Artificial Intelligence in the Life Sciences 4 (2023) 100079

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

An industrial evaluation of proteochemometric modelling: Predicting drug-target affinities for kinases

Astrid Stroobants[a,*], Lewis H. Mervin[b], Ola Engkvist[c], Graeme R. Robb[a]

[a] Chemistry, Oncology, R&D, AstraZeneca, Cambridge, UK
[b] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[c] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Cheminformatics
Machine-learning
Virtual screening
Kinase selectivity
Proteochemometric modelling
PCM
DTA
Drug target affinity

ABSTRACT

Deep learning proteochemometric (PCM) models have been reported to achieve excellent performances on public benchmarking datasets. Nevertheless, numerous papers have cast doubt on commonly used evaluation metrics, suggesting they do not reflect true prospective predictive abilities. The aim of this study is to provide a comprehensive assessment of performance of a state-of-the-art PCM model on proprietary data and evaluate its potential over other modelling approaches as a virtual screening tool for kinase inhibitors. Whilst the model has been shown to achieve an RMSE of 0.48 on a public benchmarking dataset, an impaired overall performance was observed for the proprietary dataset in this study, with an RMSE of 0.85 and a Pearson Correlation Coefficient of 0.65 using a temporal splitting strategy. We hypothesise that the more limited performance can be in part attributed to a shift in the chemical space observed over time in an industrial setting, which is not considered by the more lenient random ligand splitting strategy, more commonly used on benchmarking datasets. The overall performance of the PCM model was statistically similar to a multitask model and only slightly superior to a KNN and random forest PCM model. A comprehensive analysis of performance was performed to capture the key challenges faced in the design of competitive kinase inhibitors, which revealed the key limitations of PCM modelling. For example, the model showed poor predictive abilities for understudied targets, and a limited ability to assess ligand selectivity and promiscuity, with no improved performance over a multitask model or a random forest PCM model. Overall, these findings reveal that the PCM model assessed in this study does not provide significant benefits over less complex models such as multitask model or a random forest PCM model as a virtual screening tool for kinase inhibitors in an industrial setting. Taken together, this study highlights the need for more robust evaluations of PCM models by using stricter splitting strategies, more extensive benchmarking and more comprehensive performance analysis beyond traditional metrics.

Introduction

The application of artificial intelligence (AI) to drug discovery has been an increasingly dynamic field [1–3]. The need to accelerate the drug candidate identification process with effective high-throughput virtual screening keeps inspiring numerous AI focused publications [4]. For the prediction of drug-target affinities (DTA), proteochemometric (PCM) modelling has emerged as a comprehensive and versatile solution [5], with new approaches being published regularly. In this field, the first machine learning (ML) solutions consisted of single task quantitative structure activity relationship (QSAR) models (including random forests (RF), support vector machine and logistic regression) that relied on rule-based methods such as a fingerprint representation [6–8]. However, multi-target deep learning (DL) approaches such as multitask models were shown to outperform these single target prediction methods by leveraging larger heterogeneous data sources and cross-target information [9,10]. PCM models, in contrast to multitask models, consider not just chemical but also protein information and can thus, in principle, better leverage cross-target information to simultaneously generalize to novel ligands and targets. With the rise of deep learning (DL), more complex architectures and representations could be explored, including but not limited to language representations (e. g. SMILES, amino acid sequences) [11], graphs [12–14] and voxels [15]. On benchmarking datasets, DL PCM models have achieved excellent performances on tra-

Abbreviations: AI, artificial intelligence; DL, deep learning; DTA, drug target affinity; KNN, k-nearest neighbours; MAP, maximum achievable performance; ML, machine learning; MSE, mean square error; PCC, Pearson correlation coefficient; PCM, proteochemometric; QSAR, quantitative structure-activity relationships; RF, random forest; RMSE, root-mean-square error; SMILES, simplified molecular-input line-entry system.

* Corresponding author.
E-mail address: astrid.stroobants@astrazeneca.com (A. Stroobants).

https://doi.org/10.1016/j.ailsci.2023.100079
Received 24 January 2023; Received in revised form 18 May 2023; Accepted 23 May 2023
Available online 15 June 2023

ditional evaluation metrics. Mean square errors (MSE) between 0.2–0.3 have been repeatedly reported on the Davis dataset [11,12,14]. Nevertheless, several papers have now cast doubt on the traditional evaluation metrics, suggesting that regarding drug-target interactions as independent entities can lead to redundancies between training and validation and thus to overfitting, resulting in overly optimistic results that do not reflect the more limited prospective predictive ability of the models [16–18]. Therefore, with this study, we aim to provide a more comprehensive assessment of the predictive abilities of PCM modelling on proprietary data to assess its potential as a virtual screening tool in an industry setting over other modelling strategies.

