021).

2.1.1 Mechanism of action (MoA)

Mechanisms of action are the processes by which a molecule, or a ligand, exerts its effects on a biological system (Huang et al., 2012). Understanding these mechanisms is essential for development of bioactive compounds for multitude of purposes, especially for drug development and disease treatment (Atanasov et al., 2021). Molecular interactions between the active substance and specific biomolecular targets constitute the basis of the majority of mechanisms of action. These targets can include proteins, nucleic acids, lipids, and other essential biomolecules for cellular function.

Direct interactions include direct physical interactions between an enzyme and a substrate or ligand (Liao, 2007), while indirect interactions encompass a more general notion of mechanism of action when it is observed as a phenotypical change in affected biological system. The binding affinity between the substance and its targets is instrumental in determining the potency and selectivity of the effect (Ferguson and Gray, 2018; Liao, 2007; Changeux, 2013; Davis et al., 2011).

Mechanism of action of small compounds can be exerted in many ways, e.g. by inhibiting or activating enzymes. Kinase inhibitors and other enzyme inhibitors bind to the active site of enzymes, interfering with substrate binding or catalysis and disrupting particular metabolic or signaling pathways (Cheng et al., 2011). Enzyme activators, in contrast, increase enzyme activity, thereby enhancing catalytic function and cellular processes. The modulation of enzymatic activity is a common drug development strategy, allowing for precise regulation of disease-related biological pathways (Liao, 2007). Cell surface and intracellular receptors are crucial signal transduction mediators in biological systems. Substances can regulate a variety of cellular responses by targeting receptors, including cell proliferation, apoptosis, and immune response (Ferguson and Gray, 2018; Wang and Cole, 2014; Metz et al., 2011).

Frequently, mechanisms of action involve the modulation of signal transduction pathways, which are complex cascades of molecular events that transmit external signals to the interior of the cell, thereby influencing gene expression and cellular behavior (Roskoski Jr, 2015). Understanding the complex network of signal transduction is essential for elucidating how substances influence complex cellular processes and devising targeted therapeutic interventions (Zhao and Bourne, 2018; Liao, 2007).

Mechanisms of action can extend beyond biomolecular interactions to systemic and physiological effects. Substances can, for instance, affect cellular metabolism, ion transport, and hormone signaling, resulting in systemic alterations in organ function or whole-body responses. Understanding the systemic effects of compounds is essential for predicting adverse effects and evaluating their therapeutic potential as a whole, same as avoiding development of resistance mechanisms (Kairys et al., 2019; Zhao and Bourne, 2018).

2.1.2 Binding affinity definition

The concept of binding affinity pertains to the capacity of a ligand to interact with its protein target, resulting in the formation of a biochemical complex. This interaction can take place through several non-covalent forces, including hydrogen bonds, van der Waals forces, electrostatic interactions, and hydrophobic interactions. The higher the binding affinity, the greater the probability that the ligand will attach to the protein target and maintain its binding. On the other hand, a lower binding affinity indicates a comparatively less stable relationship (Kairys et al., 2019).

The quantification of binding affinity is commonly achieved through the utilization of

diverse metrics and approaches. There are several commonly employed methods (Jarmoskaite et al., 2020):

- The Dissociation Constant (K_d): The Dissociation Constant (K_d) quantifies the concentration of a ligand necessary to achieve interaction with its protein target at half of its maximum capacity. A lower dissociation constant (K_d) is indicative of a stronger binding affinity. The equation that governs the dissociation constant (K_d) is:

$$K_d = \frac{[L][P]}{[LP]} \tag{2.1}$$

  Where L is the concentration of the free ligand, P is the concentration of the free protein target and LP is the concentration of the ligand-protein complex.

- The inhibition constant (K_i): The inhibition constant (K_i) pertains to its use as a quantitative measure of an inhibitor's efficacy in reducing the activity of a specific enzyme or protein target. The aforementioned expression denotes the equilibrium constant associated with the formation of the complex between the enzyme and the inhibitor, and can be described by the subsequent equation.

$$K_i = \frac{k_{off}}{k_{on}} \tag{2.2}$$

  Where k_{off} is the dissociation rate constant (the rate at which the enzyme-inhibitor complex dissociates into free enzyme and free inhibitor); k_{on} is the association rate constant (the rate at which the enzyme and inhibitor bind to form the complex). A lower K_i value indicates stronger binding between the inhibitor and the protein target, signifying higher affinity and more potent inhibition.

- The change in free energy (ΔG): ΔG can be used as an alternative method to quantify binding affinity, specifically in relation to the development of the ligand-protein complex. A decrease in the value of ΔG is indicative of a higher degree of binding affinity. The relationship between the change in Gibbs free energy (ΔG) and the dissociation constant (K_d) is expressed as.

$$\Delta G = -RTln(K_d) \tag{2.3}$$

2.2 Protein kinases: Primary targets of interest

Protein kinases are an ubiquitous group of enzymes with a fundamental function in cellular signaling and regulation. They are essential components of virtually all living organisms, from unicellular organisms to complex multicellular organisms, including humans (Roskoski Jr, 2015). The term "kinase" derives from the primary function of these enzymes, which is the transfer of phosphoryl groups from adenosine-3-phosphate (ATP) to specific target proteins, a process known as phosphorylation (Goldberg et al., 2006; Roskoski Jr, 2015). This post-translational modification frequently results in conformational changes that regulate protein function, thus orchestrating an extensive array of cellular processes (Roskoski Jr, 2015).

Structurally, protein kinases share several conserved domains, such as the catalytic kinase domain, which contains the ATP-binding site and the active site responsible for substrate phosphorylation (Liao, 2007). In addition to the catalytic domain, these proteins possess regulatory domains that can either enhance or inhibit their activity in response to various cellular signals. This structural diversity contributes to the enormous array of functions performed by protein kinases (Liao, 2007; Roskoski Jr, 2015, 2022).

Protein kinases are crucial regulators of signal transduction pathways, mediating the exchange of information between extracellular signals and intracellular responses. Upon activation by various stimuli, such as growth factors, hormones, or stress signals, they transmit the information via phosphorylation cascades, resulting in the activation of downstream effectors. These effectors are proteins, enzymes, or transcription factors that orchestrate changes in cellular behavior, gene expression, and metabolism (Cheng et al., 2011). In this way, protein kinases strictly regulate essential cellular processes, such as cell growth, proliferation, differentiation, apoptosis, and responses to environmental cues (Wang and Cole, 2014).

Given their central function in cellular regulation, it is not surprising that protein kinases are associated with a wide range of diseases. Pathological conditions caused by abnormal kinase activity include cancer, neurodegenerative disorders, autoimmune diseases, and metabolic disorders (Castelo-Soccio et al., 2023; Costa-Mattioli and Walter, 2020). As a result, protein kinases have become attractive targets for drug development, and kinase inhibitors have emerged as essential therapeutics for various malignancies and other diseases.

These inhibitors seek to restore normal cellular signaling and prevent disease progression by specifically targeting dysregulated kinases (Castelo-Soccio et al., 2023; Cohen et al., 2021).

The protein kinase superfamily is extraordinarily diverse, consisting of several subfamilies based on sequence similarity and functional characteristics. One of the largest subfamilies is the serine/threonine kinases, which phosphorylate serine or threonine residues on target proteins. The tyrosine kinases, which target tyrosine residues and perform essential roles in cell growth and proliferation, constitute a further important subfamily. Dual-specificity kinases can phosphorylate both serine/threonine and tyrosine residues, thereby increasing the functional diversity of protein kinases (Metz et al., 2011; Roskoski Jr, 2015).

2.3 *In silico* modeling

2.3.1 Molecule and protein representation

Molecule and protein representation in the context of computational biology, bioinformatics, and chemoinformatics involve various methods to describe the structure, properties, and behaviors of molecules and proteins in a format that can be processed by computers (Lim et al., 2021; Nascimento et al., 2016). These representations are crucial for tasks such as molecular modeling, drug design, and understanding biological processes at the molecular level.

Molecular descriptors and sequence similarity

Molecular descriptors are numerical representations of chemical or structural properties of molecules. They play an important role in chemoinformatics, especially in QSAR modeling. These type of descriptors are usually obtained directly from SMILES (Simplified Molecular Input Line Entry System) strings, which encode the complex molecular structure into a string of symbols that are easily readable by computers (Lim et al., 2021). In mechanism of action modeling structural fingerprints have shown to achieve the state-of-the-art performance. Other types of often used molecular descriptors include constitutional, topological, geometric, electronic and quantum chemical descriptors (Oršolić et al., 2021). The choice of the type of descriptors used for modeling highly depends on the nature of the dataset.

Sequence similarity-based protein depictions are a fundamental aspect of bioinformatics and computational biology (Marti-Renom et al., 2004). It is used for understanding and prediction

of the structure, function and evolutionary relationships of protein. Similarity based approaches, same as it is the case for small molecules, rely on the principle that protein with similar amino acid sequences tend to have similar three-dimensional structures, and often, similar functions (Llinares-López et al., 2023; Marti-Renom et al., 2004).

Protein sequence alignment involves arranging the sequences to identify regions of similarity. This process is performed using two different approaches, global and local alignment (Marti-Renom et al., 2004; McClure et al., 1994). Global alignment is performed by Needleman-Wunsch algorithm, that aligns the entire sequences from end to end. This method is usefully in case when examined protein sequences are similar in length and composition. However, when the goal is to find functional domains thus the focus is on identifying smaller regions of similarity between multiple sequences, local alignment is performed by the Smith-Waterman algorithm (Nascimento et al., 2016; McClure et al., 1994).

Learning representations directly from input data

Learning representations directly from input data is a crucial idea in machine learning, particularly in deep learning. It is also known as feature learning or representation learning. This process differs from traditional ways in which features are hand-engineered. Instead, representation learning algorithms automatically discover the representations required for feature detection or classification from the raw data (Nguyen et al., 2021; Öztürk et al., 2018). This method is particularly useful in areas such as image and audio recognition, natural language processing, and bioinformatics, where manual feature engineering can be difficult and time-consuming.

Another reason this approach is a powerful tool because it reduces the need for domain expertise in feature design, and allows models to adapt to wide range of tasks.

2.3.2 Model selection and testing scenarios

In the domain of statistics and machine learning, selecting the best model from a group of candidate models for a given dataset is known as model selection. It includes weighing the trade-off between model complexity and model performance by comparing various models to see which one best reflects the underlying patterns in the data (Benavoli et al., 2017). Standard approach in model selection is to perform the null-hypothesis testing and select the best performing method, so called frequentist approach. Alternative is to perform a nonfrequentist

approach or Bayesian analysis, as it was conducted in Oršolić et al. (2021). Bayesian analysis for model comparison is particularly useful when dealing with complex models, incorporating prior knowledge, or working with limited data (Benavoli et al., 2017).

The process of creating an pipeline for modeling mechanisms of action begins with the selection of models. Following model selection, it is necessary to ascertain how the selected model can be rigorously evaluated to verify the accuracy, robustness, and generalizability of the selected approach. As proposed by the Pahikkala et al. (2015); Cichonska et al. (2017) test datasets could be carefully defined to represent various scenarios on which model could be evaluated. These scenarios simulate the most prevalent applications of trained models for predicting mechanisms of action, including drug discovery and drug re-purposing (Cichonska et al., 2017; Oršolić and Šmuc, 2023).

2.3.3 Machine learning for compound-target binding affinity modeling

Bottlenecks of compound-target binding affinity modeling include the representation of input data. Traditionally, both interacting entities are represented by a set of physicochemical features, or a set of structural fingerprints. Other way of formatting the input data is in the form of similarity computation based on aforementioned features. Recently, more promising approaches include representation learning directly from compound structures, or target protein sequences, by using graph convolutional networks (GCN) (Kipf and Welling, 2016) or convolutional neural networks (CNN) (O'Shea and Nash, 2015), respectively.

Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) are comparable to classic artificial neural networks (ANNs) in that both are made up of neurons that receive an input and perform a set of operations in order to self-optimize via learning. CNNs differ from ANNs due to the fact they are usually employed for pattern recognition within images (O'Shea and Nash, 2015). However, CNNs are also an effective tool in bioinformatics for tasks involving protein categorization, structure prediction, and function prediction. Proteins offer a distinct sequential data structure that CNNs can handle well since they are composed of amino acids arranged in linear sequence. Protein sequences are typically represented as strings of letters, with each letter corresponding to one amino acid (Gelman et al., 2021).

Graph convolutional networks (GCNs)

As a potent tool for learning on graph-structured data, Graph Convolutional Networks (GCNs) bridge the gap between conventional neural network architectures and the irregular patterns present in many real-world datasets (Kipf and Welling, 2016). Graphs are diverse structures that are used to represent entities (nodes) and their relationships (edges). They are perfect for datasets like social networks, chemical structures, and transportation networks where the relationships between the entities are just as important as the entities themselves (Kipf and Welling, 2016; Zhang et al., 2019). Considering that the drug research requires a thorough understanding of the intricate chemical structures and interactions - GCNs are a perfect tool for this domain since molecules are naturally represented as graphs, with atoms acting as nodes and bonds as edges (Nguyen et al., 2021; Öztürk et al., 2018).

2.4 Prediction validation and interpretability

2.4.1 Applicability domain paradigm

The utilization of quantitative structure-activity relationship (QSAR) modeling has proven to be a highly effective method in the fields of drug development and chemical research. The process entails constructing prediction models that establish a correlation between the chemical structure of substances and their biological or chemical actions (Kwon et al., 2019; Golbraikh et al., 2012).

The dependability of QSAR models is heavily contingent upon the idea of applicability domain (AD), which plays a critical role in determining the validity of the model's predictions over a variety of substances (Aniceto et al., 2016; Klingspohn et al., 2017; Eriksson et al., 2003). The definition of the applicability domain holds significant importance when providing guarantees of dependability and precision of QSAR models. Machine learning models are taught using a specific dataset, and their performance may be suboptimal when applied to compounds that lay beyond the scope of the training data. Accurate definition of AD aids in the identification of chemicals that fall inside the predictive scope of the model, hence mitigating the potential for inaccurate predictions (Kwon et al., 2019; Golbraikh et al., 2012; Klingspohn et al., 2017).

Numerous definitions of AD rely on chemical descriptors, which may possess limitations

in adequately portraying the intricate nature of molecular interactions. The utilization of excessively simplified descriptors may result in an inadequate consideration of crucial structural characteristics that exert an impact on biological activity (Aniceto et al., 2016; Klingspohn et al., 2017). The task of establishing a universally applicable AD that can be utilized across all quantitative structure-activity relationship (QSAR) models and datasets present a formidable challenge.

2.4.2 Inductive conformal predictor (ICP) framework for regression tasks

Inductive conformal predictors (ICP) provide a robust framework for quantifying the dependability of predictive models. This methodology is notable for its adaptability and capacity to deliver uncertainty measures for individual predictions, which is a crucial aspect in many data-driven applications (Shafer and Vovk, 2008; Papadopoulos et al., 2011).

At its core, ICP's use simple statistical techniques to estimate the confidence of the predictions generated by any machine learning model. This is accomplished through a process that involves partitioning of a datasets into distinct subsets, with one subset being utilized for model training and the other for calibrating confidence levels (Shafer and Vovk, 2008; Aniceto et al., 2016).

The so-called calibration is performed using a pre-defined calibration set. The framework operates by computing nonconformity scores that essentially measure how well or poorly each instance in a calibration set conforms to the patterns learned by the model from the training set. In practice, nonconformity scores for regression tasks are usually defined as an absolute difference between the true and the predicted value, $\alpha = y - \hat{y}$ (Shafer and Vovk, 2008; Papadopoulos et al., 2011; Vovk et al., 2018). Every time a new instance is introduced these nonconformity scores are used to assess the degree of similarity or dissimilarity between a given instance and the calibration data. Each prospective prediction is accompanied by a prediction region and a corresponding confidence level (Shafer and Vovk, 2008). Confidence level in this case is a metric that represents the probability that the actual label or value of the newly introduced instance is contained within the given set of calibration samples Papadopoulos et al. (2011).

One of the important aspects of ICP's is that they are model-agnostic, meaning that they can be integrated with any existing predictive model, including simple statistical approaches or

deep learning architectures.

3. Scientific papers

3.1 PAPER 1: Comprehensive machine learning based study of the chemical space of herbicides

Oršolić, D., Pehar, V., Šmuc, T., Stepanić, V., 2021. Comprehensive machine learning based study of the chemical space of herbicides. Sci Rep 11, 11479. https://doi.org/10.1038/s41598-021-90690-w

scientific reports

Check for updates

OPEN

Comprehensive machine learning based study of the chemical space of herbicides

Davor Oršolić[1], Vesna Pehar[2], Tomislav Šmuc[1] & Višnja Stepanić[1]✉

Widespread use of herbicides results in the global increase in weed resistance. The rotational use of herbicides according to their modes of action (MoAs) and discovery of novel phytotoxic molecules are the two strategies used against the weed resistance. Herein, Random Forest modeling was used to build predictive models and establish comprehensive characterization of structure–activity relationships underlying herbicide classifications according to their MoAs and weed selectivity. By combining the predictive models with herbicide-likeness rules defined by selected molecular features (numbers of H-bond acceptors and donors, logP, topological and relative polar surface area, and net charge), the virtual stepwise screening platform is proposed for characterization of small weight molecules for their phytotoxic properties. The screening cascade was applied on the data set of phytotoxic natural products. The obtained results may be valuable for refinement of herbicide rotational program as well as for discovery of novel herbicides primarily among natural products as a source for molecules of novel structures and novel modes of action and translocation profiles as compared with the synthetic compounds.

Herbicides are compounds of small molecular weight used for selective destruction of weeds. Because of their extensive use, the two global issues have appeared in the last two decades, an increase in weed resistance and health issues[1]. In order to circumvent development of weed resistance, herbicides with different modes of action (MoAs) are applied rotationally. Herbicides are classified according to the MoAs in ~ 25 classes within the two similar classification systems—HRAC and WSSA, set up by Herbicide Resistance Action Committee of Australia and Weed Science Society of America, respectively[2–5]. The MoAs denote the biochemical processes in weeds which herbicides modify (Table 1). Given the common name of a herbicide, the classification schemes in addition to MoA also provide the chemical family a herbicide belongs to. Sub-classification to the chemical families according to possessing common fragment(s) was made in order to refine herbicide rotation scheme and increase its efficiency against the weed resistance. The chemical sub-classification of the herbicides is, however, not unequivocal. Different number of chemical sub-groups have been defined in the HRAC and WSSA systems and recently by Forouzesh[6].

Among the MoAs, ten of them are identified with the inhibition of specific enzymes and are associated by around half of the used herbicides (Table 1). However, the precise mechanisms of action of herbicides resulting in their phytotoxic effects are rarely known [7]. For example, herbicides from the most populated and used class B are all inhibitors of the enzyme acetolactate synthase (ALS), known also as acetohydroxyacid synthase (AHAS), which catalyzes the first step in the synthesis of the branched-chain amino acids valine, leucine, and isoleucine. However, their phenotypic inhibitory effects can be different what may be due to different binding modes onto ALS/AHAS and/or their different translocation properties through weeds[7,8]. Herbicides of different MoAs have also different propensities to induce weed resistance because of not only different prevalence of their usage, but also different sites of action (SoAs) and translocation properties.

The MoA classification schemes for herbicides are examples of the application of the structure–activity relationship (SAR) analysis. The general SAR assumption is that structurally similar compounds share SoA. The sub-partition of MoA classes into chemical families is in the line with this assumption. However, such an assumption does not imply that compounds which are structurally dissimilar may not have the same SoA/MoA what may afflict the usage of the classification schemes in the rotational anti-resistance strategy. Indeed, it has been demonstrated by scaffold hopping methods in design of novel biologically active compounds that dissimilar structures can have the same MoA[9]. Furthermore, there is an open question how much compounds belonging

[1]Laboratory for Machine Learning and Knowledge Representation, Division of Electronics, Ruđer Bošković Institute, Bijenička 54, 10002 Zagreb, Croatia. [2]Croatian Defense Academy "Dr. Franjo Tuđman", Ilica 256b, 10000 Zagreb, Croatia. ✉email: stepanic@irb.hr

Legacy hrac code	hrac2020&wssa code	Number of compounds in hrac2020/ extended set	General mode of action–targeted biological process	Mode of action–targeted molecular functions
A	1	21/29	Fatty acid biosynthesis	Inhibition of acetyl-CoA carboxylase (ACCase)
B	2	58/61	Amino acid synthesis (Leu, Ile, Val)	Inhibition of acetohydroxyacid synthase/ acetolactate synthase (AHAS/ALS)
C1	5	43/53	Photosynthesis (electron transfer)	Inhibition of photosystem (PS) II protein D1 (C1/C2 Ser264; C3 His215)
C2	5	30/37		
C3	6	5/9		
D	22	4/5	Photosynthesis (electron transfer)	Inhibition of diversion of the electrons transferred by the PS I ferredoxin
E	14	29/43	Photosynthesis (heme synthesis for chlorophyll)	Inhibition of protoporphyrinogen oxidase (PPO)
F1	12	7/9	Photosynthesis (carotenoid synthesis)	Inhibition of phytoene desaturase (PDS)
F2	27	14/16		Inhibition of 4-hydroxyphenylpyruvate dioxygenase (4-HPPD)
F3	34	1/2		Inhibition of lycopene cyclase
F4	13	2/1		Inhibition of 1-deoxy-d-xylulose-5-phosphate (DOXP) synthase
G	9	1/2	Amino acid synthesis (Phe, Trp, Tyr)	Inhibition of 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase
H	10	2/4	Amino acid synthesis (Gln)	Inhibition of glutamine synthase
I	18	1/3	Tetrahydrofolate synthesis	Inhibition of dihydropteroate (DHP) synthase
K1	3	18/25	Microtubule polymerization	Inhibition of microtubule assembly
K2	23	6/9		Inhibition of microtubule organisation
K3	15	43/39[a]	Fatty acid synthesis	Inhibition of VLCFAs
L	29[b]	6/6	Cell wall synthesis	Inhibition of cellulose synthase
M	24	6/8	ATP synthesis	Uncoupling of oxidative phosphorylation
N	NA[b]	NA[b]/23	Fatty acid synthesis	Inhibition of fatty acid elongase
O	4	25/37	Regulation of auxin-responsive genes	Synthetic auxin mimics -Stimulation of transport inhibitor response protein 1 (TIR1)
P	19	2/3	Long-range hormone signaling	Auxin transport inhibitors

Table 1. HRAC classification and division of herbicides from the HRAC2020 and extended data sets across the MoA classes[a]. [a]In the HRAC2020 classification there are additional classes Q (3), R (31), S (32) and T (33), all with up to 2 members [5]. [b]Majority of herbicides from the class N are fused in the K3 (15) class. The treating 23 herbicides of the legacy N class separately, does not affect the results since this subgroup is structurally diverse from the other K3 herbicides.

to different MoA classes are mutually structurally similar and may hence act in similar way what can also impair the rotational strategy.

The other approach to circumvent weed resistance is through discovery of novel molecules with different MoA. The valuable source of such molecules is natural products (NPs)[10]. The first of the two main objectives of our computational study was to provide a formal rationale for the underlying SAR assumption of the MoA classification schemes used in confrontation with the worldwide increase in the weed resistance and to point out potential limitations of MoA labelling with using only structural similarity. In an attempt to improve herbicide characterization and thus rotational strategy, categorizations of herbicides according to their application stage and weed selectivity were also modelled for the first time as far as we are aware. By combining machine learning (ML) models with a set of herbicide-likeness rules, virtual screening platform is proposed. Another objective was to enrich the phytotoxic chemical space with molecules having novel MoA. For this purpose, the screening cascade was applied on the set of phytotoxic NPs.

Methods

Data sets. The calculations were done with the data set HRAC2020 of 346 mainly synthetic organic herbicides downloaded from the original HRAC list and its extended version of 509 herbicides with relative molecular weight within the range 84–649[5]. The extended data set contains additional 163 mostly obsolete herbicides collected from the literature and open-source online databases: Compendium of Pesticide Common Names (http://www.alanwood.net/pesticides/), PPDB: Pesticide Properties Database, PubChem and PTID: Pesticide Target Interaction Database[6,11–13]. The MoAs were assigned for 411 compounds according to the legacy HRAC system (314 herbicides from the HRAC2020 set) and on the basis of belonging to chemical families (97 herbicides forming the subset HRAC_REST) (Table 1)[5,6,14]. The remaining 98 herbicides herein referred as the Z class, were unclassified (Supplementary Table S1). The data on application stage and weed selectivity were collected for

subsets of 221 and 332 herbicides, respectively[14]. The data set of 131 phytotoxic NPs was collected from literature (Table S2) [15–24].

Molecular descriptors. The cleaned SMILES were used as inputs for the calculations of 1D and 2D molecular descriptors by the R package *rcdk*[25] and the programs DataWarrior[26] and ADMET Predictor 9.5 (Simulations Plus, Inc., USA)[27]. The *rcdk* descriptors were structural fingerprints (fp) (11 different types including extended and 166-bit MACCS fps), constitutional (17 of them), electronic (6) as well as hybrid BCUT (6) descriptors. The 141 MACCS keys which were present in more than five herbicides were used as descriptors. Physicochemical and simple structural properties which govern uptake and translocation properties of herbicides through plants[28–34] were calculated by DataWarrior (27) and ADMET Predictor 9.5 (139). The net ionization state of molecules was roughly estimated as a difference of numbers of basic nitrogen (pKa above 7.0) and acidic oxygen atoms (pKa below 7.0) calculated by DataWarrior. Prior to modelling, descriptors (except fp) were scaled as (x − mean(x))/sd(x).

Hierarchical clustering. Hierarchical clustering was performed with wardD.2 minimum variance agglomeration method and Tanimoto coefficient (TC) as a similarity index by the stratified sampling function *hclust*. The Dunn (the ratio: the cluster minimum separation/the maximum cluster diameter) and Dunn2 (the minimum average dissimilarity between two clusters/the maximum average dissimilarity within cluster) indices as well as average Silhouette (Si) width (compares the average distance to elements in the same cluster with the average distance to elements in other clusters) were used for internal clustering validation. The adjusted Rand index (ARI) was applied in order to assess the similarity of the predicted grouping with the legacy HRAC labels. The three internal validation scores are higher and better when clusters are dense and well separated. Considering external validation, more similar groupings has a positive ARI closer to 1. The clustering validation indices were calculated by the R package *fpc*.

Modelling. The multi-classification modeling in terms of subsets of various kinds of descriptors was performed by Random Forest (RF) method ('rf') available in the R package *caret* with one tunable parameter (mtry, a number of variables randomly sampled at each split) and using tenfold cross-validation (CV). The HRAC classes with less than 3 members (Table 1) were excluded from modelling and these compounds were added to the Z class. The remaining 314/419 compounds from the HRAC2020/extended set were divided into training and test sets in the 80:20 ratio, except in the case of the classes with 3–5 members, for which 50:50 ratio was applied. The splitting was done using stratified random sampling. Thus, in the case of original/ extended herbicide set, there were 257/341 training and 57/78 test compounds arranged in 16/19 classes. Analogous dividing procedure was applied for the subsets of 221/332 compounds with assigned application stage/weed selectivity.

Further, in order to optimize performance of MoA and weed selectivity models in terms of selected descriptors, the hyperparameter tuning of RF and three additional classifiers eXtreme Gradient Boosting (XGBoost), support vector machines (SVM, RBF kernel) and naive Bayes (NB) as a baseline model, all available in *caret*, were carried out by using grid search and 10 runs of tenfold CV as well as by keeping all resamples for performance comparison (Figures S1–S4). For RF and NB classifiers, parameter tuning was done by utilization of the packages *randomForest* and *klaR*, respectively. The final models were built with optimal values of tuning parameters on the entire training HRAC2020 set. The classifiers were compared mutually by analyzing resampling distributions and using Bayesian analysis (Python library *baycomp*)[35] as well as by their performance on the test test.

The model predictive capacity was assessed by counting the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class and usage of following performance metrics: sensitivity (Sensitivity or Recall = TP/(TP + FN)), precision (Precision = TP/(TP + FP)), specificity (Specificity = TN/(TN + FP)), overall predictive accuracy (Accuracy = (TP + TN)/(TP + FP + FN + TN)), F1 score (F1 Score = 2*(Recall * Precision)/(Recall + Precision)) and Cohen's unweighted kappa (Cohen's kappa = (Po − Pe)/(1 − Pe), where observed probability is Po = (TP + TN)/(TP + TN + FP + FN), and probability by chance is Pe = ((TP + FN) * (TP + FP) + (FP + TN) * (FN + TN))/(TP + TN + FP + FN)^2).

Applicability domain (AD). ADs were defined in terms of similarity with training compounds and the class probability outputs from the RF models[36]. Structural similarity between two molecules was estimated by using 141 MACCS keys and the coefficient TC as a similarity measure. Similarity in physicochemical space is assessed by applying the Euclidian distance.

Violin and PCA plots. The violin plots with relevant statistical details for comparison subgroups of herbicides in molecular properties were made by using the *ggstatsplot*. The principal component analysis (PCA) was done with *princomp*.

The R computing was done within RStudio (R version 3.6.3) environment[37].

Results and discussion

HRAC classification—descriptor and model selection. The multi- classification of herbicides according to MoAs in terms of subsets of various kinds of molecular descriptors was performed by RF modelling. The results obtained for the HRAC2020 and extended data sets were consistent. The best classification performance for the extended test set was obtained by using MACCS keys as molecular descriptors (Table S3). With other kinds of descriptors, the models somewhat deteriorated most probably because they do not contain information on specific structural arrangements of atoms within molecules. The constitutional descriptors (e.g. MW,

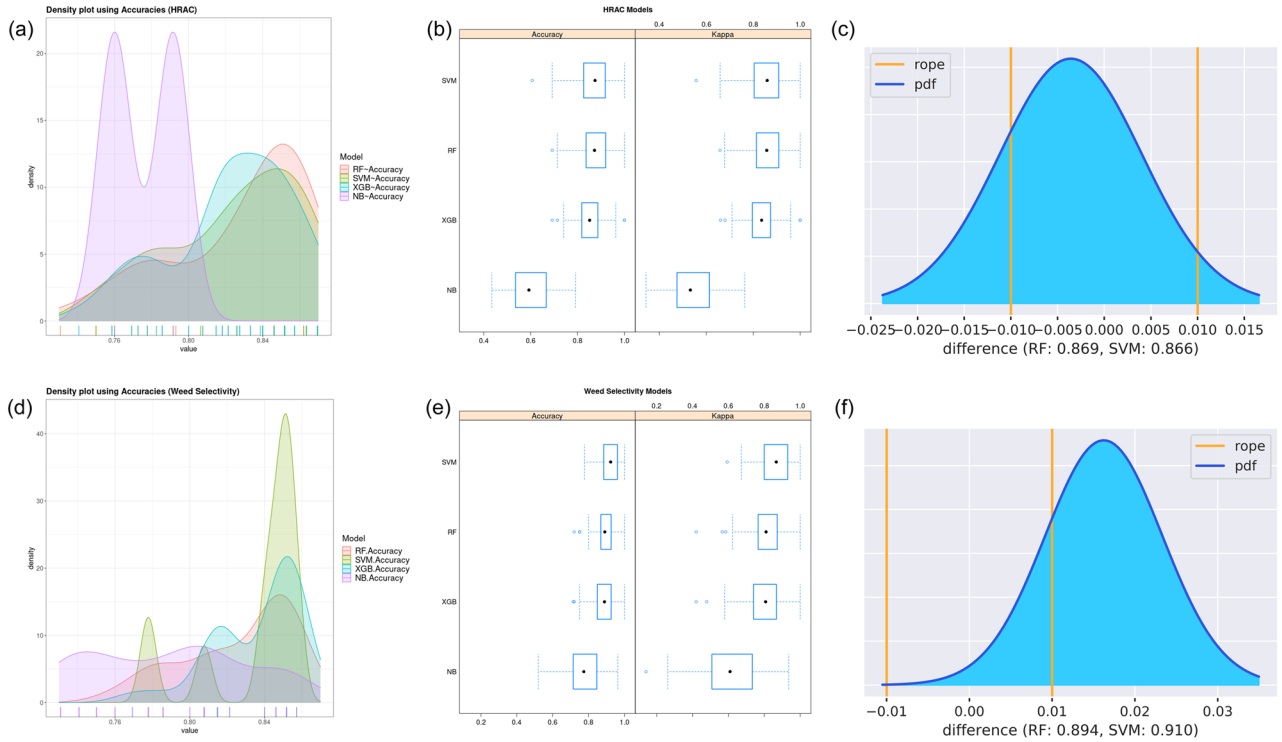


Figure 1. Comparing performance of the four ML classifiers for MoA predictions. (**a**, **d**) Density accuracy plot. (**b**, **e**) Box plots of distributions of resampled accuracies and kappa values. (**c**, **f**) Probability density plot for accuracy differences between the RF and SVM classifiers. The plots (**a**)–(**c**) described MoA classifiers (Table 2) and those (**d**)–(**f**) present comparison of weed selectivity models built with nine descriptors including log P (Table 3). The RF and SVM MoA classifiers are largely equivalent since 75.7% of posterior probability distribution is inside the region of practical equivalence (rope, the differences of accuracy are less than 1%).

numbers of atoms in the longest aliphatic chain, the largest pi system or of aromatic atoms), lipophilicity parameter and electronic descriptors (e.g. topological polar surface area (TPSA), numbers of hydrogen bond acceptor (HBA) or donor (HBD) atoms, molecular atomic and bond polarizabilities) are more general and global molecular characteristics whose values do not correlate with structural arrangement. The hybrid BCUT descriptors were also not efficient as MACCS fp in differentiation of herbicides with different MoAs although they are known for their usefulness in description of chemical diversity[38]. The MACCS structural keys better represent the scaffolds characterizing the chemical series of herbicides than other explored fp types.

The performance of the RF model was optimized by hyperparameter tuning along with 10 times tenfold CV resampling. The three additional ML classifiers XGBoost, SVM and NB were also explored and tuned in analogous way using the same seed to secure that folds between models contain the same set of compounds (Figure S1, Table S4). The Bayesian analysis for comparing performances of multiple classifier showed that RF and SVM(RBF) exhibit similar performance on the HRAC problem, dominating XGBoost while NB was clearly outperformed by the rest (Fig. 1a–c). The outputs of the RF and SVM (Table 2, Table S1) as well as MACCS keys determined as important (Table S5) for 16-class MoA categorization by both ML approaches are largely equivalent. They differ in predictions for 5 test and 12 HRAC_REST case compounds, which were all predicted with the RF class probabilities less than the cut-off value (see further).

Although SVM slightly overperformed the RF model (Table 2), we decided to perform further analysis with the RF outputs. The primary reason was possibility to use direct RF output class probabilities for definition of the model's AD. Using SVM in the context of AD definition would require additional calibration of the SVM scores, to turn them into probabilities[39].

HRAC classification and structural similarity—Chemical space analysis. The classification of herbicides into the HRAC/WSSA classes (Table 1) facilitate the rotational use of herbicides of different MoA as a strategy against the weed resistance[5]. To the best of our knowledge, the sub-classification into chemical families has been done by visual inspection[6]. Herein by applying ML approaches it is shown in an objective, formal way that dividing herbicides into chemical families and also MoA classes is based on their structural similarity.

Regardless of used descriptors (Table S3) and ML algorithm (Table 2), the MoA models were generally characterized with the higher specificity than sensitivity averaged across the classes. Such a performance points to a degree of similarity between the herbicides designated to different classes what is also supported by the clustering analysis. The herbicides were clustered primarily according to common scaffolds.

This resulted in only moderate value of ARI index signifying relatively weak agreement between the generated clusters and the HRAC classes (Fig. 2). The inter-cluster distances were also described by relatively low values

MoA	Overall[b]		Averaged across classes				
Classifier	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1	Balanced Accuracy
TEST SET							
RF	0.895	0.883	0.821	0.993	0.896	0.900	0.907
XGBoost	0.895	0.883	0.821	0.993	0.899	0.899	0.907
SVM	0.912	0.902	0.838	0.994	0.935	0.936	0.916
NB	0.561	0.500	0.332	0.969	0.663	0.604	0.651
HRAC_REST SET							
RF	0.674	0.646	0.641	0.979	0.728	0.796	0.814
XGBoost	0.663	0.633	0.594	0.978	0.670	0.771	0.790
SVM	0.696	0.667	0.631	0.980	0.673	0.797	0.809
NB	0.413	0.362	0.310	0.961	0.509	0.605	0.638

Table 2. Comparison of classification performance on the test and HRAC_REST case sets of the four optimized 16- class MoA ML models built in terms of 141 MACCS fp keys[a]. [a]Optimal values of classifiers' hyperparameters are listed in Table S4. [b]The overall accuracy and kappa values are averaged over 10 × 10-fold CV resamplings.

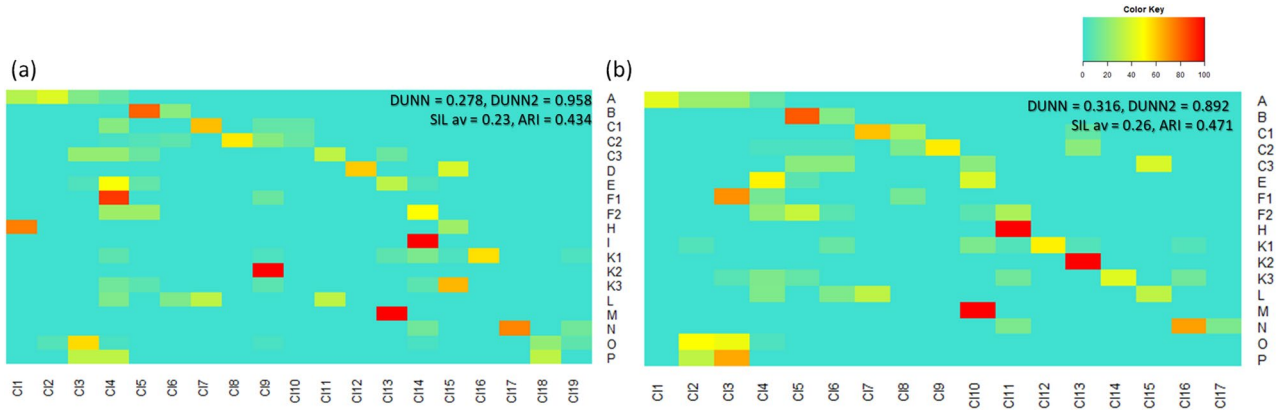


Figure 2. Heat map presentations and evaluation metrics for distributions of (**a**) HRAC2020 + HRAC_REST (411) and (**b**) HRAC2020 (314) herbicides in terms of fractions (%) of MoA classes in clusters generated by the agglomerative algorithm and MACCS fp.

of internal evaluation Dunn, Dunn2 and average Silhouette indices pointing to similarity between herbicides from different clusters in MACCS (as well other fps, results not shown) representation (Fig. 3). The unclassified Z compounds (placed in the upper right corner of the heat map in Fig. 3a) are the most structurally diverse molecules. They are structurally different mutually as well as from the rest of herbicides and thus they are unclassified. The most numerous class B (Table 1) is divided into the two relatively homogenous clusters: the 5th cluster of 49 sulfurones and sulames and the 6th cluster with 12 remaining ALS inhibitors possessing imidazoline or pyrimidinyl(thio)benzoate fragments (Fig. 1). Several herbicides with sulphonamide fragment from the other classes E, F2 and K3 are merged with the 5th cluster. The other two chemically homogenous clusters 1st and 2nd correspond to the well-known sub-groups of the ACC inhibitors of the A class—those with cyclohexanedione ring (DIMs) and those with aryloxyphenoxy-isopropionate fragment (FOPs), respectively. The five of ACC inhibitors are grouped in the 3rd cluster with the subgroup of synthetic auxins O (plant hormones), on the basis of possessing common halogenated phenoxyl fragment. In difference, the PPG oxidase (chlorophyll synthesis) inhibitors of the class E are dominant in the two heterogenous clusters (cl4 and cl13/ cl4 and cl10 in Fig. 2a/b). In the cluster cl4, they are grouped with some A, C1, C3, F1, F2, K1 and K3 herbicides, while in another cluster they are put together with all ATP synthesis inhibitors from the class M.

The obtained results illustrate that herbicides from different HRAC classes share structural fragments which may direct them to the same biological activity. Such results may point to the caution in the application of the rotational anti-resistance strategy using only MoA classification systems.

In order to apply the RF model to unclassified compounds such as Z compounds and phytotoxic NPs, the AD was defined. The AD presents the region in chemical space where the model's individual predictions are reliable. The AD boundaries were defined by the two parameters: (1) structural similarity with the training compounds and (2) the predicted RF class probability (Fig. 3c). The RF class probability has already been shown to be efficient for differentiating between reliable and unreliable predictions[36]. An RF class probability is estimated as a fraction of total number of trees which for a given compound votes for this class. It corresponds to one minus

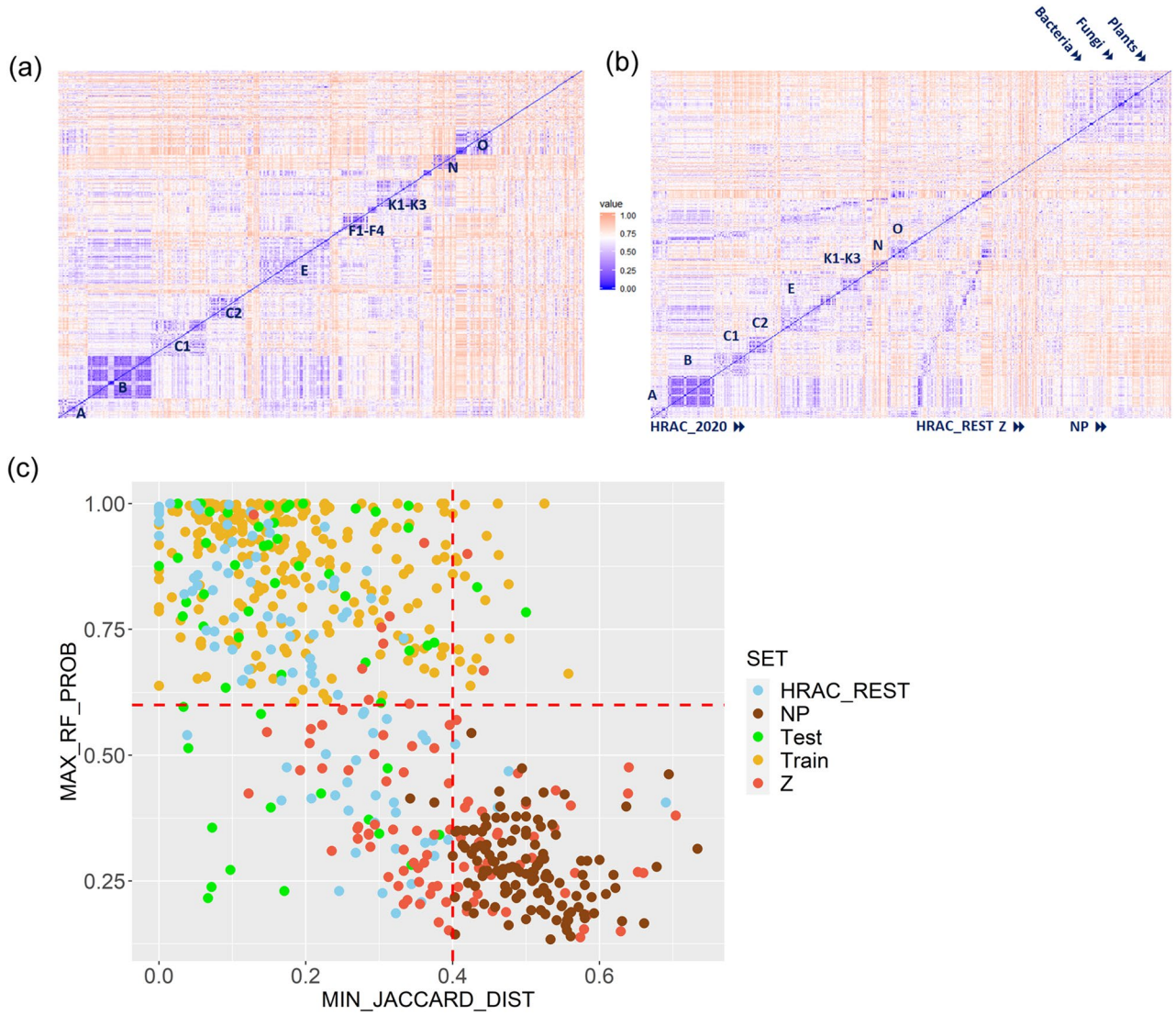


Figure 3. Heat maps for structural dissimilarity quantified by Jaccard coefficient(1-TC) calculated for all pairs of 509 synthetic herbicides (**a**) arranged into MoA classes and (**b**) divided into the subsets HRAC2020, HRAC_REST and the Z compounds with addition of the set of NPs originated from bacteria, fungi and plants. The extended, HRAC2020 and HRAC_REST compounds are ordered according to the classes A-P. More blue/red values correspond to more structurally similar/diverse compounds. (**c**) Definition of AD for the RF MoA model (Table 2): given a compound, the model's prediction is considered reliable if it is similar to at least one training herbicide with TC greater than 0.6 and the estimated class probability is greater than 0.6.

the error probability and thus provides a confidence level on the class prediction and can be used for ranking. For all training herbicides, the MoA labels were accurately predicted with the class probabilities greater than 0.6 and hence this value was taken as an AD boundary (max_rf_prob > 0.6, Fig. 3c, Table S1). For structural dissimilarity the threshold in the Jaccard index (1-TC) of 0.4 was chosen, that is an external compound should be similar to at least one of the training herbicides with a minimal TC greater than 0.6 (min_Jaccard_dist < 0.4).