We focus on a state-of-the-art PCM model, GraphDTA [12], which is a bimodal neural network based on convolution, reliant on a graph representation of the ligand and an amino acid sequence representation of the protein. The model was submitted to the IDG-DREAM Drug-Kinase Binding Prediction Challenge [19], where the model was among the ten best performing solutions. Since then, it has often been used as a benchmark for other publications. In this study we aim to further evaluate the bioactivity prediction performance of GraphDTA on the protein kinase family in an industrial setting.

Protein kinases are an important yet challenging drug target class. They are interesting drug targets in treatment areas such as oncology, neurodegenerative and viral diseases. These enzymes modulate various cellular, metabolic, and signalling pathways through phosphorylation. Abnormalities in their regulation can lead to various diseases, including cancer, diabetes, and inflammatory disorders. The first kinase inhibitor, Imatinib, received FDA approval over 20 years ago and since then more than 70 new drugs have been approved [20–22]. Most inhibitors target the ATP binding site, which is very conserved amongst the different kinase subfamilies. The similarity in binding site leads to selectivity challenges and selectivity and resistance are seen as key challenges in the development of effective kinase inhibitors [20]. Therefore, successful high-throughput virtual screening solutions for protein kinases should enable the generalisation to new targets (e.g. protein mutations) and compounds, as well as the identification of selective kinase inhibitors. These criteria, often unexplored in traditional evaluations of PCM models, are assessed in this study to form a more comprehensive performance evaluation for industrial applications.

Published PCM evaluations often rely on various public benchmarking datasets. With increasing data availability, focus has shifted on the impact of data quality on performance, considering challenges like experimental error in bioaffinity values, quality of the data source, sparsity of the drug-target matrix, and data biases [23–27]. Public datasets have a diverse quality range and are subject to experimental error [28,29]. Some databases, such as ChEMBL [30] and BindingDB [31], aggregate data from different sources with affinity data for several protein targets. The heterogenous use of public biochemical data has been shown to impact data quality, for example, ChEMBL version 27 was shown to have a median pIC_{50} standard deviation of ~0.37 across replicates [32]. Efforts have been made to facilitate and automate the curation and handling of these large datasets in order to tackle some key issues such as low quality data entries, reflecting out-of-distribution learning through data splitting, etc. [33,34]. Proprietary datasets have opportunities for standardisation, nevertheless, an AstraZeneca study evaluating biological assay variability from experiments from 2005 to 2014 still highlighted a two-fold difference in experimental reproducibility, with IC_{50} measurements having lower standard deviations compared to K_D and K_i measurements [35]. Models trained on public datasets often predict ligand affinity across multiple protein families, while models that are evaluated on smaller more focused datasets, such as the Davis [36] and Kiba [29] datasets, often focus on distinguishing affinity scores between closely related targets of the same protein family [6,37,38]. Despite a potential decrease in sparsity of the drug-target matrix, this is perhaps a more challenging task as the models need to learn more subtle differences between proteins and ligands with smaller data sources. Data biases have been shown to govern benchmarking datasets and have made it difficult to distinguish if DTA predictors can truly learn the chemical and physical features that underlie drug-target interactions or if they mainly leverage these characteristic biases [25]. This section highlights the large expected variation between experimental datapoints, and in this work, we aim to compare dataset characteristics of our in-house repository with a public benchmarking DTA dataset, which have already been shown to comprise some similarities in data distributions [10].

The challenge of data bias described above has led to the adoption of different data splitting strategies that intend to partition the data into testing and training sets while preserving a balanced representation of the interaction space [16]. The splitting strategy must enable a reliable performance evaluation which reflects model generalisability and prospective predictive ability. Despite the belief that a random compound split leads to overly optimistic reported performance, it remains among the most common approaches to assess performance on benchmarking datasets [18]. In the pharmaceutical setting, models are trained on data gathered at the time the models were made and are used to predict the activity of orphan compounds. A temporal split is thus a common strategy adopted in industry to better assess prospective prediction ability [39]. Hence, in this work we sought to evaluate the PCM model using different splitting strategies: the random ligand splitting approach, commonly used on benchmarking datasets, the temporal ligand splitting approach, commonly used on industrial datasets and a kinase splitting approach to assess generalisability to new proteins.

To summarise, we evaluate PCM modelling for industrial application using a proprietary repository for protein kinase affinity data. We first reflect on similarities and differences between our industrial dataset and a public benchmarking dataset, highlighting aspects that could limit the predictive abilities of PCM models. We then assess the performance of a state-of-the-art DL PCM model on the proprietary dataset and compare its performance to other common DTA prediction approaches. We compare overall performance of the models using the traditional ligand split, the industrial temporal split and the kinase split approach. We then extend our assessment to other important criteria for the design of protein kinase inhibitors. Our aim is to provide a more comprehensive assessment of the application of PCM modelling over other modelling approaches as an industrial virtual screening tool for protein kinase inhibitors.

Methods

Proprietary binding affinity data

Kinase-ligand affinity data from our in-house data warehouse was extracted at the start of 2021, leading to a repository of well over 5 M entries. Since the data comprised heterogenous assay types, an extensive pre-processing was required to reduce the data to high quality compound-target affinity pairs. This comprised limiting the data to exclusively contain entries that have a IC_{50} binding affinity metric measured at K_m ATP concentrations. Low precision assays such as single concentration inhibition assays, where ligands are evaluated over a panel of kinases, were excluded from the dataset. Next, we further filtered the data by clipping the data pIC_{50} values between 3 and 12. Delimited or censored data (falling outside of the assay limits) were removed, including IC_{50} values with rounding errors, thereby removing qualified or low precision data. Ligands described by invalid SMILES, SMILES containing unwanted atom types (B, Li, Be, Na, Al, Si, Ar, Ti, Fe, Zn) and compounds with a molecular weight outside the range of 200 to 900 Da were removed from the dataset. No activating protein mutations were considered in this study. Missing values were removed, and duplicate values were aggregated by taking the median pIC_{50} value. After extensive pre-processing, a subset of around 750,000 entries was retained. The final dataset had a mean pIC_{50} of 5.6 ± 1.1 and comprised around 250,000 ligands (mean pIC_{50}: 5.3 ± 1.0) and 289 kinases (mean pIC_{50}: 6.2 ± 0.8).

Models

To evaluate PCM modelling in an industrial setting, the state-of-the-art GraphDTA model was used as a reference. GraphDTA is a bimodal neural network based on convolutional neural networks. In this study, the graph isomorphism network (GIN) version of GraphDTA was applied. For further details please refer to the original paper [12]. The model was optimized on a MSE loss with Adam and was trained for 200 epochs with a learning rate of 0.0005 and a batch size of 128.

The performance of the DL PCM model was compared to several algorithmic baselines: a KNN, a multitask (ChemProp [18]) model and two PCM models using RF regression with different protein representations: one with one hot encoding of the protein (PCM-OHE) and one with Z-scales (PCM-Zscale). The RF models were generated using Scikit-learn [40] and trained on the descriptors described in the next section. The KNN model used in this study is based on the KNN in Born et al. [41]. The KNN model relied on the combination of the Tanimoto similarity between ligands and the Levenshtein distance between amino acid sequences of the proteins as a distance metric between samples, for further details refer to the original publication [41]. The ChemProp multitask model was trained using default settings and 50 epochs.

Descriptors

In the original GraphDTA publication, the ligands are described as graphs while the proteins were described by their full protein amino acid sequence. The ligand representation was not modified, for further details refer to the original publication [12]. Since the standard practice of leveraging the full primary structure has been challenged since the publication of GraphDTA (better performance was shown for PCM models when focusing on the representation of the ATP binding site [41]), the following strategy was adopted in this study: GraphDTA was adapted to only consider the sequence of key amino acids in the ATP binding site [42]. In this study the 85 key residues from the kinase–ligand interaction fingerprints and structure database (KLIFS) were used. This numbering scheme was previously developed to capture the catalytic cleft with 85 residues to facilitate the comparison of the interaction patterns of kinase-inhibitors, and thus to identify drivers of kinase-inhibitor selectivity. The 85 KLIFS amino acids were identified based on a published structurally validated multiple sequence alignment of 497 human protein kinase domains from Modi and Dunbrack [43]. All benchmarking models also relied on the reduced KLIFS representation of the targets. The PCM-Zscale RF model used a Z-scale 5 description of the 85 protein amino acids. Z-scales are principal properties derived from principal component analysis of physiochemical properties of amino acids [44]. The compound description for the PCM RF models used a 2048-bit hashed binary Morgan fingerprints of radius 3 generated from the Python RDKit module [45]. The 2048-bit encoding size was chosen after evaluating the performance of encoding sizes 512, 1024 and 2048 (see Fig. S1 for details).