The MoA class for 75.4% of the test compounds was predicted with max_rf_prob > 0.6 and for all of them the MoA was correctly predicted. In the case of the HRAC2020 set, the independent external set contains 92 herbicides (compounds assigned to the classes G, H and I were dismissed) from the HRAC_REST subset which were classified a priori on the basis of their chemical families available in the literature and online sources (Fig. 3a)[6,11–13]. Among 60 HRAC_REST compounds which lay within the AD, only ethoxyfen was predicted as A instead of E class inhibitor (Table S1) [5]. Most of these correctly predicted but obsolete herbicides are inhibitors of photosynthesis (C1, C2, E) or fatty acid synthesis (A, K3) as well as plant growth regulators (O). Although for the majority (29) of the rest of 32 compounds the minimal TC was greater than 0.6, their class probabilities were less than the cutoff 0.6 and they were hence left unclassified. Considering Z compounds, although 55 of them are structurally similar to the training compounds with TC > 0.6, only 12 of them lye within the AD and MoA might be assigned. This illustrates that structural similarity estimated on the presence of the common structural fragment(s) in MACCS representation is not sufficient condition for conclusion upon sharing the

RF /SVM[b]	Per classes				
9 descriptors with logP	Sensitivity	Specificity	Precision	F1	Balanced Accuracy
Class: BL	0.944/0.917	0.690/0.690	0.791/0.786	0.861/0.846	0.817/0.803
Class: G	0.739/0.696	0.952/0.929	0.895/0.842	0.810/0.762	0.846/0.812
Class: NS	0.500/0.667	1.000/1.000	1.000/1.000	0.667/0.800	0.750/0.883
141 MACCS					
Class: BL	1.000/1.000	0.793/0.828	0.857/0.878	0.923/0.935	0.897/0.914
Class: G	0.783/0.826	1.000/1.000	1.000/1.000	0.878/0.905	0.891/0.913
Class: NS	0.833/0.833	1.000/1.000	1.000/1.000	0.909/0.909	0.917/0.917

Table 3. Comparison of performance metrics on the test set of 3-class RF and SVM models built for prediction of BL, G or NS weed selectivity of herbicides in terms of subset of nine simple molecular and physicochemical descriptors including lipophilicity coefficient logP or 141 MACCS keys[a]. [a]The nine descriptors are logDiff, logSw, Shapeindex, Cat, sp3At, TPSA, HBA, HBD plus logP. [b]The RF and SVM models with 9 descriptors including log P/141 MACCS keys correspond to the models **1** and **7**/**3** and **9**, respectively, in Table S6. The models were trained and applied with using tuned hyperparameters' values (Figures S2–S4).

common MoA. The more complex representation is necessary for similarity based AD definition than provided by MACCS(-like) fingerprint—one that is inherently captured by more complex models such as those provided by RF or SVM algorithms.

Weed selectivity and application stage—descriptor and model selection. Adding descriptors which are known to describe uptake and distribution of compounds through plants, reduced the sensitivity of the MoA classification models (Table S3)[28–34]. The increase in number of FNs indicated that there are common molecular characteristics between members of different MoA classes. Herbicides are also classified according to their application stage and selectivity toward different types of weeds. The phytotoxic effectiveness greatly depends upon herbicide application timing and environmental conditions. Correct application timing maximizes weed control and limits crop injury. There are pre-emergent (here denoted as PRE) herbicides that control seedling growth of weeds and post-emergence (POST) ones which control actively growing tissue of young weeds in a way to be applied directly onto weeds and away from a crop. There are also compounds which can be applied in both regimes (BOTH). The analyzed subset of synthetic herbicides included 221 herbicides of which 49/90/82 are applied in PRE/POST/BOTH regime (Table S1)[14]. The 3-class models for the complex application stage variable built by using MACCS keys, physicochemical and/or simple molecular features of compounds without considering environmental variables, had, in general, lower predictive power (test set: accuracy ~ 0.62, kappa ~ 0.40) than the predictive models for MOAs (Table 2) and weed selectivity (Table 3). Hence, we did not pursue further model analysis and interpretation.

Herbicides may be divided into the three classes with regard to weed selectivity: herbicides which act selectively against broadleaf (BL) or grass (G) weeds and those which are non-selective (NS) and act on broad spectrum of weeds[40]. The BL or G herbicides clear away only certain weeds by acting on processes that are more important for those types of weeds, while the NS herbicides act on processes that are important in all plants. Although the weed resistance is observed for herbicides regardless of their weed selectivity class, the rotational change of herbicides with different selectivity may reduce weed resistance caused by change in herbicide translocation profile [8]. In the data subset of 332 herbicides, 181 BL selective herbicides are from MoA classes C1, C2 and E associated with the photosynthesis inhibition and the class O of growth regulators. The 118 G selective herbicides are from the classes A, K1, K3 and N and are mostly inhibitors of fatty acid synthesis. The most of 33 collected NS herbicides are mainly from the classes B, D and P. The most prone to weed resistance are inhibitors from the classes B, C1-C3, A and G[5].

The 3-class RF models were built by dividing 332-data set into 267 training and 65 test compounds represented by MACCS keys and more than 160 other molecular properties. By employing the later set of descriptors, the nine conceptually clear and whole molecular features were identified among most important and efficient for herbicide differentiation according to weed selectivity (Table 3). Adding or using other descriptors did not change predictive power of models significantly. These are partition (logP) or distribution (logD at pH 7.4) coefficient, native solubility Sw in pure water at 25 °C (transformed to log(Sw/mol L^{-1})), diffusion coefficient in water (Hayduk-Laudie formula, log(Diff × 10^{-5}/ (cm^2/s)), TPSA as well as numbers HBA and HBD all calculated by ADMET Predictor[27], as well as ShapeIndex (spherical < 0.5 < linear) and numbers of sp^3-hybridized (sp3At) and all carbon (Cat) atoms within molecule calculated by DataWarrior[26].

Among explored ML classifiers the most competitive were RF and SVM models (Fig. 1e, Table 3, Table S6). The RF and SVM predictions differ mutually for one/three test compounds and 36/24 case compounds described in terms of MACCS fp /nine whole molecular features including logP without taking AD criteria into regard. Although classification of synthetic herbicides into BL, G and NS classes was somewhat better in terms of MACCS fp (Table 3), we decided to promote the set of whole molecular descriptors. The later descriptors provide simple and meaningful interpretation to the potential end users including chemists interested in discovery and development of not only novel herbicides but also molecular probes for investigation of biological processes in

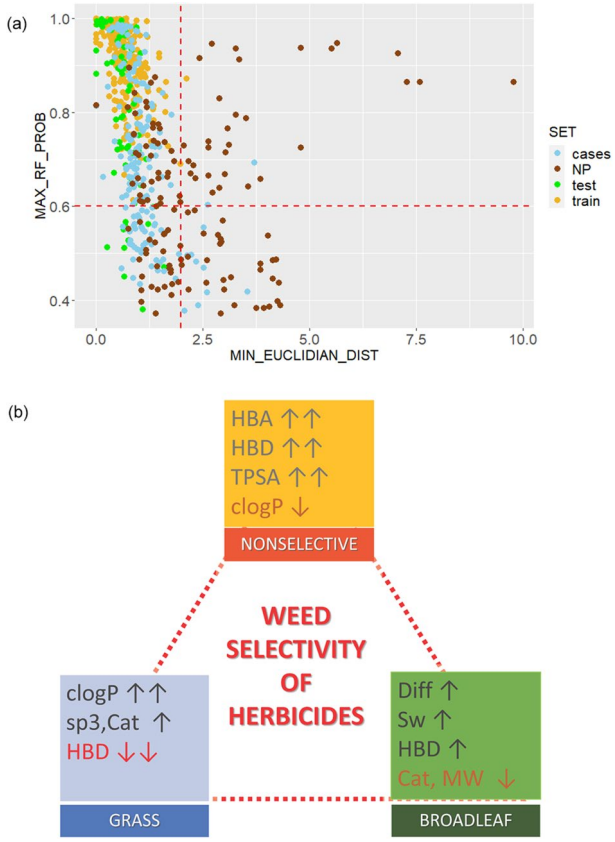


Figure 4. (**a**) The AD for the RF weed selectivity model (**1** in Tables 3 and S6). Given a compound, the prediction can be considered credible for the class probability above 0.6 and the Euclidian distance less than 2.0. (**b**) The most distinguishing molecular features of the broad-leaved or grass selective and non-selective herbicides.

plants. Additionally, in comparison with the models built in terms of MACCS fp keys, the models built in terms of physicochemical and whole molecular descriptors are more general and may not be limited to structurally similar compounds as it is demonstrated by comparison of the ADs in Fig. 3c vs Fig. 4a. The use of either logP or logD did not impact predictive power of the RF models considerably (Table S6). Since logP coefficients are more readily calculated, the further analysis is focused on the RF model with logP.

Weed selectivity—physicochemical space analysis. The AD for the RF model (**1** in Tables 3 and S6) is defined by the use of its class probability outputs and Euclidean similarity with the training compounds in the physicochemical space spanned by the nine descriptors (Fig. 4a). All training compounds were predicted with the class probability above 0.6. The RF model predicts correctly weed selectivity for more than 3/4 of 65 test synthetic herbicides using the thresholds of 0.6 for class probability and 2.0 for Euclidian distance (Fig. 4a). The half of the rest of the test compounds was either left unclassified (class probability < 0.6) or were wrongly assigned in spite of their similarity with the training compounds in the physicochemical space.

Considering 177 external case compounds, 135 were within the AD and for them weed selectivity was assigned using the probability cutoff of 0.6 (Table S1, Fig. 4a). Most of these synthetic herbicides were predicted to be BL by all classifiers (Table S1).

The nine physicochemical and simple molecular properties are, in general, associated with uptake and translocation of compounds through plants[41,42]. However, this observed dependence of the weed type selectivity may also be related to the specific sub- cellular/plastid location of target proteins (pathways) and/or to different characteristics of binding sites of herbicides on targets. As compared with the BL and G selective compounds, the NS herbicides are more polar molecules possessing larger polar surfaces TPSA and more HBA (> 5) and HBD (mostly 2) heteroatoms and hence they are more hydrophilic (smaller logP/logD values and more soluble in water) (Figs. 4b and S5). In opposite, the G selective herbicides are molecules with the smallest number of HBD atoms and the smallest relative polar surface. Majority of BL herbicides have one HBD atom. While most of the broad-spectrum NS herbicides have logP lower than 2, most of selective herbicides particularly of the G type has logP greater than 3.0. The BL selective herbicides have the smallest number of sp3 hybridized atoms,

molecular weight and molecular volume what may be reflected in their distinguishing diffusion and distribution properties in comparison with herbicides from the other two selectivity classes[43].

Assessing the potential of phytotoxic natural products. Natural products are a treasured source for novel biologically active compounds, including those with phytotoxic effect[15,18]. So far NPs have had a relatively small impact on the discovery and development of novel herbicides as compared with insecticides and fungicides. Less than 10% of active ingredients registrations for weed management have been of natural origin[16]. However, in ten of the HRAC classes either a NP, a semisynthetic derivative or synthetic herbicide inspired by a natural scaffold are present[18]. Importantly, most of NPs have different modes of phytotoxic activity than synthetic organic herbicides[16,19,21].

The data set of 131 phytotoxic NPs, with MW less than 650, was collected from the literature[15,16,19]. They are mainly of bacterial (39.6%), fungal (35.1%) or plant (17.9%) origin (Table S2). Although coming from different sources, these natural compounds are structurally more similar mutually than to the synthetic herbicides (Fig. 3b). Since phytotoxic NPs are structurally different, they fall outside the ADs of the models based on the MACCS structural keys of the synthetic herbicides (Fig. 3c). In comparison, more than half of NPs are similar to the training compounds within space defined by the nine descriptors, having Euclidian distance less than 2.0 (Fig. 4a, Table S2). However, only 1/3 of the whole NP set fall within the AD RF model. This analysis indicated that NPs may differ from synthetic herbicides not only in structural space and MoAs, but also in space of the physicochemical and simple molecular features which are often associated with uptake and translocation properties (Fig. 5a and Figure S6)[28–34].

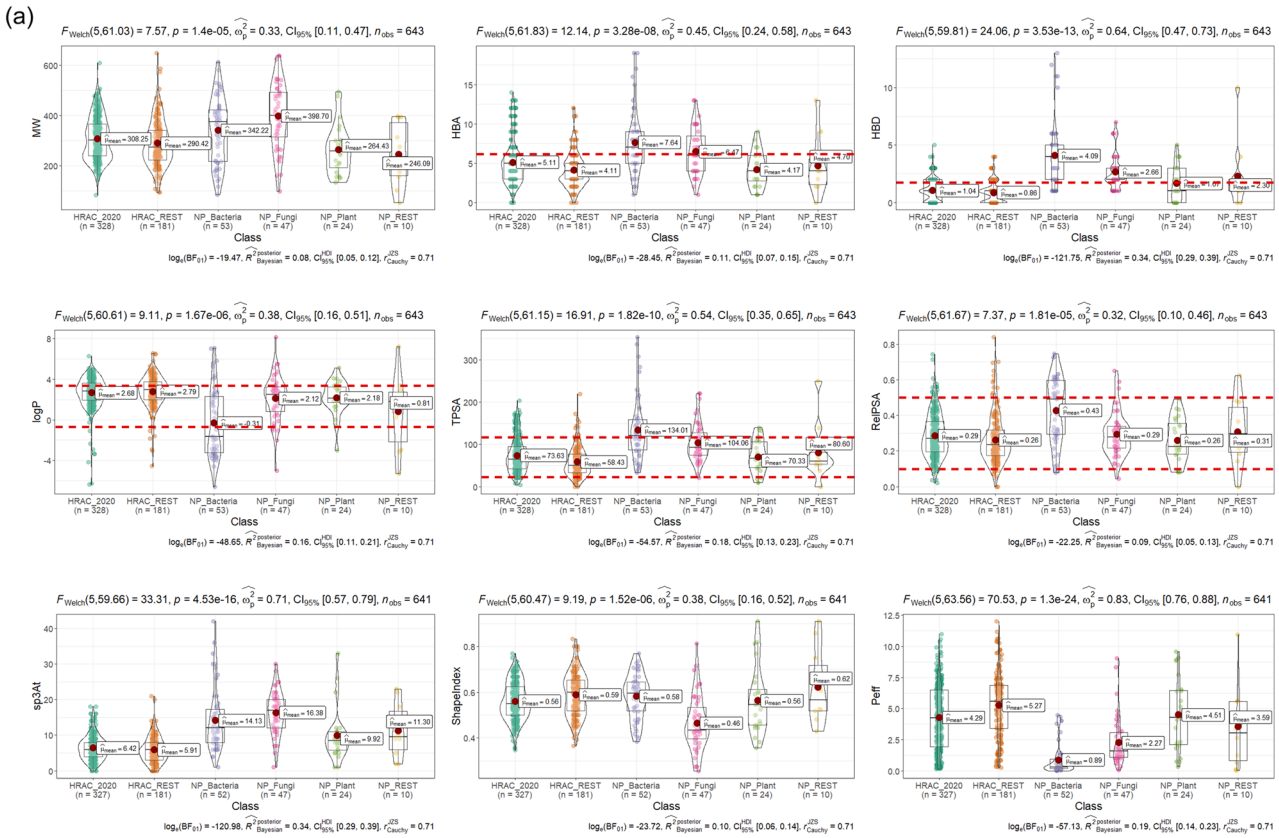
Herbicide-like properties. For synthetic herbicides distributions of physicochemical and simple molecular properties have already been reported[28–34]. These simple molecular properties and physicochemical features largely influence the mass distribution of herbicides across plants and plant cell compartments and hence may be applied for characterizing herbicide-likeness of compounds[41,42]. The phytotoxic effect of a herbicide largely depends upon its translocation through plants to its site of action analogously as pharmacological effects of drugs are considerably influenced by their absorption and distribution throughout the human body[44]. Drug-likeness filters are commonly used in early drug discovery process to eliminate compounds out of the sets aimed for biological activity screening. In analogous way, herbicide-likeness features may be used as a first-pass filter for eliminating compounds from the analyzed compound data sets and libraries which are less probable to show biological activity in weeds. The proposed herbicide-like features obtained by analyzing the extended set of 509 synthetic organic herbicides with MW less than 650 Da, are listed in Table 4. They were applied on the data set of NPs.

Phytotoxic molecules produced by plants are found to be the most similar to the synthetic herbicides both in structural and physicochemical spaces (Fig. 5a). In difference, fungal and particularly bacterial NPs vary in the physicochemical space from the rest of studied compounds (Figures S6 and S7). They are richer in H-bond interacting atoms similarly as many other types of NPs[45]. The bacterial phytotoxic compounds are relatively more polar, hydrophilic and charged molecular species. The fungal products have more sp3-hybridized atoms and are also more spherical compounds what may imply their different translocation capacity and features. The most of bacterial and fungal phytotoxic compounds were estimated to have lower permeation rates (Peff (cm/s x 10^4) in Fig. 5a) across lipophilic membranes as compared with the plant NPs and synthetic organic herbicides. The lower membrane permeability is generally associated with compounds having lower lipophilicity and larger number of H-bond interacting atoms, particularly larger number of HBD atoms and may also be caused by the membrane retention[42,45]. However, the uptake and translocation of a small dissolved phytotoxic NPs can be determined not only by their passive permeation across membranes, but also by the active translocation by transport proteins[8]. The translocation propensity of bacterial and some fungal compounds can also be affected by the presence of ionized carboxyl group(s)[46].

In silico screening platform. The comprehensive modelling carried out on the set of synthetic herbicides and application of the models and herbicide-likeness filter on phytotoxic NPs encouraged us to propose the in silico screening platform which can be applied on any set /library of compounds for characterization of their herbicide-likeness and possibly phytotoxic ways of action (Fig. 5b). Considering the data set of 131 NPs, 81 molecules satisfy 4 or more herbicide-likeness criteria (Table 4), and 35 of them lay within the AD of the RF weed selectivity model (Fig. 4a), while all are outside the AD of the MoA and other models built in terms of specific structural fp keys. This result suggests further experimental studies that might reveal new MoAs for these compounds, which in turn may lead to new herbicides, potentially also adding more robustness to the current rotational strategies for minimizing weed resistance, based on available classes of herbicides.

Conclusions

There are two main ways to minimize weed resistance, the application of herbicides according to the rotation strategy which is well-accepted by the end users and to discover and develop novel phytotoxic compounds. The developed predictive classifiers to a large extent confirm MoAs assignation for the HRAC herbicides based on structural similarity and additionally enables MoA assignment for herbicides, mainly obsolete due to their side effects and thus lying outside the HRAC list. However, the performed modelling points out limitations of using only structural similarity for MoA classification and further for selection of herbicides for rotation strategy. The conducted ML modelling of weed selectivity reveals that it is largely determined by simple molecular and physicochemical features which also influence uptake and distribution of small molecules through plants. Since



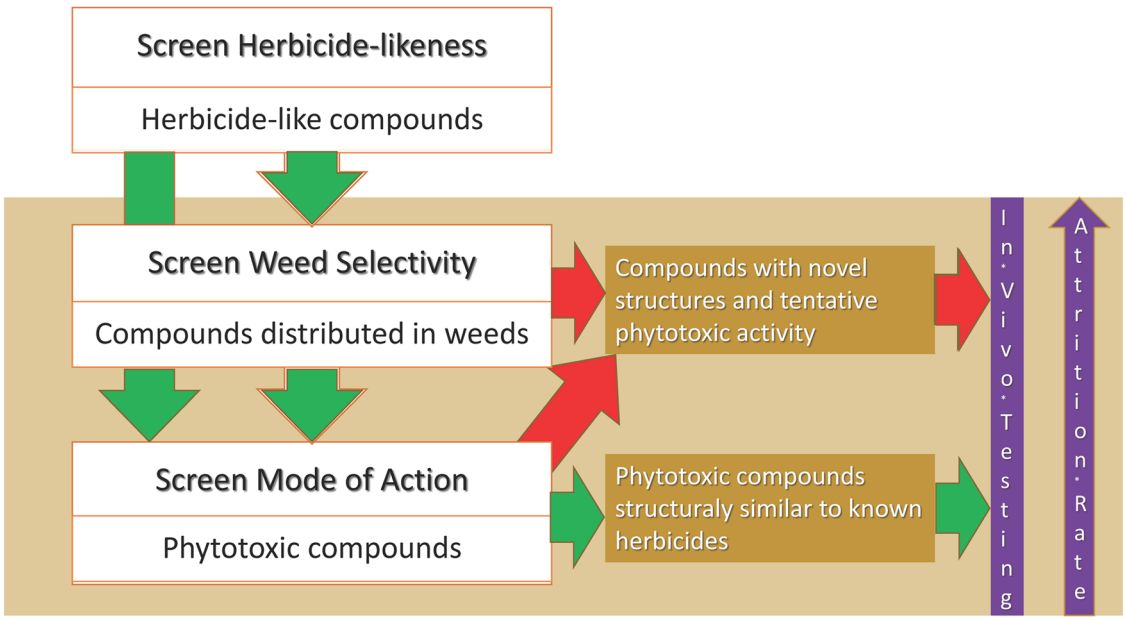


Figure 5. (**a**) The comparison of six subgroups of phytotoxic molecules according to selected molecular properties. Herbicide-like boundaries (Table 4) are denoted by red dash lines. (**b**) Virtual screening platform proposed for preselecting phytotoxic compounds. Its proof-of-concept should be carried out by in vivo testing.

similarity in uptake and translocation properties of herbicides may lead to the similar mechanisms of induction of weed resistance, the weed selectivity categorization is suggested as an additional rotational criterion.

The additional output of the study is the proposal of in silico stepwise screening platform for detecting herbicide-like molecules with selectivity for weed types and possibly with pre-specified mode of action, from any chemical library or database (Fig. 5b). Application of the platform to the data set of pyhtotoxic natural products reveals that they lie outside the space of synthetic herbicides considering not only molecular structure, but also

Descriptor	Range	% of 509 synthetic herbicides	% of 131 NPs
HBD (OH/NH)	≤ 2	95	51.9
HBA (O/N)	≤ 6; ≤ 7	66.7; 80.0	58.8; 65.6
clogP[a]	0.5 < clogP ≤ 3.5;0.5 < clogP ≤ 4.5	66.7; 80.0	47.3; 53.4
TPSA	20 Å^2 < TPSA ≤ 120 Å^2	80	63.4
Relative PSA	0.1 < RelPSA ≤ 0.5	80	81.7
Net charge[b]	≤ 0	95	65.6

Table 4. Herbicide-like chemical space defined in terms of common molecular descriptors (Fig. 5). [a]Regardless logP values were calculated by ADMET Predictor or DataWarrior. [b]More than 95% of synthetic organic herbicides are either neutral molecules (around 2/3) or anions (30%) (Figure S7).

physicochemical properties guiding weed selectivity. Therefore, natural products might represent worthy source of novel phytotoxic scaffolds with new/different modes of action, thus contributing to more effective and weed-resistance robust use of herbicides.

The proposed herbicide-likeness and screening cascade can be used for prioritization of the in vivo experiments.

Data availability

The R scripts and data sets for model performance are available at GitHub (https://github.com/mlkr-rbi/Herbicide-Classification.git). Data sets analyzed and/or generated during the current study are available in Supplementary information.

Received: 9 October 2020; Accepted: 17 May 2021
Published online: 01 June 2021

References

1. Lushchak, V. I., Matviishyn, T. M., Husak, V. V., Storey, J. M. & Storey, K. B. Pesticide toxicity: a mechanistic approach. *EXCLI. J.* **17**, 1101–1136 (2018).
2. Retzinger, E. J. & Mallory-Smith, C. Classification of herbicides by site of action for weed resistance management strategies. *Weed Technol.* **11**, 384–393 (1997).
3. Meene, H. & Kocher H. HRAC classification of herbicides and resistance development. In *Modern Crop Protection Compounds*. Vol. 1, 2nd (eds. Kramer, W., Schrimer, U., Jeschke, P., Witschel, M.) 5–28 (Wiley-VCH: Weinheim, Germany, 2012).
4. http://wssa.net/ (2019)
5. https://www.hracglobal.com/ (2020)
6. Forouzesh, A., Zand, E., Soufizadeh, S. & Forushani Samadi, S. Classification of herbicides according to chemical family for weed resistance management strategies—an update. *Weed Res.* **55**, 334–358 (2015).
7. Zhou, Q., Liu, W., Zhang, Y. & Liu, K. K. Action mechanisms of acetolactate synthase-inhibiting herbicides. *Pestic. Biochem. Physiol.* **89**, 89–96 (2007).
8. Menendez, J., Rojano-Delgado, M. A. & De Prado R., Differences in herbicide uptake, translocation, and distribution as sources of herbicide resistance in weeds. In *Retention, Uptake, and Translocation of Agrochemicals in Plants* (eds. Myung, K., Norbert M., Satchivi, N. M., Kingston C. K.) 141–157 (ACS Symposium Series. **1171**, 2014).
9. Lamberth, C. Agrochemical lead optimization by scaffold hopping. *Pest Manag. Sci.* **74**, 282–292 (2018).
10. Duke, S. O. Why have no new herbicide modes of action appeared in recent years?. *Pest Manag. Sci.* **68**, 505–512 (2012).
11. Lewis K.A., Green A., Tzilivakis J., Warner D. The pesticide properties database (PPDB) developed by the Agriculture & Environment Research Unit (AERU). University of Hertfordshire; 2015, pp 2006e2015
12. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
13. Gong, J. *et al.* PTID: an integrated web resource and computational tool for agrochemical discovery. *Bioinformatics* **29**, 292–294 (2013).
14. Gandy, M. N., Corral, M. G., Mylne, J. S. & Stubbs, K. A. An interactive database to explore herbicide physicochemical properties. *Org. Biomol. Chem.* **13**, 5586–5590 (2015).
15. Dayan, F. E. & Duke, S. O. Natural Compounds as next-generation herbicides. *Plant Physiol.* **166**, 1090–1105 (2014).
16. Cantrell, C. L., Dayan, F. E. & Duke, S. O. Natural products as sources for new pesticides. *J. Nat. Prod.* **75**, 1231–1242 (2012).
17. Duke, S. O. & Dayan, F. E. Modes of action of microbially-produced phytotoxins. *Toxins (Basel).* **3**, 1038–1064 (2011).
18. Gerwick, C. B. & Sparks, T. C. Natural products for pest control: an analysis of their role, value and future. *Pest Manag. Sci.* **70**, 1169–1185 (2014).
19. Dayan, F. E., Owens, D. K. & Duke, S. O. Rationale for a natural products approach to herbicide discovery. *Pest Manag. Sci.* **68**, 519–528 (2012).
20. Seiber, J. N., Coats, J., Duke, S. O. & Gross, A. D. Biopesticides: State of the art and future opportunities. *J. Agric. Food Chem.* **62**, 11613–11619 (2014).
21. Peng, J. *et al.* Marine natural products as prototype agrochemical agents. *J. Agric. Food Chem.* **51**, 2246–2252 (2003).
22. Duke, S. O., Dayan, F. E., Romagni, J. G. & Rimando, A. M. Natural products as sources of herbicides: Current status and future trends. *Weed Res.* **40**, 99–111 (2000).
23. Dayan, F. E., Cantrell, C. L. & Duke, S. O. Natural products in crop protection. *Bioorgan. Med. Chem.* **17**, 4022–4034 (2009).
24. Sparks, T. C., Hahn, D. R. & Garizi, N. V. Natural products, their derivatives, mimics and synthetic equivalents: Role in agrochemical discovery. *Pest Manag. Sci.* **73**, 700–715 (2016).
25. Guha, R. Chemical informatics functionality in R. *J. Stat. Softw.* **18**, 1–18 (2007).
26. Sander, T., Freyss, J., von Korff, M. & Rufener, C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **55**, 460–473 (2015).
27. Lowless, M. S., Waldman, M., Franczkiewicz, R. & Clark, R. D. Using chemoinformatics in drug discovery. In *New Approaches to Drug Discovery, Handbook of Experimental pharmacology* (eds. Nielsch, U., Fuhrmann, U., Jaroch, S.) 139–170 (Springer International Publishing AG: Switzerland **232**, 2016).

28. Tice, C. M. Selecting the right compounds for screening: does Lipinski's Rule of 5 for pharmaceuticals apply to agrochemicals?. *Pest Manag. Sci.* **57**, 3–16 (2001).
29. Tice, C. M. Selecting the right compounds for screening: use of surface-area parameters. *Pest Manag. Sci.* **58**, 219–233 (2002).
30. Clarke, D. E. & Delaney, J. S. Physical and molecular properties of agrochemicals: An analysis of screen inputs, hits, leads, and products. *Chimia* **57**, 731–734 (2003).
31. Avram, S. *et al.* Quantitative estimation of pesticide-likeness for agrochemical discovery. *J. Cheminform.* **6**, 42 (2014).
32. Rao, H. *et al.* Physicochemical profiles of the marketed agrochemicals and clues for agrochemical lead discovery and screening library development. *Mol. Inform.* **34**, 331–338 (2015).
33. Zhang, Y. *et al.* Physicochemical property guidelines for modern agrochemicals. *Pest Manag. Sci.* **74**, 1979–1991 (2018).
34. Pehar, V., Oršolić, D. & Stepanić, V. Drug-likeness, herbicide-likeness and toxicity of herbicidal compounds – in silico analysis. In *Proceedings: 17th Ružička Days Today Science – Tomorrow Industry*, (eds. Tomas, S., Ačkar Đ.) 112–123 (Josip Juraj Strossmayer University of Osijek, Faculty of Food Technology Osijek and Croatian Society of Chemical Engineers (CSCE), Osijek, 2019).
35. Benavoli, A., Corani, G., Demšar, J. & Zaffalon, M. Time for a change: a tutorial for comparing multiple classiers through Bayesian analysis. *J. Mach. Learn. Res.* **18**, 1–36 (2017).
36. Klingspohn, W., Mathea, M., Ter Laak, A., Heinrich, N. & Baumann, K. Efficiency of different measures for defining the applicability domain of classification models. *J. Cheminformatics* **9**, 44. https://doi.org/10.1186/s13321-017-0230-2 (2017).
37. RStudio Team (2015). RStudio: Integrated development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/
38. Pearlman, R. S. & Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **39**, 28–35 (1999).
39. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). Association for Computing Machinery, New York, NY, USA, 694–699, 2002.
40. Kraehmer, H. *et al.* Herbicides as weed control agents: State of the art: II recent achievements. *Plant Physiol.* **166**, 1132–1148 (2014).
41. Hofstetter, S., Beck, A., Trapp, S. & Buchholz, A. How to design for a tailored subcellular distribution of systemic agrochemicals in plant tissues. *J. Agric. Food Chem.* **66**, 8687–8697 (2018).
42. Trapp, S. Plant uptake and transport models for neutral and ionic chemicals. *Environ. Sci. Pollut. Res. Int.* **11**, 33–39 (2004).
43. Partington, J. *Fundamental Principles: The properties of gases.* An Advanced Treatise on Physical Chemistry, Vol. 1, Fundamental Principle: The Properties of Gases. Longmans Green: New York, 1949.
44. Ursu, O., Rayan, A., Goldblum, A. & Oprea, T. I. Understanding drug-likeness. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 760–781 (2011).
45. Stepanić, V. *et al.* Physicochemical profile of macrolides and their comparison with small molecules. *Eur. J. Med. Chem.* **47**, 462–472. https://doi.org/10.1016/j.ejmech.2011.11.016 (2012).
46. Briggs, G. G., Rigitano, R. L. O. & Bromilow, R. H. Physico-chemical factors affecting the uptake by roots and translocation to shoots of weak acids in barley. *Pestic. Sci.* **19**, 101–112 (1987).

Acknowledgements
The authors like to thank Croatian Government and the European Union (European Regional Development Fund—the Competitiveness and Cohesion Operational Program), for funding this study through the project Bioprospecting of the Adriatic Sea (KK.01.1.1.01.0002), granted to The Scientific Centre of Excellence for Marine Bioprospecting—BioProCro.

Author contributions
V.P. formed the data sets and calculated descriptors. D.O., V. S. and T. Š. designed the experiments. D.O. performed all modelling. V. S. and V. P. analyzed and interpreted the data. T. Š. reviewed the manuscript. V.S. conceived the presented idea and wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-90690-w.

Correspondence and requests for materials should be addressed to V.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 PAPER 2: Crowdsourced mapping of unexplored target space of kinase inhibitors

Cichońska, A., Ravikumar, B., Allaway, R.J., Wan, F., Park, S., Isayev, O., Li, S., Mason, M., Lamb, A., Tanoli, Z., Jeon, M., Kim, S., Popova, M., Capuzzi, S., Zeng, J., Dang, K., Koytiger, G., Kang, J., Wells, C.I., Willson, T.M., The IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium, User oselot, Tan, M., Team N121, Huang, C.-H., Shih, E.S.C., Chen, T.-M., Wu, C.-H., Fang, W.-Q., Chen, J.-Y., Hwang, M.-J., Team Let_Data_Talk, Wang, X., Ben Guebila, M., Shamsaei, B., Singh, S., User thinng, Nguyen, T., Team KKT, Karimi, M., Wu, D., Wang, Z., Shen, Y., Team Boun, Öztürk, H., Ozkirimli, E., Özgür, A., Team KinaseHunter, Lim, H., Xie, L., Team AmsterdamUMC-KU-team, Kanev, G.K., Kooistra, A.J., Westerman, B.A., Team DruginaseLearning, Terzopoulos, P., Ntagiantas, K., Fotis, C., Alexopoulos, L., Team KERMIT-LAB - Ghent University, Boeckaerts, D., Stock, M., De Baets, B., Briers, Y., Team QED, Luo, Y., Hu, H., Peng, J., Team METU_EMBLEBI_CROssBAR, Dogan, T., Rifaioglu, A.S., Atas, H., Atalay, R.C., Atalay, V., Martin, M.J., Team DMIS_DK, Jeon, M., Lee, J., Yun, S., Kim, B., Chang, B., Team AI Winter is Coming, Team hulab, Turu, G., Misák, Á., Szalai, B., Hunyady, L., Team ML-Med, Lienhard, M., Prasse, P., Bachmann, I., Ganzlin, J., Barel, G., Herwig, R., Team Prospectors, **Oršolić, D.**, Lučić, B., Stepanić, V., Šmuc, T., Challenge organizers, Oprea, T.I., Schlessinger, A., Drewry, D.H., Stolovitzky, G., Wennerberg, K., Guinney, J., Aittokallio, T., 2021. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun 12, 3307. https://doi.org/10.1038/s41467-021-23165-1

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23165-1 OPEN

Crowdsourced mapping of unexplored target space of kinase inhibitors

Anna Cichońska[1,2,3,51], Balaguru Ravikumar[1,51], Robert J. Allaway[4,51], Fangping Wan[5], Sungjoon Park[6], Olexandr Isayev[7], Shuya Li[5], Michael Mason[4], Andrew Lamb[4], Ziaurrehman Tanoli[1], Minji Jeon[6], Sunkyu Kim[6], Mariya Popova[7], Stephen Capuzzi[8], Jianyang Zeng[5], Kristen Dang[4], Gregory Koytiger[9], Jaewoo Kang[6], Carrow I. Wells[10], Timothy M. Willson[10], The IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium*, Tudor I. Oprea[11], Avner Schlessinger[12], David H. Drewry[10], Gustavo Stolovitzky[13], Krister Wennerberg[14,52]✉, Justin Guinney[4,52]✉ & Tero Aittokallio[1,2,15,16,17,52]✉

Despite decades of intensive search for compounds that modulate the activity of particular protein targets, a large proportion of the human kinome remains as yet undrugged. Effective approaches are therefore required to map the massive space of unexplored compound–kinase interactions for novel and potent activities. Here, we carry out a crowdsourced benchmarking of predictive algorithms for kinase inhibitor potencies across multiple kinase families tested on unpublished bioactivity data. We find the top-performing predictions are based on various models, including kernel learning, gradient boosting and deep learning, and their ensemble leads to a predictive accuracy exceeding that of single-dose kinase activity assays. We design experiments based on the model predictions and identify unexpected activities even for under-studied kinases, thereby accelerating experimental mapping efforts. The open-source prediction algorithms together with the bioactivities between 95 compounds and 295 kinases provide a resource for benchmarking prediction algorithms and for extending the druggable kinome.

A full list of author affiliations appears at the end of the paper.

Only 11% of the human proteome can be currently targeted by small molecules or drugs, whereas one in three proteins remains understudied[1]. Despite many years of target-based drug discovery, chemical agents inhibiting single protein targets are still rare[2]. Most approved drugs have multiple targets, suggesting their therapeutic efficacy as well as adverse side-effects originate from polypharmacological effects[3]. Systematic mapping of the target binding profiles is therefore critical not only to explore the therapeutic potential of promiscuous agents, but also to better predict and manage possible adverse effects within early stages of drug development process to mitigate future risks and costs. Comprehensive understanding of the polypharmacological effects of approved drugs could also uncover novel off-target potencies to extend their therapeutic application area via off-label use or repurposing[4]. However, due to the massive size of the chemical universe, an exhaustive experimental mapping of compound-target activities is infeasible, even with automated high-throughput profiling assays.

To accelerate the mapping efforts, we hosted the IDG-DREAM Drug-Kinase Binding Prediction Challenge, a crowd-sourced competition that evaluated the power of machine learning (ML) models as a systematic and cost-effective means for predicting yet unexplored compound-target potencies. The Challenge focused on predicting quantitative target activities of kinase inhibitors, since kinases are implicated in a wide range of diseases, such as cardiovascular disorders and cancers. However, protein kinase domains are inherently similar in their structure and sequence, and most kinase inhibitors bind to conserved ATP-binding pockets, leading to extensive target promiscuity and polypharmacological effects[5–8]. Such multi-target activities require methods for effective target deconvolution, including multi-target ML approaches, that leverage the information extracted from similar kinases and compounds to predict the activity of so far unexplored compound-kinase interactions[9,10].

The specific questions this Challenge sought to address were: (i) What are the best computational modeling approaches for predicting quantitative compound-target activity profiles?; (ii) What are the best molecular, chemical, and protein descriptors for maximal prediction accuracy?; and (iii) What are the most informative bioactivity assays for dose-response bioactivity prediction? Models submitted to the Challenge were quantitatively evaluated using bioactivity data contributed by—and in partnership with—the Illuminating the Druggable Genome (IDG) consortium (https://druggablegenome.net/). IDG is a NIH Common Fund program aimed at improving our understanding of understudied proteins within three drug-targeted protein families: G-protein coupled receptors, ion channels, and protein kinases[1]. Specifically, it seeks to improve the druggability of dark kinases by kinome-wide profiling small-molecule agents, with the goal of extending the activity information for the understudied human kinome.

Here, we describe the benchmarking results of the Challenge, as well as the post-Challenge analysis of top-performing models to identify so far unexplored kinase inhibitor activities. The benchmarking results include a total of 268 predictions from 212 active Challenge participants, covering a wide range of ML approaches, including linear regularized regression, deep and kernel learning algorithms, and gradient boosting decision trees.

Results

Challenge implementation and training datasets. To develop regression models for prediction of quantitative bioactivities, participants were encouraged to utilize a wide variety of bioactivity data for model training and cross-validation through open databases such as ChEMBL[11], BindingDB[12], and IDG Pharos[13] (Fig. 1). For training data collection, integration, management and harmonization, the Challenge made use of an open-data platform, DrugTargetCommons (DTC)[14]. DTC is a community platform that provides a comprehensive and standardized interface to retrieve compound-target profiles and related information to support predictive activity modeling (Supplementary Fig. 1). The Challenge infrastructure was built on the Synapse collaborative science platform[15], which supported receiving, validating and scoring of the teams' predictions as well as long-term management of the test bioactivity data and submitted Challenge models as a benchmarking resource (Fig. 1).

Challenge test datasets of kinase inhibitors. The blinded evaluation of the model predictions was based on unpublished kinase activity data generated by the IDG Consortium, with a focus to investigate especially understudied yet readily screenable human kinome, so-called dark kinases[13], and those lacking small-molecule activities in ChEMBL[11], but with a robust assay readily available through commercial vendors[16]. The Challenge was conducted over a series of rounds based on availability of test datasets (Supplementary Fig. 3). Round 1 test dataset was generated based on the two-step screening approach[6,7,16], where the quantitative dose-response measurement of the dissociation constant (K_d) activities was carried out across 430 interactions between 70 inhibitors and 199 kinases that had inhibition >80% in the single-dose kinome activity scan (see Methods). An additional set of completely new K_d data was generated for Round 2, consisting of 394 multi-dose assays between 25 inhibitors and 207 kinases with single-dose inhibition >80%. Together, these 824 K_d assays spanned a total of 95 compounds and 295 kinases, covering 57% of the human kinome (Fig. 2a, b). The Challenge test data consisted both of promiscuous compounds targeting multiple kinases at low concentrations, compounds with narrow target profiles, as well as compounds with no potent targets among the tested kinases (Supplementary Fig. 2).