Splitting strategy

Three splitting strategies were considered in this study: a random ligand split, a kinase split and a temporal split. The ligand split consists of a ten-fold cross-validation where the ligands are split between folds without stratification for the protein. Instead, the split was stratified by the mean pIC_{50} per ligand. This splitting strategy remains among the most common splits adopted on benchmarking datasets. The kinase split consists of a five-fold cross-validation where the kinases are split between folds with stratification by the mean pIC_{50} per kinase. We also adopted a temporal split which consists of an eight-fold cross-validation, further details are provided in Fig. 1. A temporal split is a common strategy used in industry to estimate the prospective predictive ability. The compound registration date was used as a timestamp and a two-year time bin was selected, roughly equivalent to a project duration. The training data in the first temporal split was used for model hyperparameter optimisation to avoid leakage to the validation and test sets in the temporal split.

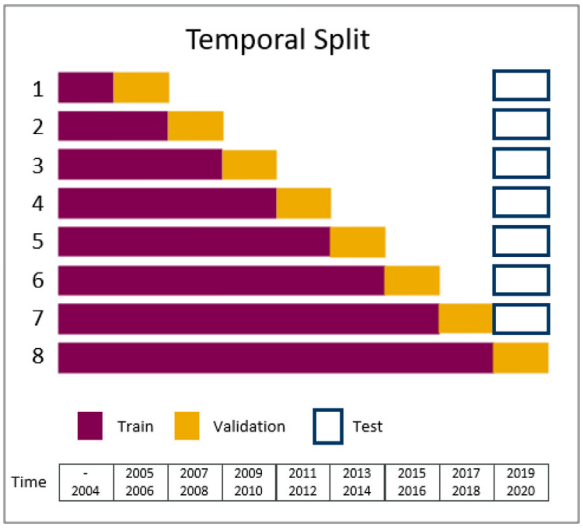

Fig. 1. Visual representation of the temporal data splitting strategy adopted in this study. In an industrial setting, a temporal split is a common strategy. The eight-fold cross-validation takes place on a rolling basis. For Fold 1, a small subset of data with ligands registered before 2005 is used for training purposes. While model performance is forecasted for compounds registered between 2005 and 2006. This validation set is then included as part of the training set for the next fold and performance is then forecasted for the next two years. The performance for the first 7 folds is tested on a communal test set (the validation set of Fold 8), to assess how model performance varies due to the progressive accumulation of training data and the expansion of the chemical space explored in training.

Hyperparameter optimisation

The hyperparameters of the RF models were optimised using a Scikit-learn grid search of 10 different parameter settings focusing on maximum depth and number of estimators. The hyperparameter optimisation of the KNN resulted in the selection of 12 nearest neighbours, please refer to Fig. S2.

Evaluation metrics

In most publications, the performance of the PCM models is assessed using a random ligand split and is reported using basic metrics such as Pearson correlation coefficient (PCC) or root-mean-square-error (RMSE). We therefore first evaluate overall cross-validation performance of the different modelling approaches as a function of these two metrics on three data slitting strategies (random ligand split, random kinase split and temporal split). Within the same splitting framework, we contrast performance in relation to four arithmetic baselines: a KNN, multitask, ML PCM model with Z-scale encoding of the protein (PCM-Zscale) and ML PCM model with one-hot encoding of the protein (PCM-OHE); and several reference points: a maximum achievable performance (MAP) reference, a median regressor and random prediction models. MAP was derived from the PCC between replicate compound-target pairs (see Fig. S3 for details). A ‘median per kinase regressor’ (MKR) assigned the median potency value of a given kinase for a given training set to each test compound for that kinase. This control follows a recent suggestion for QSAR modelling by Janela et al. [46] but instead of taking the overall training set median, it was adapted to PCM modelling by taking the median per target. For targets that were not in the training set, the overall median was taken. Random prediction models for GraphDTA and the multitask model were obtained by random scrambling of the potency values for each target across training, validation and