Round 1 enabled the teams to carry out the initial testing of various model classes and data resources, whereas Round 2, implemented 6 months later once the new K_d data became available, was used to score the final prediction models and to select the top-performing teams. None of the K_d values were available in the public domain, and the Round 1 test data remained blinded in Round 2. Round 1 and 2 test datasets had very similar pK_d distributions (Fig. 2c), which provided comparable binding affinity outcome data to monitor the improvements made by the teams between the two rounds. The tested kinase inhibitors in the two test sets were mutually exclusive between the rounds (Fig. 2a), with Round 2 including less selective inhibitors with broader target profiles (Fig. 2d), and therefore fewer inactive compound-kinase pairs ($pK_d = 5$). Round 1 and 2 kinase targets were partly overlapping, and covered all the major kinase families and groups (Fig. 2b, e). Taken together, these two test datasets provided a standardized and sufficiently large quantitative bioactivity resource to evaluate the accuracy of predicting on- and off-target kinase activities, using pharmacologically realistic and computationally rather challenging compound and target spaces of multi-targeting kinase inhibitors.

Predictive performance of the Challenge models. The competition phase challenged the participants to predict blinded K_d profiles between 95 inhibitors and 295 kinases. Since the goal of this Challenge was to encourage regression model development that would exceed state-of-the-art, we selected as baseline model a recently published and experimentally validated kernel

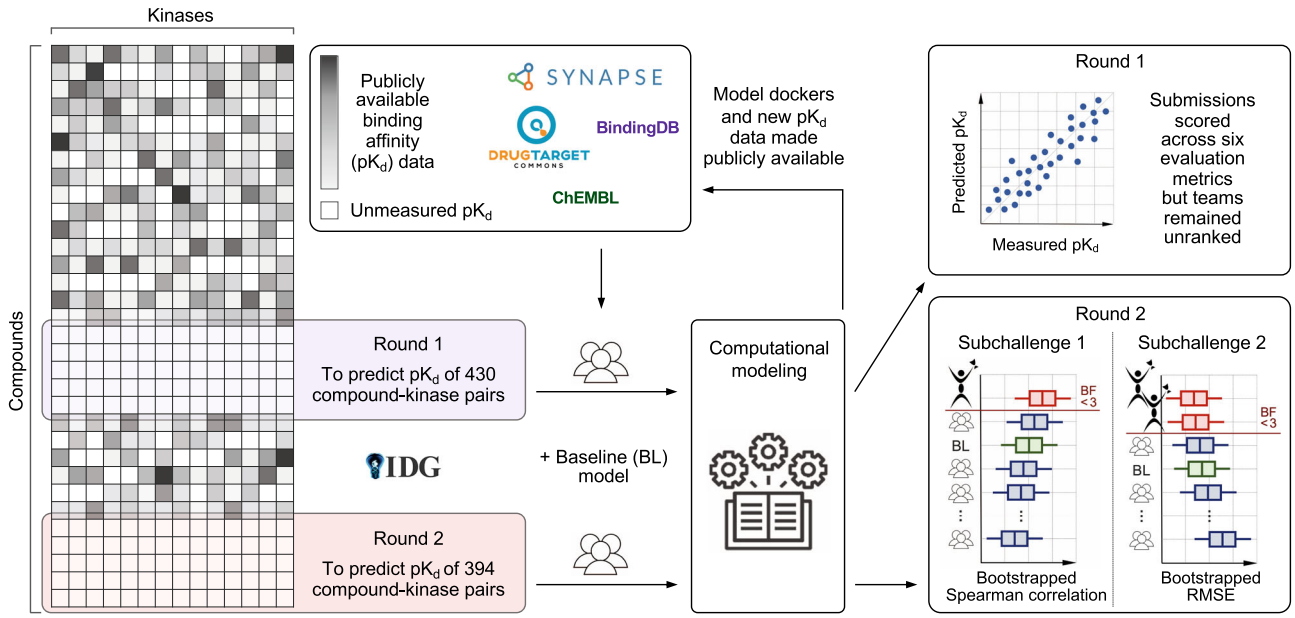


Fig. 1 Implementation of the IDG-DREAM Drug-Kinase Binding prediction Challenge. The participants had access to publicly available large-scale target profiling training data, and the quantitative predictions from regression models were then validated in two unpublished and blinded test datasets profiled by the Illuminating the Druggable Genome (IDG) program (Round 1 and Round 2 datasets). Heatmap on the left is for illustrative purposes only (see Supplementary Fig. 2 for the actual test data matrices, and Supplementary Fig. 3 for the Challenge timeline). All the models, new bioactivity data, and benchmarking infrastructure are openly available to support future target prediction and benchmarking studies. BF Bayes factor; RMSE Root Mean Square Error.

regression approach for compound-kinase activity prediction[17]. The performance of the Challenge model predictions improved from Round 1 to Round 2 submissions as measured by Spearman correlation (two-sample Wilcoxon test, $P < 0.005$; Fig. 3a) and Root Mean Square Error (RMSE, $P < 10^{-6}$; Fig. 3c). Comparison against the baseline model indicated that the Round 2 dataset was marginally easier to predict (Supplementary Fig. 4), partly due to a smaller proportion of inactive pairs in Round 2 ($pK_d = 5$, Fig. 2c). To take into account this shift, we compared the submissions against a set of random predictions. Using Spearman correlation, we observed that 48% of the submissions were better than random in Round 1, compared to 61% in Round 2 (Fig. 3b). Using RMSE, 71% of the submissions in Round 1 were better than random, compared to 76% in Round 2 (Fig. 3d).

The 20 teams that participated in both rounds improved their K_d predictions ($P < 0.05$ and $P < 0.001$ for Spearman correlation and RMSE, respectively, paired Wilcoxon signed-rank test), but when comparing against the baseline model, the overall improvements became insignificant ($P > 0.05$). However, there were individual teams (like Zahraa Sobhy) that were able to improve their predictions considerably between the two rounds. The practical upper bound of the model predictions was defined based on experimental replicates of K_d measurements (Fig. 3b, d). The predictive accuracy of the top-performing models in Round 2 was relatively high based on both of the winning metrics, Spearman correlation for ranked pairs predictions and RMSE for quantitative activity predictions; these metrics showed less-correlated performance over the less-accurate models in Round 2 (Fig. 3f). The tie-breaking metric, averaged area under the receiver operating characteristic (ROC) curve, provided complementary information on prediction accuracy when compared to RMSE but not to Spearman correlation (Supplementary Fig. 5). Overall, the models based on deep learning algorithms did not perform better than other learning algorithms submitted in Round 2 (Fig. 3f).

Selection of the top-performing Challenge models. The top-performing models were selected in Round 2 based on 394 pK_d predictions between 25 compounds and 207 kinases. Only those participants who submitted their Dockerized models, method write-ups, and method surveys were qualified to win the two subchallenges (see Supplementary Table 1 for all submissions in Round 2 from the participants who submitted method surveys, together with their model features and training data). To select the top performers, we conducted a bootstrap analysis of each participant's best submission, and then calculated a Bayes factor (K) relative to the bootstrapped overall best submission for each winning metric (Supplementary Fig. 6). Considering Spearman correlation, the top performer was team Q.E.D ($K < 3$; Fig. 4a). For the RMSE metric, the top-performing teams were AI Winter is Coming (AIWIC) and DMIS_DK ($K < 3$), with AIWIC having a marginally better tie-breaking metric (average AUC of 0.773; Fig. 4b). Only two non-qualifying participants (Gregory Koytiger and Olivier Labayle) showed comparable performance. Overall, these five teams performed the best across the 54 teams and the 99 total submissions in Round 2 (Supplementary Fig. 7).

Notably, the top-performing models were based on rather different ML approaches, including deep learning, graph convolutional networks, gradient boosting decision trees, kernel learning and regularized regression (Table 1). To study whether combining predictions from the multiple ML approaches could further improve prediction accuracy, we constructed an ensemble model by simple mean aggregation of an increasing number of top-performing models in Round 2. A combination of the four best performing models resulted in the peak Spearman correlation (Fig. 4c), demonstrating a complementary value of these models and their features. After adding more models, the ensemble prediction accuracy decreased rapidly in terms of Spearman correlation and RMSE (Fig. 4d). Combinations of four random models resulted in a decreased performance compared to the top-model ensemble (empirical $P = 0.0$, Supplementary Fig. 8). This suggests that combination of best performing

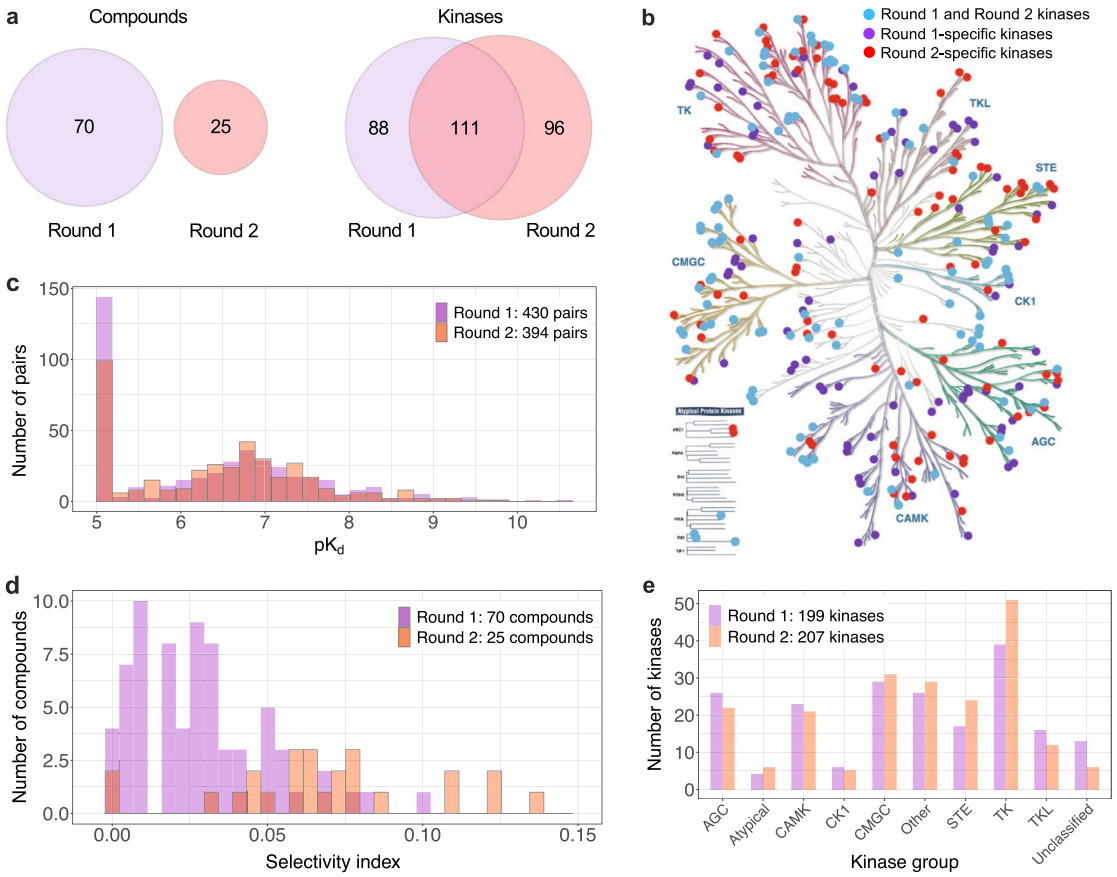


Fig. 2 Challenge test datasets. a The overlap between Round 1 and Round 2 kinase inhibitors and kinase targets, and their distributions in the kinome tree (**b**), and across various kinase groups (**e**). **c** The quantitative dissociation constant (K_d) of compound-kinase activities was measured in dose-response assays (see Methods), presented in the logarithmic scale as $pK_d = -\log_{10}(K_d)$. The higher the pK_d value, the higher the inhibitory ability of a compound against a protein kinase (Supplementary Data 1 includes the compounds and kinases in Round 1 and Round 2 test datasets). The frequent values of $pK_d = 5$ originate from inactive pairs (maximum tested concentration of 10 μM in the multi-dose activity profiling). **d** The selectivity index of kinase inhibitors was calculated based on the single-dose activity assay (at 1 μM concentration) across the full compound-kinase matrices before the Challenge. The kinome tree figure was created with KinMap, reproduced courtesy of Cell Signaling Technology, Inc. Source data are provided as a Source Data file[54].

approaches using an ensemble model leads to accurate and robust predictions of kinase inhibitor potencies across multiple kinase families.

Analysis of the Q.E.D and ensemble models. To better understand how the amount and diversity of training data contribute to the Q.E.D model accuracy, we removed training bioactivity data based on compound structural similarity (Fig. 5a). Surprisingly, we found that the structural similarity of the training and test compounds was relatively unimportant in predicting the activity of the test compounds, indicating that the Q.E.D model made use of other, structurally diverse set of compounds in the test compound activity predictions (Fig. 5a). At the lower similarity cutoffs (Tanimoto similarity <0.7), the model performance decreased substantially, likely due to an increased disparity in chemistry between the test and training compounds, as well as an overall decrease in the training dataset size. We also performed a similar experiment to test the importance of high- and low-potency compounds on model accuracy (Fig. 5b), by removal of training data compounds with high pK_d, low pK_d, or both. As anticipated, we observed that removal of high pK_d compound-kinase pairs (pK_d values larger than 8) reduced performance of the model. This is likely a consequence of both loss of the overall number of training data and loss of rare extreme activities. However, removal of the small number of compound-kinase pairs with the most extreme pK_d values (training on pK_d values between 4 and 10) had no effect on accuracy.

We further systematically investigated the relative contributions of various chemical and protein descriptors to the predictive performance of the Q.E.D model. These results showed that whilst several different chemical fingerprints performed similarly well (Supplementary Fig. 10), the choice of protein descriptor had a more notable impact on the model prediction accuracy (Fig. 6a). Especially the protein kernel based on amino acid subsequences of ATP-binding pockets resulted in a poor performance (adjusted $P < 10^{-10}$, Pearson and Filon test), compared to the full amino acid sequences, which can at least partly be explained by the missing subsequences for several kinases that reduced the training dataset size and also led to some activity predictions of zero (Supplementary Fig. 11; we note that this is also the case for kinase domain sequences). We also re-trained the Q.E.D model with different combinations of training bioactivity data types to investigate which types contributed most to the high prediction accuracy. We observed that while K_d alone or in combination with other bioactivity data types, especially with K_i, systematically resulted in rather accurate K_d predictions, the other types led to significantly worse prediction performances (Fig. 6b). Especially the rather abundant EC_{50} and IC_{50} bioactivities alone led to poor pK_d prediction accuracy (Supplementary Fig. 12). This result can be explained by the fact that, in contrast to K_d affinity assay, EC_{50}

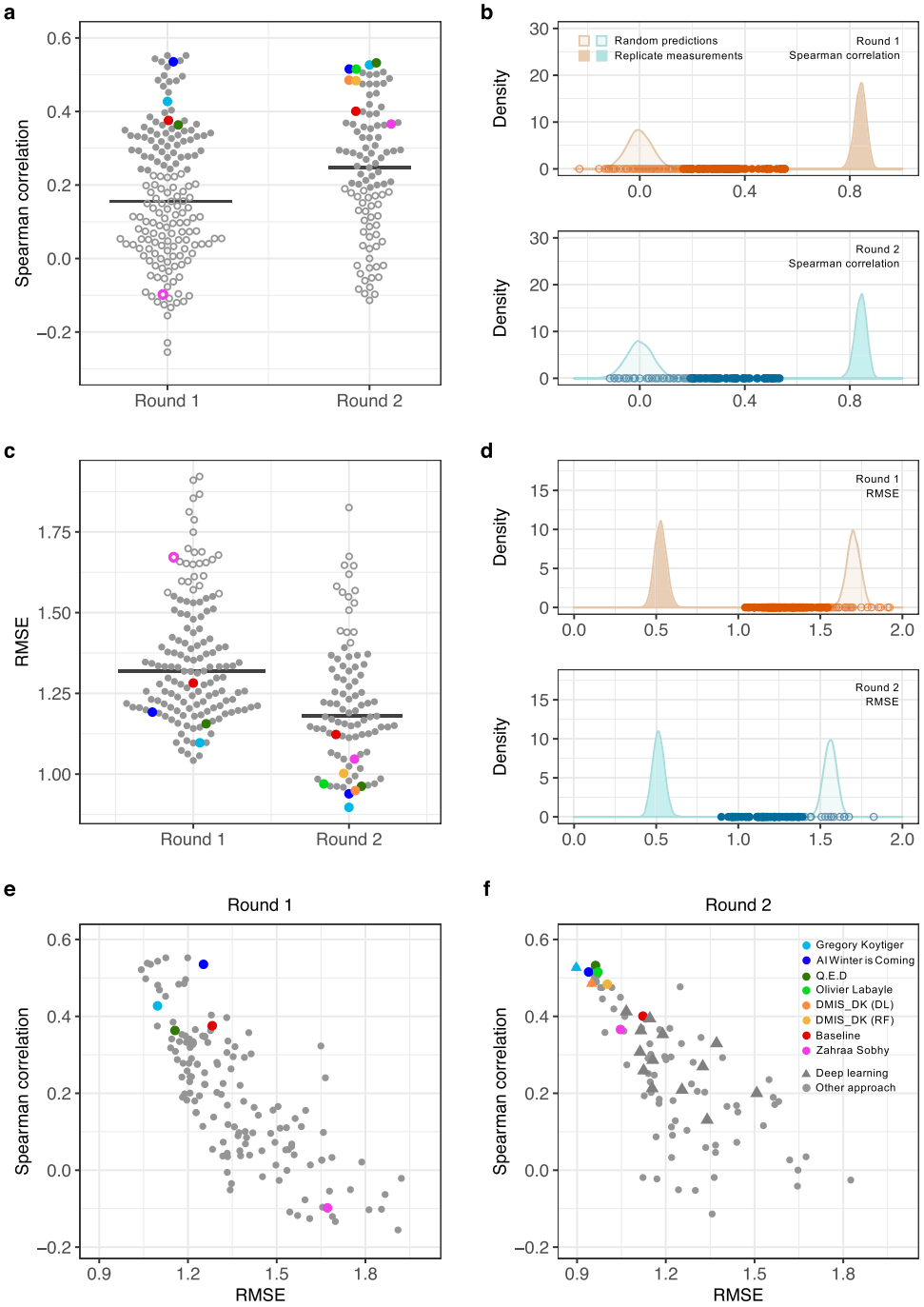


Fig. 3 Overall performance of the Challenge submissions. a, **c** Performance of the submissions in terms of the two winning metrics in Round 1 ($n = 169$ submissions) and Round 2 ($n = 99$ submissions). The horizontal lines indicate median correlation and the colors mark the baseline model and the top-performing participants in Round 2 (see the color legend of **f**). The empty circles mark the submissions that did not differ from random predictions (the open pink circle indicates the Round 1 submission of Zahraa Sobhy as an example). The baseline model[17] remained the same in both of the rounds. **b**, **d** Distributions of the random predictions (based on 10,000 permuted pK_d values) and replicate distributions (based on 10,000 subsamples with replacement of overlapping pK_d pairs between two large-scale target activity profiling studies[5,6]) in Round 1 (top panel) and Round 2 (bottom). The points correspond to the individual submissions. **e**, **f** Relationship of the two winning metrics across the submissions in Round 1 and Round 2. The triangle shape indicates submissions based on deep learning (DL) in Round 2 (**f**). For instance, team DMIS_DK submitted predictions based both on random forest (RF) and DL algorithms in Round 2, where the latter showed slightly better accuracy. A total of 33 submissions with Root Mean Square Error (RMSE) >2 are omitted in the RMSE results (**c**, **e**, **f**). Source data are provided as a Source Data file[54].

and IC_{50} values are dependent on the pre-specified target protein concentration of the assay.

We also investigated how well the Challenge models predicted various kinase classes to study their applicability ranges. We first ranked the compound-kinase pairs based on their absolute errors (AEs), and then systematically explored whether any kinase group or family would be enriched among the best or worst-predicted pairs (see Methods). When considering 90 out of 99 Challenge submissions in Round 2 (with average AE < 2), the compound-kinase pairs involving mitogen-activated protein (MAP) and

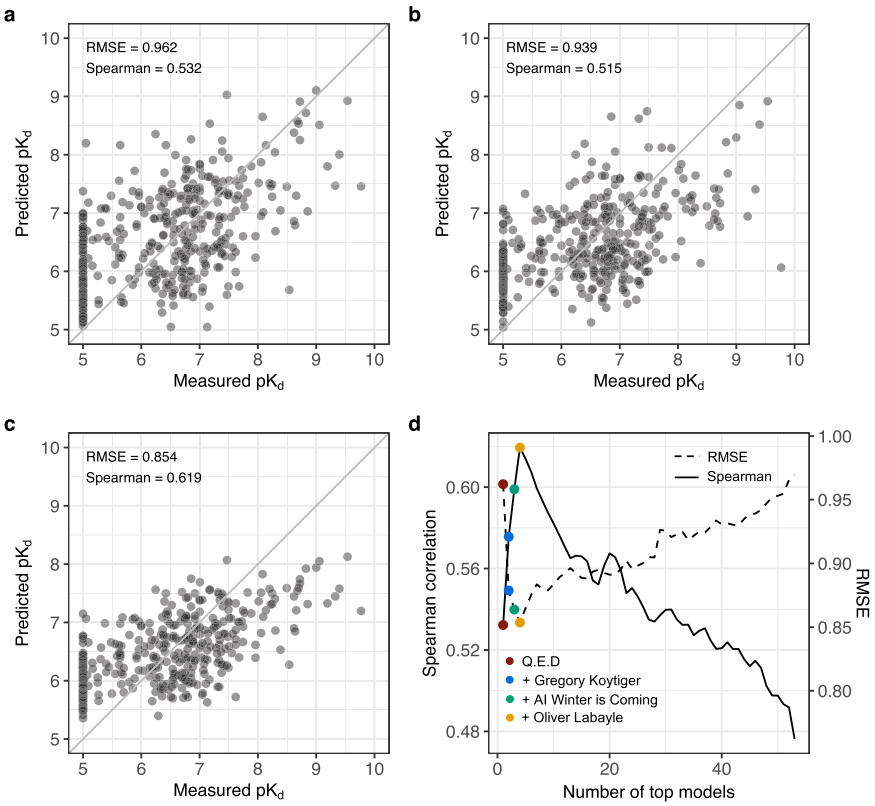


Fig. 4 The top-performing Challenge models and their ensemble combination. a Spearman correlation sub-challenge top performer in Round 2 (Q.E.D). **b** RMSE sub-challenge top performer in Round 2 (AI Winter is Coming). The points correspond to 394 pairs between 25 compounds and 207 kinases. **c** Ensemble model that combines the top four models selected based on their Spearman correlation in Round 2. **d** The mean aggregation ensemble model was constructed by adding an increasing number of top-performing models based on their Spearman correlation (the solid curve), until the ensemble correlation dropped below 0.45. The peak performance was reached after aggregating four teams (marked in the legend; see Supplementary Fig. 9 for all the teams. Note: ensemble prediction from a total of 21 best teams had a significantly better Spearman correlation compared to the Q.E.D model alone). The right-hand *y*-axis and the dotted curve show the Root Mean Square Error (RMSE) of the ensemble model as a function of an increasing number of top-performing models. Source data are provided as a Source Data file[54].

platelet-derived growth factor receptor kinases showed poorer accuracies compared to other kinase families ($P = 0.001$, Kruskal–Wallis test), but these families were better predicted using the Q.E.D and the top-ensemble models (Supplementary Fig. 13). For MAP kinases, the higher prediction error (adjusted $P = 0.016$, Kolmogorov–Smirnov test) could be attributed to the fact that most of the inhibitors targeting MAP kinases are noncompetitive allosteric inhibitors[18]. Similarly, pairs in the CMGC kinase group, including e.g. cyclin-dependent kinases, showed an increased error for bulk of the submissions (adjusted $P = 0.030$, Kolmogorov–Smirnov test), but again both the ensemble and Q.E.D models made better predictions also in this kinase group (Supplementary Fig. 14).

Comparison against single-dose activity assays. We next investigated how well the top-performing prediction models compare against the single-dose activity assays in terms of reducing the number of false positives and negatives when selecting most potent compound-kinase activities for more detailed, multi-dose K_d profiling. Such two-step screening approach is widely used in large-scale kinase-profiling studies[5–7,16], where K_d profiling is carried out only for compound-kinase pairs with an inhibition above 80% in the single-dose assays. For this classification task, we defined the ground truth activity classes based on the measured K_d values, which provide a more practical prediction outcome, compared to the rank correlation analyses that already demonstrated predictive rankings with the top-performing models (Fig. 4). Using the activity cut-off of measured $pK_d = 6$ and a single-dose inhibition cut-off of 80%, similar to previous studies[7,16,19], the positive predictive value (PPV) and the false discovery rate (FDR) of the single-dose assay were PPV = 0.66 and FDR = 0.44, respectively, in the Round 2 dataset. When using the mean aggregation ensemble from the top-performing models and the same cut-off of $pK_d = 6$ for both the predicted and measured activities, we observed an improved precision of PPV = 0.76 and FDR = 0.24.

We repeated the activity classification experiment with multiple pK_d activity cut-offs, and ranked the Round 2 pairs both using the model-predicted pK_d values and the measured single-dose inhibition assay values, and then compared these rankings against the true activity classes based on the measured dose-response assay (with either $pK_d > 6$ or 7 indicating true positive activity). These analyses demonstrated an improved activity classification accuracy using the mean ensemble of the top-performing models (Fig. 7a), especially when focusing on the most potent compound-kinase activities with the highest specificity. This improvement in both sensitivity and specificity was achieved without making any additional activity measurements, and it became even more pronounced with the precision-recall (PR) analysis, which showed that the precision of the ensemble model remained above PPV = 75% level even when the recall (sensitivity) level exceeded 75% (Fig. 7b). The top-performing model (Q.E.D) also showed improved performance when compared to the single-dose activity assay. As expected, the prediction

Table 1 Model classes, compound and kinase descriptors and training data used by the Round 2 top-performing teams and the baseline model[17].

Team	Algorithm type	Algorithm name	Combined models	Training strategy
DMIS_DK	Deep learning, multi-target learning	Multi-task graph convolutional neural networks	12	Train test split
AI Winter is Coming	Gradient boosting decision trees	XGboost	5 per target	K-fold nested cross validation, boosting
Q.E.D	Kernel learning	CGKronRLS	440	Boosting
Gregory Koytiger	Deep learning, artificial neural network	Not applicable	6	Fixed hold out
Olivier Labayle	Ridge regression	Not applicable	Not applicable	K-fold cross validation
Baseline	Kernel learning	CGKronRLS	1	K-fold nested cross validation

Team	Training data sources	Compound-protein pairs	Bioactivity types	Protein features	Chemical features
DMIS_DK	DrugTargetCommons, BindingDB	953521	K_d, K_i, IC_{50}	None	Molecular graphs
AI Winter is Coming	DrugTargetCommons, ChEMBL	600000	K_d, K_i, IC_{50}, EC_{50} %inh, %activity	None	ECFP5, ECFP7, ECFP9, ECFP11
Q.E.D	DrugTargetCommons, ChEMBL, UniProt	60462	K_d, K_i, EC_{50}	Amino acid sequences	ECFP4, ECFP6
Gregory Koytiger	ChEMBL	250000	K_d, K_i, IC_{50}	Amino acid sequences	SMILES strings
Olivier Labayle	DrugTargetCommons, ChEMBL, UniProt	18200	K_d	K-mer counting	ECFP
Baseline	DrugTargetCommons	44186	K_d	Amino acid sequences	Path-based fingerprints

Even if the teams chose to combine predictions from multiple models, they had to submit only one prediction per compound-kinase pair for scoring against the measured activities. Supplementary Table 1 provides further details of all the models submitted together with method surveys and model performances in Round 2.

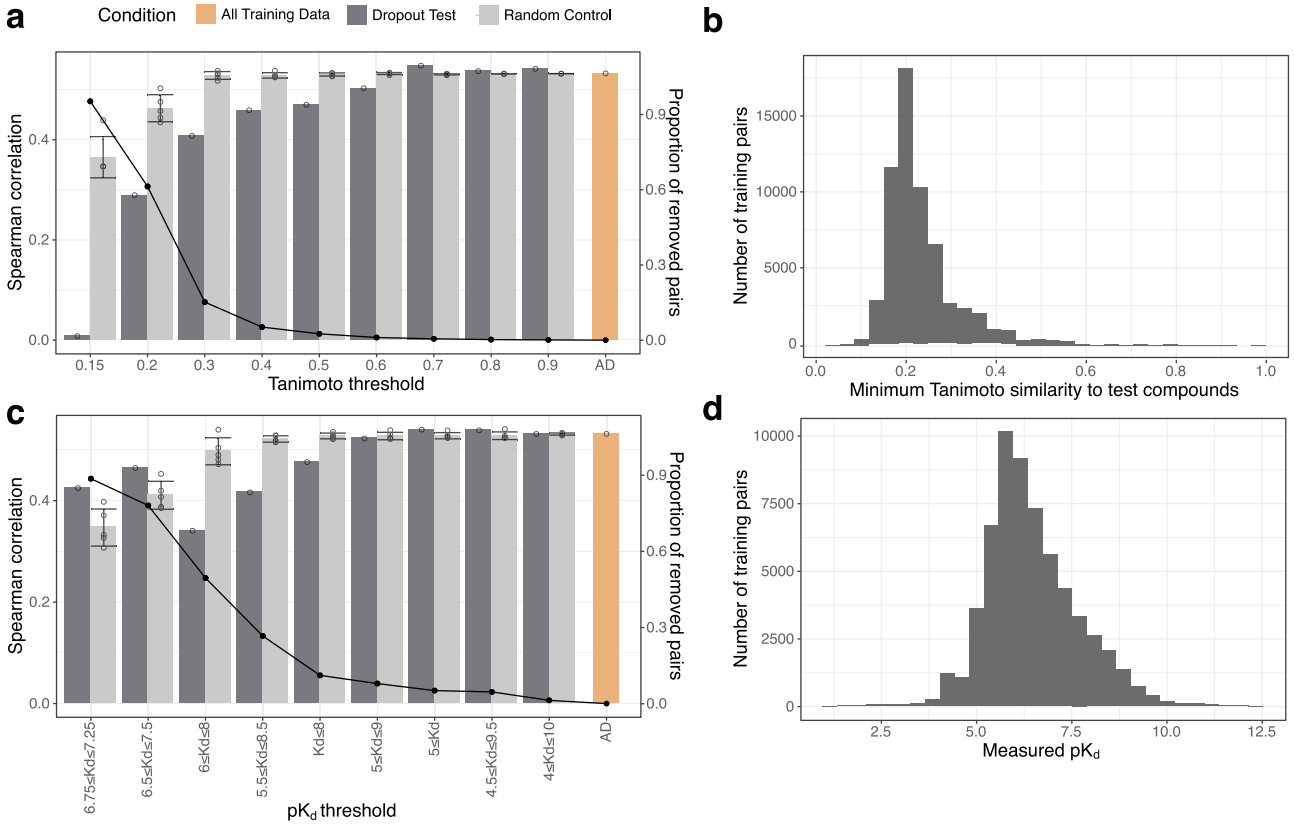


Fig. 5 The Q.E.D model performance as a function of training data size and scope. a The drop-out experiment removed increasing numbers of training compounds (as measured by maximum Tanimoto similarity with ECFP4 fingerprint between each training compound and all Round 2 test set compounds), retrained the Q.E.D model, and tested the performance. AD stands for all data. A noticeable decrease in performance begins to appear only at around 0.6 Tanimoto similarity suggesting that highly similar compounds in the training dataset are not necessarily required for accurate model performance. As a control, identical numbers of random compound-kinase pairs were removed, repeated 5 times to assess the variability of random removal. The error bars indicate the standard deviation of these replicates. Black points indicate proportions of removed compound-kinase pairs. **b** A histogram describing the full training dataset used to generate the results in **a**. **c** Model performance with multiple training datasets and varying pK_d levels, where the ranges in the *x*-axis labels refer to the compound-kinase pairs that were included for the model training. AD stands for all data. Random dropout control was repeated 5 times. The error bars indicate the standard deviation of these replicates. **d** A histogram describing the full training dataset used to generate the results in **c**. Source data are provided as a Source Data file[54].

accuracies decreased when using a more stringent measured activity cut-off of $pK_d > 7$ (Supplementary Fig. 15), since these rare extreme activities are more challenging to predict.

Model-based kinase predictions and their validation. To further investigate both the sensitivity and specificity of the model predictions, we experimentally profiled 81 additional compound-kinase pairs, which were not part of Round 1 or 2 datasets, selected based on the pK_d predictions from the top-performing models. These post-Challenge experiments were carried out in an unbiased manner, regardless of the compound classes, kinase families, or inhibition levels, to investigate the accuracy of predictive models to identify potent inhibitors of kinases with less than 80% single-dose inhibition; this activity cut-off is often used when selecting pairs for multi-dose K_d testing[7,16,19] but it may miss the more challenging compound-kinase dose-response relationships. Most of the measured pK_d values of these 81 pairs were distributed as expected, according to the expected single-dose inhibition function (Fig. 8a, black trace). However, the model-based approach also identified a large number of unexpected activities ($pK_d > 6$) that had been missed based on the single-dose inhibition assay alone (inhibition <80%); selected examples are discussed below.

As an example of a potent activity missed by the single-dose assay, the ensemble of the top-performing models predicted PYK2 (PTK2B) as a high-affinity target of a PLK inhibitor TPKI-30 (Fig. 8a). The new multi-dose pK_d measurements carried out after the Challenge validated that TPKI-30 indeed has an activity against PYK2 close to its potency towards PLK2 (Fig. 8b, left panel). Neither PYK2 or FAK would have been predicted as potent targets based on the single-dose testing alone, which led to multiple false negatives (Fig. 8b, right panel). In general, the single-dose testing had a relatively low predictivity of the actual TPKI-30 potencies, since kinases other than PLKs with high single-dose activity were confirmed as non-potent targets based on the dose-response K_d testing, resulting in many false positives. In contrast, the top-performing ensemble model predictions turned out to be relatively accurate, except for a few receptor tyrosine kinases (Fig. 8b, left panel). This example shows how the predictive models identify so far unexplored compound-kinase activities missed by standard methods (see also next section).

Another unexpected kinase activity was predicted for GSK1379763 that showed a novel chemotype for inhibition of DDR1 based on the subsequent K_d assays, exceeding that of the AURKB (Fig. 8c, left panel). The single-dose testing suggested that this compound would have potency neither against DDR1 or AURKB (Fig. 8c, right panel), whereas the multi-dose assays

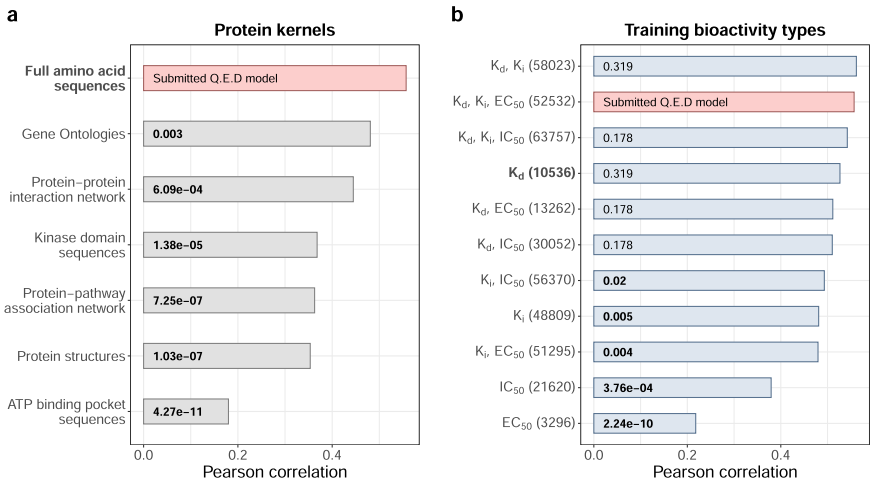


Fig. 6 The effect of protein descriptors and bioactivity types on Q.E.D model accuracy. The bars show Pearson correlations between the measured and Q.E.D model-predicted pK_d's calculated over the 394 Round 2 compound-kinase pairs based on different **a** protein kernels and **b** training bioactivity data types. The total number of training bioactivity data points is written in parentheses. The original, submitted Q.E.D model based on the full amino acid sequence-based protein kernel and using K_d, K_i, and EC_{50} bioactivities in the training dataset is marked with red. No other changes were introduced to the submitted Q.E.D model, which is an ensemble of the regressors with different regularization hyperparameter values and eight compound kernels, but where each regressor is built upon the same protein kernel based on full amino acid sequences. The protein kernel and training bioactivity type used in the baseline model are marked in boldface. The numbers inside the bars are Benjamini–Hochberg adjusted two-sided *P* values calculated with the Pearson and Filon test for comparing the correlation of the submitted Q.E.D model and each of its re-trained variants. Since the two correlations under comparison are calculated on the same set of data points and they have one variable in common (measured pK_d), the dependence between pK_d's predicted by the submitted Q.E.D model and the new model variant is taken into account in the statistical test. Significant *P* values (adjusted $P < 0.05$) are written in boldface. Source data are provided as a Source Data file[54].

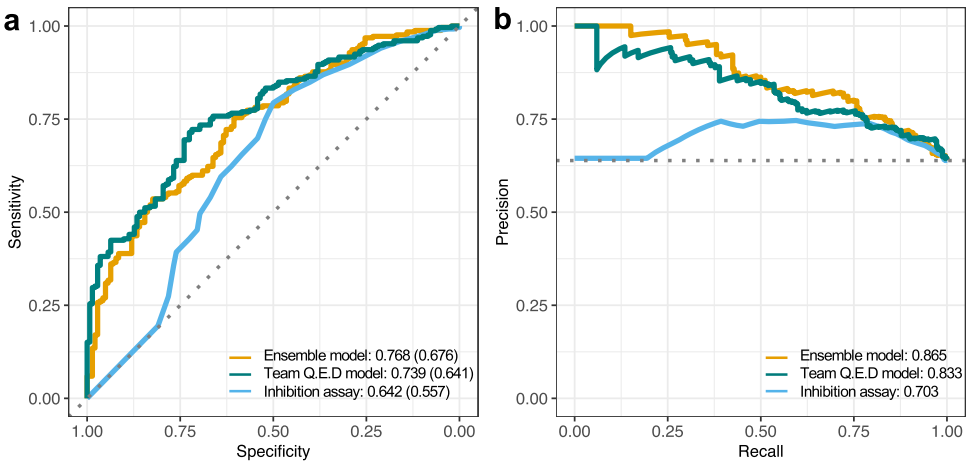


Fig. 7 Top-performing model predictions compared against single-dose assays. a Receiver operating characteristic (ROC) curves when ranking the 394 compound-kinase pairs in Round 2 using the pK_d predictions either from the ensemble of the top-performing models (average predicted pK_d from Q.E.D, DMIS_DK and AI Winter is Coming), or only from the Q.E.D model, compared against the experimental single-dose inhibition assays (the pairs with higher inhibition% are ranked first). The true positive activity class contains pairs with measured $pK_d > 6$ (see Supplementary Fig. 15 for $pK_d > 7$). The area under the ROC curve values are shown after the predictors (and the balanced accuracy is marked in the parentheses), and the diagonal dotted line shows the random predictor with an accuracy of AU-ROC = 0.50. **b** Precision-recall (PR) curves for the same activity classification analysis as shown in **a**. The area under the PR curve values are shown after the predictors and the horizontal dotted line indicates the random predictor with a precision of 0.64. Note: Round 2 K_d measurements were pre-selected to include mostly pairs with single-dose inhibition >80%, which makes Round 2 pairs optimal for systematic analysis of false positive predictions, and hence sensitivity (recall) and PPV (precision). However, these 394 pairs pre-selected for K_d profiling were less optimal for a comprehensive analysis of false negative predictions, and the evaluation of specificity. Source data are provided as a Source Data file[54].

confirmed potency towards DDR1 at a similar level as the Round 2 highest affinity target MEK5 (MAP2K5). A novel activity was predicted also between PFE-PKIS14 and CSNK2A2, a dark kinase nominated by the IDG consortium, which was missed by the Round 2 single-dose assay (inhibition = 78%; Fig. 8d, right panel). The single-dose assay led also to a number of other false positive and false negative activities for PFE-PKIS14, whereas the ensemble model demonstrated again a good predictive accuracy (Fig. 8d, left panel). Arguably, however, this interaction and the ensemble-predicted activity between AKI00000050a and FLT1 could have been identified based on their relatively high single-dose activity, even if less than 80% (Fig. 8a).

Comparison with other target prediction methods. To study whether standard target prediction methods could identify the

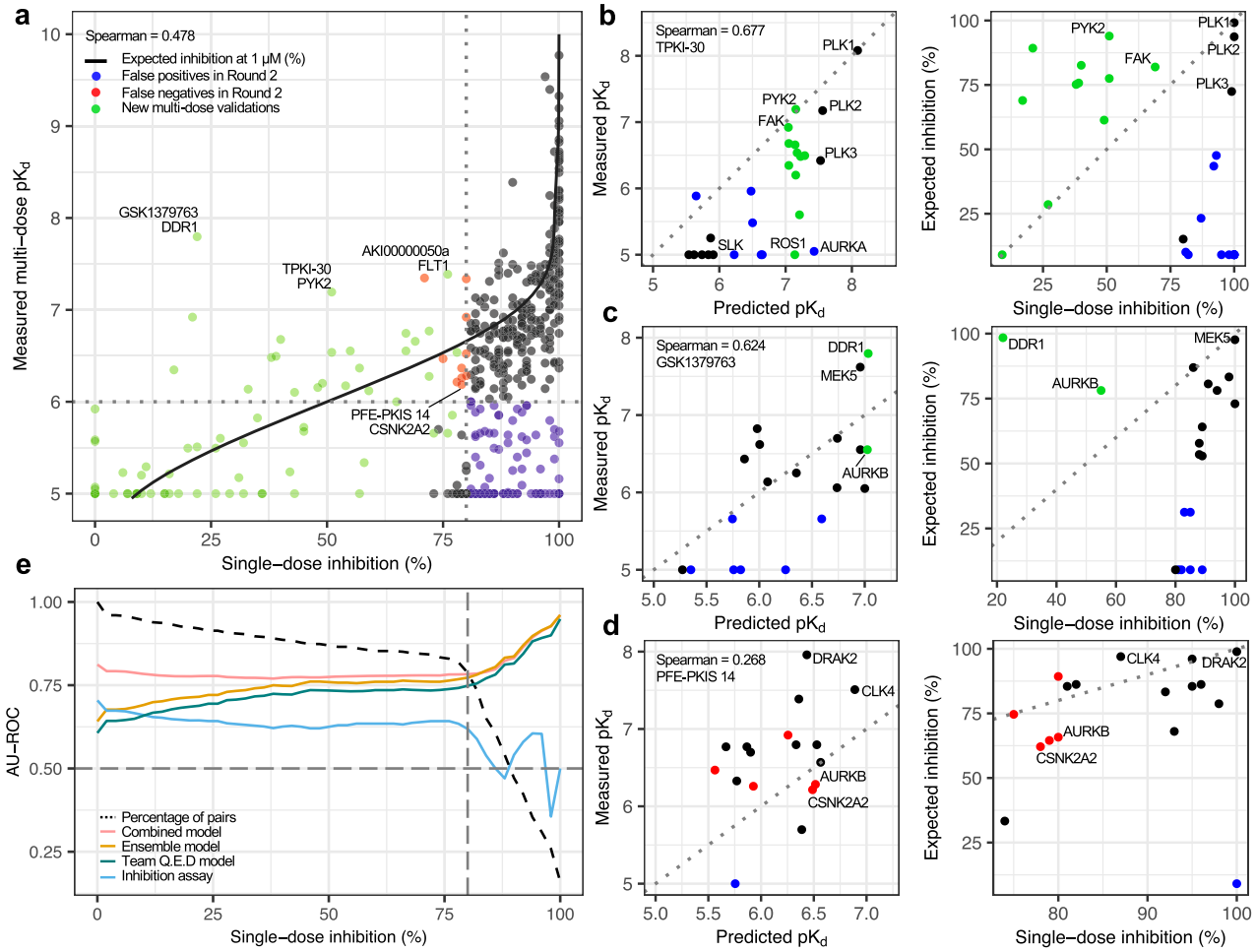


Fig. 8 Machine learning-based kinase activity predictions. a Comparison of single-dose inhibition assay (at 1 μM) against multi-dose K_d assay activities across 475 compound-target pairs (395 from Round 2 and 81 from the post-Challenge experiments). The red points indicate false negatives and blue points false positives when using the cut-offs of $pK_d = 6$ and inhibition = 80% among the 394 Round 2 pairs (including 75 pairs with inhibition >80% but that showed no activity in the dose-response assays, i.e, $pK_d = 5$). The green points indicate the new 81 pairs profiled post-Challenge solely based on the ensemble model predictions, regardless of their inhibition levels. The black trace is the expected %inhibition rate based on measured pK_d's, estimated using the maximum ligand concentration of 1 μM both for the single-dose and dose-response assays (see Methods). **b**–**d** Multi-dose (left) and single-dose (right) assays for kinases tested with TPKI-30, GSK1379763, and PFE-PKIS14. Green points indicate the new experimental validations based on the ensemble model predictions, whereas black points come from Round 2 data. Blue points indicate false positive predictions based either on predictive models or single-dose testing. **e** Predictive accuracy of the top-performing ensemble model (average predicted pK_d), top-performing Q.E.D model and single-dose assay (at 1 μM), when classifying subsets of the 475 pairs into the true activity classes with measured pK_d less or higher than 6. The *y*-axis indicates the area under the receiver operating characteristic (ROC) curve (AU-ROC) as a function of the single-dose inhibition% levels, *x*-axis the pairs with inhibition >*x*%, and the dashed black curve the percentage of all pairs that passed that single-dose activity threshold. The combined model trace corresponds to the average of measured and expected inhibition values, where the latter was calculated based on the mean ensemble of the top-performing model pK_d predictions (Q.E.D, DMIS_DK and AI Winter is Coming). See Supplementary Fig. 16 for the corresponding analysis with precision-recall (PR) metric, and Supplementary Fig. 17 for the ROC and PR curves for all the 475 pairs. Source data are provided as a Source Data file[54].

selected compound-target activities predicted by the top-performing ensemble model (Fig. 8), we used the similarity ensemble approach (SEA), a popular target classification method that relates proteins based on chemical similarity among their ligands[20]. Strikingly, the SEA method did not identify target activity among any of the three selected kinases and their confirmed inhibitors (Supplementary Table 2). For instance, the highest scoring hit from SEA for compound TPKI-30 was FAK (PTK2), which belongs to the same subfamily of kinases as PYK2, that was confirmed as potent target of TPKI-30, but their sequence identity is only ~43%. To further model the ligand-receptor interaction between TPKI-30 and PYK2, in the absence of 3D chemical structures, we carried out an in-silico docking procedure. As expected, the protein structure-based docking approach was not informative enough for predicting the dose-response activity relationships between TPKI-30 and PYK2, but its results supported a potent binding between TPKI-30 and PYK2, with a similar binding affinity compared to the known active ligands that bind to the same binding pocket of PYK2 (Supplementary Fig. 18).

Based on the observation that the single-dose assays and model-based pK_d predictions were overall only weakly correlated (Supplementary Fig. 19), and that they showed opposite trends for the pK_d prediction accuracy when increasing the inhibition cut-off level (Fig. 8e), we finally studied whether the single-dose measurements and the ensemble-based pK_d predictions could be combined for improved kinase activity predictions. Specifically, for each compound-kinase pair, we calculated the average of its

measured and expected inhibition values based on the single-dose assay and ensemble model predictions, respectively. This combined predictor showed improved activity classifications beyond that of the ensemble model predictions, across various inhibition levels, and it identified an extended number of potent compound-kinase interactions at lower single-dose activity, compared to the standard 80% cut-off (Fig. 8d, dotted line). In the full set of all the 475 pairs, the combined model improved both the sensitivity and specificity of the pK_d predictions (Supplementary Fig. 17a), and especially the precision of the top-activity predictions that are prioritized for further validation (Supplementary Fig. 17b). Based on the wider availability of single-dose activity data, this integrated method provides a generally applicable and cost-effective approach for future target activity profiling studies.

Discussion

While experimental mapping of target activities is critical for understanding compounds' mode of action, biochemical target activity profiling experiments are both time consuming and costly. The enormous size of the chemical universe, spanned by up to 10^{20} molecules with potential pharmacological properties[21,22], makes the experimental bioactivity mapping of the full compound and target space quickly infeasible in practice. The IDG-DREAM Drug Kinase Binding Prediction Challenge was designed to benchmark algorithms capable of predicting and prioritizing compound-kinase activities, and therefore to guide data-driven decision making and reduce the high failure rates. The model-guided approach has the potential to help both phenotype-based drug discovery (e.g., mapping of the activity space of lead compounds), and target-based drug discovery (e.g., identification of candidate compounds that selectively inhibit a particular disease-related kinase). As an example, the ensemble of the top-performing models led to a surprising result that the PLK inhibitor TPKI-30 targets also PYK2, and with a somewhat lesser potency also its paralog, FAK (Fig. 8b). Another selected example, CSNK2A2, belongs to the dark kinases nominated by the IDG consortium[23], suggesting that the prediction models can identify potent inhibitors even for the currently understudied kinases. The two other highlighted kinases, PYK2 and DDR1, were neither among the most-studied kinases in terms of the number of dose-response bioactivity data points in the public domain for the model training (Supplementary Fig. 20).

There is an increasing number of studies published each year that introduce new computational algorithms to predict compound-target activities (Supplementary Fig. 21a). Although previous studies have demonstrated the potential of ML algorithms to help fill in the gaps in compound-target interaction maps[17,24], and to accelerate several phases of drug discovery[25,26], to date there has been no systematic and unbiased evaluations of quantitative prediction models for target activity on a blinded and large-enough dataset, such as the one used in the present benchmarking. Participants of this Challenge made use of various ML approaches, which led to relatively wide performance differences (Supplementary Figs. 6 and 7), and covered the most popular ML approaches used for compound-target activity prediction, especially when considering the supervised regression problem (Supplementary Fig. 21b–d; Supplementary Table 1). Only the k-nearest neighbors (kNN) and Bayesian methods were not part of the Challenge submissions. Recently, many advanced deep learning (DL) algorithms have been proposed for compound-target interaction prediction[27–29], and a previous comparative work that used nested cross-validation on bioactivity data from ChEMBL found out that DL methods outperformed other methods, including kNN, support vector machines, random forests, naive Bayes and SEA, as representative target prediction methods[24]. In contrast, our Challenge results did not support the overall superiority of the DL methods compared to the other learning approaches (Fig. 3f).

Among the 31 teams that answered our survey at the end of the Challenge, none of the method classes had a very strong contribution to the prediction accuracy (Supplementary Fig. 22a, b), similarly as has been seen also in other DREAM challenges[30–32]. A striking observation from the survey was that there was a tendency for improved K_d prediction accuracies by teams that used other types of multi-dose bioactivity data (e.g., K_i, IC_{50}, EC_{50}), compared to using K_d data alone (Supplementary Fig. 22c, d). This provides a further opportunity for ML models such as DL that require relatively large training datasets, as these bioactivity types are among the most common in multi-dose target profiling (Supplementary Fig. 22e). Single-dose bioactivity measurements (e.g., potency% and other activity assays) are most abundant in the open bioactivity databases, making their use an exciting option for predicting dose-response activities such as K_d. In the Challenge, single-dose %inhibition and %activity data were utilized by one of the top-performing models, AIWIC, whereas the other top performer Q.E.D missed the most abundant multi-dose IC_{50} bioactivities in the model training (Table 1). However, we showed how the integrated use of the other multi-dose bioactivity types, especially K_i, compensated for the lack of IC_{50} data and led to the top-performance of the Q.E.D model (Fig. 6b). In contrast, our results based on the Q.E.D model showed that the use of other than kinase proteins and kinase inhibitors in the training data led to a decreased prediction performance compared to the original Q.E.D model with kinases only (Supplementary Fig. 23).

To further study whether the individual models complement each other and could yield an overall better result, we aggregated the top-performing models as a mean ensemble model. Many previous DREAM Challenges have demonstrated that such wisdom of the crowds may improve the predictive power of the individual models through combining models as meta-predictors or ensemble models[30–32]. The ensemble model constructed in this Challenge made use of the various modeling approaches and features of the top-performing models, after which adding more models led to rapid decrease in accuracy (Fig. 4d). In our post-Challenge analyses, the combination of the top-performing ML models improved both the sensitivity and specificity, compared to single-dose target activity assays, without requiring any additional experiments (Fig. 7). We also observed that the combination of the top-performing models using an ensemble model led to accurate and robust predictions of kinase inhibitor potencies across multiple kinase families and groups (Supplementary Figs. 13 and 14). Subsequent target profiling experiments carried out based on the ensemble model predictions demonstrated that the ML models facilitate experimental mapping efforts, both for well-studied and understudied kinases (Fig. 8). Interestingly, combining the single-dose inhibition measurements with the top-performing ML models led to even higher prediction accuracy than using either one alone, while identifying an increased number of potent compound-kinase activities compared to that using the standard 80% inhibition cut-off (Fig. 8e).

The Spearman correlation sub-challenge top performer (Q.E.D) used the same kernel-based regression algorithm as the baseline model[17], yet showed markedly better performance (Fig. 3f). The two models, however, differ in several aspects. The Q.E.D model integrated multiple bioactivity types in their training data, as opposed to using K_d only as was done in the baseline model, and this integrative approach led to significant differences in the prediction accuracy (Supplementary Fig. 12). Although the training dataset sizes of both models had similar numbers of bioactivity values (baseline 44,186 vs. Q.E.D 60,462), Q.E.D used

bioactivity data points for many more compounds than the baseline approach (1968 vs. 13,608 compounds). This increased the diversity of the training dataset, which is often more important than its actual size, especially when majority of the test compounds have no multi-dose bioactivity data available for model training. Furthermore, while both models used the same protein kernel based on Smith–Waterman amino acid sequence alignment, Q.E.D implemented an ensemble model of 440 individual regressors based on various model hyperparameters and eight compound kernels, which resulted in an effective integration of several different compound representations and an improved prediction performance (Supplementary Fig. 24). However, we found that many combinations of the widely used kinase and chemical descriptors led to relatively high prediction accuracies (Fig. 6a; Supplementary Fig. 10), which should make the ensemble approach practical for future applications, also beyond kinases. We also observed that full amino acid sequences used as protein kernels performed significantly better than those based on kinase domain sequences (Fig. 6a). This observation is most likely due to a number of missing kinase domain sequences in the Q.E.D model, which resulted in several pK_d predictions of zero (7%), and reduced training dataset size.

Rather surprisingly, the number of training bioactivity data did not strongly contribute to the prediction accuracies of the top-performing Q.E.D model (Supplementary Fig. 25), provided the training data had sufficient structural diversity for the kinase families being predicted (Fig. 5a). Our training data drop-out analyses have substantial implications for the application of supervised ML in predicting the activity of kinase inhibitors, as they demonstrated that the predictions are reasonably robust even when only limited numbers of structurally similar training data exist (Fig. 5). This observation is also evident from the fact that the top-performing models used a rather different number of training bioactivity values from different multi-dose assays when predicting the pK_d profiles (Table 1). This suggests that the number of training data is not the strongest factor for the predictive performance, rather the way the model is constructed has a much larger contribution to the prediction accuracy, which has implications especially for so-far understudied kinases. Given that the currently available bioactivity data are still rather limited and come in various types, it was comforting to note that the top-performing models made use of the various data types in the training phase (Table 1). This can be considered as another form of 'wisdom of the crowds', and suggests that beyond the community effort for target activity predictions, there is a need for also crowdsourced collection, annotation, and harmonization of different types of bioactivity data to further improve the accuracy and coverage of the predictive models.

To enable the community to apply the predictive models benchmarked in the Challenge to various drug development applications, we have made available the top-performing models as containerized source code. The Docker models enable continuous validation of the model predictions whenever new experimental kinase-profiling data will become available, as well as make it possible to run the best performing models on private data that would otherwise remain closed and unavailable to the research community[33]. The current test data covers ca. 57% of the human protein kinome, and future screening efforts are warranted to extend it to additional interactions with remaining kinases and other important target families. Future applications should select the model class that best fits the specific needs. All the top-performing teams used ML models that leverage information extracted from similar kinases and/or inhibitors to predict the activity of so far unexplored interactions (see Table 1 and Supplementary Table 1). Most of the top-performing models also used amino acid sequences or K-mer counting as target-based features in their class-agnostic prediction models, and two of the top performers did not utilize any type of protein features. Furthermore, none of the top-models required 3D or other detailed chemical information, making the ML models straightforward to apply for various compound classes. We therefore believe the Challenge models and the current benchmarking results will provide useful information for constructing predictive models also beyond kinases inhibitors.

In conclusion, we envision that the IDG-DREAM Challenge will provide a continuously updated resource for the chemical biology community to benchmark, prioritize, and experimentally test new kinase activities toward accelerating many drug discovery and repurposing applications.

Methods

Challenge infrastructure and timeline. The Challenge was organized and run on the collaborative science platform Synapse. All prediction files were submitted using the Challenge feature of this platform to track which teams and individuals submitted files, and to track the number of submissions per team. Challenge infrastructure scripts including code for calculating the scoring metrics are available at https://github.com/Sage-Bionetworks/IDG-DREAM-Drug-Kinase-Challenge and archived at https://doi.org/10.5281/zenodo.4648011. Teams were permitted to submit three predictions for Round 1, and two predictions for Round 2 (Supplementary Fig. 3). In Round 2, we selected the best of the two submissions for each scoring metric. This led to a selection of 54 final prediction sets for each of the Round 2 scoring metrics (Spearman correlation and RMSE, see 'Scoring of the model predictions' below) from the 99 total submissions in Round 2. For Rounds 1 and 2, we used a common workflow language-based challenge infrastructure to perform the following tasks: (1) validate a prediction file to ensure that it conformed to the correct file structure and had numeric pK_d predictions and return an error email to participants if invalid, (2) run a python script to calculate the performance metrics for a submitted prediction, and (3) return the score to the Synapse platform. For Round 1b, in which we permitted 1 submission a day for 60 days, we implemented a modified Ladderboot[34] protocol to prevent model overfitting. This was done by modifying step (2) above as follows: the scoring infrastructure receive a submitted prediction, check for a previous submission from the same team and run an R script to bootstrap the current and previous submission 10,000 times, calculate a Bayes factor (K) between the two submissions; the scoring harness would then only return an updated score if it was substantially better ($K > 3$) than the previous submission.

New bioactivity data for model testing. To generate unpublished test bioactivity data for scoring of predictions, we sent kinase inhibitors to DiscoverX (Eurofins Corporation) for the generation of new dose-response dissociation constant (K_d) values, as a measure of a binding affinity. In order to give a better sense of the relative compound potencies, K_d is represented in the logarithmic scale, as $pK_d = -\log_{10}(K_d)$, where K_d is given in molars [M]. The higher the pK_d value, the higher the inhibitory ability of a compound against a protein kinase. A two-step screening approach was adopted[5–7], where the dose-response K_d values were generated for a range of compound-kinase pairs that had inhibition >80% in the primary single-dose screen using the DiscoverX KINOMEscan protocol (https://www.discoverx.com/services/drug-discovery-development-services/kinase-profiling/kinomescan). KINOMEscan employs a competitive binding assay to estimate K_d, wherein the immobilized ligands and the test compound compete for the same binding pocket of the assayed kinase. The compounds were supplied as 10 mM stocks in DMSO, and the top screening concentration was 10 µM in the graded-dose profiling (with one technical replicate). The single-dose assays used a single fixed concentration of 1 µM (no replicates).

A total of 25 of the axitinib-kinase pairs generated for Round 2 were already profiled in previous published studies[7,16], and were therefore excluded from the Round 2 test dataset. The Spearman correlation between these newly measured pK_d's and those available from DTC was 0.701 (Supplementary Fig. 26a), providing the experimental consistency of the K_d measurements for axitinib. We note this 25 pK_d's is a rather limited set for such analysis of consistency, and therefore we extracted a larger set of 416 K_d measurements that overlapped with the Round 2 kinases from two comprehensive target profiling studies[5,6], including 104 pairs where $pK_d = 5$ in both of the studies. The Spearman correlation of these replicate pK_d measurements was 0.842 (Supplementary Fig. 26b), demonstrating a relatively good reproducibility for the large-scale binding affinity measurements. These replicate measurements were also used for determining a practical upper limit of the predictive accuracy of machine learning models in the scoring of their predictions (see below).

The selected kinase targets are a part of the SGC-UNC screening initiative, the Kinase Chemogenomic Set[16]. The primary selection criterion was to investigate the readily screenable human kinome, i.e., kinases with a robust assay readily available through commercial vendors. An additional focus point was to include those screenable kinase targets that are either understudied and/or targets with a Gene

Ontology information available but lacking associated small-molecule activities in ChEMBL[11], called as dark kinases (Tdark) and Tbio targets, respectively[13]. Out of the 392 wild-type human kinases subjected to the screening study by the KGCS Consortium, a subset of 295 kinases were used in our IDG-DREAM Challenge during the Rounds 1 and 2. The 95 kinase inhibitors used in the Challenge (70 for Round 1 and 25 for Round 2) were a part of the kinase inhibitor collection at the SGC-UNC for which we already had the single-dose inhibition screening done at DiscoverX across their large kinase panel (scanMax^{SM}).

To subsequently test the top-performing model predictions in additional compound-kinase pairs that were not part of Round 1 or 2 datasets, we selected a set of 88 pairs that showed most potency based on the average predicted pK_d of the top-performing models (Q.E.D, DMIS-DK, and AIWIC), regardless of their single-dose inhibition levels. These 88 pairs were actually scattered across the whole spectrum of single-dose inhibition levels, ranging from 0 to 78% (Supplementary Fig. 19; note: pairs with inhibition >80% were K_d-profiled already in Round 2). One of the compounds (TPKI-35) was not available from IDG, so the predicted 7 kinase targets for that compound could not be tested experimentally, resulting in a dataset of total of 81 compound-kinase pairs that were shipped to DiscoverX for multi-dose K_d profiling. One of the compounds (GW819776) was shipped separately in a tube, whereas the other 14 compounds were supplied as 10 µM stocks in DMSO, and the K_d profiling was done using the same KINOMEscan competitive binding assay protocol as for the Round 1 and Round 2 pairs.

Estimating the expected inhibition levels. The KINOMEscan assay protocol utilized for both the single-dose and dose-response assays is based on competitive binding assays, where the maximum compound concentration tested was 1 µM and 10 µM respectively. For a given compound-kinase pair, the K_d values calculated from the dose-response assay (excluding pairs with activity ≥10 µM) were then used to estimate the expected single-dose %inhibition level (at 1 µM of compound) using the conventional ligand occupancy formula:

$$\text{Ligand occupancy}\,(\%) = \frac{\text{Maximum ligand concentration(M)}}{\text{Maximum ligand concentration(M)} + \text{Measured}K_d\text{(M)}} \quad (1)$$

In Eq. (1), the maximum ligand concentration is 1 µM in the kinase assay. Therefore, a measured $pK_d = 3$ (i.e. $K_d = 10^{-3}$ M) results in the expected inhibition of 0%, $pK_d = 4$ and 5 in 1% and 10% expected inhibitions, respectively, and $pK_d = 9$ (i.e. $K_d = 10^{-9}$ M) results in expected inhibition of 100%. The single-dose %inhibition assays were not optimized to accurately estimate the activity values of any specific compound-kinase interaction, leading to a variability in Fig. 8.

Scoring of the model predictions. In the Challenge phase, we used the following six metrics to score the quantitative pK_d predictions from the participants:

- Root mean square error (RMSE): square root of the average squared difference between the predicted pK_d and measured pK_d, to score continuous activity predictions.
- Pearson correlation: Pearson correlation coefficient between the predicted and measured pK_d's, which quantifies the linear relationship between the activity values.
- Spearman correlation: Spearman's rank correlation coefficient between the predicted and measured pK_d's, which quantifies the ability to rank pairs in correct order.
- Concordance index (CI)[35]: probability that the predictions for two randomly drawn compound-kinase pairs with different pK_d values are in the correct order based on measured pK_d values.
- F1 score: the harmonic mean of the precision and recall metrics. Interactions were binarized by their measured pK_d values into true positive class ($pK_d > 7$) and true negative class ($pK_d \leq 7$).
- Average area under the curve (AUC): average area under ten receiver operating characteristic (ROC) curves generated using ten interaction thresholds based on the measured pK_d interval [6, 8] to binarize pK_d's into true class labels.

The submissions in Round 1 were scored across the six metrics but the teams remained unranked. The Round 2 consisted of two sub-challenges, the top performers of which were determined based on RMSE and Spearman correlation, respectively. Spearman correlation evaluated the predictions in terms of accuracy at ranking of the compound-kinase pairs according to the measured K_d values, whereas RMSE considers the AEs in the quantitative binding affinity predictions. The tie-breaking metric for both Rounds was the averaged AUC metric in the ROC analyses that evaluated the accuracy of the models to classify the pK_d values into active and inactive classes based on multiple K_d cutoffs.

In the post-Challenge activity classification analyses, we used two additional metrics that take into account potentially unbalanced class distributions (see also Activity classification analyses):

- PR: area under the PR curve, where precision (PPV) is the fraction of true actives among positive predictions and recall equals to sensitivity.
- Balanced accuracy: the arithmetic mean of the precision and recall metrics. Interactions were binarized into true active class and true inactive class based on the measured pK_d values.

Two different activity cut-offs were applied (measured $pK_d > 6$ or 7) to study how the ground truth class balance affects the results (see Fig. 7, and Supplementary Figs. 15–17). The same cut-off value was used for the predicted pK_d to calculate the balanced accuracy.

Statistical evaluation of the predictions. Determination of the top performers was made by calculation of a Bayes factor relative to the top-ranked submission in each category. Briefly, we bootstrapped all submissions (10,000 iterations of sampling with replacement), and calculated RMSE and Spearman correlation to the test dataset to generate a distribution of scores for each submission. A Bayes factor was then calculated using the challengescoring R package (https://github.com/sage-bionetworks/challengescoring) for each submission relative to the top submission in each sub-challenge. Submissions with a Bayes factor $K \leq 3$ relative to the top submission were considered to be tied as top performers. Tie breaking for both sub-challenges was performed by identifying submission with the highest average AUC. To create a distribution of random predictions, we randomly shuffled the 430/394 K_d values across the set of 430/394 compound-kinase pairs in the Round 1/Round 2 datasets, and repeated the permutation procedure 10,000 times. Then we compared the actual Round 1/Round 2 prediction scores to Spearman and RMSE calculated from the permuted K_d data. We defined a prediction as better than random if its score was higher than the maximum of the 10,000 random predictions (empirical $P = 0.0$, non-parametric permutation test).

Statistical comparison of the predictions in terms of the two winning metrics was performed using either two-sample or paired Wilcoxon tests (non-parametric tests), depending whether groups of participants or the same participants were compared between the two Challenge scoring rounds. We compared the magnitudes of Pearson correlations between the measured and predicted pK_d's from two different models using Pearson and Filon test for two overlapping correlations implemented in cocor[36] R package. Specifically, since the two correlations under comparison were calculated on the same set of compound-kinase pairs and have one variable in common (measured pK_d), the correlation between pK_d's predicted by two different models is taken into account in the statistical test. Parametric test was applied in these comparisons due to the large number of compound-target pairs in Round 2 (394 pairs). When analysing the questionnaire's results, statistical significance was assessed using the non-parametric Kruskal–Wallis test, adjusted for multiple comparisons with Benjamini–Hochberg control of FDR. All the measurements corresponded to distinct participants or teams in the Challenge.

To determine the maximum possible performance practically achievable by any computational models, we utilized replicate K_d measurements from distinct studies that applied a similar biochemical assay protocol. We used the DrugTargetCommons to retrieve 863 and 835 replicated K_d values for kinases or compounds that overlapped with the Round 1 and 2 datasets, respectively. These data originated from two comprehensive screening studies[5,6]. To better represent the distribution of pK_d values in the test data, we subset the DTC data to contain 35% (Round 1) and 25% (Round 2) $pK_d = 5$ values, approximately matching the proportion of $pK_d = 5$ values in Round 1 and Round 2 test sets. For Round 1, we used 317 replicated K_d's, including 111 randomly selected pairs where $pK_d = 5$. For Round 2, we used 416 replicated K_d's, including 104 randomly selected pairs where $pK_d = 5$. We randomly sampled the replicate measurements of these compound-kinase pairs (with replacement), calculated the Spearman correlation and RMSE between the pK_d's of the two studies for each 430 and 394 sub-sampled sets for Round 1 and 2, respectively, and repeated this procedure for a total of 10,000 samplings.

The baseline prediction model. We used a recently published and experimentally validated kernel regression framework as a baseline model for compound-kinase binding affinity prediction[17]. Our training dataset consisted of 44,186 pK_d values (between 1968 compounds and 423 human kinases) extracted from DTC. Median was taken if multiple pK_d measurements were available for the same compound-kinase pair. We constructed protein kinase kernel using normalized Smith–Waterman alignment scores between full amino acid sequences, and four Tanimoto compound kernels based on the following fingerprints implemented in rcdk R package[37]: (i) 881-bit fingerprint defined by PubChem (pubchem), (ii) path-based 1024-bit fingerprint (standard), (iii) 1024-bit fingerprint based on the shortest paths between atoms taking into account ring systems and charges (shortestpath), and (iv) extended connectivity 1024-bit fingerprint with a maximum diameter set to 6 (ECFP6; circular). We used CGKronRLS as a learning algorithm (implementation available at https://github.com/aatapa/RLScore)[38]. We conducted a nested cross-validation in order to evaluate the generalization performance of CGKronRLS with each pair of kinase and compound kernels as well as to tune the regularization hyperparameter of the model. In particular, since the majority of the compounds from the Challenge test datasets had no bioactivity data available in the public domain, we implemented a nested leave-compound-out cross-validation to resemble the setting of the Challenge as closely as possible. The model comprising protein kernel coupled with compound kernel built upon path-based fingerprint (standard) achieved the highest predictive performance on the training dataset (as measured by RMSE), and therefore it was used as a baseline model for compound-kinase binding affinity prediction in both Challenge Rounds.

Top-performing models. Supplementary write ups provide details of all qualified models submitted to the Challenge[39]. The key components of the top-performing models are listed in Table 1 and summarized below.

Team Q.E.D model. To enable a fine-grained discrimination of binding affinities between similar targets (e.g., kinase family members), the team Q.E.D explicitly introduced similarity matrices of compounds and targets as input features into their regression model. The regression model was implemented as an ensemble version (uniformly averaged predictor) of 440 CGKronRLS regressors (CGKronRLS v0.81)[38,40], but with different choices of regularization strengths [0.1, 0.5, 1.0, 1.5, 2.0], training epochs [400, 410, …, 500], and similarity matrices: the protein similarity matrix was derived based on the normalized striped Smith–Waterman alignment scores[41] between full protein sequences (https://github.com/mengyao/Complete-Striped-Smith-Waterman-Library). Eight different alternatives of compound similarity matrices were computed using both Tanimoto and Dice similarity metrics for different variants of 1024-bit Morgan fingerprints[42] ('radius' [2, 3] and 'useChirality' [True, False], implementation available at https://github.com/rdkit/rdkit). Unlike the baseline method, which used only the available pK_d values from DTC for training, the team Q.E.D model extracted 16,945 pK_d, 53,894 pK_i, and 3301 pEC_{50} values from DTC. After merging the same compound-kinase pairs from different studies by computing their medians, 60,462 affinity values between 13,608 compounds and 527 kinases were used as the training data.

Team DMIS_DK model. Team DMIS_DK built a multi-task Graph Convolutional Network (GCN) model based on 953,521 bioactivity values between 474,875 compounds and 1474 proteins extracted from DTC and BindingDB. Three types of bioactivities were considered, that is, pK_d, pK_i, and pIC_{50}. Median was computed if multiple bioactivities were present for the same compound-protein pair. Multi-task GCN model was designed to take compound SMILES strings as an input, which were then converted to molecular graphs using RDKit python library (http://www.rdkit.org). Each node (i.e. atom) in a molecular graph was represented by a 78-dimensional feature vector, including the information of atom symbol, implicit valence, aromaticity, number of bonded neighbors in the graph, and hydrogen count. No protein descriptors were utilized. The final model was an ensemble of four multi-task GCN architectures described in the Supplementary writeups[39]. For the Challenge submission, the binding affinity predictions from the last K epochs were averaged, and then the average was taken over the 12 multi-task GCN models (four different architectures with three different weight initializations). Hyper-parameters of the multi-task GCN models were selected based on the performance on a hold-out set extracted from the training data. The GCN models were implemented using PyTorch Geometric (PyG) library[43].

Team AI Winter is Coming model. Team AI Winter is Coming built their prediction model using Gradient Boosted Decision Trees (GBDT) implemented in XGBoost algorithm (xgboost v0.90, scikit-learn v0.20.3)[44]. Training dataset included 600,000 pK_d, pK_i, pIC_{50}, and pEC_{50} values extracted from DTC and ChEMBL (version 25), considering only compound-protein pairs with ChEMBL confidence score of 6 or greater for 'binding' or 'functional' human kinase protein assays. For a given protein target, replicate compounds with different bioactivities in a given assay (differences larger than one unit on a log scale) were excluded. For similar replicate measurements, a single representative assay value was selected for inclusion in the training dataset. Chemical data was standardized using the ChemAxon Standardizer v18 (https://www.chemaxon.com), and further processed with OpenEye chemistry toolkit (Software Inc, https://www.eyesopen.com/oechem-tk). Each compound was characterized by a 16,000-dimensional feature vector being a concatenation of four ECFP fingerprints (as implemented in RDKit) with a length set to 5, 7, 9, and 11. No protein descriptors were used in the XGBoost algorithm[44]. A separate model for each protein target was trained using nested cross-validation (CV), where inner loops were used to perform hyperparameter optimization and recursive feature elimination. The final binding affinity prediction was calculated as an average of the predictions from the cross-validated models based on five outer CV loops.

Training data dropout experiments. We developed Docker containers using the Team Q.E.D model that accepted input parameters for minimum Tanimoto similarity to the test dataset (similarity calculated using the ECFP4 fingerprint), or pK_d cutoff values, to eliminate training data based on various thresholds (see Data and Code Availability). For each condition, training data were dropped out, the model was trained on the remaining data, and the trained model generated predictions for the Round 2 test compound-kinase pK_d values. The predicted pK_d values for each training condition were then scored by calculating the Spearman correlation in the test dataset. We trained and tested each experimental condition once. As a control for each experimental condition, we randomly removed an equivalent number of training compounds, repeated 5 times per condition.

Ensemble model construction. Ensemble models were generated by combining the best-scoring Round 2 predictions from each team. We iteratively combined models starting from the highest scoring Round 2 prediction (e.g., ensemble #1—highest scoring prediction, ensemble #2—second highest scoring, ensemble #3—third highest scoring, and so on) for all 54 Round 2 submitting teams. Three types of ensembles were created using arithmetic mean, median, and rank-weighted summarization approaches. The rank-weighted ensemble was calculated by multiplying each set of predictions by the total number of submissions plus 1 minus the rank of the prediction file, summing these weighted predictions, and then dividing by the sum of the multiplication factors. The 54 ensemble predictions for each of the three summary metrics were bootstrapped and Bayes factors were calculated as described in the 'Statistical evaluation of the predictions' Methods section to determine which models were substantially different from the top-ranked submission. We also randomly sampled 1000 sets of 4 models among the Challenge submissions, ensembled the predictions in each set, and scored each set. These combinations of four random-performance models could not match or supersede the performance of an ensemble of the top four models (i.e., an empirical $P = 0.0$, Supplementary Fig. 8).

Activity classification analyses. To compare the top-performing prediction models and their ensemble against the single-dose activity assay, the standard confusion matrix was constructed using the measured pK_d values to define the true positive and true negative classes for the 394 pairs in Round 2, using either $pK_d > 6$ or $pK_d > 7$ for indicating true positive activity. The predicted positive and negative classes for the pairs were defined based on either the single-dose activity measurement, using inhibition cut-off of 80%[7,16,19], or the model-predicted pK_d values, using the same activity thresholds as with the measured pK_d values (i.e., either $pK_d = 6$ or $pK_d = 7$). PPV and FDR were calculated as the classification performance scores. The lower threshold of measured $pK_d = 6$ was used in the classification evaluations to have more balanced true positive and negative classes. To carry out a more systematic analysis of the model prediction accuracies, the 394 pairs in Round 2 were ranked both using the model-predicted pK_d values and the measured single-dose %inhibition values, and then these rankings were compared against the ground-truth activity classification based on the dose-response measurements (using again either $pK_d > 6$ or $pK_d > 7$ for indicating the true positive activity). The results were visualized using both ROC and PR curves, implemented in the pROC and pRROC R-packages, respectively[45,46]. The area under the ROC curve (AU-ROC) and PR curve (PR-AUC) were calculated as summary classification performance metrics.

Class enrichment analyses. For each of the 394 compound-kinase pairs from the Round 2 test set, we calculated an AE (i.e., residual errors between predicted and measured pK_d values) considering (i) 90 out of all 99 submissions with average AE below 2, (ii) Spearman correlation-based mean aggregation ensemble model, and (iii) the best submission from the top-performing Q.E.D team. We computed median AE across 90 submissions and, in each case (i–iii), we ranked all the compound-kinase pairs according to their AE (from highest to lowest AE). To explore whether any of the pre-defined kinase classes were enriched among the predictions with the highest or lowest AE, we applied the enrichment analysis implemented in the clusterProfiler R package[47]. In this tool, the enrichment P values were calculated based on a weighted Kolmogorov–Smirnov-like statistic, similar to gene set enrichment analysis (GSEA). We considered the classes defined based on kinase families and kinase groups.

PubMed literature scan. A total of 959 abstracts of drug-target interaction prediction publications were extracted from PubMed (on 16 February 2021) using easyPubMed R package[48] with the following query: ((“compound target”) OR (“target affinity”) OR (“drug target”) OR (“binding affinity”)) AND ((“prediction”) OR (“algorithm”)) AND (“computational”) NOT (review[Publication Type]) NOT (news[Publication Type]) NOT (newspaper article[Publication Type]) NOT (systematic review[Publication Type]) NOT (editorial[Publication Type]). textmineR[49] and SnowballC[50] R packages were used to convert all words in the abstracts to lowercase, remove punctuation, numbers and stop words, as well as perform stemming. Next, 4847 n-grams of size up to three and occurring in at least five abstracts were extracted and manually curated to keep only n-grams related to machine learning methods (e.g., deep_neural, deep_learn, kernel_base) and problem classes (e.g., classif_model, regress_model, supervis_learn). Finally, the resulting n-grams were grouped (e.g., both deep_neural and deep_learn bigrams represent deep learning methods), and the various modeling approaches used by the Challenge teams were mapped into the approaches based on the literature scan. A co-occurrence graph of the problem classes and machine learning methods was created using the igraph[51] R package.

Existing target prediction methods. We applied the online SEA web-application (http://sea.bkslab.org/search) to make target predictions for the three compounds highlighted in the revised manuscript, TPKI-30, GSK1379763 and PFE-PKIS14, for which Q.E.D model-predicted strong activity against DDR1, PYK2 (PTK2B) and CSNK2A2 ($pK_d > 6$), and which were experimentally validated post-Challenge. In the SEA method, we used the ECFP4 fingerprints that were also used by the top-performing prediction models in the Challenge (see Table 1).

To model the interaction between TPKI-30 and PYK2 (PDB entry 5TO8 [https://doi.org/10.2210/pdb5TO8/pdb]), we carried out binding affinity predictions of various active ligands with docking study in terms of their measured

pK_d/pK_i activity values. The docking was done with AutoDock Vina[52]. The X-ray crystal structure of protein PYK2 (PDB entry 5TO8 [https://doi.org/10.2210/pdb5TO8/pdb]) was obtained from RCSB[53], and a collection of 26 compounds (including TPKI-30), with potent activity towards PYK2 (i.e., $pK_d/pK_i > 6$) from ChEMBL[11], BindingDB[12], and DTC[14], were used as ligands in the docking procedure.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Challenge Round 1 and Round 2 pK_d datasets are available from DrugTargetCommons (https://drugtargetcommons.fimm.fi/), and in Supplementary Data 1. The pK_d values of additional compound-target pairs selected for post-Challenge DiscoverX profiling are available in Supplementary Data 2. Source data underlying the figures and display items are provided on Zenodo[54] (subdirectory source_data) and with this paper as a Source Data file. The study made use of the following publicly available databases: Druggable Genome (IDG) consortium (https://druggablegenome.net/), ChEMBL (https://www.ebi.ac.uk/chembl/), BindingDB (https://www.bindingdb.org), IDG Pharos (https://pharos.nih.gov/), DrugTargetCommons (https://drugtargetcommons.fimm.fi/), Synapse (https://www.synapse.org/). The crystal structure to model the interaction between TPKI-30 and PYK2 was obtained from the RCSB PDB (https://www.rcsb.org/) with the PDB code 5TO8 [https://doi.org/10.2210/pdb5TO8/pdb]. Source data are provided with this paper.

Code availability

The Docker containers of the top-performing teams are available on Synapse[55]. Please refer to the Synapse.org documentation (https://docs.synapse.org/articles/docker.html) for guidance on using the Synapse Docker repository. The codes for reproducing the results and figures are available at GitHub (https://github.com/Sage-Bionetworks/IDG-DREAM-Challenge-Analysis/) and archived in Zenodo[54]. Key R packages used beyond those mentioned elsewhere in Methods include tidyverse[56] and the Synapse Python Client (https://github.com/Sage-Bionetworks/synapsePythonClient); packages used and their versions are listed in the renv lockfile in the Github and Zenodo repositories.

Received: 20 June 2020; Accepted: 15 April 2021;
Published online: 03 June 2021

References

1. Oprea, T. I. et al. Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).
2. Arrowsmith, C. H. et al. The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536–541 (2015).
3. Santos, R. et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
4. Dar, A. C., Das, T. K., Shokat, K. M. & Cagan, R. L. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486**, 80–84 (2012).
5. Fabian, M. A. et al. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23**, 329–336 (2005).
6. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
7. Elkins, J. M. et al. Comprehensive characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **34**, 95–103 (2016).
8. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682–690 (2008).
9. Schlessinger, A. et al. Multi-targeting Drug Community Challenge. *Cell Chem Biol* **24**, 1434–1435 (2017).
10. Azencott, C.-A. et al. The inconvenience of data of convenience: computational research beyond post-mortem analyses. *Nat. Methods* **14**, 937–938 (2017).
11. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
12. Gilson, M. K. et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
13. Nguyen, D.-T. et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).
14. Tang, J. et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell Chem. Biol.* **25**, 224–229.e2 (2018).
15. Omberg, L. et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat. Genet.* **45**, 1121–1126 (2013).
16. Drewry, D. H. et al. Progress towards a public chemogenomic set for protein kinases and a call for contributions. *PLoS One* **12**, e0181585 (2017).
17. Cichonska, A. et al. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLOS Comput.Biol.* **13**, e1005678 (2017).
18. Zhao, Y. & Adjei, A. A. The clinical development of MEK inhibitors. *Nat. Rev. Clin. Oncol.* **11**, 385–400 (2014).
19. Wells, C. I., Kapadia, N. R., Couñago, R. M. & Drewry, D. H. In depth analysis of kinase cross screening data to identify chemical starting points for inhibition of the nek family of kinases. https://doi.org/10.1101/137968.
20. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206 (2007).
21. Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **43**, 374–380 (2003).
22. Reymond, J.-L. & Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **3**, 649–657 (2012).
23. Berginski, M. E. et al. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Res.* **49**, D529–D535 (2021).
24. Mayr, A. et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).
25. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
26. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
27. Wen, M. et al. Deep-learning-based drug–target interaction prediction. *J. Proteom. Res.* **16**, 1401–1409 (2017).
28. You, J., McLeod, R. D. & Hu, P. Predicting drug-target interaction network using deep learning model. *Comput. Biol. Chem.* **80**, 90–101 (2019).
29. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329–3338 (2019).
30. Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nature Methods* **9**, 796–804 (2012).
31. Eduati, F. et al. Prediction of human population responses to toxic compounds by a collaborative competition. *Nature Biotechnol.* **33**, 933–940 (2015).
32. Saez-Rodriguez, J. et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Rev. Genet.* **17**, 470–486 (2016).
33. Guinney, J. & Saez-Rodriguez, J. Alternative models for sharing confidential biomedical data. *Nature Biotechnol.* **36**, 391–392 (2018).
34. Neto, E. C. et al. Reducing overfitting in challenge-based competitions. arXiv [stat.AP] (2016).
35. Pahikkala, T. et al. Toward more realistic drug-target interaction predictions. *Brief. Bioinform.* **16**, 325–337 (2015).
36. Diedenhofen, B. & Musch, J. cocor: a comprehensive solution for the statistical comparison of correlations. *PLOS ONE* **10**, e0121945 (2015).
37. Guha, R. Chemical Informatics Functionality inR. *J. Stat. Softw.* **18**, 1–16 (2007).
38. Airola, A. & Pahikkala, T. Fast Kronecker Product Kernel Methods via Generalized Vec Trick. *IEEE Trans Neural Netw Learn Syst* **29**, 3374–3387 (2018).
39. Allaway, R. The IDG-DREAM Drug Kinase Binding Prediction Challenge Community. The IDG-DREAM drug kinase binding prediction challenge community method writeups. (2019) https://doi.org/10.7303/SYN21445941.1.
40. Pahikkala, T. & Airola, A. RLScore: Regularized Least-Squares Learners. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
41. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: An SIMD Smith-Waterman C/C Library for Use in Genomic Applications. *PLoS ONE* **8**, e82138 (2013).
42. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inform. Model.* **50**, 742–754 (2010).
43. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. arXiv [cs.LG] (2019).
44. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794 (Association for Computing Machinery, 2016).
45. Robin, X. et al. pROC: an open-source package for R and S to analyze and compare ROC curves. BMC Bioinform. **12**, 77 (2011).
46. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
47. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

48. Fantini, D. easyPubMed: Search and retrieve scientific publication records from PubMed. R package version 2.13. (2019). https://cran.rproject.org/package=easyPubMed.
49. Jones, T. & Doane, W. textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.4. (2019). https://cran.rproject.org/package=textmineR.
50. Bouchet-Valat, M. SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. R package version 0.7.0. (2020). https://cran.rproject.org/package=SnowballC.
51. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter J. Complex Syst.* **1695**, 1–9 (2006).
52. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
53. Burley, S. K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).
54. Cichońska, A. et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. Zenodo. https://doi.org/10.5281/ZENODO.4648011 (2021).
55. The IDG-DREAM Drug-Kinase Binding Prediction Challenge Community. IDG-DREAM drug-kinase binding prediction challenge. Synapse. https://doi.org/10.7303/SYN15667962 (2018).
56. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

Acknowledgements

The authors thank the IDG Kinase Data and Resource Generation Center for generating new sets of target activity data for the Challenge Rounds 1 and 2, Olle Hansson (FIMM) for technical assistance with DrugTargetCommons platform, Tianduanyi Wang (FIMM) for his help with the baseline submissions, Anna Goldenberg (University of Toronto, Canada) and Chloe-Agathe Azencott (Institut Curie, France) for organizing the DREAM Idea Challenge, and Barbara Rieck and Ladan Naghavian for the bioactivity profiling at DiscoverX (Eurofins Corporation). T.A. acknowledges support from the Academy of Finland (grants 310507, 313267, 326238), Cancer Research UK and the Brain Tumour Charity (grant REF: C42454/A28596), and Helse Sør-Øst (grant No. 2020026). C.W., T. W., D.D. acknowledge support from the National Institutes of Health (1U24DK116204-01). The SGC is a registered charity that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (ULTRA-DD 115766), Wellcome Trust, Janssen, Merck Kga, Merck Sharp & Dohme, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, and Takeda. O.I. acknowledges support from the National Science Foundation (NSF CHE-1802789 and CHE-2041108), and Eshelman Institute for Innovation (EII) awards. O.I. thanks the OpenEye Free Academic Licensing Program for providing a free academic license for their chemistry toolkit. M.P. acknowledges support from The Molecular Sciences Software Institute (MolSSI) Software Fellowship and NVIDIA Graduate Fellowship. We gratefully acknowledge the support and hardware donation from NVIDIA Corporation. J.G. acknowledges support from the National Institutes of Health (U54OD020353). T.I.O. acknowledges support from the National Institutes of Health (U24CA224370; U24TR002278; U01CA239108).