test sets. For each kinase, the potency value of each compound was randomly assigned to another, thus generating random structure–potency relationships per target for training. Models were trained on these random permutations and were applied to predict a randomized test set. This reference was applied as a previous study highlighted an unrealistically limited margin between random and best prediction models [46]. Statistical significance assessment of the relative performance of all models compared to the MKR was achieved using the non-parametric Wilcoxon test [47]. Statistical significance between the performance of GraphDTA and other modelling approaches was also assessed with this method. The alpha threshold was set to 0.05. The *p*-values were compared with alpha (*p* below 0.05) and the null hypothesis was rejected/accepted.

In a following section we assessed the performance for the temporal split on a communal test set (data obtained between the date range 2019–2020). This evaluation aimed to highlight how model performance varies due to the progressive accumulation of training data and the expansion of the chemical space explored in training.

Next, the overall cross-validation and test performance was also evaluated for a subset of kinases to assess the ability of the models to predict ligand affinity for understudied kinases. We focused on the 18 understudied kinases which have less than 1000 datapoints in the training set in the last temporal fold (Fold 8).

The performance was also evaluated on a per ligand basis by calculating the concordance index for each ligand and the model sensitivity and specificity; thereby assessing if PCM modelling can be used to assess ligand promiscuity. The concordance index (CI) was calculated using the *lifelines* python package [48]. The assessment of model sensitivity and specificity was achieved via transforming the affinities to a binary representation indicating whether a ligand was active (pIC_{50} equal or above 6) or inactive (pIC_{50} smaller than 6) for a specific target. A pIC_{50} threshold of 6 is believed to be adequate for separating hits and weaker off-target interactions, similar pIC_{50} cut-offs have been proposed in prior kinase inhibitor classification studies [49]. Hence, this becomes a classification problem for which the sensitivity and specificity can be calculated for each of the 927 ligands that were tested across more than 25 kinases. This analysis was carried out over the validation results since only two ligands in the test were tested across more than 25 kinases. Statistical significance was assessed using the non-parametric Wilcoxon test. Due to the large test number (927 ligands), the alpha threshold was corrected using the Bonferroni correction. The *p*-values were compared with the corrected alpha (*p* below $5.\ 10^{-5}$) and the null hypothesis was rejected/accepted. In addition, to assess selectivity, the models' ability to predict differences in affinities per ligand across similar targets was assessed by concentrating on four selectivity examples: selectivity between (1) CDK1 and CDK2, (2) SIK2 and SIK3, (3) IGF1R and FGFR1, (4) CLK2 and CLK3. To achieve this, we sought to relate the PCC between the delta in experimental affinities and the delta in predicted affinities for each pair of targets.

Results & discussion

Dataset characteristics and limitations

In this first section we sought to analyse the characteristics of the proprietary dataset and compare them to the BindingDB [31] dataset used in a recent PCM paper focused on the kinase target class [41]; an overview is provided in Fig. 2. Prior work has shown that benchmarking datasets can be biased towards active compounds [50], presumably due to a drive to publish positive results. Nevertheless, the overall data distribution profile of the curated in-house dataset largely resembles the distribution of the BindingDB kinase subset, with a median pIC_{50} of 5.4 and 5.3 respectively. The in-house dataset contains five times more data (742,533 vs. 158,900). In the in-house dataset, 53 kinases have more than 5000 ligands tested against them, versus 9 for public data, respectively. This analysis highlights that certain kinases dominate the dataset while others are under-represented with fewer than 50 ligands tested. After the extensive cleaning described in the methods, only 21 ligands were screened over a large set of kinases (more than 50), highlighting a sparse representation of the interaction landscape, which is an observation reported for public datasets as described in the Introduction section. A bias in the BindingDB dataset was highlighted for kinases, since those screened against 1000 or more ligands tended to have a higher average affinity [41]. In the in-house database comprised an opposite but equally strong bias ($PCC = -0.38$), where kinases screened against more ligands tended to have a lower average affinity. We hypothesise this may be due to the repeated safety screening of kinases when evaluating selectivity or promiscuity. The mean pIC_{50} for kinases with few datapoints (less than 10,000 ligands) is more deviant (mean pIC_{50} 6.2 ± 0.8) compared to kinases screened against more ligands (mean pIC_{50} 5.5 ± 0.4).