Author contributions

Conceptualization: A.C., B.R., R.J.A., K.D., A.S., D.H.D., G.S., K.W., J.G., T.A.; data curation: A.C., B.R., R.J.A., A.L., C.I.W., T.M.W., D.H.D.; formal analysis: A.C., B.R., R.J.A., F.W. S.P., O.I., S.L., M.M., Z.T., M.J., S.K., M.P., S.C., J.Z., T.A.; funding acquisition: O.I., M.P., C.W., T.M.W., T.I.O., D.H.D., K.W., J.G., T.A.; investigation: A.C., B.R., R.J.A., C.I.W., D.H.D., T.A.; methodology: A.C., B.R., R.J.A., F.W., S.P., O.I., S.L., M.M., A.L., Z.T., M.J., S.K., M.P., S.C., J.Z., K.D., G.K., J.K., T.A.; project administration: T.I.O. D.H.D., G.S., J.G., T.A.; resources: A.C., B.R., R.J.A., A.L.; software: A.C., B.R., R.J.A., A.L.; supervision: K.W., J.G., T.A.; validation: A.C. B.R., R.J.A., D.H.D., K.W., T.A.; visualization: A.C., B.R., R.J.A., T.A.; writing—original draft: A.C., B.R., R.J.A., F.W., S.P., O.I., S.L. M.M., A.L., Z.T., M.J., S.K., M.P., S.C., J.Z., K.D., G.K., J.K., C.I.W. T.M.W., T.I.O, A.S., D.H.D., G.S., K.W., J.G., T.A.; writing—review and editing: A.C., B.R., R.J.A., O.I., T.I.O., A.S., K.W., J.G., TA.

Competing interests

The SGC is a registered charity that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (ULTRA-DD 115766), Wellcome Trust, Janssen, Merck Kga, Merck Sharp & Dohme, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, and Takeda. T.I.O. has received honoraria or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth. J.Z. is founder and CTO of Silexon AI Technology Co. Ltd. and has an equity interest. The rest of the authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23165-1.

Correspondence and requests for materials should be addressed to K.W., J.G. or T.A.

Peer review information *Nature Communications* thanks Jeffrey Peterson and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland. [2]Department of Computer Science, Helsinki Institute for Information Technology (HIIT), Aalto University, Espoo, Finland. [3]Department of Computing, University of Turku, Turku, Finland. [4]Computational Oncology, Sage Bionetworks, Seattle, WA, USA. [5]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [6]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea. [7]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA. [8]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA. [9]Immuneering Corporation, Cambridge, MA, USA. [10]Structural Genomics Consortium, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA. [11]Translational Informatics Division and Comprehensive Cancer Center, University of New Mexico School of Medicine, Albuquerque, NM, USA. [12]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [13]IBM T J Watson Research Center, IBM, Yorktown Heights, NY, USA. [14]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. [15]Department of Mathematics and Statistics, University of Turku, Turku, Finland. [16]Institute for Cancer Research, Oslo University Hospital, Oslo, Norway. [17]Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Oslo, Norway. [18]Dept. of Computer Engineering, TOBB University of

Economics and Technology, Ankara, Turkey. [19]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [20]Department of Biomedical Engineering, University of California, Davis, CA, USA. [21]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [22]Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, OH, USA. [23]Department of Computer Engineering, VIIT, Pune, India. [24]Applied Artificial Intelligence Institute, Deakin University, Geelong, VIC, Australia. [25]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. [26]Microsoft, One Microsoft Way, Redmond, WA, USA. [27]The University of Pennsylvania, Philadelphia, PA, USA. [28]Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA. [29]The University of Texas at Austin, Austin, TX, USA. [30]Department of Computer Engineering, Bogazici University, Istanbul, Turkey. [31]Department of Chemical Engineering, Bogazici University, Istanbul, Turkey. [32]Department of Biochemistry, Graduate Center, The City University of New York, New York, NY, USA. [33]Department of Computer Science, Hunter College, The City University of New York, New York, NY, USA. [34]Department of Neurosurgery, Cancer Center Amsterdam CCA, De Boelelaan 1117, Amsterdam, The Netherlands. [35]Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark. [36]BioSys Lab, National Technical University of Athens, Athens, Greece. [37]Department of Biotechnology, Ghent University, Ghent, Belgium. [38]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium. [39]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [40]School of Medicine, Tsinghua University, Beijing, China. [41]Department of Computer and AI Engineering, Hacettepe University, Ankara, Turkey. [42]Department of Computer Engineering, METU, Ankara, Turkey. [43]KanSiL, Department of Health Informatics Graduate School of Informatics, METU, Ankara, Turkey. [44]European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK. [45]Semmelweis University, Department of Physiology, Budapest, Hungary. [46]Max-Planck-Institute for Molecular Genetics, Department Computational Molecular Biology, Berlin, Germany. [47]University of Potsdam, Department of Computer Science, Potsdam, Germany. [48]MicroDiscovery GmbH, Berlin, Germany. [49]Division of Electronics, Ruđer Bošković Institute, Zagreb, Croatia. [50]NMR Centre, Ruđer Bošković Institute, Zagreb, Croatia. [51]These authors contributed equally: Anna Cichońska, Balaguru Ravikumar, Robert J. Allaway. [52]These authors jointly supervised this work: Krister Wennerberg, Justin Guinney, Tero Aittokallio. *A list of authors and their affiliations appears at the end of the paper. ✉email: krister.wennerberg@bric.ku.dk; jguinney@gmail.com; tero.aittokallio@helsinki.fi

The IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium

User oselot Mehmet Tan[18]

Team N121 Chih-Han Huang[19], Edward S. C. Shih[19], Tsai-Min Chen[19], Chih-Hsun Wu[19], Wei-Quan Fang[19], Jhih-Yu Chen[19] & Ming-Jing Hwang[19]

Team Let_Data_Talk Xiaokang Wang[20], Marouen Ben Guebila[21], Behrouz Shamsaei[22] & Sourav Singh[23]

User thinng Thin Nguyen[24]

Team KKT Mostafa Karimi[25,26], Di Wu[25,27], Zhangyang Wang[28,29] & Yang Shen[25]

Team Boun Hakime Öztürk[30], Elif Ozkirimli[31] & Arzucan Özgür[30]

Team KinaseHunter Hansaim Lim[32] & Lei Xie[33]

Team AmsterdamUMC-KU-team Georgi K. Kanev[34], Albert J. Kooistra[35] & Bart A. Westerman[34]

Team DruginaseLearning Panagiotis Terzopoulos[36], Konstantinos Ntagiantas[36], Christos Fotis[36] & Leonidas Alexopoulos[36]

Team KERMIT-LAB - Ghent University Dimitri Boeckaerts[37], Michiel Stock[38], Bernard De Baets[38] & Yves Briers[37]

Team QED Fangping Wan[5], Shuya Li[5] & Yunan Luo[39], Hailin Hu[40], Jian Peng[39] & Jianyang Zeng[5]

Team METU_EMBLEBI_CROssBAR Tunca Dogan[41], Ahmet S. Rifaioglu[42], Heval Atas[43], Rengul Cetin Atalay[43], Volkan Atalay[42] & Maria J. Martin[44]

Team DMIS_DK Sungjoon Park[6], Minji Jeon[6], Sunkyu Kim[6] & Junhyun Lee[6], Seongjun Yun[6], Bumsoo Kim[6], Buru Chang[6] & Jaewoo Kang[6]

Team AI Winter is Coming Mariya Popova[7], Stephen Capuzzi[8] & Olexandr Isayev[7]

Team hulab Gábor Turu[45], Ádám Misák[45], Bence Szalai[45] & László Hunyady[45]

Team ML-Med Matthias Lienhard[46], Paul Prasse[47], Ivo Bachmann[48], Julia Ganzlin[47], Gal Barel[46] & Ralf Herwig[46]

Team Prospectors Davor Oršolić[49], Bono Lučić[50], Višnja Stepanić[49] & Tomislav Šmuc[49]

Challenge organizers Anna Cichońska[1,2,3,51], Balaguru Ravikumar[1,51], Robert J. Allaway[4,51], Michael Mason[4], Andrew Lamb[4], Ziaurrehman Tanoli[1], Kristen Dang[4], Carrow I. Wells[10], Timothy M. Willson[10], Tudor I. Oprea[11], Avner Schlessinger[12], David H. Drewry[10], Gustavo Stolovitzky[13], Krister Wennerberg[14,52]✉, Justin Guinney[4,52]✉ & Tero Aittokallio[1,2,15,16,17,52]✉

3.3 PAPER 3: Dynamic applicability domain (dAD): compound-target binding affinity estimates with local conformal prediction

Oršolić, D., Šmuc, T., 2023. Dynamic applicability domain (dAD): compound–target binding affinity estimates with local conformal prediction. Bioinformatics 39, btad465. https://doi.org/10.1093/bioinformatics/btad465

Bioinformatics, 2023, **39(8)**, btad465
https://doi.org/10.1093/bioinformatics/btad465
Advance access publication 18 August 2023
Original Paper

OXFORD

Systems biology

Dynamic applicability domain (dAD): compound–target binding affinity estimates with local conformal prediction

Davor Oršolić [1] **and Tomislav Šmuc** [1,*]

[1]Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, Zagreb 10000, Croatia

*Corresponding author. Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia. E-mail: tomislav.smuc@irb.hr (T.Š)

Associate Editor: Pier Luigi Martelli

Abstract

Motivation: Increasing efforts are being made in the field of machine learning to advance the learning of robust and accurate models from experimentally measured data and enable more efficient drug discovery processes. The prediction of binding affinity is one of the most frequent tasks of compound bioactivity modelling. Learned models for binding affinity prediction are assessed by their average performance on unseen samples, but point predictions are typically not provided with a rigorous confidence assessment. Approaches, such as the conformal predictor framework equip conventional models with a more rigorous assessment of confidence for individual point predictions. In this article, we extend the inductive conformal prediction framework for interaction data, in particular the compound–target binding affinity prediction task. The new framework is based on dynamically defined calibration sets that are specific for each testing pair and provides prediction assessment in the context of calibration pairs from its compound–target neighbourhood, enabling improved estimates based on the local properties of the prediction model.

Results: The effectiveness of the approach is benchmarked on several publicly available datasets and tested in realistic use-case scenarios with increasing levels of difficulty on a complex compound–target binding affinity space. We demonstrate that in such scenarios, novel approach combining applicability domain paradigm with conformal prediction framework, produces superior confidence assessment with valid and more informative prediction regions compared to other 'state-of-the-art' conformal prediction approaches.

Availability and implementation: Dataset and the code are available on GitHub (https://github.com/mlkr-rbi/dAD).

1 Introduction

The fast growth of experimental results published in the scientific literature and through repositories makes modelling of binding affinity between compounds and protein targets expanding and interesting from both scientific and industrial aspects. Methods for *in silico* modelling and screening of large chemical compound spaces are often computational pipelines based on feature generation tools and machine-learning algorithms (Cichonska *et al.* 2017, Öztürk *et al.* 2018, Cichońska *et al.* 2021, Nguyen *et al.* 2021). Compound and target spaces can be described with a multitude of different descriptors, including those that describe their structural or physicochemical properties (Lim *et al.* 2021). In Öztürk *et al.* (2018), authors used convolutional neural networks to learn small molecule and protein representations from 1D sequences. On the other hand, to improve the predictive power of the model with a more realistic representation of molecules, the GraphDTA method was proposed in Nguyen *et al.* (2021). This approach is based on a graph convolutional block that learns compound representations from a molecular graph (Kipf and Welling 2016). In QSAR modelling, predictive models must not only aim for high accuracy on unseen samples, but they must also be accompanied by estimations of the prediction region with a certain degree of confidence. Conventional QSAR modelling uses applicability domain (AD) to improve prediction credibility (Gadaleta *et al.* 2016). AD of the QSAR model represents a bounded chemical space within which the model is guaranteed a well-defined and reliable performance on average (Aniceto *et al.* 2016, Mathea *et al.* 2016, Klingspohn *et al.* 2017). However, the concept of AD does not provide an apparatus that would determine how reliable certain model predictions are (Aniceto *et al.* 2016). When using AD, the user determines the portion of external data that falls within the established boundaries, without assessing the AD's ability to differentiate between 'acceptable' and 'unacceptable' new predictions (Aniceto *et al.* 2016, Mathea *et al.* 2016). Intuitively, the AD increases the confidence of the model's predictions, but this is not directly quantified. According to Aniceto *et al.* (2016), AD is set using training sample similarity thresholds or class probability estimates, etc. This approach treats AD as a space between the defined limits typically overlooking the possibility of localized holes in the chemical space where the model's predictions may be unreliable (Klingspohn *et al.* 2017). Furthermore, a well-balanced AD formulation would need to include information on distribution of both entities when dealing with interacting pairs.

Conformal prediction (CP) framework was introduced by Gammerman *et al.* (1998), with the intention of providing confidence for classification predictions made by the support vector machines. First attempt at CP was made using transductive conformal predictors, which required retraining of the model for each individual prediction and were therefore

Received: August 30, 2022. **Revised:** April 26, 2023. **Editorial Decision:** July 16, 2023

computationally expensive (Shafer and Vovk 2008). As a result, the inductive conformal prediction (ICP) framework was developed (Papadopoulos 2008). The ICP framework uses a calibration set, that typically accounts for a smaller portion of the training samples, to calibrate the trained model and compute confidence levels. The disadvantage of taking such an approach for regression problems is that the prediction regions computed based on the calibration set for any particular level of confidence are fixed, meaning that they do not alter from test sample to test sample (Johansson *et al.* 2014). In the CP framework, AD related measures (Klingspohn *et al.* 2017) can also be used for conformity scoring to locate the AD or the calibration set, for each individual test sample. We call this set a conformity region of sample x.

The typical non-conformity function that is utilized for regression tasks is the absolute difference between the true label and the predicted label for a particular sample, $|y_i - \hat{y}_i|$ (Shafer and Vovk 2008, Papadopoulos and Haralambous 2010, Papadopoulos *et al.* 2011, Johansson *et al.* 2014). This approach focuses solely on the non-conformity in the label space, and assumes that training and calibration sets must be exchangeable (Shafer and Vovk 2008). For a predefined confidence level prediction regions are fixed, meaning they remain the same for any new test sample tested which prevents them from being as informative as we would like them to be. In order to alleviate this, later studies introduced normalization steps in the context of non-conformity function definition (Papadopoulos and Haralambous 2010, Papadopoulos *et al.* 2011).

In this article, we introduce the CP framework that is better suited for the interaction data character of the compound–target binding affinity modelling task. We rationalize that binding affinity modelling task, or for that matter, any dataset containing interactions between two entities, requires special treatment when defining non-conformity in the space of interacting pairs. For binding affinity modelling specifically, binding affinities are conditioned on input data that comes from two distinct distributions: one of compounds and one of targets. Conformal predictors based solely on predictions, i.e. distribution of errors in the label space, cannot reflect the true non-conformity of the sample in the space of predictor variables. We thus aim to expand the CP framework for this type of problem by introducing the concept of the dynamic calibration set, the calibration set that is specific for the particular compound–target pair being tested. The 'localised' calibration based on the dynamic calibration set should provide more precise non-conformity scores for tested samples and also better performance in cases when testing samples are found outside the boundaries of AD or out-of-distribution.

We test our method and compare it against other 'state-of-the-art' approaches over several benchmark compound–target binding affinity datasets, as well as a specifically designed small compound–protein kinase binding affinity dataset that is constructed to allow the testing of the conformal predictor frameworks in settings that are representative of the real use-case scenarios.

2 Proposed CP framework

The initial step in a conventional ICP framework is to divide the training set into a proper training set and a calibration set (Shafer and Vovk 2008, Alvarsson *et al.* 2021). The calibration set must reflect the distribution of the training samples, satisfying the assumption that the data are independent and identically distributed, or a more relaxed assumption that they are exchangeable (Shafer and Vovk 2008, Papadopoulos *et al.* 2011). Contrary to the conventional calibration set definition, which is stationary and represents the overall training space, we define a dynamic calibration set for each individual test sample by locating the most conforming samples of the training landscape to the sample that is being tested. Let us denote the training set of compound–target pairs as:

$$Z = (x_1, y_1), \ldots, (x_m, y_m), \tag{1}$$

where x_i is a compound–target pair and y_i is a measured binding affinity for that pair. Let the set of compounds and targets be denoted by C and T, respectively, and the corresponding set of training compound–target pairs by $X = (C, T)$.

Calibration set (Z^c) of a new test sample is dynamically constructed from training samples that have maximum Tanimoto similarity coefficients (s) towards the tested compound–target pair. Let $x = (c, t)$ be the new test sample for which we choose a subset of X by retrieving $C \subset C^t$, $|C| = k$, such that $s(c^{(i)}, c) \geq s(c^{(j)}, c)$ holds $\forall c^{(i)} \in C$ and $\forall c^{(j)} \in C^t$ C. Equivalently, we define $T \subset T^t$, a subset of targets such that $|T| = q$. k and q are tuneable hyperparameters, determining the neighbourhood of the tested $x = (c, t)$ pair in the training set Z. Dynamic calibration set for the new test instance x is then defined as:

$$Z^c = \{(x^{(ij)}, y^{(ij)}) : x^{(ij)} \subset (C, T) \, and \, \exists y^{(ij)} \subset Y^t\}, \tag{2}$$

where each $x^{(ij)}$ is actually a tuple $(c^{(i)}, t^{(j)})$. The dynamical calibration set defined in this manner represents the most conforming part of the training set bioactivity space with respect to the tested compound–target pair. To allow forming dynamic calibration sets from training samples, we train the model over the entire dataset by applying 10×10-fold cross-validation and calculate non-conformity scores for each training sample as a difference of the mean of cross-validation predictions and the true labels (Vovk *et al.* 2018). We call the proposed approach dynamic applicability domain (dAD). In the following section, we define the two alternative non-conformity scores.

2.1 Definition of calibration and test non-conformity scores

In this work, we define and test two alternative formulations of non-conformity scores for the compound–target pairs from the dynamic calibration set. The first variant, dAD (NN), non-conformity score α_i^{nn} (3) is calculated by taking the difference between the experimental binding affinity of each pair in the Z^c and mean label value for all pairs in Z^c. In alternative formulation, dAD (CV), non-conformity score α_i^{cv} (4) is based on a difference between the experimental binding affinity from the mean of 10×10-fold cross-validation predictions for each pair. For the test sample, the putative non-conformity scores are defined as the difference of the predicted label ($\hat{y}$) and each experimental compound–target pair binding affinity in the dynamic calibration set, α^x (5). Thus, for every new test instance x we get the corresponding calibration non-conformity vectors, S^{cv} or S^{nn} and its own vector of non-conformity scores, S^x.

Definitions of non-conformity scores associated with the test sample and its dynamic calibration set are given below:

$$\alpha_i^{cal} = \alpha_i^{nn} = |y_i^{cal} - \hat{y}_i^{nn}|, \alpha_i^{nn} \in S^{nn}, \quad (3)$$

$$\alpha_i^{cal} = \alpha_i^{cv} = |y_i^{cal} - \hat{y}_i^{cv}|, \alpha_i^{cv} \in S^{cv}, \quad (4)$$

$$\alpha_i^{x} = |y_i^{cal} - \hat{y}^{x}|, \alpha_i^{x} \in S^{x}, \quad (5)$$

where y_i^{cal} is a true value for i'th sample in calibration set; $\hat{y}_i^{nn}$ and $\hat{y}_i^{cv}$ are predictions for the i'th sample in dAD (NN) and dAD (CV) approaches, respectively; $\hat{y}^x$ is a prediction for test sample x.

In the next step, we have to find the true prediction region for a predefined confidence level for the test sample x. For that purpose, minimal value from the calibration non-conformity scores α_i^{cal} is found, for which the expression below holds as in Shafer and Vovk (2008) and Johansson *et al.* (2014):

$$\frac{\#\{z_i \in Z^c | \alpha^x \leq \alpha_i^{cal}\}}{N_{Z^c}} \geq 1 - \delta, \quad (6)$$

where $z_i \in Z^c$ are samples from the calibration set. We annotate minimal value of α_i^{cal} for which above expression (6) holds, α_δ^{min}. Then, the prediction region for any new example is defined as $\Gamma_x^\delta = \hat{y} \pm \alpha_\delta^{min}$. As one can notice, for any predefined level of confidence, dAD produced prediction region varies between test samples as the calibration sets are sample specific.

Chosen α_δ^{min} represents the partition of samples in the $S = S^{cv}$ or $S = S^{nn}$ that have higher non-conformity scores than any given $\alpha_i^x \in S^x$. Since tentative labels for each tested pair x are based on dynamic calibration set samples, and reflect the local model performance, no normalization step is necessary and individual prediction regions are directly inferred from α_δ^{min}. The pseudocode of the algorithm for the construction of dynamic calibration set and calculation of non-conformity scores is given in Supplementary Algorithm S1.

2.2 Normalized non-conformity measures

Approaches like Papadopoulos (2008), Johansson *et al.* (2014) and Alvarsson *et al.* (2021), rely on fixed calibration sets and for that reason prediction regions these methods output are not sample specific and do not reflect the local non-conformity of test samples. In order to achieve more informative prediction regions for an individual test sample Papadopoulos *et al.* (2011) and Papadopoulos and Haralambous (2010) introduced different normalization measures.

In Shafer and Vovk (2008), non-conformity score (α) is calculated as the absolute difference of the predicted and the true value, Equation (7) in Table 1. Normalization approach as in Equation (8) requires an extra model to estimate the accuracy of individual predictions, μ_i. On the other hand, in Equation (9), non-conformity scores are normalized by dividing it with a factor representing normalized distances of nearest neighbours (λ_i^k)—or by ξ_i^k, factor representing normalized standard deviation of nearest neighbours of sample x, Equation (10).

3 Data

Validity of the proposed approach, is tested over several publicly available databases as shown in Table 2. Furthermore,

Table 1. Non-conformity measure (α) as used per four different reference studies, with included normalization coefficients.[a]

Reference	Non-conformity measure	Normalization coefficient	Eq.
Shafer and Vovk (2008)	$\alpha_i = \lvert y_i - \hat{y}_i \rvert$		(7)
Papadopoulos and Haralambous (2010)	$\alpha_i = \frac{\lvert y_i - \hat{y}_i \rvert}{\exp(\mu_i)}$	$\exp(\mu_i)$	(8)
Papadopoulos *et al.* (2011)	$\alpha_i = \lvert \frac{y_i - \hat{y}_i}{\gamma + \lambda_i^k} \rvert$	λ_i^k	(9)
Papadopoulos *et al.* (2011)	$\alpha_i = \lvert \frac{y_i - \hat{y}_i}{\gamma + \xi_i^k} \rvert$	ξ_i^k	(10)

[a] $\exp(\mu_i)$ represents model accuracy estimate ensuring always a positive value; λ_i^k coefficient is based on distances; and ξ_i^k is based on standard deviation of sample x_i and its k nearest neighbours. γ is a sensitivity parameter in control of the sensitivity to changes in both λ_i^k and ξ_i^k measures.

Table 2. Publicly available databases used for compound–target binding affinity prediction, including Davis (Davis *et al.* 2011), KIBA (Tang *et al.* 2014), BindingDB (Gilson *et al.* 2016), ChEMBL (Gaulton *et al.* 2012), and DTC (Tang *et al.* 2018) and a SCKBA dataset, we constructed from the mentioned databases for the purposes of a unified representation of this specific bioactivity space.

Dataset	#cmpds	#trgts	#int
Davis	68	379	27 621
KIBA	2068	229	118 036
BindingDB	10 968	311	25 674
ChEMBL	11 637	235	51 360
DTC (GPCR)	1681	119	17 245
DTC (SSRI)	3640	49	19 046
SCKBA	7860	210	43 433

we combine all mentioned datasets into a single dataset covering larger space of kinase inhibitors over human kinome space. Aside from the kinase inhibitors, we also retrieve the G protein-coupled receptors (GPCR) and selective serotonin reuptake inhibitor (SSRI) subsets from the Drug Target Commons (Tang *et al.* 2018) database, in order to test the performance across more diverse bioactivity spaces. Data preprocessing involved several preprocessing filters ensuring the final datasets contained no duplicates and only those bioactivity profiles measured over the human kinome superfamily narrowing the potential protein target space down to nine distinct kinase groups, as given in Fig. 2.

Acquiring larger kinase inhibitor dataset by combining available human kinome centric databases makes it possible to construct several testing scenarios. The final dataset is made sure to contain only small molecules with a molecular weight of ~900 Da and protein targets that are members of the human kinome in order to build a more consistent small compound–kinase binding affinity dataset (SCKBA) with a well-rounded representation of the bioactivity space (Table 2). To increase the number of measured bioactivities both K_d and K_i were used interchangeably, as it was shown in Cichońska *et al.* (2021) that combination of bioactivity types can increase model's overall performance. Finally, to reflect the real machine-learning use-cases with increasing levels of difficulty, the data were distributed between the training set and four different test set scenarios, as it was introduced in Cichonska *et al.* (2017) and Pahikkala *et al.* (2015), with an illustrative example in Fig. 1.

We did chemical space analysis on 7860 compounds using *t*-SNE for the generation of testing sets in this manner (Supplementary Fig. S1). Training compounds, together with

their bioactivity profiles, were collected to assure high coverage of diverse compound scaffolds available in the overall compound set. The quality of the resulting compound clustering was determined by inspecting the maximum common substructures of dense clusters with a threshold of 0.7, which means that the substructure must be present in at least 70% of compounds in the chosen region. We also chose two arbitrary 'soft clusters', in the compound space's middle cloud to provide us a frame of reference when deciding on specificity of common substructures. As shown in Fig. 1, first test scenario (S1) includes samples already seen in the training set, while for the second test scenario (S2), we retrieve compounds from high density regions, data points colour-coded in orange (Supplementary Fig. S1), mostly including compounds with fewer bioactivity profiles, but with many similar compounds retained in the training set. Third test scenario (S3) includes bioactivity profiles chosen based on the kinome space with fewer experimentally measured bioactivities in the overall dataset, consequently selecting compounds mostly contained within the middle cloud of compound samples, colour-coded in green (Supplementary Fig. S1). To avoid the severe reduction of the training set, for the S3 we selected only three well-known protein targets with over 200 measured binding affinities each (Supplementary Fig. S2). These protein targets include phosphoinositide-dependent kinase 1 (PDK1), checkpoint kinase 2 (CHK2), and tropomyosin receptor kinase A (TRKA), belonging to the AGC, Ca^{2+}/calmodulin-dependent protein kinase (CAMK), and tyrosine kinase (TK) groups, respectively. In the fourth scenario (S4), test set contains both compounds and targets that are not present in the training set, ultimately comprising of compounds from a cluster well separated from the rest of chemical space, colour-coded in red (Supplementary Fig. S1). For similar reasons as in S3, in S4 only one protein target is retained as shown in Supplementary Fig. S2, NF-$\kappa\beta$-inducing kinase from the STE group, with binding affinities measured over 600 different compounds. Additionally, the trained model is tested on a full test set (S0) that comprises of all samples from S1 to S4 test scenarios.

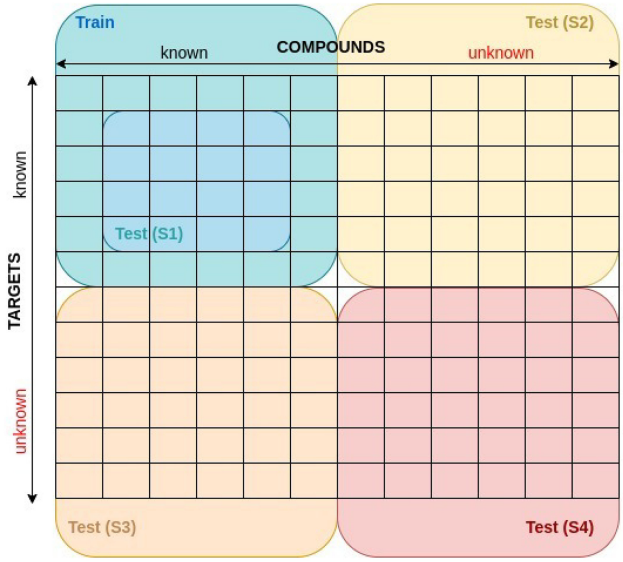


Figure 1. Illustration of test set construction with four difficulty levels, where S1 contains new compound–target pairs and reflects a standard way of testing and evaluating model performance by using stratified sampling and making predictions over known compounds and targets, as it was applied for all datasets in Table 2; S2 contains new compound–target pairs with compounds not available in the training set; S3 contains new compound–target pairs with targets not available in the training set; and S4 contains never seen compounds nor targets. S2–S4 scenarios were only applied for SCKBA dataset.

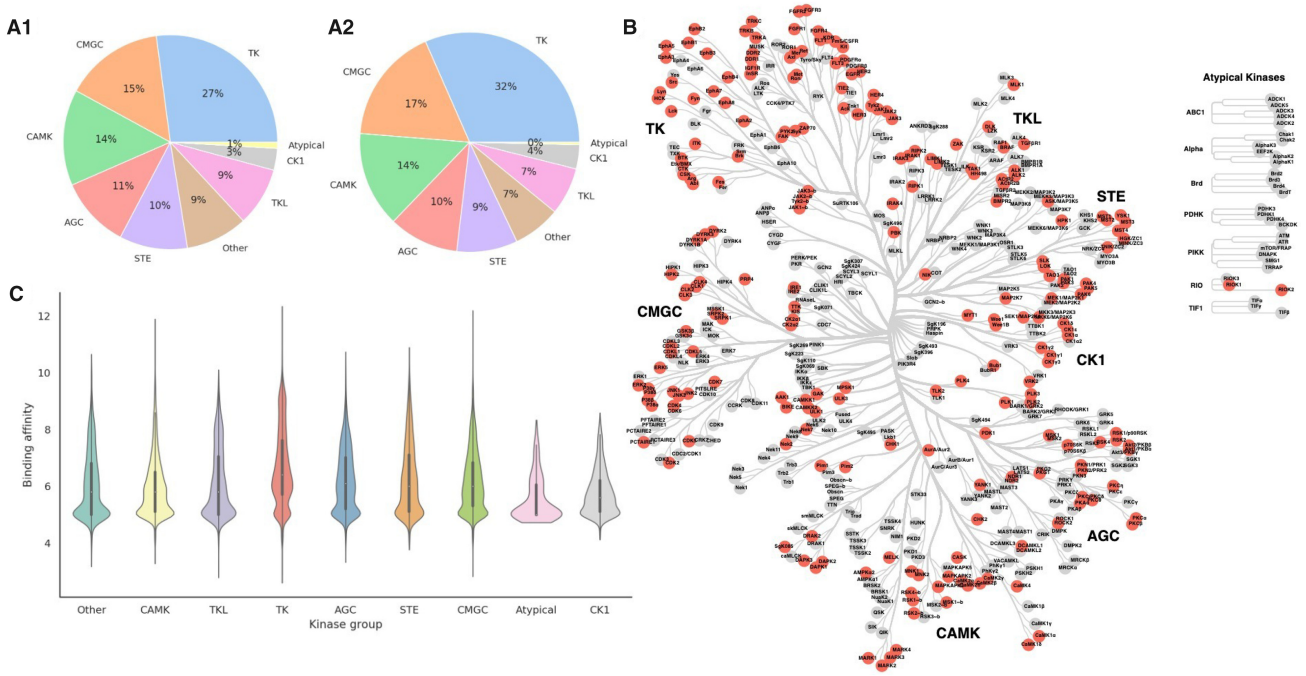


Figure 2. Human kinome targets available in SCKBA dataset. (A1) shows the percentage of protein targets from every kinome group available in the dataset; (A2) similarly to (A1), gives a number of compounds with measured binging affinities across different kinase groups. Complementary to the first two, (B) shows all the selected kinases and affiliated groups (Manning *et al.* 2002, Metz *et al.* 2018), and (C) shows how the experimental binding affinities (pKd and pKi) are distributed per kinase group.

3.1 Compound space representation

Based on the obtained SMILES structures of the chemical space Tanimoto similarity scores are calculated by performing pairwise comparison of the entire compound space based on the 2048-bit Morgan fingerprints with selected radii of 2.

For the training of GCN on the chemical space, molecules are represented as adjacency matrices between atoms, and each atom is represented as a vector of properties. Instead of using one-hot-encoded vector representation (Nguyen *et al.* 2021), we use *rdkit* library (Landrum *et al.* 2020) in Python to compute atomic attributes that include the atomic number, charge, hybridization state, number of radical electrons, number of hydrogen atoms bound, chirality, and ring membership. Deep learning approach is implemented using PyTorch (Paszke *et al.* 2019) and PyTorch Geometric (Fey and Lenssen 2019).

3.2 Target space representation

The human kinome consists of ∼530 enzymes clustered into 10 smaller groups or super-families that share a common evolutionary origin (Manning *et al.* 2002, Roskoski 2015), which is best shown in Fig. 2B with selected kinases given in red. Most importantly, they catalyse phosphorylation and are included in the most important regulatory mechanisms in all living organisms (Liao 2007, Roskoski 2015). To get a better feeling for the protein targets available in the SCKBA dataset, we can check their placement in the human kinome phylogenetic tree (Fig. 2B), same as the number of kinases in any of nine dedicated kinase groups in the collected dataset (Fig. 2A1) or number of measured binding affinities per kinase group (Fig. 2A2), which tells us just how prevalent certain kinase groups are as targets, e.g. TK group, CMGC group including mostly proline-directed serine/threonine kinases, and CAMK group comprise over 50% of all protein targets in our dataset. Moreover, Fig. 2C depicts the distribution of experimentally measured binding affinities across kinase groups.

Local similarities between these protein targets are computed by applying the Smith–Waterman algorithm to the protein kinase (PK) sequences with default parameters of the *protr* library in *R* (Xiao *et al.* 2015) (gap.opening = 10, gap.extension = 4) and 'BLOSUM62' substitution matrix. Same approach was performed for computation of sequence similarities of the GPCR and SSRI datasets, retrieved from the DrugTargetCommons database (Tang *et al.* 2018). Specifically for the SCKBA dataset, we compute local similarities only for the PK domain, because it is highly conserved and is a major focus for the small molecule inhibitor design, with most of the approved drugs targeting exactly ATP-binding cleft or the surrounding regions (Manning *et al.* 2002, Liao 2007, Roskoski 2015). For the training of deep learning architecture (GCN–CNN) on SCKBA dataset, the same strategy is adopted as in Nguyen *et al.* (2021), treating protein targets as sequences of characters within a CNN block, with the difference of learning sequence representations by feeding the PK domains into the network, instead of using the whole protein sequences, for reasons explained above. We identified the longest PK domain sequence in our dataset and applied zero padding to other targets to match the length of the identified sequence.

4 Materials and methods

XGBoost is trained on all datasets from Table 2, where binding affinities are expressed as the negative logarithm of equilibrium of dissociation (K_d) or inhibition (K_i) constants, whereas GCN–CNN architecture is applied only for the SCKBA dataset.

4.1 XGBoost

It is a boosting ensemble method used mostly with decision tree algorithm that has proven to be fast and highly effective in achieving 'state-of-the-art' results on many problems (Chen and Guestrin 2016). There are several hyperparameters determining the quality and performance of the XGBoost model for a given task; we utilized grid-search approach to find the set of hyperparameters providing best performing model. Since hyperparameter tuning in this manner can become expensive, smaller subset of the SCKBA dataset was used for this purpose.

4.2 GCN–CNN

For the GCN–CNN approach, we started from GraphDTA architecture proposed in Nguyen *et al.* (2021). We updated it by changing node representations of compound graphs from one-hot-encoded vectors to the physicochemical atomic feature representations. Furthermore, NN architecture is customized by implementing early stopping for no significant improvement over 20 consecutive epochs, with the decrease in learning rate for every 10 epochs, in order to avoid overfitting the model on the training data (Supplementary Fig. S7). GCN–CNN approach is trained for the total of 260 epoch with starting learning rate of 0.0005.

4.3 CP

Computation of calibration scores for any of the baseline approaches, Table 1, we use *Python* library 'nonconformist'. We tune a gamma sensitivity parameter for each dataset individually in the standard CP framework. Using the narrowest median prediction region (α_δ) as a reference, we picked an appropriate value for γ under the restriction that the mean error rate does not exceed the mean of the maximum error rates while still retaining the validity of the prediction regions for each confidence level (Supplementary Figs S9 and S10).

As it is shown in Fig. 3, in order to assess the validity of the confidence levels (75%, 80%, 85%, 90%, 95%, and 99%), the trained model is subjected to four different levels of testing difficulty. In order to compare proposed approach with baseline studies, we implement normalized non-conformity scoring as proposed in Papadopoulos and Haralambous (2010) and Papadopoulos *et al.* (2011). For normalizing the non-conformity scores by additional error model, instead of training NN (Papadopoulos and Haralambous 2010) or random forest (Johansson *et al.* 2014), in this work, we train an XGBoost model with the same hyperparameters as the model trained to predict binding affinities. dAD uses model trained on all samples and defines a dynamic calibration set such that for each test sample compound–target pair x, k, and q of the most similar compounds and protein targets, respectively, are selected from the training space, with respect to the tested compound–target pair. Both hyperparameters, k and q, were tuned manually for the SCKBA test (S1) dataset, with the aim of inspecting the impact on the validity and size of the prediction regions (Supplementary Fig. S8). For the majority of cases with larger number of different compounds and targets,

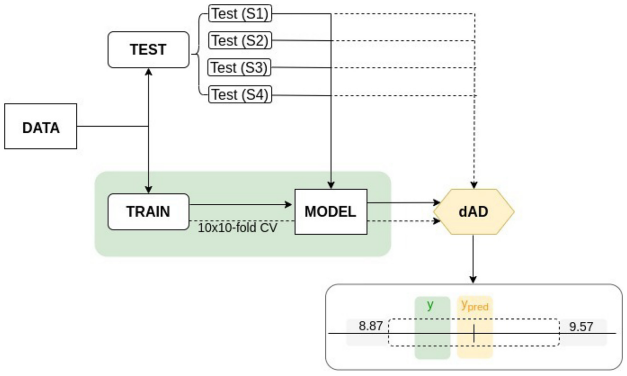


Figure 3. Illustration of the proposed dAD workflow from training of the underlying model to CP and validity evaluation. Data are split on train and test set(s), with all training samples being used for the modelling and allocation of the conformity region for each individual sample in the test set resulting in individual prediction regions. For a given confidence level error rate is defined by the number of samples for which the true labels (y) are not in the CP defined prediction region of y_{pred} (single CP prediction region is illustrated above as [8.87, 9.57]). Conformal predictor is valid if its error rate does not exceed 1-confidence.

we used $k = 250$ for nearest neighbours in the compound space and $q = 25$ for nearest neighbours in the target space, with the exception of Davis (Davis *et al.* 2011) and SSRI (Tang *et al.* 2018) datasets. Davis consists of 68 compounds in total and SSRI consists of only 49 protein targets, so smaller values for k and q were used, $k = 25$ and $q = 10$, respectively.

4.4 CP evaluation

Conformal Predictor framework is a model agnostic approach and addresses the uncertainty of the predictions. It calibrates model predictions in terms of some predefined confidence level using the calibration set of samples, and provides uncertainty assessment on an individual prediction level: for each model prediction, it provides prediction region where the true value should lie (Shafer and Vovk 2008). The CP diagnostics typically addresses validity and efficiency (informativeness) of the conformal predictor over different levels of confidence. This is also the basis of our assessment of dAD framework and comparisons against other CP frameworks. Validity of conformal predictors is assessed through error rate defined as the ratio of tested samples whose predictions do not fall into prediction region for certain confidence level, where expected error is ∼1-confidence. Conformal predictor is said to be valid if observed error rate is smaller or equal the expected error. Informativeness is another measure used to assess conformal predictors. It is related to the size of the prediction region for certain level of confidence. For the regression problem and CP assessment the size of the prediction region for the particular sample is defined by the value of non-conformity score (α) (see Section 2.1). To test how well the conformal regressor can recognize true binding affinities and produce valid prediction regions, the share of falsely classified compound–target pairs is determined for each confidence level of the four test scenarios in SCKBA dataset (Table 4). Due to the variable size of dynamic calibration sets and the difference in distribution of calibration and test sample non-conformity scores, dAD does not guarantee the extraction of α_{δ}^{min} scores for every sample at any confidence level. This 'abstaining from prediction' property is a direct result of introducing putative test non-conformity scores, as described in Section 2.1. For that reason, we denote the level of coverage (% of samples with prediction) for every confidence level as a value in addition to the reported error rates. Also, for a more direct comparison between the methods, we report the results only for those compound–target pairs and confidence levels that dAD (NN) and dAD (CV) methods were able to produce (Supplementary Tables S1, S2 and S4). Ultimately, dAD is compared to the previous studies by inspecting the median α range and mean error rates over all confidences levels. How well the calibration set represents the test sample is completely up to the representation of samples in the training set regarding the distance of nearest neighbours and number of experimentally measured binding affinities. Accordingly, following the work of Kuleshov *et al.* (2018) and Levi *et al.* (2022), we compare how well calibrated our proposed approach is for different testing scenarios and when compared to the baseline (Supplementary Figs S12 and S13).

5 Results

5.1 Choosing the prediction algorithm and model

We tested two different algorithms and problem representations to produce prediction models for testing and comparison of CP approaches over all datasets. As a conventional and faster approach, XGBoost was used as a baseline method. In addition, GCN–CNN architecture is adapted and applied, as a 'state-of-the-art' deep representation learning approach. We start with the architecture from Nguyen *et al.* (2021) as explained in Section 4. In Table 3, we can see that XGBoost approach is comparable, if not better in performance than the more complex convolutional network. This outcome may be unique to this dataset, where extended circular fingerprints capture chemical structure variation in the compound space sufficiently well for the model to generalize to real-world scenarios (S2 and S3). Furthermore, it is possible that the number of samples required for a robust NN model was not large enough, so by training over larger datasets with pretrained embeddings, GCN–CNN could be boosted in terms of scoring metrics, but this was beyond the scope of this study.

5.2 Comparison over different difficulty scenarios

As shown in Table 4, the dAD approach exhibits lower error rates per confidence level, in comparison to the standard approaches. The performance gap is most pronounced in the testing scenarios S2 and S3, which include compounds and targets that were not seen during the training phase, respectively. Poorer performance in these two settings could be due to the fact that using one-fits-all calibration approach confines the hypothetical prediction value between strict upper and lower boundaries that do not generalize well when it comes to unseen samples. Significant difference can also be observed on S0 test scenario, where both Equations (8) and (9) produce valid prediction regions for every confidence level, but with slightly larger prediction regions on average than dAD variants.

In the testing scenario S1, all approaches seem to be equally effective in keeping the number of incorrectly classified samples within proper limits. In contrast, in the testing scenario S4, neither approach is effective, and all of them show a high number of incorrectly classified samples. Considering that protein targets in S3 and S4 test sets belong to the kinase groups with many representatives in the training set, we assume that the skewed performance is due the variability in the

compound space, especially in the S4 setting where the *t*-SNE analysis shows that all compounds belong to the cluster separate from the rest of the training samples. Supplementary Fig. S3 depicts how median of the prediction region (α) varies for different CP approaches; Shafer and Vovk (7) method is represented simply as a point, due to the fact that prediction region is constant for all test samples at given confidence level. Median prediction regions for dAD (NN) and dAD (CV) are comparable, and range from one to three units depending on the confidence level. Variation across methods, in terms of prediction regions remains similar even when only dAD covered test samples are taken into account (Supplementary Fig. S4). However, we cannot make the assumption that one method is superior to another simply by looking at the error rates or the validity of a conformal predictor; but rather also consider the width of the prediction regions for each of these test cases. A good conformal predictor should strike a balance between both of these measures, staying valid while ensuring that prediction regions are as narrow as possible. Large prediction regions, while ensuring validity for any test scenario, are not useful. It is well illustrated in Fig. 4A (Supplementary Fig. S11A), where all six CP approaches are compared by examining the relationship of the mean error rates with the median α_δ scores of paired and unpaired datasets (Supplementary Table S1).

Table 3. XGBoost and GCN–CNN trained on a SCKBA train set with 40 578 interactions between 6325 compounds and 206 PKs.[a]

Test				XGBoost		GCN–CNN	
SX	#int	#cmpds	#trgts	MSE	CI	MSE	CI
S0	2855	2170	181	1.24	0.74	1.49	0.70
S1	655	411	169	0.28	0.83	0.43	0.82
S2	935	935	59	0.61	0.79	0.95	0.75
S3	655	439	3	0.60	0.72	0.61	0.74
S4	600	600	1	3.97	0.53	3.74	0.53

[a] Models are tested and compared over (S0) test set, including four difficulty scenarios (S1–S4) obtained by segmentation of the S0 dataset.

Table 4. Comparison of baseline methods with proposed dAD approach on a combined SCKBA dataset over four difficulty scenarios, with sensitivity parameter γ for Equations (9) and (10) in S1 scenario being $\gamma_{(\lambda)}$=0.2 and $\gamma_{(\xi)}$=0, respectively, and 0 for the rest.[a]

SCKBA

		Median		Error rates per confidence level (%)					
Approach	SX	α_δ	#calib	75%	80%	85%	90%	95%	99%
Shafer & Vovk (7)	S0	0.86	4000	44.69	40.39	35.90	29.81	22.87	12.26
	S1	0.86	4000	21.83	16.34	11.91	6.56	3.66	0.76
	S2	0.86	4000	40.64	35.08	30.27	21.60	12.41	2.57
	S3	0.86	4000	35.49	31.73	26.02	18.95	11.43	3.91
	S4	0.86	4000	86.17	84.50	81.83	80.00	72.83	49.17
Papadopoulos (8)	S0	1.94	4000	27.18	23.64	20.49	16.04	9.88	1.61
	S1	2.45	4000	5.65	4.12	3.05	2.14	1.53	0.31
	S2	1.65	4000	22.35	18.93	15.72	10.48	5.99	1.39
	S3	2.38	4000	11.13	7.82	5.11	3.61	2.56	0.30
	S4	1.62	4000	76.00	69.83	64.00	53.67	33.17	4.83
Papadopoulos (9)	S0	1.64	4000	25.43	19.58	15.13	9.56	5.15	1.02
	S1	1.02	4000	19.69	15.73	10.84	7.33	5.50	1.68
	S2	1.13	4000	30.16	24.39	19.68	13.69	7.49	1.39
	S3	1.55	4000	22.71	18.35	14.74	10.68	5.86	1.20
	S4	3.23	4000	33.83	25.17	17.17	6.33	2.67	0.00
Papadopoulos (10)	S0	0.85	4000	45.25	40.56	36.50	31.17	25.08	14.43
	S1	0.85	4000	22.60	16.49	12.98	7.79	4.12	1.22
	S2	0.85	4000	43.10	37.75	32.09	25.67	17.11	5.67
	S3	0.95	4000	32.18	27.37	22.71	16.54	10.53	3.01
	S4	0.78	4000	87.83	85.83	84.33	81.50	76.5	55.17
dAD (NN)	S0	1.77	259	13.14 (0.63)	17.97 (0.64)	14.00 (0.67)	15.36 (0.72)	12.38 (0.75)	3.69 (0.69)
	S1	1.85	315	1.87 (0.73)	1.65 (0.74)	0.79 (0.77)	1.00 (0.76)	0.62 (0.74)	0.00 (0.63)
	S2	1.78	253	8.76 (0.74)	6.72 (0.70)	3.54 (0.63)	2.94 (0.58)	1.28 (0.59)	0.36 (0.60)
	S3	1.65	279	11.90 (0.69)	10.11 (0.71)	8.07 (0.73)	6.03 (0.77)	4.76 (0.79)	1.03 (0.73)
	S4	1.79	232	70.70 (0.26)	68.40 (0.38)	60.47 (56)	52.27 (0.84)	39.22 (0.98)	12.69 (0.87)
dAD (CV)	S0	1.57	259	14.90 (0.54)	16.62 (0.55)	16.85 (0.54)	18.98 (0.55)	14.59 (0.46)	2.81 (0.28)
	S1	1.57	315	2.04 (0.60)	2.13 (0.57)	1.13 (0.54)	1.37 (0.45)	0.47 (0.33)	0.00 (0.19)
	S2	1.55	253	10.37 (0.58)	7.47 (0.56)	4.87 (0.48)	3.54 (0.42)	1.41 (0.42)	0.77 (0.28)
	S3	1.46	279	12.97 (0.60)	11.68 (0.60)	9.24 (0.55)	7.36 (0.55)	5.90 (0.41)	0.00 (0.18)
	S4	1.74	232	71.23 (0.24)	68.60 (0.34)	59.67 (0.51)	53.47 (0.72)	42.96 (0.70)	8.15 (0.45)

[a] The SX denotes the testing scenario; $\alpha_{(\delta)}$ is the median prediction region of the test set; *#calib* is number of samples in the calibration set or median number of samples for the dAD method with varying calibration sizes. Error rates represent the percent of samples with labels outside of prediction regions. Values next to the dAD (CV) and dAD (NN) error rates denote the coverage of the test set, with 0 meaning that the proposed approach was not able to produce prediction regions for a given confidence and 1 meaning that it produced a prediction region for every test sample.

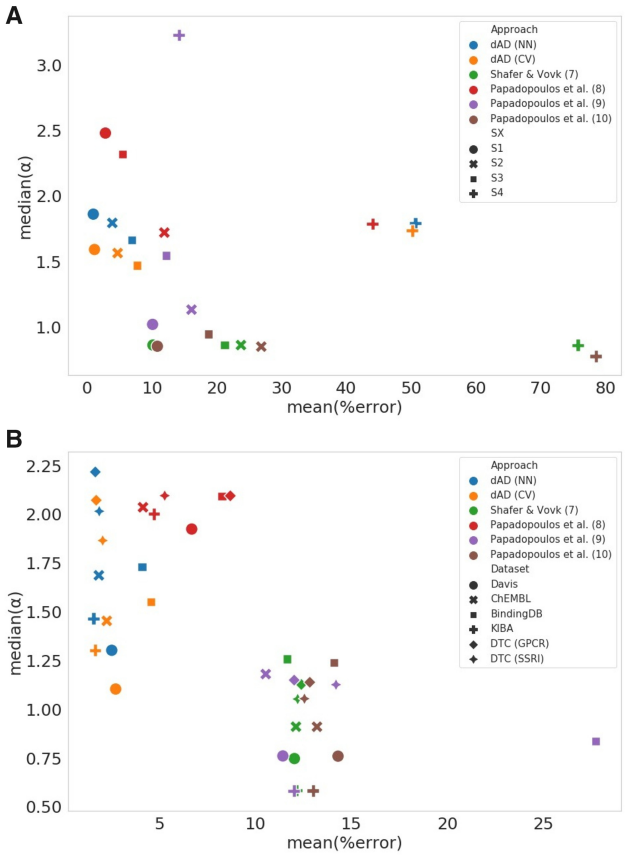


Figure 4. Comparison of the proposed dAD approach with the baseline studies by showing the relationship of mean (%) error and median non-conformity scores for (A) four different testing scenarios (S1–S4) of SCKBA dataset and (B) over six compound–target datasets, Davis (Davis *et al.* 2011), KIBA (Tang *et al.* 2018), ChEMBL database (Gaulton *et al.* 2012), BindingDB (Gilson *et al.* 2016), DTC (GPCR) (Tang *et al.* 2018), and DTC (SSRI) (Tang *et al.* 2018) represented with different shapes.

Results of both dAD (NN) and dAD (CV) seem to be comparable. Similar case is with the Shafer and Vovk approach and Papadopoulos (10), showing somewhat tighter prediction regions but higher error rates on average than the proposed dAD approach. In Fig. 4A, we visualize performance of all six methods over four different testing scenarios (S1–S4) by plotting the relationship between the mean error rates and median prediction region width (α). For S1 test setting, denoted as points in Fig. 4A, all methods besides Papadopoulos (8) in red, give reasonably tight prediction regions with maximum median prediction region lower than two units and with low error rates, which is to be expected due to the nature of that testing scenario comprising of compound and PK targets already seen during the training phase. At the other end of the spectrum, there is an S4 test setting where all trained models show almost random performance, as we see in Fig. 4A where the performance over the S4 test set marked as pluses occupies the right end of the scatter plot. Papadopoulos (9) is an exception, achieving very low mean error rate on the forth scenario due to the very wide median prediction region. Both proposed dAD approaches show low mean error rates with relatively tight prediction regions, which is especially important for more realistic test scenarios, S2 and S3.

Calibration quality of dAD is inspected in terms of expected and observed confidence levels, and in comparison to baseline studies. We additionally investigate whether and how calibration quality changes with the addition of a normalization measure similar to Equation (9). In terms of calibration, we can say that dAD exhibits underconfident results, with the observed confidence always being higher than the expected one, which is opposite of how the baseline approaches behave. Even if in real use-case scenarios, an underconfident estimator would be preferred over an overconfident one—we inspect if the both observed and expected confidence could be brought to a closer agreement. Introducing a normalization measure leads to a reduction in the width of selected prediction regions (Supplementary Table S5). As a result, the observed confidence decreases, causing the dAD to become overconfident for higher confidence levels in S2 and S3. Moreover, this normalization makes the approach even less reliable in cases where it was initially overconfident, such as S4. It is best shown in Supplementary Fig. S12, where the effect of normalization evidently draws observed and expected confidence level closer to a certain point. Furthermore, Supplementary Fig. S13 reveals that baseline approaches exhibit good calibration on average for the S1 testing scenario. However, in the other testing scenarios, they tend to yield higher expected confidence than the observed confidence, leading to overconfident predictions.

5.3 Comparison over standard benchmark datasets

Table 5 and Supplementary Tables S3 and S4 compare the four conformal predictor approaches with two proposed dAD variants. Shafer and Vovk approach (Shafer and Vovk 2008) shows lower validity of predicted regions when compared to the normalization based CPs or the proposed approaches. This behaviour is expected, especially since it produces fixed prediction regions (α scores) given the confidence level. Having this in mind, even if the mean of the prediction region is lower than in any other approach, true difference is shown in the validity of their predictions.

On the other hand, normalization using the underlying error model and λ_i^k coefficient gives comparable results to the proposed dAD method for 95% and 99% confidence levels on the Davis dataset. The difference is better depicted in boxplots in Supplementary Figs S5 and S6 showing the distribution of prediction regions for every x in the test set, and only over samples for which dAD methods were capable to define prediction regions, respectively.

True difference in the performance of any of mentioned methods is shown by scatter plot giving relation of the mean percentage of wrongly classified samples to the median α score for confidence levels 75%–99% (Fig. 4B and Supplementary Fig. S11B), where the better performing CPs should be closer to zero on both axes. When compared in this sense, approaches (7), (9), and (10) have the narrowest prediction regions when tested on all six datasets, consequently with higher mean error rates ranging above 10%. When taking into consideration both the prediction regions and mean error rates, dAD (NN) and dAD (CV) produce more optimal prediction regions in relation to the mean error rate (Fig. 4B), with median prediction regions exceeding the two units only for the DTC (GPCR) and DTC (SSRI) datasets.

5.4 Demonstration of the direct application of dAD

The dAD, as defined in this study, provides range-to-point predictions with a certain level of confidence. However, how does it help us condense the wide field of possible

Table 5. Comparison of baseline methods with proposed dAD approach on compound–kinase binding affinity datasets from Table 2, with sensitivity parameter γ of Equations (9) and (10) for the Davis and KIBA datasets being $\gamma_{(\lambda)} = 0.7$ and $\gamma_{(\xi)} = 0$; for BindingDB dataset $\gamma_{(\lambda)} = 0$ and $\gamma_{(\xi)} = 0$; and for the ChEMBL dataset $\gamma_{(\lambda)} = 0.3$ and $\gamma_{(\xi)} = 0$.[a]

Benchmark datasets (KI)

		Median		Error rates per confidence level (%)					
Dataset	Approach	α_δ	#calib	75%	80%	85%	90%	95%	99%
Davis	Shafer & Vovk (7)	0.75	1500	23.86	19.28	14.43	9.77	4.13	0.76
	Papadopoulos (8)	1.92	1500	13.12	10.36	7.63	5.32	2.76	0.91
	Papadopoulos (9)	0.76	1500	23.13	18.46	13.61	9.03	3.73	0.65
	Papadopoulos (10)	0.76	1500	26.08	21.77	17.61	12.78	6.28	1.36
	dAD (CV)	1.10	502	3.56 (0.34)	3.33 (0.52)	3.30 (0.71)	3.04 (0.80)	2.21 (0.77)	0.87 (0.46)
	dAD (NN)	1.30	502	3.43 (0.33)	3.25 (0.52)	3.24 (0.71)	2.81 (0.83)	1.87 (0.91)	0.48 (0.91)
KIBA	Shafer & Vovk (7)	0.58	3000	23.96	19.08	14.73	9.41	4.79	0.94
	Papadopoulos (8)	2.00	3000	8.65	6.92	5.55	3.82	2.34	1.00
	Papadopoulos (9)	0.58	3000	23.62	18.96	14.51	9.42	4.66	0.91
	Papadopoulos (10)	0.58	3000	24.85	20.01	15.73	10.23	5.77	1.46
	dAD (CV)	1.30	1661	3.77 (0.73)	2.62 (0.80)	1.99 (0.87)	1.04 (0.91)	0.39 (0.84)	0.09 (0.46)
	dAD (NN)	1.47	1661	3.60 (0.72)	2.46 (0.81)	1.85 (0.90)	0.96 (0.96)	0.39 (0.98)	0.1 (0.93)
BindingDB	Shafer & Vovk (7)	1.26	3000	23.36	28.85	13.08	8.84	5.00	0.84
	Papadopoulos (8)	2.09	3000	13.48	11.37	8.84	6.9	5.13	3.85
	Papadopoulos (9)	0.84	3000	37.41	34.79	31.12	27.28	23.01	12.88
	Papadopoulos (10)	1.24	3000	25.52	21.6	16.42	11.78	7.45	1.83
	dAD (CV)	1.55	133	9.06 (0.58)	6.72 (0.55)	5.45 (0.49)	3.54 (0.44)	1.86 (0.39)	0.80 (0.27)
	dAD (NN)	1.33	133	8.64 (0.73)	6.08 (0.71)	4.58 (0.68)	3.09 (0.67)	1.54 (0.65)	0.61 (0.48)
ChEMBL	Shafer & Vovk (7)	0.91	3000	24.62	19.27	13.92	9.40	4.51	0.99
	Papadopoulos (8)	2.04	3000	8.01	6.43	4.62	3.28	1.83	0.69
	Papadopoulos (9)	1.18	3000	19.79	16.01	12.13	8.77	4.91	1.66
	Papadopoulos (10)	0.91	3000	25.41	20.49	15.13	10.63	5.86	1.77
	dAD (CV)	1.45	253	5.18 (0.74)	3.70 (0.68)	2.35 (0.62)	1.47 (0.53)	0.62 (0.40)	0.00 (0.15)
	dAD (NN)	1.69	253	4.38 (0.86)	3.11 (0.86)	1.85 (0.86)	1.13 (0.86)	0.43 (0.83)	0.15 (0.57)

[a] Column definitions are the same as in Table 4.

interactions? We demonstrate this on the S2 test set, one of the more difficult scenarios, by permuting all S2 test set compounds across the protein targets in the training set. This yields the heatmap depicted in Fig. 5. In the upper portion of Fig. 5, a heatmap depicts all predicted binding affinities with pKd $\geq$ 5.5, as generated by the trained model, while five heatmaps represent the outcomes of dAD confidence filters. Considering that binding affinities with measured pKd around six units are considered significant for the drug–kinase interaction problem and taking into account highly unbalanced data, the original filter is set up to retain only those interactions with pKd $\geq$ 5.5 and with prediction regions for a certain confidence level that do not exceed the lower boundary of activity. Results are displayed for confidence levels of 75%, 80%, 85%, 90%, and 95%, and for each level, only 13%, 7%, 3%, 1%, and 0.08% of samples, respectively, are filtered through as potentially significant, drastically reducing the noise in the interaction space and condensing the interaction landscape of interest. The filter criteria can be determined based on the available data and the acceptable prediction region width as deemed appropriate by the end-user.

6 Discussion

This work merges concepts of an AD and conformal predictors to provide improved estimates for the bioactivity prediction tasks, involving complex interaction data spaces. 'State-of-the-art' CP frameworks rely on fixed calibration sets based on their overall distributional similarity to the label space of the training data. This is a limiting factor, especially considering the compound–target binding affinity problem is defined by duality, independently distributed chemical and biological spaces—and models produced over such datasets may perform differently over distinct subspaces of compounds and target families depending on their distribution in the training set. Moreover, in drug discovery and repurposing, machine-learning models must predict novel compounds or targets for which exchangeability is only marginally satisfied. This was a motivation for developing novel conformal predictor framework. We solve these problems by defining calibration sets separately for each tested sample, taking into account data distribution from both ends, and then retrieving experimentally measured binding affinities for existing interaction pairs. Consequently, this method produces prediction regions that are specific for the particular test sample, given the confidence level of interest.

We prove in this work that our CP approach more accurately reflects the performance of the model in the area close to the tested sample, providing more robust prediction region estimates for any given confidence. On the standard testing scenario (S1) our experiments show that this approach provides similar performance in terms of validity and size of prediction regions as other 'state-of-the-art' CP approaches (Papadopoulos and Haralambous 2010, Papadopoulos *et al.* 2011). However, for more difficult scenarios (S2–S4), involving test samples at or beyond the borders of the training data space, proposed dAD approach proved to be more effective, providing strong validity with reasonable sizes of prediction regions. These findings imply that a dynamically defined CP calibration strategy more precisely reflects biases of the trained models in the neighbourhood of tested points. In terms of practical impact of these findings for biochemical

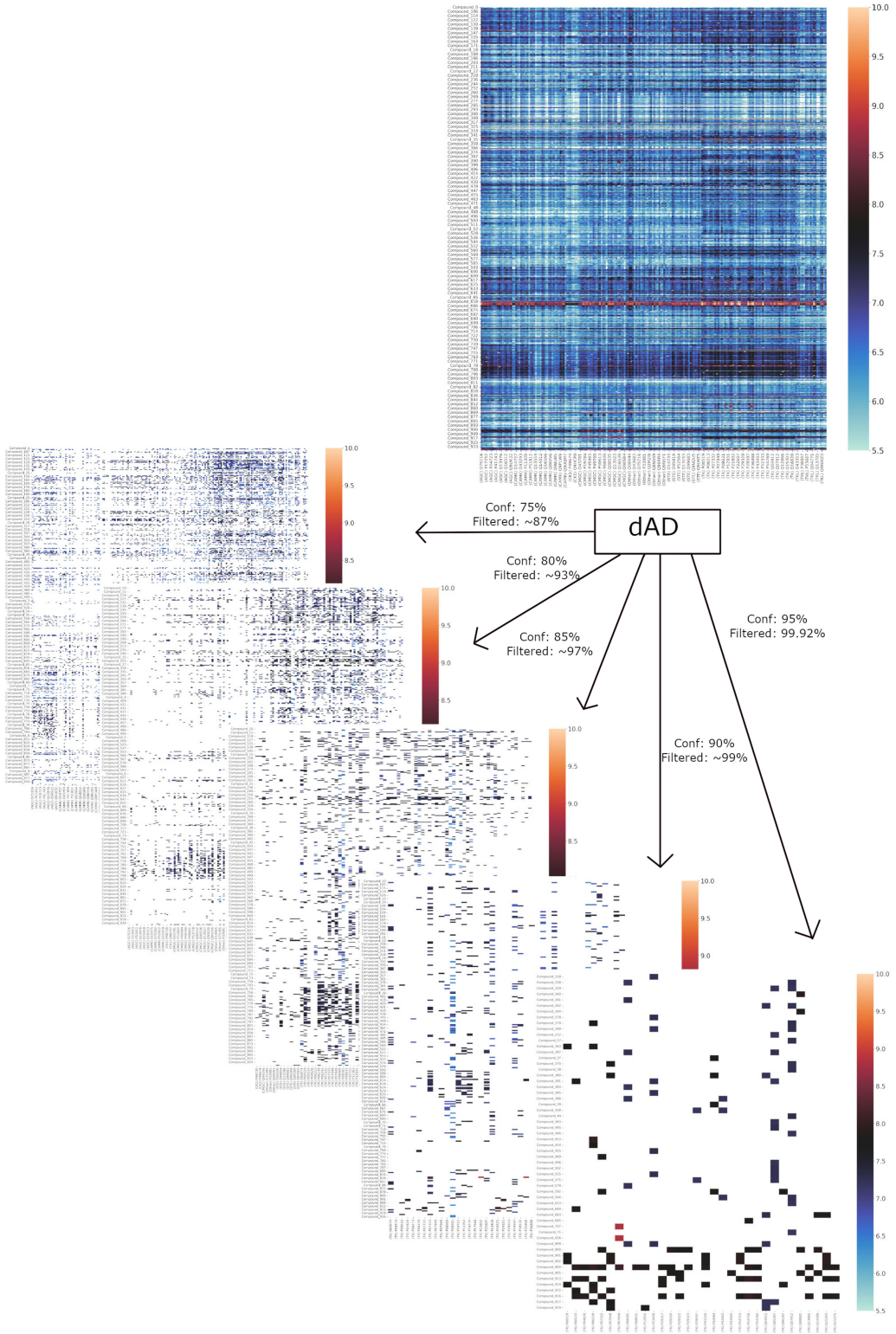


Figure 5. dAD as a screening filter for model predictions: first heatmap represents a subset of model predictions for all permutations of S2 compounds and protein targets in the training set, with pKd predictions ranging from 5.5 to 10 units. After applying the dAD, interaction landscape is condensed based on the chosen confidence level—there are five versions of the same heatmap depending on the dAD confidence filter applied; for confidence of 75%—13% original samples remains; 80%—7% of original samples remains; confidence of 85%—3% of original test samples remains in the heatmap, etc.

research, this methodology should lead to more efficient experimental research by reducing number of false positives when screening for novel prospective drugs. As the dAD framework provides range predictions with more accurate uncertainty assessment, it is crucial for more efficient prioritization of (expensive) experiments (e.g. selection of the most

promising hit molecules from *in silico* screening or drug repurposing experiments). Specifically, we demonstrate that it is more effective in realistic modelling use-cases in which predictions are made at the model's AD boundary. We should acknowledge several caveats related to dAD approach. While providing better uncertainty calibration and prediction region estimates for realistic use-cases, due to the smaller sizes of calibration sets and novel algorithm for the determination of non-conformity scores of tested samples, dAD approach can abstain from prediction for some of the tested samples at required confidence level. This property is the consequence of our definition of non-conformity scores for test samples and reduced size of the calibration sets, and is not observed for the other CP approaches. In Supplementary Fig. S14, we demonstrate this effect and analyse the joint and individual impacts of the non-conformity definition and reduced calibration set size. Supplementary Fig. S14A illustrates joint impact of non-conformity definition and reduced calibration set size on coverage (% of test samples with CP), showing that an increased calibration set size results in increased coverage across all confidence levels. Supplementary Fig. S14B depicts the coverage of test samples using the standard definition of non-conformity (i.e. prediction regions for the test sample are based on a calibration set sample with predefined level of confidence), thus seeing the impact of reduced calibration set size on coverage, without dAD non-conformity.

The 'abstaining from prediction' property of dAD results in underconfident behaviour (error rate is smaller than predicted) for the samples that obtained dAD CP at certain level of confidence. This is especially characteristic for more realistic scenarios (S2 and S3) and in contrast to other approaches which are overconfident in their prediction (error rates are higher than predicted) (Supplementary Figs S12 and S13 and Supplementary Table S1). Abstaining from prediction is a common property of the conformal predictors for classification problems. We consider it to be a positive feature, which results in lower rate of false positives in more demanding prediction settings. Another caveat is related to smaller calibration sets, which limits the use of the method to problems with larger number of samples available for training the model and subsequent dynamic calibration. However, when this requirement is satisfied (as is the case for SCKBA dataset in this work), we have shown that the method exhibits stable performance for rather broad span of calibration set sizes. Future avenues related to the dAD approach may focus on explainability aspects of individual interaction predictions by exploiting localized calibration of predictions and, by extending the approach to other biological interaction type of problems.

Acknowledgements

We thank Dr Mathieu Dutour Sikiric and Miha Keber for helping us through useful discussions and suggestions related to earlier versions of the manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported in part by the Research Cooperability Program of the Croatian Science Foundation, funded by the European Union from the European Social Fund under the Operational Programme Efficient Human Resources 2014–2020, through the Grant 8525: Augmented Intelligence Workflows for Prediction, Discovery, and Understanding in Genomics and Pharmacogenomics; and by the Croatian Government and the European Union under the European Regional Development Fund—the Competitiveness and Cohesion Operational Program, through the project Bioprospecting of the Adriatic Sea [KK.01.1.1.01.0002], granted to The Scientific Centre of Excellence for Marine Bioprospecting—BioProCro.

References

Alvarsson J, Arvidsson McShane S, Norinder U *et al.* Predicting with confidence: using conformal prediction in drug discovery. *J Pharm Sci* 2021;**110**:42–9.

Aniceto N, Freitas AA, Bender A *et al.* A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *J Cheminform* 2016;**8**:1–20.

Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–94, 2016.

Cichonska A, Ravikumar B, Parri E *et al.* Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Comput Biol* 2017;**13**:e1005678.

Cichońska A, Ravikumar B, Allaway RJ *et al.*; IDG-DREAM Drug-Kinase Binding Prediction Challenge Consortium. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat Commun* 2021;**12**:3307–18.

Davis MI, Hunt JP, Herrgard S *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51.

Gadaleta D, Mangiatordi GF, Catto M *et al.* Applicability domain for QSAR models: where theory meets reality. *IJQSPR* 2016;**1**:45–63.

Gammerman A, Vovk V, Vapnik V. Learning by transduction. In: *Fourteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1998, 148–55.

Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.

Gilson MK, Liu T, Baitaluk M *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**:D1045–53.

Fey M, Lenssen JE. Fast graph representation learning with pytorch geometric. In: *7th International Conference on Learning Representations*. New Orleans, US: Ernest N. Morial Convention Center. 2019.

Johansson U, Boström H, Löfström T *et al.* Regression conformal prediction with random forests. *Mach Learn* 2014;**97**:155–76.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *5th International Conference on Learning Representations, Toulon, France*. 2017.

Klingspohn W, Mathea M, Ter Laak A *et al.* Efficiency of different measures for defining the applicability domain of classification models. *J Cheminform* 2017;**9**:44–17.

Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. In: *Proceedings of the 35th International Conference on Machine Learning*, pp. 2796–804. PMLR, 2018.

Landrum G, Tosco P, Kelley B *et al.* rdkit/rdkit: 2019_09_03 (Q3 2019) Release. 2020. https://doi.org/10.5281/zenodo.3603542. https://zenodo.org/record/3603542 (August-November, 2020, date last accessed).

Levi D, Gispan L, Giladi N *et al.* Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors* 2022;**22**:5540.

Liao JJ-L. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J Med Chem* 2007; **50**:409–24.

Lim S, Lu Y, Cho CY *et al.* A review on compound-protein interaction prediction methods: data, format, representation and model. *Comput Struct Biotechnol J* 2021;**19**:1541–56.

Manning G, Whyte DB, Martinez R *et al.* The protein kinase complement of the human genome. *Science* 2002;**298**:1912–34.

Mathea M, Klingspohn W, Baumann K *et al.* Chemoinformatic classification methods and their applicability domain. *Mol Inform* 2016; **35**:160–80.

Metz KS, Deoudes EM, Berginski ME *et al.* Coral: clear and customizable visualization of human kinome data. *Cell Syst* 2018;7:347–50.e1.

Nguyen T, Le H, Quinn TP *et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 2021; **37**:1140–7.

Öztürk H, Özgür A, Ozkirimli E *et al.* DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 2018;**34**:i821–9.

Pahikkala T, Airola A, Pietilä S *et al.* Toward more realistic drug–target interaction predictions. *Brief Bioinform* 2015;**16**:325–37.

Papadopoulos H. *Inductive Conformal Prediction: Theory and Application to Neural Networks*. Rijeka: INTECH Open Access Publisher, 2008.

Papadopoulos H, Haralambous H. Neural networks regression inductive conformal predictor and its application to total electron content prediction. In: Diamantaras K, Duch W, Iliadis LS (eds) *International Conference on Artificial Neural Networks*. US: Springer 2010, 32–41.

Papadopoulos H, Vovk V, Gammerman A *et al.* Regression conformal prediction with nearest neighbours. *JAIR* 2011;**40**: 815–40.

Paszke A, Gross S, Massa F *et al.* PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, Vol. **32**. 2019.

Roskoski R. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol Res* 2015;**100**: 1–23.

Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res* 2008;**9**:371–421.

Tang J, Szwajda A, Shakyawar S *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model* 2014;**54**:735–43.

Tang J, Tanoli Z-U-R, Ravikumar B *et al.* Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell Chem Biol* 2018;**25**:224–9.e2.

Vovk V, Nouretdinov I, Manokhin V *et al.* Cross-conformal predictive distributions. In: Gammerman A, Vovk V, Luo Z (Eds) *Conformal and Probabilistic Prediction and Applications*. PMLR, 2018, 37–51.

Xiao N, Cao D-S, Zhu M-F *et al.* protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 2015;**31**:1857–9.

1 Supplementary

Algorithm 1 dAD (CV/NN)

Input: x=(c_x, t_x), δ, C, T
Output: S^{cal}, α_δ^{min}, Γ_x^δ

Step 1. Selecting k closest compounds
for each $c_j \in C$ **do**
  Compute Tanimoto similarity coefficients for c_x towards all compounds in the training set, $c_i \in C$
  Rank all compounds based on their similarity scores in descending order
  Select top k compounds from the training set (C)
end for

Step 2. Selecting q closest targets
for each $t_j \in T$ **do**
  Compute Smith-Waterman similarity coefficients for t_x towards all targets in the training set, $t_j \in T$
  Rank all targets based on their similarity scores in descending order
  Select top q targets from the training set (T)
end for

Step 3. Find experimentally measured binding affinities, $y_{ij} \in Y$, between the top ranking samples
Retrieve all (c_i, t_j, y_{ij}) tuples and construct a calibration set, such that $c_j \in C^{cal}$ and $t_j \in T^{cal}$ and $y_{ij} \in Y^t$

Step 4. Determine the prediction region of sample x, Γ_x^δ
Compute nonconformity scores for x, α^{cal} and α^x, according to Eq. (**??-??**)
for each $\alpha_i^{cal} \in S^{cal}$ **do**
  conf = **countif** $\alpha_j^x \in S^x \leq \alpha_i^{cal}$
  if (conf $\geq$ 1-δ) **and** ($\alpha_i^{cal} < \alpha_\delta^{min}$) **then**
    $\alpha_\delta^{min} = \alpha_i^{cal}$; $\Gamma_x^\delta = \hat{y} \pm \alpha_\delta^{min}$
  else
    α_δ^{min} = *undefined*
  end if
end for

Table 1: Comparison of baseline methods with proposed dynamic applicability domain (dAD) approach on combined drug-kinase binding affinity dataset over four difficulty scenarios. The *SX* denotes the testing scenario; $\alpha_{(\delta)}$ is the median prediction region of the test set; #*calib* is number of samples in the calibration set or median number of samples for the dAD method with varying calibration sizes. Error rates represent the percent of samples with labels outside of prediction regions. Coverage values next to the dAD (CV) and dAD (NN) error rates are not reported since all results here are matched to the indices of samples where both dAD approaches were able to produce prediction regions.

SKCBA (paired)									
		Median		Error rates per confidence level (%)					
Approach	SX	α_δ	#calib	75%	80%	85%	90%	95%	99%
Shafer & Vovk (7)	S0	0.78	4000	36.69	34.06	33.08	34.09	32.85	19.94
	S1	0.78	4000	20.66	13.91	11.33	5.83	3.15	0.00
	S2	0.78	4000	35.68	29.55	22.65	16.99	11.83	2.72
	S3	0.78	4000	36.60	29.14	26.83	22.35	16.10	1.77
	S4	0.94	4000	86.50	84.19	80.00	80.62	76.29	48.2
Papadopoulos (8)	S0	1.90	4000	18.79	18.80	16.18	17.21	11.27	0.82
	S1	1.98	4000	5.06	4.45	2.52	1.59	1.15	0.00
	S2	1.71	4000	17.14	15.05	8.97	7.82	5.16	1.06
	S3	1.93	4000	14.45	9.91	5.90	3.92	3.26	0.00
	S4	1.99	4000	73.68	67.00	52.35	44.13	24.11	1.14
Papadopoulos (9)	S0	1.34	4000	24.14	20.15	16.48	9.43	5.20	0.63
	S1	0.89	4000	19.11	14.44	9.70	7.93	6.37	0.00
	S2	0.88	4000	30.65	25.00	19.77	12.50	6.23	1.48
	S3	1.37	4000	26.05	22.08	18.38	11.24	5.79	0.00
	S4	3.04	4000	25.76	25.87	24.45	7.95	4.76	0.00
Papadopoulos (10)	S0	0.84	4000	35.36	33.38	33.29	33.95	34.23	22.44
	S1	0.77	4000	21.20	15.53	12.29	7.93	3.85	0.00
	S2	0.88	4000	34.39	34.43	23.58	17.62	13.91	9.38
	S3	0.85	4000	33.07	25.91	24.51	17.00	12.83	0.85
	S4	0.86	4000	86.47	84.73	82.89	80.52	79.24	52.24
dAD (NN)	S0	1.54	259	16.61	15.77	16.85	19.11	17.48	5.10
	S1	1.55	315	2.36	2.18	1.14	1.38	0.96	0.00
	S2	1.52	253	10.47	7.86	4.54	3.63	1.48	0.89
	S3	1.42	279	13.39	11.40	10.03	8.22	1.17	0.00
	S4	1.68	232	69.92	68.47	61.75	54.23	45.11	13.06
dAD (CV)	S0	1.55	259	14.54	15.56	16.30	18.63	16.42	3.35
	S1	1.56	315	2.09	2.18	1.14	1.38	0.48	0.00
	S2	1.55	253	10.09	7.47	4.76	3.63	1.48	0.92
	S3	1.46	279	13.65	11.40	9.19	7.65	6.04	0.00
	S4	1.73	232	70.68	67.98	59.73	53.05	42.96	8.21

Table 2: Comparison of baseline methods with proposed dynamic applicability domain (dAD) approach on benchmark datasets involving compound-kinase binding affinities. $\alpha_{(\delta)}$ is the median prediction region of the test set; *#calib* is number of samples in the calibration set or median number of samples for the dAD method with varying calibration sizes. Error rates represent the percent of samples with labels outside of prediction regions. Coverage values next to the dAD (CV) and dAD (NN) error rates are not reported since all results here are matched to the indices of samples where both dAD approaches were able to produce prediction regions.

Benchmark datasets (KI) [paired]									
		Median		Error rates per confidence level (%)					
Dataset	Approach	α_δ	#calib	75%	80%	85%	90%	95%	99%
Davis	Shafer & Vovk (7)	0.75	1500	23.86	19.28	14.43	9.77	4.13	0.76
	Papadopoulos (8)	2.40	1500	1.39	1.23	1.23	1.64	1.35	1.11
	Papadopoulos (9)	0.84	1500	8.93	6.24	5.50	4.56	2.75	0.83
	Papadopoulos (10)	0.82	1500	14.86	10.55	9.05	7.53	4.23	1.11
	dAD (CV)	1.11	502	3.54	3.26	3.29	3.04	2.19	0.87
	dAD (NN)	1.11	502	3.43	3.26	3.27	2.9	2.21	0.94
KIBA	Shafer & Vovk (7)	0.58	3000	23.96	19.08	14.73	9.41	4.79	0.94
	Papadopoulos (8)	2.09	3000	7.04	5.20	4.28	3.15	2.24	1.15
	Papadopoulos (9)	0.56	3000	20.02	15.58	12.28	8.38	4.65	1.26
	Papadopoulos (10)	0.57	3000	21.93	16.98	13.42	9.02	5.42	1.79
	dAD (CV)	1.31	1661	3.71	2.59	1.97	1.03	0.39	0.09
	dAD (NN)	1.32	1661	3.68	2.53	1.91	1.01	0.43	0.15
BindingDB	Shafer & Vovk (7)	1.26	3000	23.36	28.85	13.08	8.84	5.00	0.84
	Papadopoulos (8)	2.05	3000	11.27	10.19	8.07	6.4	5.00	5.8
	Papadopoulos (9)	0.57	3000	38.89	36.64	34.29	29.61	23.51	21.49
	Papadopoulos (10)	1.19	3000	22.26	18.65	13.94	8.32	6.32	3.00
	dAD (CV)	1.52	133	8.85	6.61	5.35	3.49	1.85	0.75
	dAD (NN)	1.52	133	9.28	6.84	5.35	3.49	1.79	1.03
ChEMBL	Shafer & Vovk (7)	0.91	3000	24.62	19.27	13.92	9.40	4.51	0.99
	Papadopoulos (8)	1.88	3000	7.04	5.6	4.18	2.93	2.14	0.22
	Papadopoulos (9)	0.96	3000	20.48	17.08	13.48	9.98	6.44	2.46
	Papadopoulos (10)	0.83	3000	23.86	18.89	13.92	9.34	5.45	1.09
	dAD (CV)	1.45	253	5.06	3.68	2.31	1.46	0.63	0.22
	dAD (NN)	1.44	253	5.01	3.71	2.27	1.53	0.76	0.22

Table 3: Comparison of baseline methods with proposed dynamic applicability domain (dAD) approach on DTC (GPCR) and DTC (SSRI) datasets from Table 2, with sensitivity parameter γ for Eq. 9 and 10 for GPCR dataset being $\gamma_{(\lambda)}$=0.7 and $\gamma_{(\xi)}$=0; and for SSRI dataset $\gamma_{(\lambda)}$=0 and $\gamma_{(\xi)}$=0. Column definitions are the same as in Table 4.

DTC (GPCR; SSRI)									
		Median		Error rates per confidence level (%)					
Dataset	Approach	α_δ	#calib	75%	80%	85%	90%	95% CI	99%
GPCR	Shafer & Vovk (7)	1.13	1500	25.02	18.85	14.29	10.00	5.16	1.07
	Papadopoulos (8)	2.09	1500	13.62	11.24	9.02	7.46	6.02	4.82
	Papadopoulos (9)	1.15	1500	24.15	18.53	14.29	9.19	5.13	0.87
	Papadopoulos (10)	1.14	1500	25.28	18.93	14.93	10.44	6.18	1.28
	dAD (CV)	2.14	874	3.69 (.84)	2.98 (.81)	1.84 (.77)	1.12 (.70)	0.54 (.59)	0.1 (.58)
	dAD (NN)	2.25	874	3.72 (.94)	2.82 (.93)	1.77 (.90)	1.09 (.85)	0.45 (.78)	0.08 (.75)
SSRI	Shafer & Vovk (7)	1.05	1500	24.7	19.13	14.59	9.53	4.25	1.05
	Papadopoulos (8)	2.09	1500	10.21	8.17	6.15	3.83	2.06	1.24
	Papadopoulos (9)	1.13	1500	25.83	21.09	17.26	12.06	6.6	2.41
	Papadopoulos (10)	1.05	1500	24.57	19.69	15.33	9.87	4.7	1.21
	dAD (CV)	1.86	234	4.17 (.68)	3.15 (.63)	4.47 (.53)	1.56 (.47)	0.66 (.40)	0.23 (.21)
	dAD (NN)	2.02	234	3.98 (.89)	2.99 (.85)	2.09 (.81)	1.38 (.74)	0.63 (.67)	0.11 (.49)

Table 4: Comparison of baseline methods with proposed dynamic applicability domain (dAD) approach on benchmark datasets involving subsets of the DTC dataset, including GPCR and SSRI datasets. $\alpha_{(\delta)}$ is the median prediction region of the test set; #*calib* is number of samples in the calibration set or median number of samples for the dAD method with varying calibration sizes. Error rates represent the percent of samples with labels outside of prediction regions. Coverage values next to the dAD (CV) and dAD (NN) error rates are not reported since all results here are matched to the indices of samples where both dAD approaches were able to produce prediction regions.