Despite the extensive pre-processing, the quality of the in-house dataset suffers from reproducibility limits as highlighted by the correlation between replicate compound-target pairs (see Fig. S3 for details), and hence the hypothetical maximal performance of any model is thus limited to a PCC of ~0.8.

Overall, the datasets differ in size, but similar distributions are observed for both datasets. Strong biases dominate both datasets and are likely to impact the predictive abilities of models. The experimental reproducibility of the proprietary dataset highlights the importance of relating this with the performance limits for any model trained herein.

Performance using standard metrics

In this section we sought to assess the overall performance of GraphDTA, a DL PCM model, on proprietary data. We explore performance in relation to several reference points and arithmetic baselines: a low complexity KNN, a multitask model, a ML PCM model with Z-scale encoding of the protein (PCM-Zscale), a ML PCM model with one-hot encoding of the protein (PCM-OHE), a maximum achievable performance (MAP) baseline calculated from experimental error, random prediction baselines and a 'median per kinase regressor' (MKR). The MKR represents the baseline performance that can be achieved when solely learning the pIC_{50} bias per kinase, which, as highlighted in the previous section, is an important bias that dominates the dataset. In the same line, random prediction models for GraphDTA and the multitask model were also considered as a baseline by random scrambling of the potency values per target. Results for this analysis based on the predictions across all ligand-target pairs for the ligand, kinase and temporal splits are shown in Fig. 3.

Considering the random ligand split, GraphDTA achieved a performance of 0.58 RMSE and a PCC of 0.86 for the proprietary dataset. Despite the increased data availability for the in-house dataset, the performance is slightly inferior to the performance reported in the original paper, where the model achieved 0.229 MSE (0.48 RMSE) on the Davis dataset with a similar splitting strategy [12]. Nevertheless, the model achieves a good overall predictive performance that is well above that of the MKR ($p = 0.002$).

The random ligand split, with a mean PCC of 0.86 for GraphDTA, suggests performances close to the estimate MAP (PCC = 0.83) can be obtained, portraying an overly optimistic performance compared to the temporal split, where considerably lower performances are observed (mean PCC of 0.65 for GraphDTA). In the temporal split, the performance of all models is lower but remains significantly above the MKR baseline and the GraphDTA and multitask models achieved significantly better performance than their respective y-scrambled model ($p = 0.008$). The temporal splitting delta in PCC of 0.18 to the maximal achievable performance shows that this strategy provides a more realistic assessment of true prospective performance. The performance window between MAP and the performance obtained using median regression is 1.5 times bigger on the PCC scale in the temporal split versus the random ligand split, thus potentially better enabling model comparison. The random kinase split appears the most challenging split: GraphDTA