DTC (GPCR; SSRI) [paired]									
		Median		Error rates per confidence level (%)					
Dataset	Approach	α_δ	#calib	75%	80%	85%	90%	95%	99%
GPCR	Shafer & Vovk (7)	1.13	1500	25.02	18.85	14.29	10.00	5.16	1.07
	Papadopoulos (8)	2.26	1500	13.81	10.83	8.14	7.71	5.11	4.56
	Papadopoulos (9)	1.25	1500	23.92	18.46	14.6	9.48	4.94	0.78
	Papadopoulos (10)	1.25	1500	24.32	18.77	14.97	10.49	5.19	1.26
	dAD (CV)	2.21	874	3.64	3.11	1.67	1.14	0.51	0.11
	dAD (NN)	2.39	874	3.62	3.04	1.69	1.06	0.42	0.12
SSRI	Shafer & Vovk (7)	1.05	1500	24.7	19.13	14.59	9.53	4.25	1.05
	Papadopoulos (8)	1.91	1500	10.02	8.4	7.06	3.98	2.27	1.94
	Papadopoulos (9)	0.84	1500	30.19	25.74	23.02	16.3	8.62	3.37
	Papadopoulos (10)	0.96	1500	24.27	19.42	15.18	9.76	4.53	1.23
	dAD (CV)	1.82	234	4.25	3.18	2.53	1.55	0.72	0.31
	dAD (NN)	1.80	234	4.53	3.32	2.68	1.67	1.01	0.17

Table 5: Error rates (%) and median α scores for dAD (NN) and dAD (CV) with applied normalisation measure as in Eq. 9

SCKBA [dAD normalised]									
		Median		Error rates per confidence level (%)					
Approach	SX	α_δ	#calib	75%	80%	85%	90%	95%	99%
dAD (NN) [norm]	S1	1.81	315	5.82 (.73)	4.32 (.74)	3.16 (.77)	3.21 (.76)	2.05 (.74)	1.21 (.63)
	S2	1.32	253	19.83 (.74)	18.17 (.7)	15.01 (.63)	13.97 (.58)	13.16 (.59)	8.2 (.6)
	S3	1.36	279	21.86 (.69)	16.42 (.71)	15.11 (.73)	12.26 (.77)	9.52 (.79)	5.76 (.73)
	S4	1.01	232	83.44 (.26)	83.12 (.39)	77.29 (.56)	76.33 (.84)	70.46 (.98)	54.62 (.87)
dAD (CV) [norm]	S1	1.52	315	6.85 (.6)	5.32 (.57)	3.97 (.54)	4.7 (.45)	3.27 (.33)	0.0 (.19)
	S2	1.24	253	20.37 (.58)	19.66 (.56)	15.89 (.48)	14.68 (.42)	11.52 (.38)	9.27 (.28)
	S3	1.16	279	23.4 (.61)	20.35 (.6)	18.06 (.56)	14.13 (.55)	10.99 (.41)	6.61 (.18)
	S4	0.95	232	83.56 (.24)	83.09 (.34)	78.69 (.51)	75.23 (.72)	72.55 (.7)	49.63 (.45)

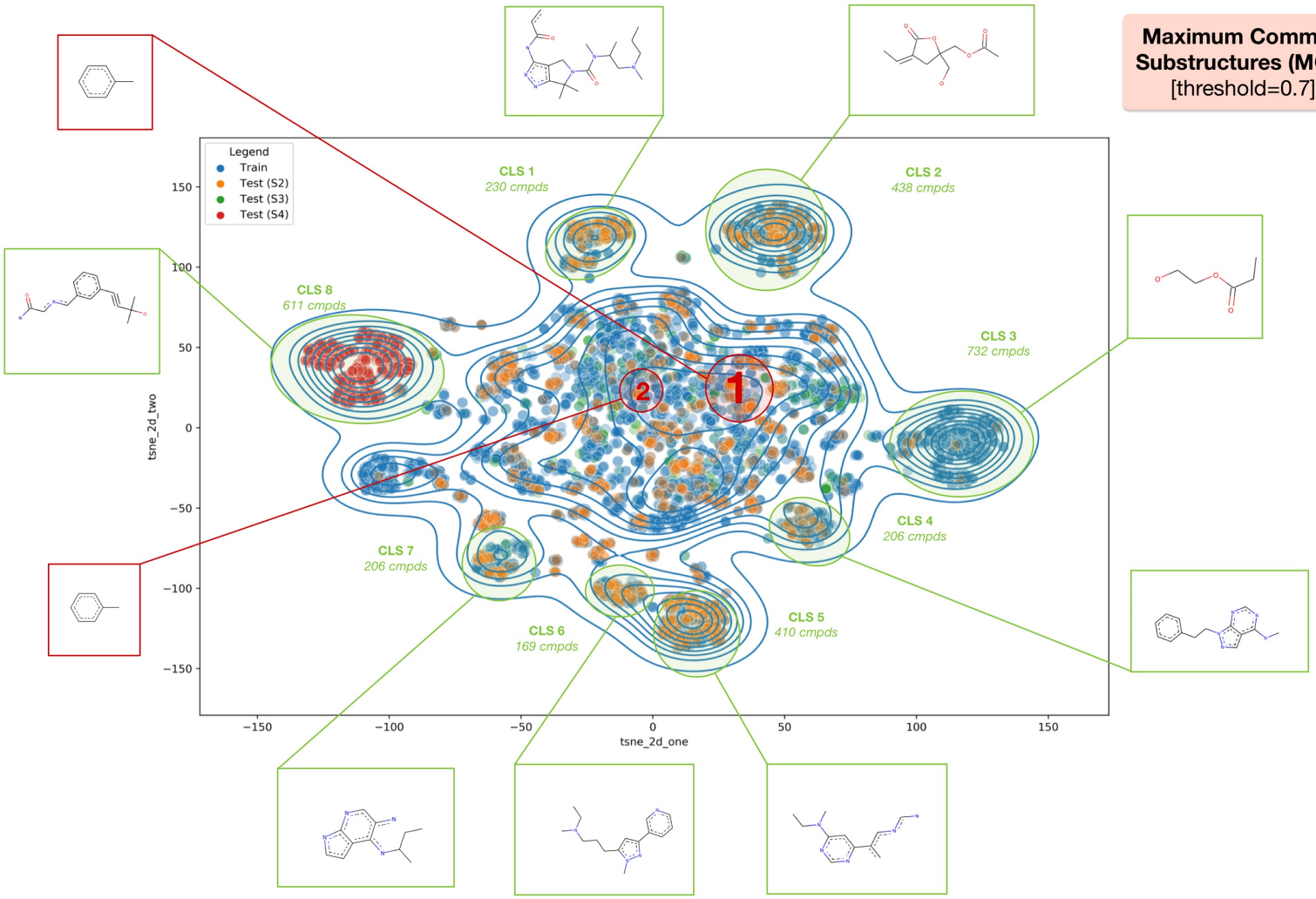


Figure 1: Results of t-SNE analysis performed over the whole chemical space of SCKBA dataset, consisting of the 7860 compounds. Compounds from the training set are shown in blue, test (S2-S4) compounds are shown in orange, green and red, respectively. Green circles show high density clusters on edges of the compound space, and for every cluster there is a maximum common substructure obtained with threshold of 0.7. Two red circles are arbitrarily assigned to the middle cloud, representing soft "clusters" with less separation between the compounds.

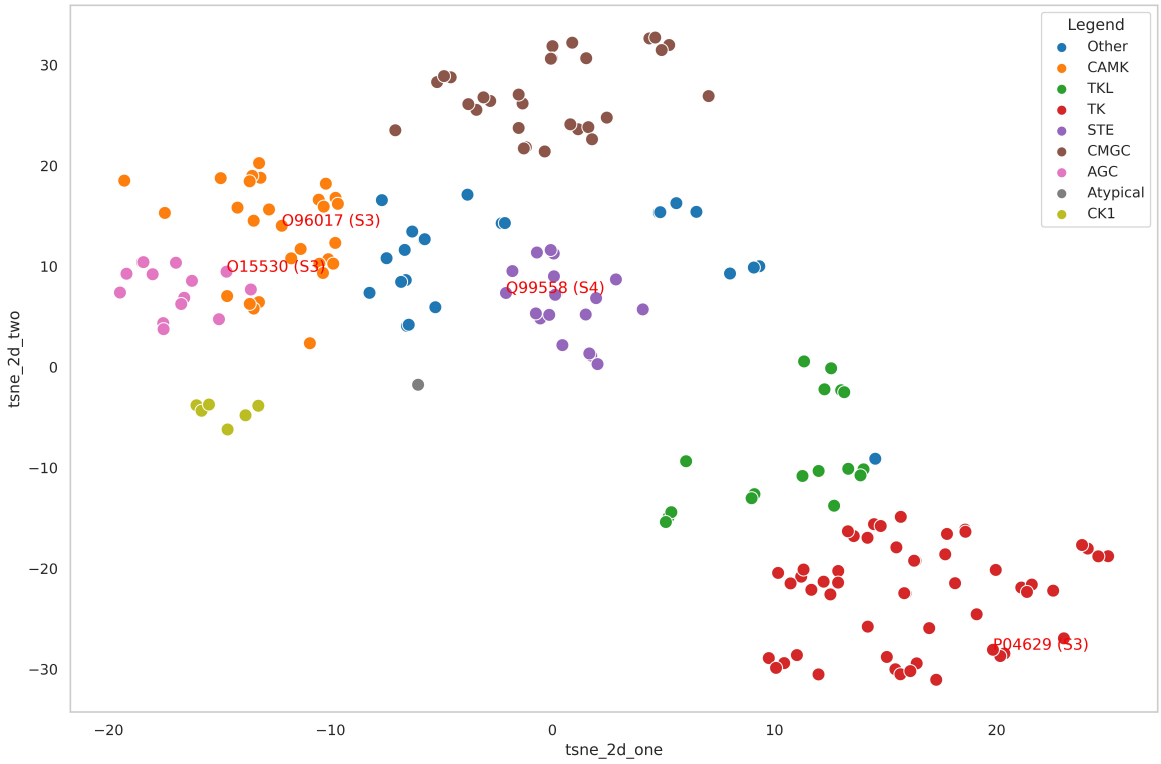


Figure 2: Results of t-SNE analysis performed over the 210 human kinases used in SCKBA dataset, with S3 and S4 targets denoted with Uniprot labels.

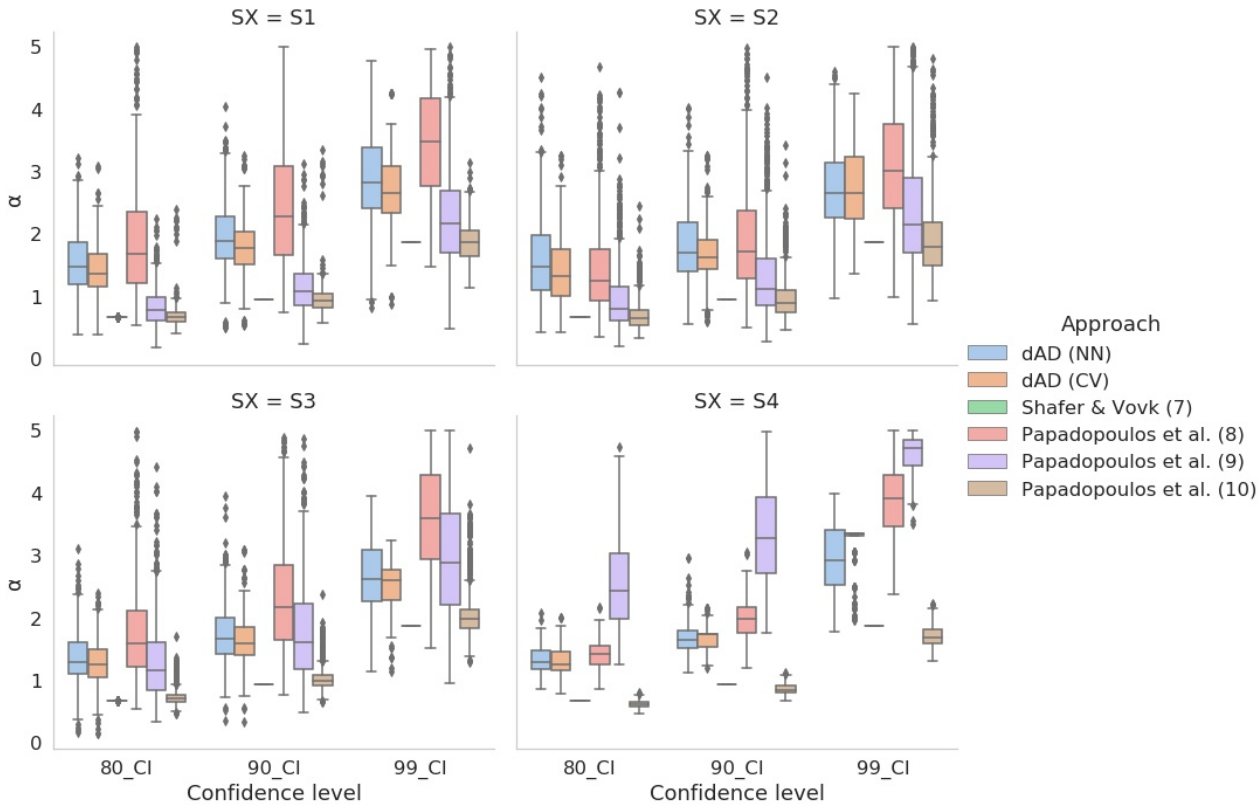


Figure 3: Comparison of original study by (Shafer and Vovk, 2008), studies with normalisation measures (Papadopoulos and Haralambous, 2010; Papadopoulos et al., 2011), and two proposed dAD variants (NN and CV) over the SCKBA dataset with four testing scenarios (S1-S4).

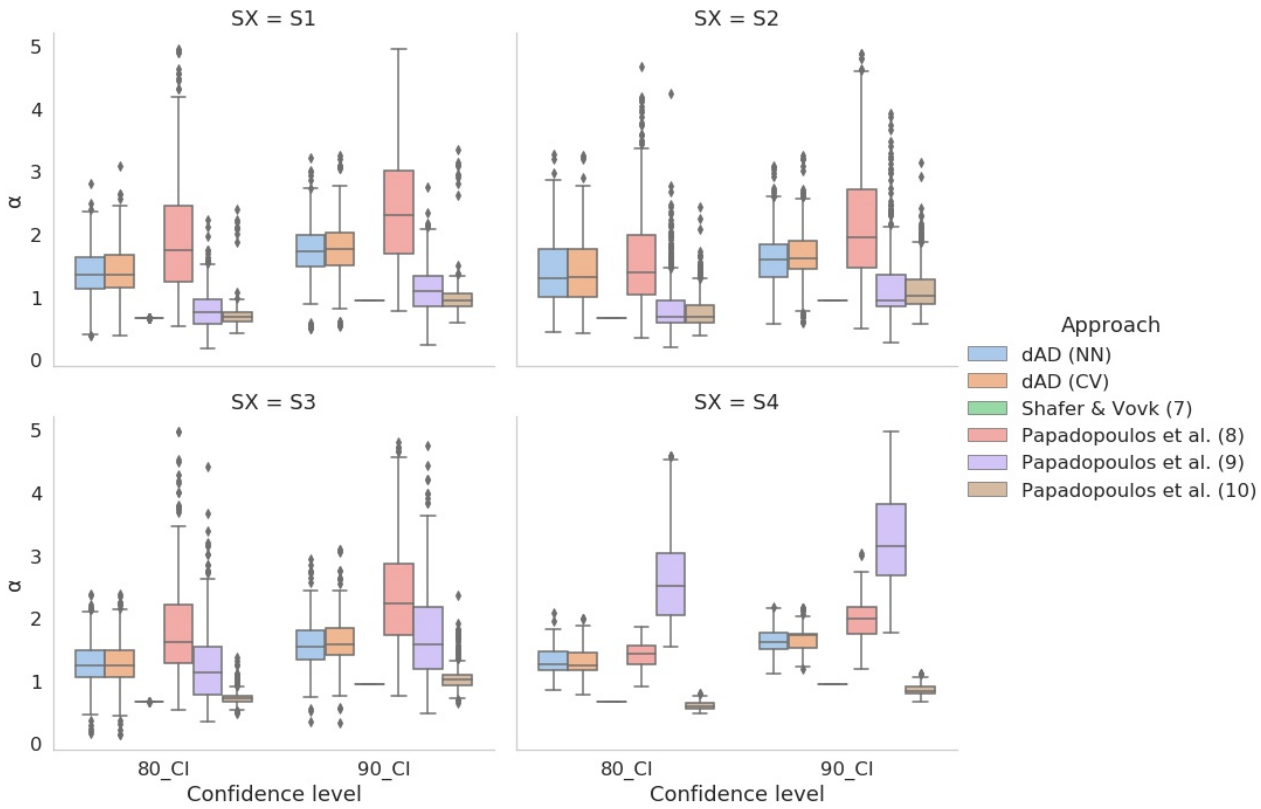


Figure 4: Comparison of original study by Shafer and Vovk (2008), studies with normalisation measures (Papadopoulos and Haralambous, 2010; Papadopoulos et al., 2011), and two proposed dAD variants (NN and CV). Results from all six of mentioned approaches are paired with indices of samples for which both dAD (NN) and dAD (CV) were able to produce prediction regions and compared over SCKBA dataset with four testing scenarios (S1-S4).

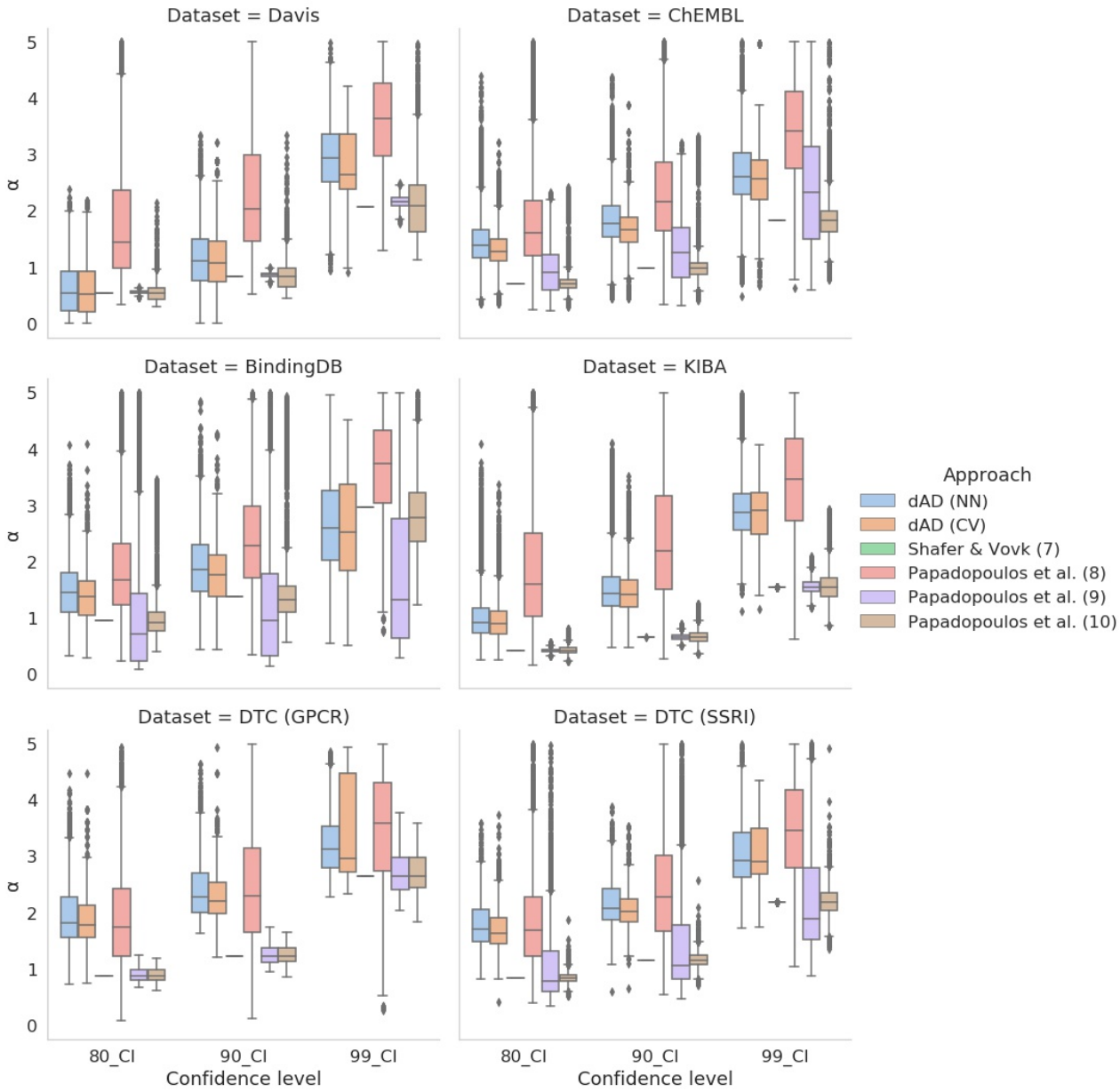


Figure 5: Comparison of original study by (Shafer and Vovk, 2008), studies with normalisation measures (Papadopoulos and Haralambous, 2010; Papadopoulos et al., 2011), and two proposed dAD variants (NN and CV) over six benchmark datasets.

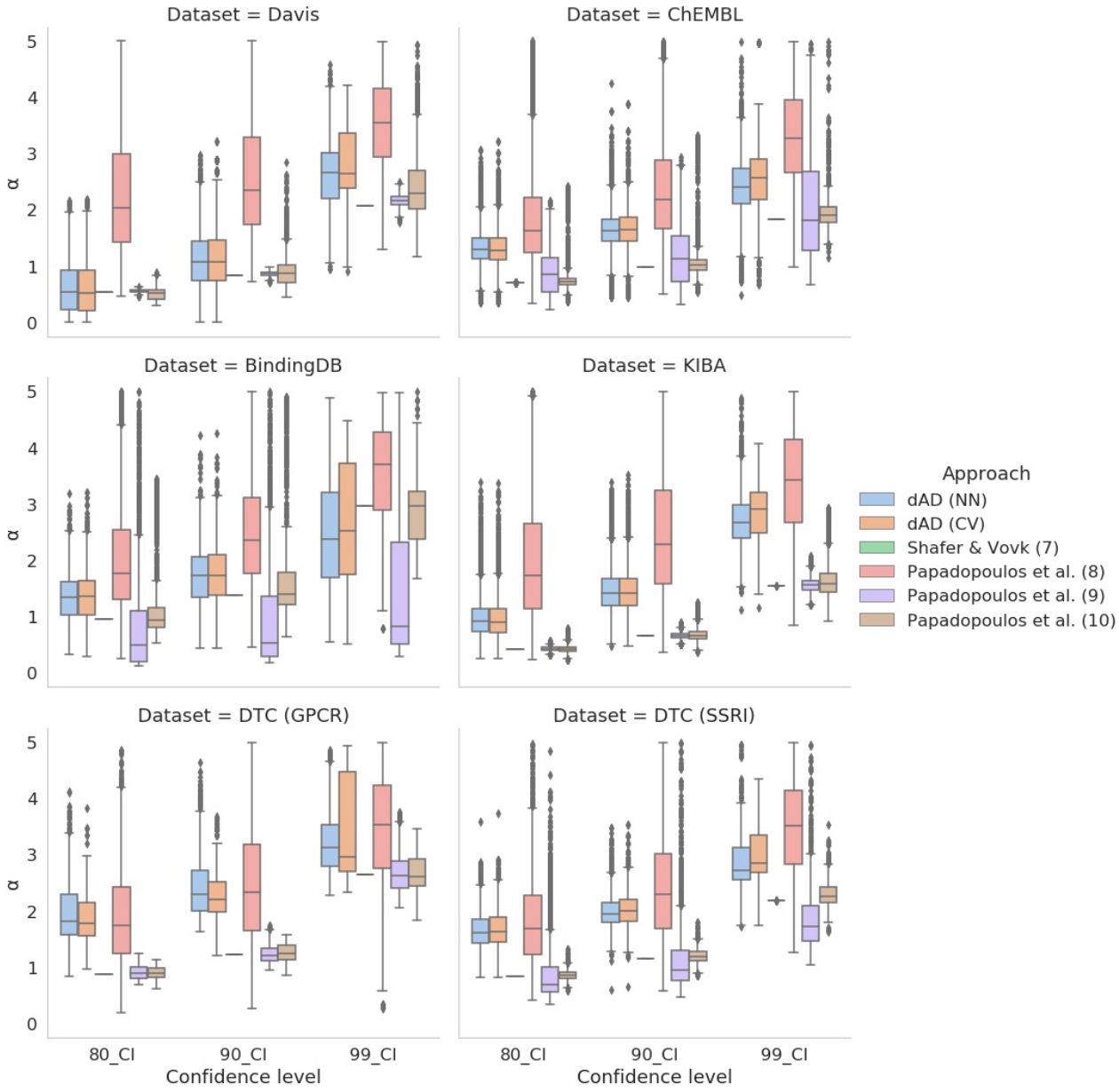


Figure 6: Comparison of original study by (Shafer and Vovk, 2008), studies with normalisation measures (Papadopoulos and Haralambous, 2010; Papadopoulos et al., 2011), and two proposed dAD variants (NN and CV). Results from all approaches over six available datasets are paired with indices of samples where for any confidence interval both dAD methods were able to produce prediction regions.

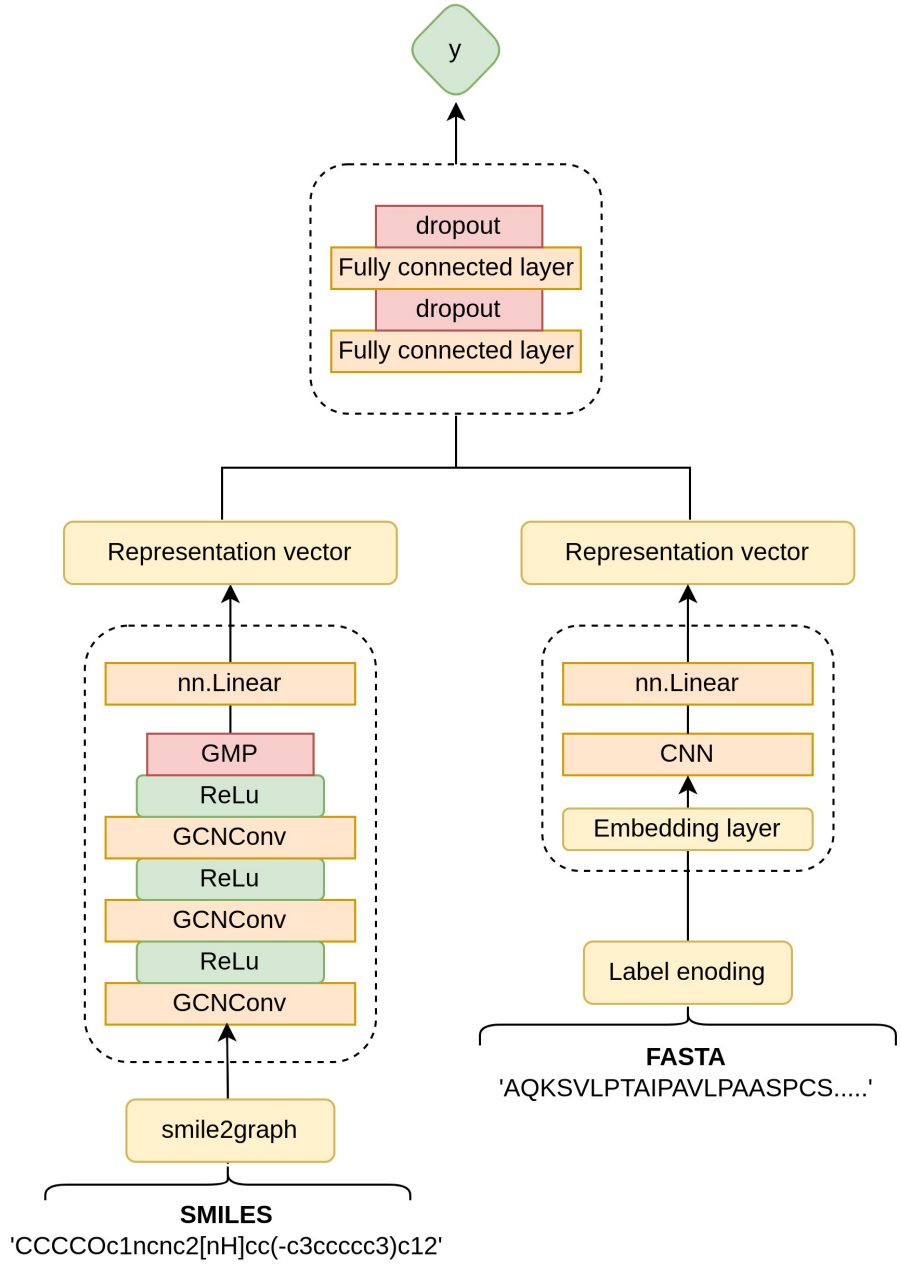


Figure 7: Schematic representation of the GCN-CNN architecture, with graph convolutional block taking SMILES as input and learning the molecular representation from graph features and convolutional block learning sequence representations from protein targets in FASTA format.

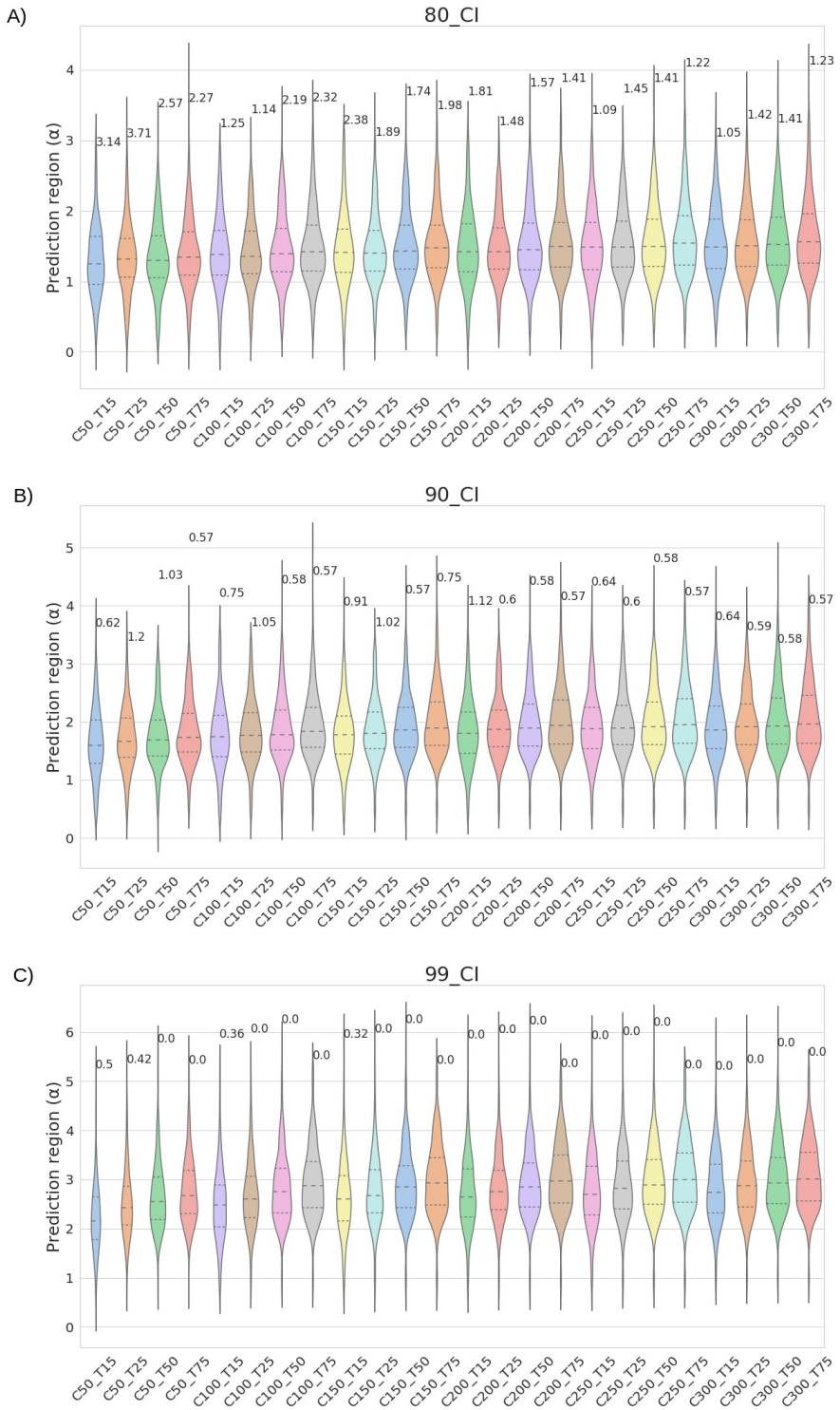


Figure 8: Figure shows the results in terms of prediction region distributions and error rates (values above each violin) for predefined confidence levels = [80%, 90%, 99%] for different number of nearest neighbours for compounds (k) and targets (q). Both hyperparameters were tuned for test (S1) scenario of SCKBA dataset.

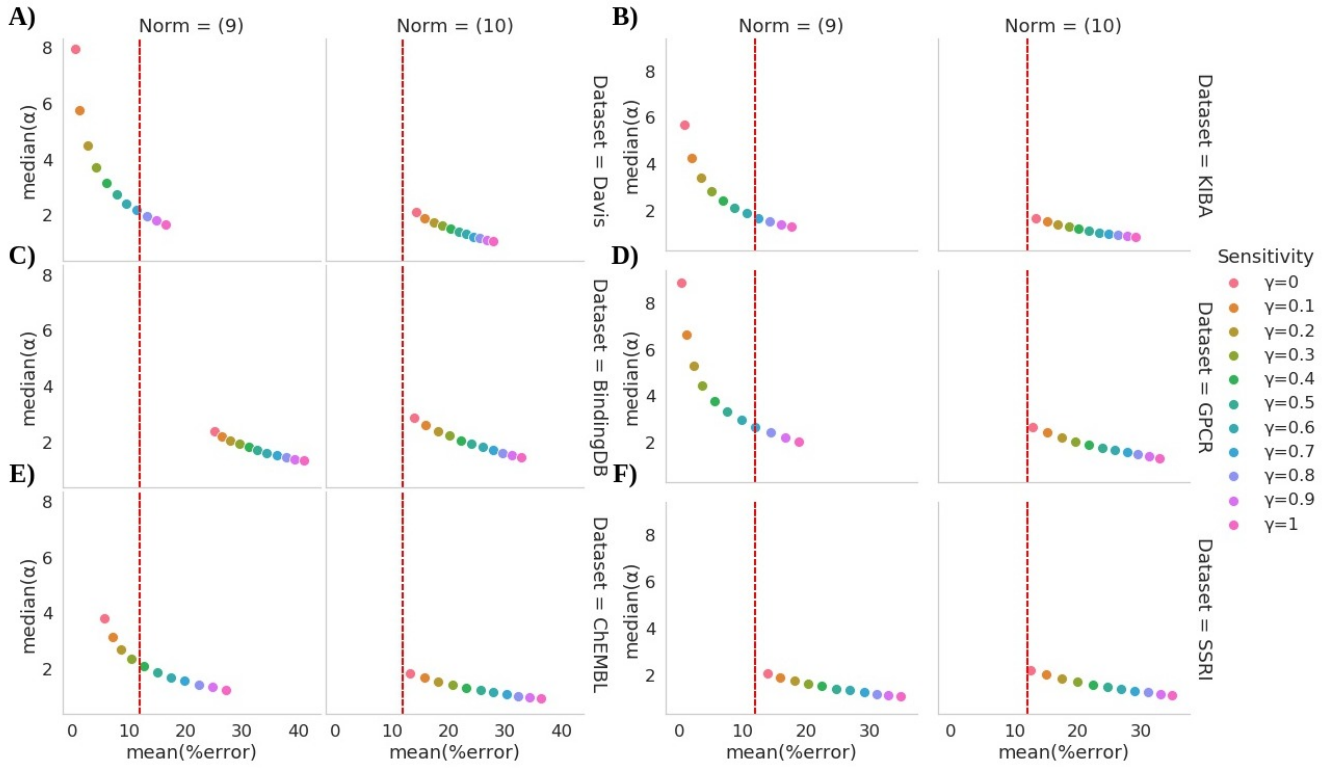


Figure 9: Figure shows the ratio between the mean error rates and median prediction region for different value of sensitivity parameter $\gamma = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$ for both Papadopoulos (9) and (10) approaches, tuned for the A) Davis, B) KIBA, C) BindingDB, D) DTC (GPCR), E) ChEMBL and F) DTC (SSRI) datasets.

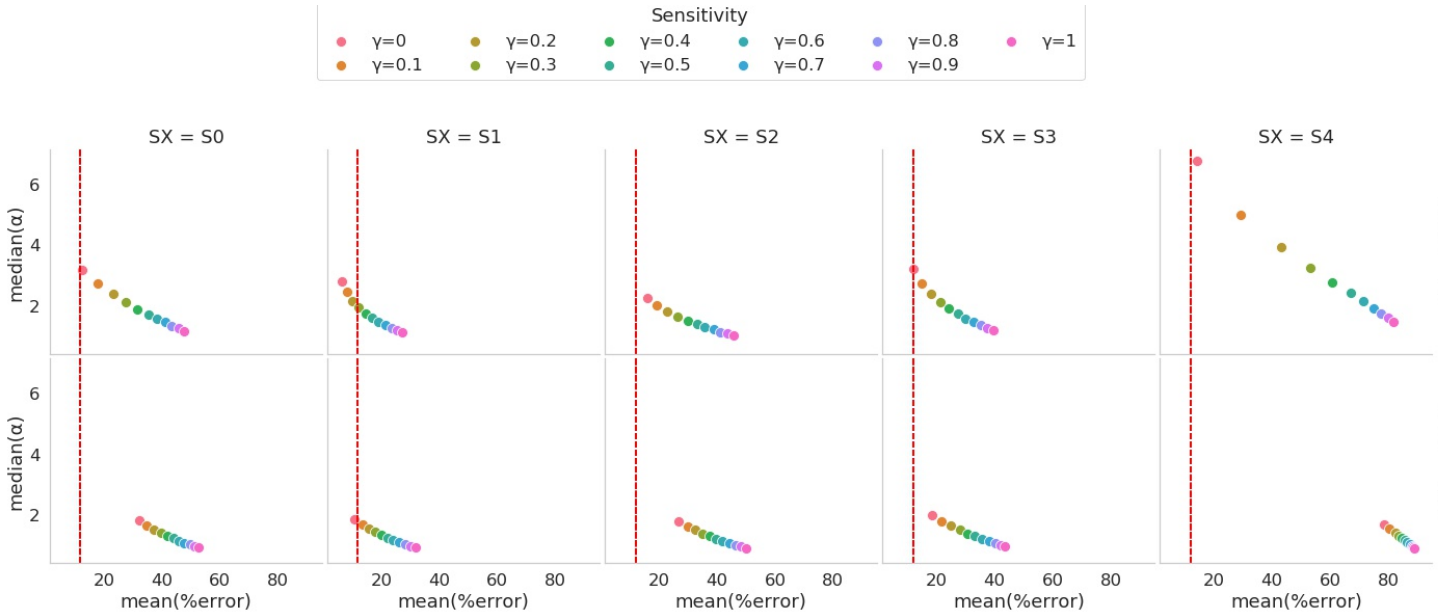


Figure 10: Figure shows the ratio between the mean error rates and median prediction region for different value of sensitivity parameter $\gamma = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$ for both Papadopoulos (9) and (10) approaches, tuned for all four test scenarios of the SCKBA dataset (S1-14).

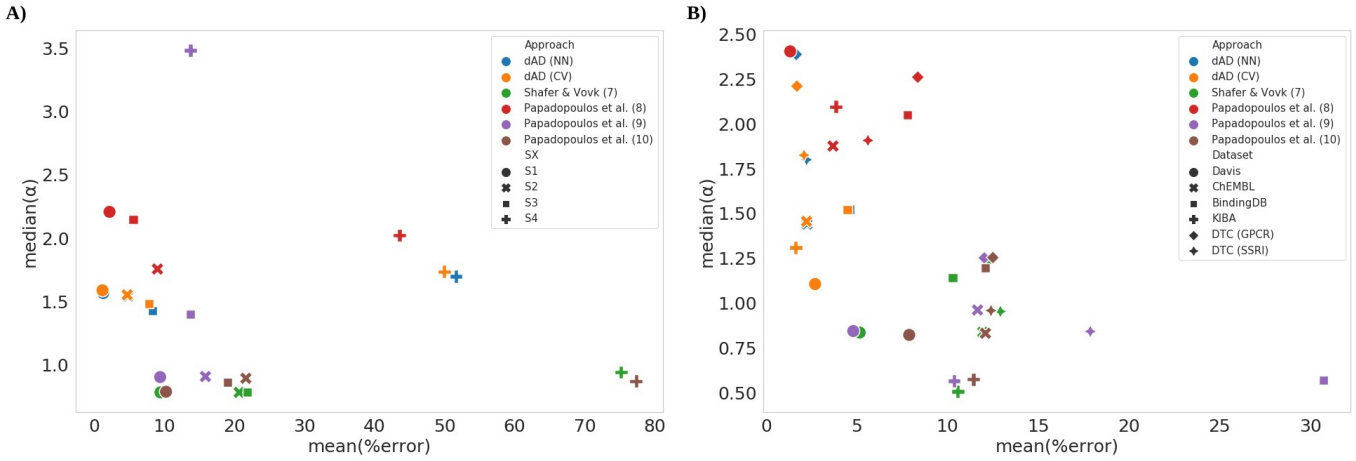


Figure 11: Comparison of the proposed dAD approach with the baseline studies by showing the relationship of mean (%) error and median nonconformity scores for A) four different testing scenarios (S1-S4) of SCKBA dataset and B) over six compound target datasets, Davis (Davis et al., 2011), KIBA (Tang et al., 2018), ChEMBL database (Gaulton et al., 2012), BindingDB (Gilson et al., 2016), DTC (GPCR) (Tang et al., 2018) and DTC (SSRI) (Tang et al., 2018) represented with different shapes. Results from all five of mentioned approaches are paired with indices of samples for which both dAD (NN) and dAD (CV) were able to produce prediction regions and compared over SCKBA dataset with four testing scenarios (S1-S4).

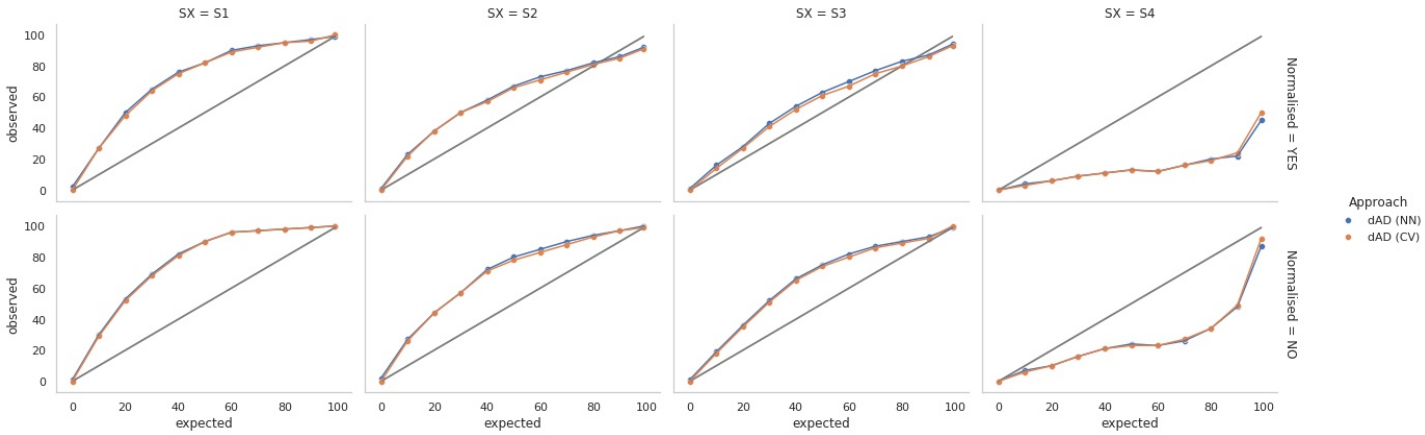


Figure 12: Linear relationship of expected and observed confidence levels for both normalised and non-normalised instances of dAD (NN), dAD (CV).

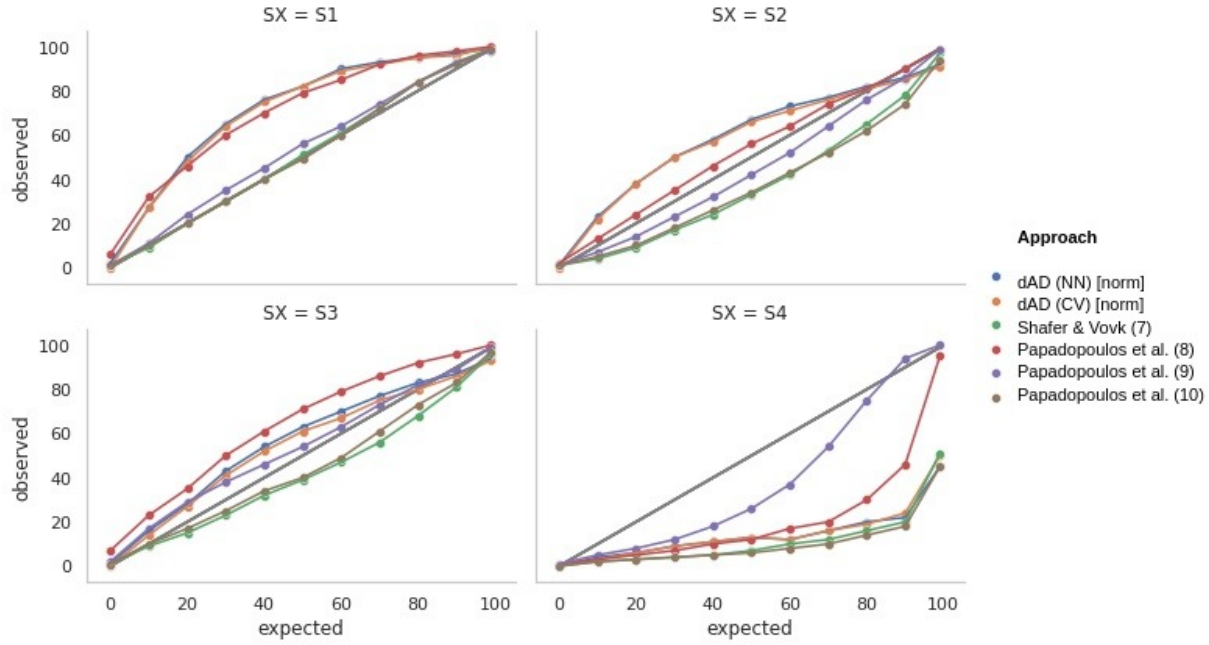


Figure 13: Comparison of the proposed dAD approach with the baseline studies over A) four different testing scenarios (S1-S4) of SCKBA dataset with paired indices across approaches and B) over six compound target datasets (Table 1) represented with different shapes, and also paired over indices to match in number of samples.

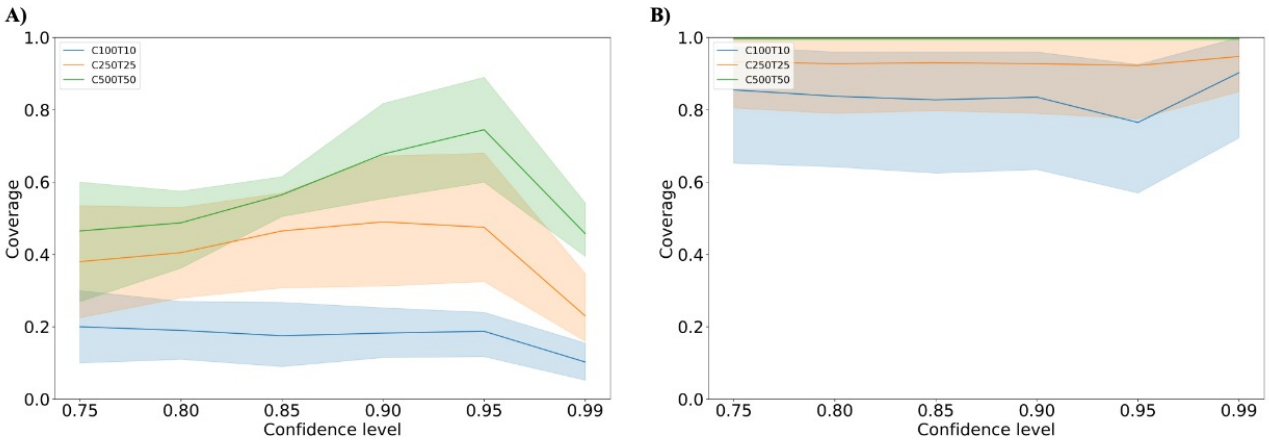


Figure 14: Demonstration of calibration size influence on the coverage of the proposed dAD approach with putative test noncoformity scores (A), and dAD coverage depending solely on calibration scores for prediction region estimation (B). Labels depict the compounds (C) and targets (T) with number of neighbours included in the conformity region of a test sample, e.g. C250T25 represents a conformity region defined by the 250 nearest neighbours in the compound space and 25 nearest neighbours in the target space of the training set.

References

M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1): D1100–D1107, 2012.

M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1): D1045–D1053, 2016.

H. Papadopoulos and H. Haralambous. Neural networks regression inductive conformal predictor and its application to total electron content prediction. In *International Conference on Artificial Neural Networks*, pages 32–41. Springer, 2010.

H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

J. Tang, B. Ravikumar, Z. Alam, A. Rebane, M. Vähä-Koskela, G. Peddinti, A. J. van Adrichem, J. Wakkinen, A. Jaiswal, E. Karjalainen, et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell chemical biology*, 25(2):224–229, 2018.

4. General discussion

Modeling molecular mechanisms of action is a well defined problem specifically established for drug mechanism prediction with application in drug discovery and re-purposing. The fundamental basis of most mechanisms of action is rooted in the molecular interactions that occur between the active substance and specific biomolecular targets. The targets encompass a variety of biomolecules, such as proteins, nucleic acids, lipids, and other essential cellular components. Interactions between molecules can manifest in two distinct manners: i) direct interactions, exemplified by the binding of an enzyme to its substrate, and ii) indirect interactions, which encompass phenotypic changes or biological readout in the affected organisms. The determination of potency and selectivity of the effect is heavily influenced by the binding affinity and specificity between the substance and its target.

The mechanisms of action have the potential to extend beyond the scope of specific biomolecular interactions, encompassing broader systemic and physiological effects. For instance, various substances have the potential to impact cellular metabolism, ion transport, and hormone signaling, resulting in systemic alterations in organ function or overall physiological responses. Understanding the systemic ramifications of substances is imperative in order to anticipate adverse reactions and evaluate their comprehensive therapeutic efficacy, as well as their cross effects, and long-term effects on the ecological niche. Depending on the level of abstraction, in this thesis, mechanism of action prediction modeling is defined on both fronts - one targeting specifically direct physical interactions represented in the form of binding affinities, alluding to the strength of the interaction between interacting entities, as defined in Paper 2 and 3. And the other relating to the mechanism of action encompassing perturbation effects of compounds resulting in an observable change in the affected cells, Paper 1.

In light of this, modeling the mechanism of action as the bioactivity of chemical compounds characterized by phenotype change in response to perturbation is a simpler

approach that is highly dependent on the similarity-activity relationship of the investigated compounds. As is the case with herbicide activity predictions, trained models presume that similar compounds will have the same effect on the affected weeds. In addition, because the model is trained exclusively on the compound representations, all training samples are drawn from the same distribution, which simplifies validation and the definition of the applicability domain. Consequently, the herbicide activity dataset is first used to examine the effects of compound representation. Herbicide space is described by a set of physicochemical characteristics and a set of structural signatures, as available in the **R** library **rcdk**.
The right side of Figure 1 in Paper 1 illustrates the comparative performance of all tested algorithms. The posterior distribution frequency is represented by the blue line, while the region of practical equivalence (ROPE), is indicated by the space between the yellow lines. Based on the comparative analysis and the overall interpretability, the random forest algorithm was selected as the preferred approach for conducting further modeling of herbicidal activity. Various representations of compound space were examined and assessed in relation to the performance of random forest (RF) and hierarchical clustering (HC), using several representation metrics for the two types of problems modeled by RF and HC. Figure 4.1A presents performance measures for four physicochemical descriptor sets, along with the average performance of structural fingerprints. On the other hand, Figure 4.1B displays the performance of each available structural fingerprint representation in the rcdk library. The most effective model employs MACCS fingerprints, a 166-bit binary vector consisting of a series of queries with responses represented as 0 or 1. The set of descriptors is relatively straightforward and interpretable to a high degree. This implies that for a homogeneous set of synthetic compounds, such as the HRAC set of herbicides, structural signatures reflect the behavior of these compounds more accurately than more complex and abstract physicochemical features.

Accordingly, Figure 4.1 indicates that simpler structural depictions of chemical structures produce a more accurate and robust predictive model, as they are better in line with the rationale of how these compounds were grouped in dedicated classes in the first place, using manual curation and visual inspection of biological readouts, as explained in Paper 1. Figure 4.1C contains an additional table containing a random sample of predictions generated through the use of a random forest with MACCS signatures. This section seeks to provide an illustration of how this method contributes to the evaluation of prediction accuracy. Upon observation, it

is clear that the first four compounds possess a relatively high probability. In addition, these compounds are part of extremely homogeneous clusters, which strengthens the validity of these predictions. The samples with a probability greater than 0.60 and a homogeneity greater than 80% were color-coded green. In the case of trifludimaxazin, the probability is measured at 0.5 and the homogeneity of the dedicated cluster is only 35%, which indicates that a substantial portion of samples within dedicated clusters are from classes other than their own.

As shown in Figure 4.1, representation of the chemical space of compounds bears great importance in terms of model predictive power, with structural depictions explaining most of the variability in the dataset of herbicides. Due to the simplicity of this dataset, the definition of the applicability domain for this approach is more straightforward, thus allowing for definition of direct boundaries in the two dimensional space (Figures 3-4 in Paper 1).

A)

	Random Forest		Hierarchical Clustering					
			Internal Validation		External Validation			
	10-CV Acc	Kappa	Si	Dunn	ARI	VI	Purity	Entropy
hybrid	*0.74*	*0.7*	*0.2*	*0.44*	*0.14*	*3.59*	*0.34*	*0.61*
constitutional	*0.76*	*0.73*	*0.14*	*0.39*	*0.19*	*2.81*	*0.47*	*0.49*
electronic	*0.62*	*0.58*	*0.16*	*0.20*	*0.19*	*3.06*	*0.42*	*0.52*
topological	*0.85*	*0.84*	*-0.01*	*0.52*	*0.01*	*3.00*	*0.86*	*0.16*
fingerprints	*0.85±0.03*	*0.83±0.03*	-	-	*0.29±0.05*	*2.56±0.19*	*0.50±0.03*	*0.44±0.03*

B)

	Random Forest		Hierarchical Clustering			
	10-CV Acc	Kappa	ARI	VI	Purity	Entropy
standard	*0.85*	*0.84*	*0.32*	*2.46*	*0.51*	*0.43*
extended	*0.87*	*0.85*	*0.29*	*2.51*	*0.51*	*0.44*
graph	*0.86*	*0.85*	*0.19*	*3.03*	*0.42*	*0.52*
hybridization	*0.87*	*0.85*	*0.3*	*2.6*	*0.5*	*0.45*
maccs	*0.88*	*0.87*	*0.35*	*2.45*	*0.5*	*0.42*
estate	*0.77*	*0.75*	*0.28*	*2.69*	*0.48*	*0.46*
pubchem	*0.85*	*0.84*	*0.31*	*2.58*	*0.52*	*0.45*
kr	*0.83*	*0.81*	*0.33*	*2.42*	*0.51*	*0.44*
shortestpath	*0.84*	*0.83*	*0.29*	*2.45*	*0.51*	*0.42*
circular	*0.84*	*0.82*	*0.23*	*2.41*	*0.49*	*0.41*

C)

	HRAC	Random Forest		Hierarchical Clustering	
		Class	Prob	Cluster	Homogeneity
sulfallate	Rest	N	0.8	N	94.00%
ethametsulfuron	Rest	B	0.98	B	89.00%
methyldymron	Z	C2	0.97	C2	82.00%
phenobenzuron	Rest	C2	0.58	C2	82.00%
trimeturon	Rest	C2	0.61	C2	82.00%
difenozoquat	Z	D	0.53	D	75.00%
dichlormate	Rest	K2	0.53	K2	60.00%
chloropon	Rest	N	0.59	Z1/N	63.00%
alorac	Rest	N	0.57	Z1/N	63.00%
trifludimaxazin	Rest	E	0.52	E	35.00%

Figure 4.1: Exploring the feature space and predictive power of trained models, including a set of physicochemical features and structural fingerprints, as included in the R library, rcdk (A,B). Table C depicts a sample of random forest results trained with a 166-bit MACCS fingerprints as best performing set of features, over a sample of compounds belonging to the Z and Rest classes, as defined in Paper 1, with additional confirmation of model predictions by applying hierarchical clustering and determining the cluster quality via internal (silhouette index - SI, Dunn index) and external validation (Adjusted Rand Index - ARI, purity, entropy, homogeneity).

The next level of abstraction, as described in Paper 2, pertains to the direct physical interactions that occur between compounds and protein targets. This phenomenon is particularly evident in the field of drug development, where there is a focus on identifying specific protein targets that belong to a certain protein family or are associated with a specific condition or disease of interest. The assessment of herbicidal activity for compounds may be

Table 4.1: Comparison of GCN-CNN approach with different protein kinase representations, one applying representation learning from whole protein sequences (WSEQ) that can range from 1000 to over 1430 amino acids, and another learning directly from protein kinase domain sequences ranging from 300 to 606 amino acids in length. SX denotes any of the testing scenarios described in Paper 3 excluding the S4 test scenario. GCN-CNN approach was trained for 200 epochs, with learning rate of 0.0005 and batch size of 128.

			WSEQ				DSEQ			
Model	SX	Compounds	MSE	Pearson	AUC	CI	MSE	Pearson	AUC	CI
GCN-CNN	S1	graph	0.351	0.871	0.785	0.860	0.279	0.901	0.801	0.876
	S2		0.806	0.645	0.698	0.713	0.820	0.639	0.729	0.705
	S3		0.998	0.496	0.664	0.654	0.576	0.693	0.799	0.709

considered a limited perspective, as it primarily concentrates on the structural features of the compounds being tested. However, when predicting the bioactivity of compounds intended for human use, a more comprehensive understanding of both interacting entities, the bioactive compound and the target protein, is required.

The targeting of the human kinome has been the subject of extensive research spanning several decades, resulting in a substantial amount of data that continues to be created to this day. In recent years, the advancement of sophisticated statistical frameworks and machine learning techniques has led to a proliferation of proposed algorithms in the academic literature aimed at addressing the challenge of predicting binding affinity (Cichonska et al., 2017). It is noteworthy that the increase in complexity of suggested methodologies does not consistently correspond to improved model predictions, in terms of both accuracy and interpretability. However, it has been demonstrated that validating a trained model on stratified test sets does not accurately reflect the model's capacity to generalize on unknown data. In fact, this approach typically results in overfitting on the training data (Pahikkala et al., 2015; Cichonska et al., 2017). This issue poses a considerable challenge when attempting to use these methodologies on empirical data. One of the primary objectives of this thesis was to develop a compound-kinase binding affinity predictor that exhibits improved generalization capabilities when applied to novel samples. These samples may consist of either new compound-target combinations or new compounds and targets considered individually.

To accomplish this, it is also important to gain a deeper understanding of how the model leverages the given data. As part of the analysis from Paper 2, several views of this problem are outlined. The first perspective to consider is the representation of protein kinases, which

Table 4.2: Examining impact of individual representation learning blocks, with GCN performing direct representation learning from graph structure of compounds and CNN learning from the protein kinase domain sequences (dseq).

Model	Test (SX)	Compounds	Kinases	MSE	Pearson	AUC	CI
GCN	S1	graph	-	0.640	0.746	0.698	0.784
	S2			0.929	0.573	0.628	0.717
	S3			0.859	0.466	0.691	0.666
CNN	S1	-	dseq	1.013	0.546	0.617	0.692
	S2			1.378	0.193	0.517	0.582
	S3			1.322	-0.054	0.500	0.502
GCN-CNN	S1	graph	dseq	0.339	0.875	0.839	0.865
	S2			0.782	0.672	0.703	0.728
	S3			0.569	0.694	0.708	0.729

can be examined either as complete protein sequences ranging from 1000 to 1430 amino acids or as protein kinase domains with a length of 300 to 606 amino acids. As demonstrated in the case of herbicides, the manner in which chemical space is represented can greatly influence the performance of models, particularly in the context of unsupervised learning. Table 4.1 suggests that when employing identical hyperparameters, the utilization of solely domain sequences results in superior prediction accuracy across all evaluation metrics in all three testing scenarios. Furthermore, from a physical standpoint, it is reasonable to represent protein kinase sequences in a more succinct manner, specifically by excluding extraneous regions that do not significantly contribute to the physical interaction between inhibitors and their binding sites. As demonstrated in Table 4.2 models trained on drug-kinase interacting pairs assign greater weight to the chemical space of compounds in their decision-making process compared to protein kinases. The optimal performance is observed when both input spaces are combined into a single space, but learning from the chemical space alone results in substantially superior performance than learning from the protein kinase space alone. One of the possible causes for this behavior could be the abundance of information used to train the model, with chemical space representing an abundant space with a large number of samples and a high level of scaffold diversity. In contrast, the protein kinase space is significantly more conserved and presents only a small fraction of its interacting partner. When contrasted to the conserved space of protein kinases, this may lead to the conclusion that the compound space accounts for the majority of the data's variability.

Table 4.3: Performance measures for GCN-CNN and XGBoost approach over four testing scenarios, as described in Paper 3

A)

Model	SX	MSE	Pearson	AUC	CI
GCN-CNN	S1	0.287	0.872	0.795	0.865
	S2	0.719	0.756	0.784	0.771
	S3	0.935	0.600	0.596	0.723
	S4	3.544	0.075	0.499	0.526

B)

Model	SX	MSE	Pearson	AUC	CI
XGBoost	S1	0.240	0.891	0.811	0.850
	S2	0.617	0.792	0.788	0.796
	S3	0.934	0.574	0.600	0.686
	S4	4.00	-0.120	0.500	0.440

In contrast to the herbicide activity task, the two-entity interaction problem entails a special challenge in which the applicability domain is not easily confined to two dimensions without excessively simplifying the intricate nature of interacting pairs. Casting model prediction and average error rates for every testing scenario, as given in Figure 4.2 gives insight into the importance of compound representations and model behavior in different sub-regions of compound space. t-SNE analysis of the compound space is based on Morgan fingerprints with radii of 2 units. Figure 4.2A relates to the first testing scenario (S1), which involves a stratified sample for testing, and produces high quality predictions with most of the average error rates lower or equal to 1 unit. Given that compounds in S2 test are not encountered during the training phase, it is not surprising that the S1 test scenario performs significantly better on the initial comparison. The majority of the first testing scenario compounds populate the low density cloud in the middle, but due to their presence in the training set, the model still obtains a high level of accuracy. The S2 testing scenario is more difficult, but the model still yields low average errors, particularly in regions of high density in neighboring clusters. High quality performance in the S2 scenario is reflected in the fact that S2 compounds, even though they are not part of the training space, contain many similar counterparts with experimentally measured binding affinities in the training set.

Interestingly, the third testing scenario (S3), according to both Table 4.3 and Figure 4.2, demonstrates a minor lag in model performance, which we attribute to the majority of compounds occupying the low density middle cloud. The fourth testing scenario (S4) is comprised of a completely out-of-distribution cluster of compounds and exhibits nearly random behavior. Interestingly enough, although the S3 test set contains known compounds, its efficacy is somewhat inferior to that of the more difficult S2 test set. This appears to be a result of the close proximity of the compounds chosen for the test set. Compounds in the S2 test set originate from dense regions of training samples and thus have many similar neighbors,

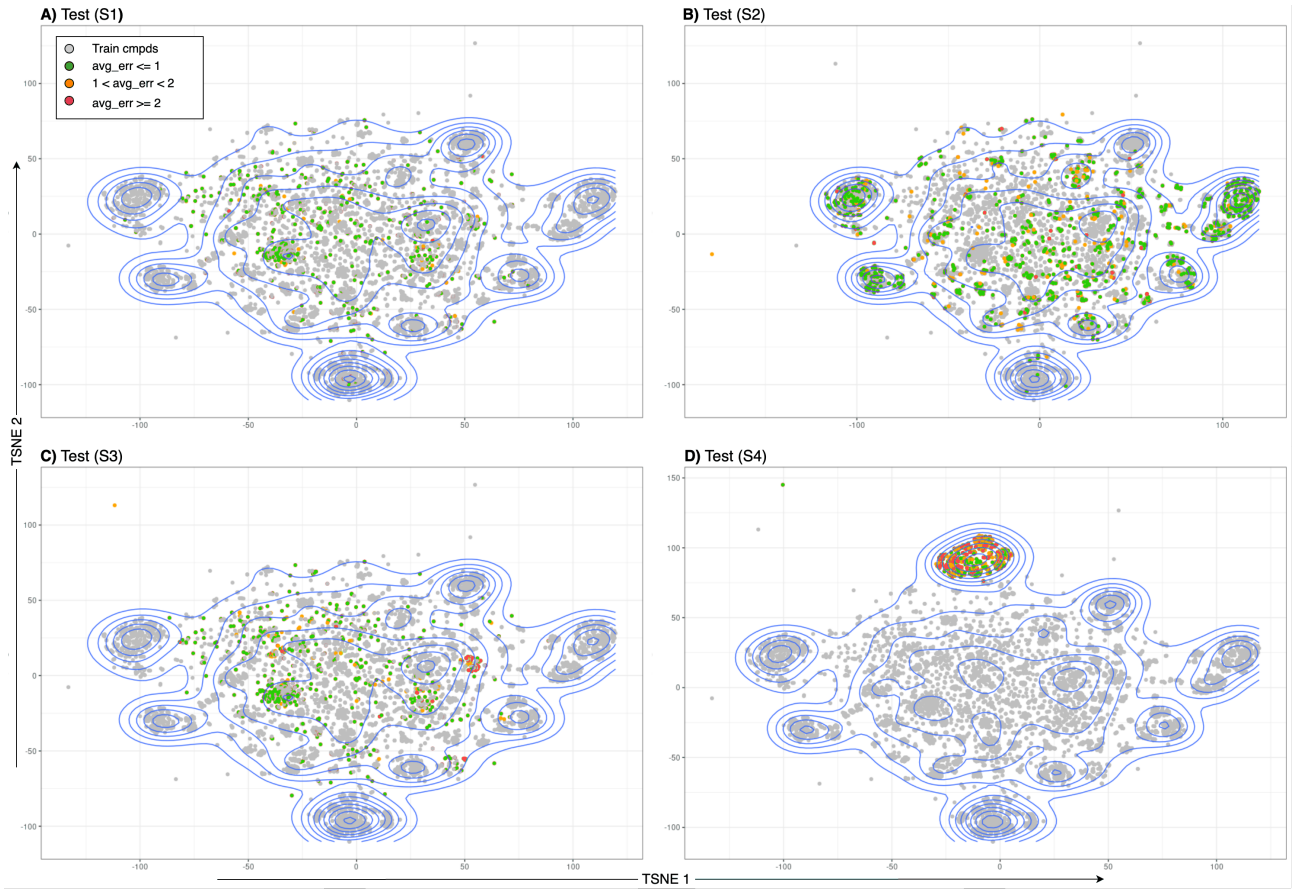


Figure 4.2: Model evaluation according to various testing scenarios. Test compounds are color-coded based on their average absolute error rate, with red representing an average error rate of less than 1 unit, orange representing an average error rate of between 1 and 2 units, and green representing an average error rate of greater than or equal to 2 units. Which are projected over training compounds in gray and combined with a 2D kernel density estimation in blue.

whereas compounds in the S3 test set come from a cloud with higher chemical diversity.

Since, there is no direct method to define the applicability domain for this task, similar logic to the herbicide task was used. The applicability domain is defined in the context of the compound space due to its higher impact on model performance and sheer diversity compared to the human kinome, Figure 4.3. As given in Figure 4.3, it shows only one perspective of model application, and there is another potential caveat when observing the density of training samples, denoted with by yellow line in Figure 4.3. Following the density estimate over the training space, it is noticeable that the training space comprises several high density clusters and a more spread out middle cloud with a few low density regions. This problem is best described in Aniceto et al. (2016), and the potential ramifications it can have on defining the applicability domain in a highly dimensional spatial context.

In order to account for the discrepancy in the compound space and incorporate both

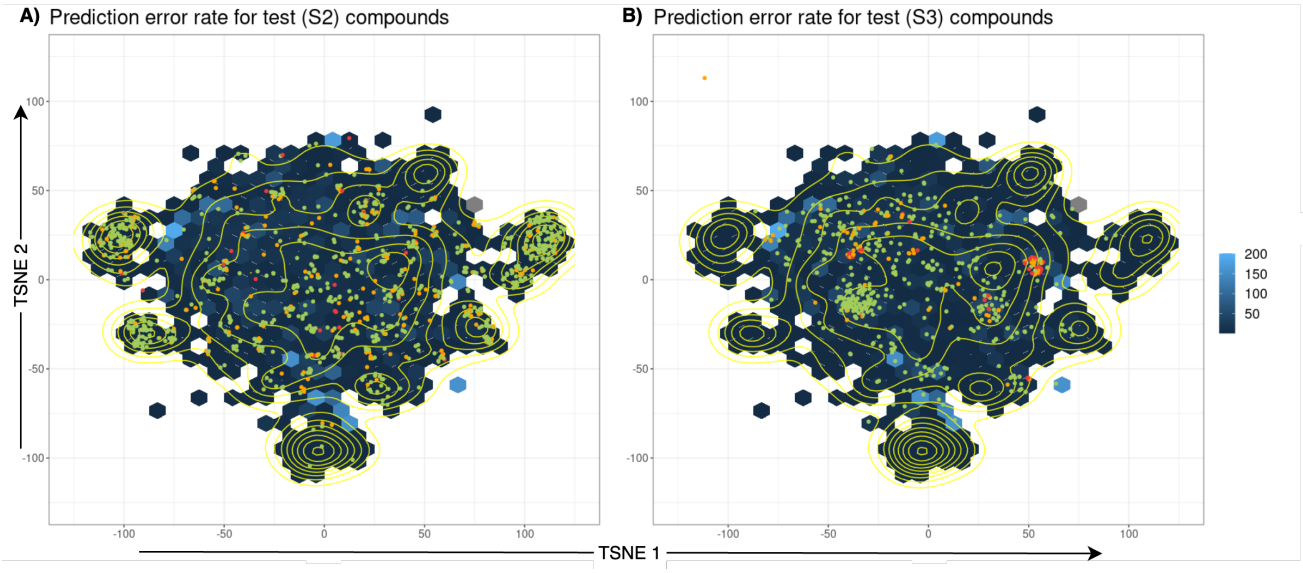


Figure 4.3: Definition of applicability domain for S2 (A) and S3 (B) testing scenarios with compounds plotted in the t-SNE space. Blue hexagons represent training compounds with a color gradient respective to the interaction count of accounted area. Test S2 and S3 compounds are color coded based on the average error rates, with error rates <= 1 shown in green >1<=3 in orange, and >2 unites in red. Yellow lines over compound space depict the kernel density of training samples for that designated area. The blue gradient on the right denotes the number of experimentally measured interactions.

interacting entities with inherently different distributions, the new concept is proposed in Paper 3. The application of the dynamic applicability domain (dAD) is illustrated in Figure 4.4. The concept of dAD can be described as an extension of the conventional application domain, using the principles of inductive conformal predictors. The study conducted in Paper 3 integrates the notions of applicability domain and conformal predictors to provide improved assurances for bioactivity prediction tasks and extends this to interaction-type problems in the context of drug discovery. This extension is particularly relevant for binding affinity prediction. The problem of compound-target binding affinity is characterized by the presence of chemically and biologically distinct spaces that are independently distributed. Models developed using these datasets may demonstrate varying levels of effectiveness across different subsets of compounds and target families. This variability is influenced by factors such as the distribution of compounds and target families in the training set, the distribution of measured affinities, and the properties of the model itself.

The proposed method is equally applicable for the chemical space of herbicides for mode of action prediction or for prediction of other chemical properties - as it also allows for the extension of a traditional quantitative structure-activity relationship modeling approach to more complex interaction tasks. The dynamic applicability domain, as defined in Paper 3 has

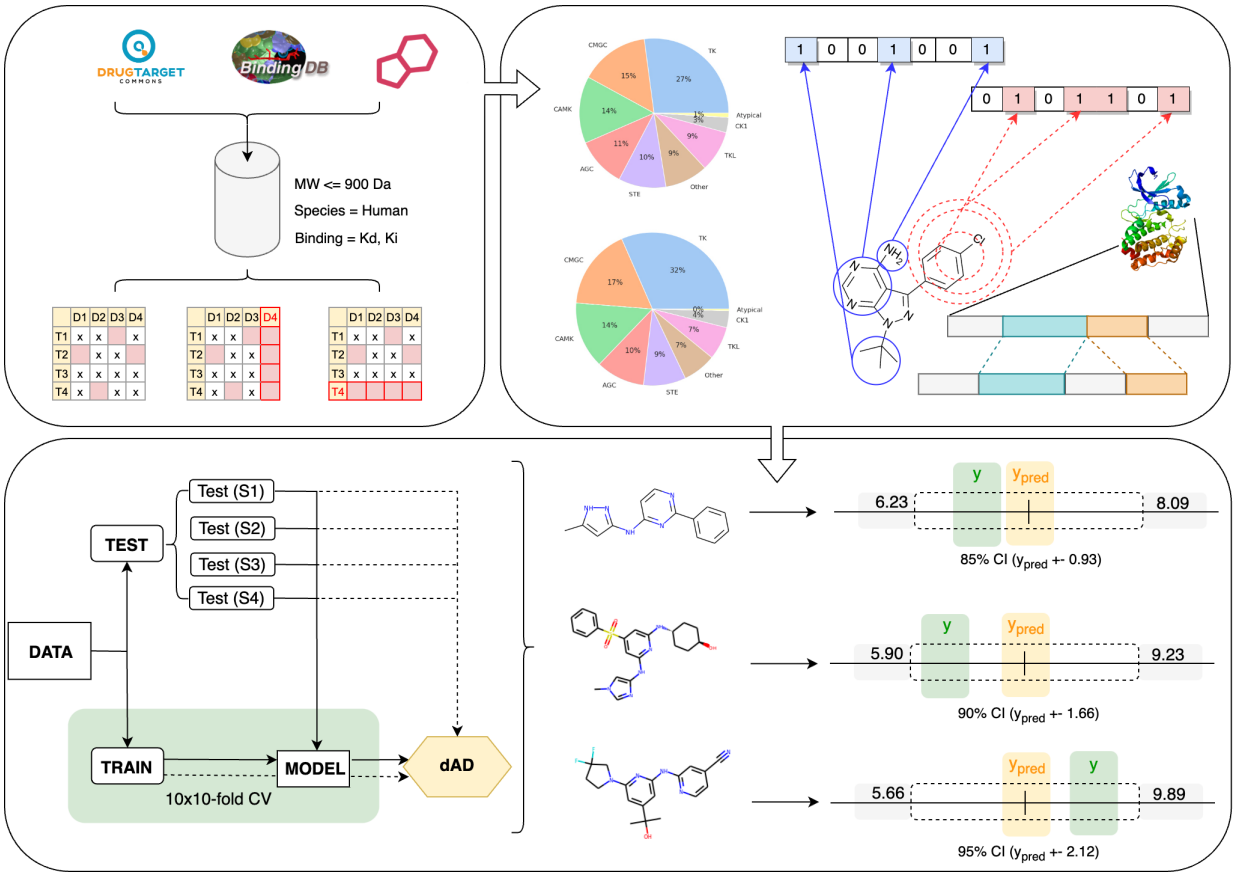


Figure 4.4: Dynamic applicability domain workflow from data processing, compound and protein target space representation, to prediction region estimates for a predefined confidence level.

been rigorously tested over several different interaction datasets and has exhibited very promising results on different difficulty scenarios. On the other hand, an emergent feature of dAD includes refraining from prediction, which is a common property of conformal predictors for classification problems. Regardless of whether it is related to the molecular mechanism of action prediction in terms of phenotype change or direct physical interactions, this property reduces the rate of false positives in more challenging prediction contexts, primarily for S2 and S3 testing scenarios. Another property of dAD relates to extraction of smaller calibration sets, which restricts the method's application to problems with a larger number of samples available for model training and subsequent dynamic calibration. However, when this requirement is met, Paper 3 demonstrates that the method's performance is stable over a relatively large range of calibration set sizes. Furthermore, this approach extends upon the concept of applicability domain as defined in Paper 1 and Figure 4.3, by defining the conformity regions for individual entities, thus defining the applicability domain in the subspace of the training set avoids the limitations of defining direct boundaries in the two-dimensional space. The dAD could be applied to one or more entity interaction problems with the goal of providing the explainability

of individual predictions by utilizing localized calibration of predictions, possibly extending the application to other biological interaction-type tasks.

5. Conclusions

- This thesis has made a few steps forward in advancing the application of machine learning in drug discovery and re-purposing, specifically in the context of modeling mechanisms of action. One of the key contributions includes the definition of a systematized machine learning framework tailored for the classification of a wide range of compounds and the prediction of mechanisms of action for the underrepresented group of compounds with phytotoxic activity. This is essential given that this compound class has seen a large growth in use over the last decade, yet the risk of abuse can have serious effects on ecological niches that were not the original aim of its activity. A designed compound classification pipeline focusing on phytotoxic properties has successfully categorized a select number of natural compounds according to their specific mechanisms of action, as defined by the Herbicide Resistance Action Committee (HRAC). This systematic approach has enhanced our understanding of how these natural substances function at a molecular level in plant systems.

- When training models on large datasets dealing with direct physical interactions between compounds and targets, it is very hard to distinguish the impact of individual constituents. In this study, there was a large focus on explaining the model behavior with respect to diverse compound spaces and a conserved protein kinase group. We show that the compound space, due to the sheer diversity of compound scaffolds available, highly impacts the model performance, implying that the focus of feature optimization should be shifted to capturing this diversity when computing compound representations or learning them directly from compound structures. Regardless if it is the case of herbicides or kinase inhibitors, structural representations have shown to explain most of the variability in these large chemical spaces compared to more nuanced physicochemical descriptors.

- This work provides a new method for robust evaluation of individual predictions, providing binding affinity estimates with local conformal prediction. It merges the principles of applicability domain (AD) and conformal predictors, leading to a more comprehensive understanding of the model's applicability domain across distinct areas of the interaction space. The implementation of these principles assists in circumventing the pitfalls associated with the conventional approach to defining the applicability domain. Furthermore, it enables the adoption of this framework for practical use-case scenarios that closely resemble real-world situations.

- Lastly, the integration of a binding affinity prediction model with a proposed dynamic applicability domain (dAD) framework enhances the reliability of model predictions and allows for broader application making it a valuable tool for discovery and re-purposing of bioactive molecules.

References

L. Altucci and M. G. Rots. Epigenetic drugs: from chemistry via biology to medicine and back, 2016.

N. Aniceto, A. A. Freitas, A. Bender, and T. Ghafourian. A novel applicability domain technique for mapping predictive reliability across the chemical space of a qsar: reliability-density neighbourhood. *Journal of cheminformatics*, 8:1–20, 2016.

A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, and C. T. Supuran. Natural products in drug discovery: Advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.

H. J. Beckie, R. Busi, F. J. Lopez-Ruiz, and P. A. Umina. Herbicide resistance management strategies: how do they compare with those for insecticides, fungicides and antibiotics? *Pest Management Science*, 77(7):3049–3056, 2021.

A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

I. Bloch, H. Haviv, I. Rapoport, E. Cohen, R. S. B. Shushan, N. Dotan, I. Sher, Y. Hacham, R. Amir, and M. Gal. Discovery and characterization of small molecule inhibitors of cystathionine gamma-synthase with in planta activity. *Plant Biotechnology Journal*, 19(9): 1785–1797, 2021.

L. Castelo-Soccio, H. Kim, M. Gadina, P. L. Schwartzberg, A. Laurence, and J. J. O'Shea. Protein kinases: drug targets for immunological disorders. *Nature Reviews Immunology*, pages 1–20, 2023.

C. M. A. Cefalo, F. Cinti, S. Moffa, F. Impronta, G. P. Sorice, T. Mezza, A. Pontecorvi, and A. Giaccari. Sotagliflozin, the first dual sglt inhibitor: current outlook and perspectives. *Cardiovascular diabetology*, 18(1):1–14, 2019.

J.-P. Changeux. The concept of allosteric interaction and its consequences for the chemistry of the brain. *Journal of Biological Chemistry*, 288(38):26969–26986, 2013.

H.-C. Cheng, R. Z. Qi, H. Paudel, and H.-J. Zhu. Regulation and function of protein kinases and phosphatases. *Enzyme research*, 2011, 2011.

A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, and T. Aittokallio. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS computational biology*, 13(8):e1005678, 2017.

A. Cichońska, B. Ravikumar, R. J. Allaway, F. Wan, S. Park, O. Isayev, S. Li, M. Mason, A. Lamb, Z. Tanoli, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature communications*, 12(1):3307, 2021.

P. Cohen, D. Cross, and P. A. Jänne. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature reviews drug discovery*, 20(7):551–569, 2021.

M. Costa-Mattioli and P. Walter. The integrated stress response: From mechanism to disease. *Science*, 368(6489):eaat5314, 2020.

M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, and P. Gramatica. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based qsars. *Environmental health perspectives*, 111(10):1361–1375, 2003.

F. M. Ferguson and N. S. Gray. Kinase inhibitors: the road ahead. *Nature reviews Drug discovery*, 17(5):353–377, 2018.

S. K. Garg, R. R. Henry, P. Banks, J. B. Buse, M. J. Davies, G. R. Fulcher, P. Pozzilli, D. Gesty-Palmer, P. Lapuerta, R. Simó, et al. Effects of sotagliflozin added to insulin in patients with type 1 diabetes. *New England Journal of Medicine*, 377(24):2337–2348, 2017.

S. Gelman, S. A. Fahlberg, P. Heinzelman, P. A. Romero, and A. Gitter. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 118(48):e2104878118, 2021.

A. Golbraikh, X. S. Wang, H. Zhu, and A. Tropsha. Predictive qsar modeling: methods and applications in drug discovery and chemical risk assessment. *Handbook of computational chemistry*, pages 1309–1342, 2012.

J. M. Goldberg, G. Manning, A. Liu, P. Fey, K. E. Pilcher, Y. Xu, and J. L. Smith. The dictyostelium kinome—analysis of the protein kinases from a simple model organism. *PLoS genetics*, 2(3):e38, 2006.

S.-M. Huang, J. J. Lertora, and A. J. Atkinson Jr. *Principles of clinical pharmacology*. Academic Press, 2012.

I. Jarmoskaite, I. AlSadhan, P. P. Vaidyanathan, and D. Herschlag. How to measure and evaluate binding affinities. *Elife*, 9:e57264, 2020.

V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis, and E. Kazlauskas. Binding affinity in drug design: experimental and computational techniques. *Expert opinion on drug discovery*, 14(8):755–768, 2019.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

W. Klingspohn, M. Mathea, A. Ter Laak, N. Heinrich, and K. Baumann. Efficiency of different measures for defining the applicability domain of classification models. *Journal of cheminformatics*, 9:1–17, 2017.

S. Kwon, H. Bae, J. Jo, and S. Yoon. Comprehensive ensemble in qsar prediction for drug discovery. *BMC bioinformatics*, 20(1):1–12, 2019.

J. J.-L. Liao. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *Journal of medicinal chemistry*, 50(3):409–424, 2007.

S. Lim, Y. Lu, C. Y. Cho, I. Sung, J. Kim, Y. Kim, S. Park, and S. Kim. A review on compound-protein interaction prediction methods: data, format, representation and model. *Computational and Structural Biotechnology Journal*, 19:1541–1556, 2021.

F. Llinares-López, Q. Berthet, M. Blondel, O. Teboul, and J.-P. Vert. Deep embedding and alignment of protein sequences. *Nature Methods*, 20(1):104–111, 2023.

M. A. Marti-Renom, M. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Science*, 13(4):1071–1087, 2004.

M. A. McClure, T. K. Vasi, and W. M. Fitch. Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biology and Evolution*, 11(4):571–592, 1994.

J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, and P. J. Hajduk. Navigating the kinome. *Nature chemical biology*, 7(4):200–202, 2011.

A. C. Nascimento, R. B. Prudêncio, and I. G. Costa. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics*, 17:1–16, 2016.

T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

D. Oršolić and T. Šmuc. Dynamic applicability domain (dad): compound-target binding affinity estimates with local conformal prediction. *Bioinformatics*, page btad465, 2023.

D. Oršolić, V. Pehar, T. Šmuc, and V. Stepanić. Comprehensive machine learning based study of the chemical space of herbicides. *Scientific reports*, 11(1):11479, 2021.

K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16 (2):325–337, 2015.

H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

S. B. Powles and Q. Yu. Evolution in action: plants resistant to herbicides. *Annual review of plant biology*, 61:317–347, 2010.

R. Roskoski Jr. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacological research*, 100:1–23, 2015.

R. Roskoski Jr. Properties of fda-approved small molecule protein kinase inhibitors: A 2023 update. *Pharmacological research*, page 106552, 2022.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

B. R. Stockwell. Exploring biology with small organic molecules. *Nature*, 432(7019):846–854, 2004.

P. C. Taylor, E. C. Keystone, D. Van Der Heijde, M. E. Weinblatt, L. del Carmen Morales, J. Reyes Gonzaga, S. Yakushin, T. Ishii, K. Emoto, S. Beattie, et al. Baricitinib versus placebo or adalimumab in rheumatoid arthritis. *New England Journal of Medicine*, 376(7): 652–662, 2017.

V. Vovk, I. Nouretdinov, V. Manokhin, and A. Gammerman. Cross-conformal predictive distributions. In *conformal and probabilistic prediction and applications*, pages 37–51. PMLR, 2018.

Z. Wang and P. A. Cole. Catalytic mechanisms and regulation of protein kinases. *Methods in enzymology*, 548:1–21, 2014.

M. Zhang, O. J. Sul, J. Fu, and Q. Wang. Natural compounds regulating epigenetics for treating chronic inflammatory diseases. *Frontiers in Pharmacology*, 13:1121165, 2023.

S. Zhang, H. Tong, J. Xu, and R. Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

Z. Zhao and P. E. Bourne. Progress with covalent small-molecule kinase inhibitors. *Drug discovery today*, 23(3):727–735, 2018.

Curriculum vitae

Davor Oršolić was born in Brčko, Bosnia and Herzegovina, in 1991. He obtained his Masters degree in molecular biotechnology at the Faculty of Food Technology and Biotechnology, University of Zagreb. In 2018, he began his professional career as a research assistant in the Laboratory for Machine Learning and Knowledge Representations at the Ruđer Bošković Institute (RBI), under the supervision of Dr. Tomislav Šmuc. During this period, he volunteered as a research assistant representative in the Council of Young Scientists and as an active member of the Union of Science and Education Initiative.

During the course of his studies, he completed a one-month internship in the Chemoinformatics and Computational Metabolomics group at the Friedrich Schiller University in Jena, Germany, and a two-month internship in the Genome Data Science group at the Institute for Biomedical Research in Barcelona, Spain. From 2018. to 2023., he published four papers in peer-reviewed journals.