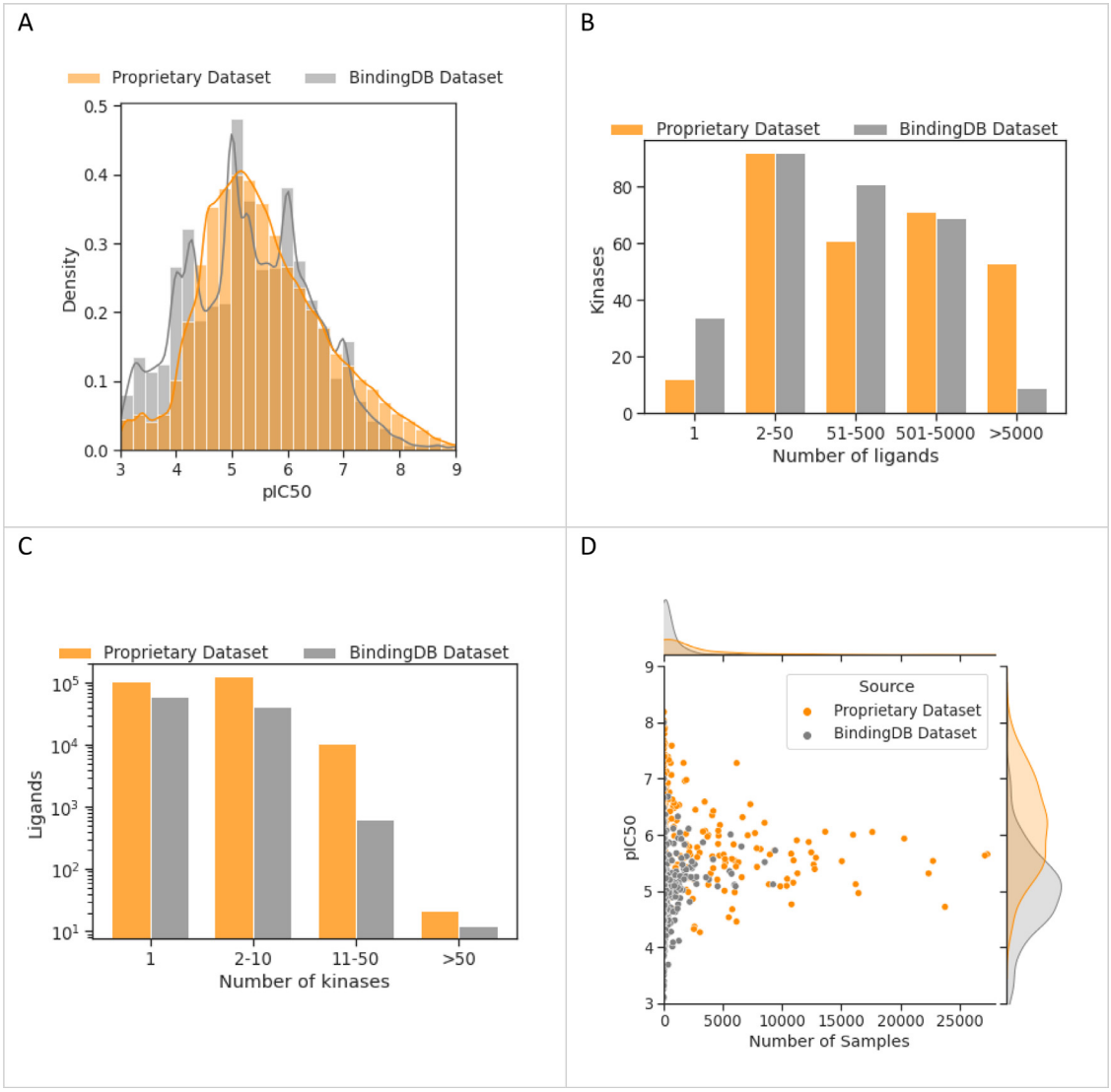

Fig. 2. Dataset characteristics comparison of kinase inhibitor data from a proprietary source and BindingDB. A) Distribution of pIC50 scores in the proprietary dataset (N = 742,533) and BindingDB (N = 185,988). **B)** Histogram of the number of ligands screened per kinase. More kinases have been screened for over 5000 ligands in the proprietary dataset. **C)** Histogram of the number of kinases screened per ligand. More ligands have been screened for over 10 kinase in the proprietary dataset. **D)** Understudied kinases tend to have more active samples in the proprietary dataset and more inactive samples in BindingDB.

achieves an RMSE of 0.94 and a PCC of 0.59 which is a weaker performance compared to results obtained for the temporal split.

The difference in performance between the ligand splitting strategies (random and temporal) is likely to be in part due to the temporal split also evaluating protein generalisability as progressively data becomes available for new kinases over time. Additionally, the difference in performance can also be attributed to an increased similarity between the ligands in the training and validation sets in the ligand split, while a larger shift in chemical space between training and validation is expected for a temporal split. Indeed, the average Tanimoto similarity across folds in the temporal split is 0.50 ± 0.02 while in the random ligand split the average score is 0.68 ± 0.001. Outside of an industrial setting, other ligand splitting strategies can be adopted to better assess the performance of PCM upon bigger shifts in the chemical space, such as a scaffold split or even a clustered split where ligands are clustered based on similarity.

The variability between folds is greater in the temporal split compared to the random split, with a standard deviation of 0.05 versus 0.005, respectively, which can be partially explained by variability in the data quality of the validation set and the scale of the chemical shift between training and validation (refer to Fig. S4 for details). The variability between kinase splitting folds is similar to that of the temporal split.

As highlighted in Fig. 3 and the raw data in Tables S1–2, the relative performance ranking of the different models varies across the three different splitting strategies. For the temporal split, GraphDTA performs statistically similar to the multitask model (p = 0.31 for PCC) and only marginally better than the KNN (delta in RMSE of 0.05 and delta in PCC of 0.01; p = 0.008) and the PCM-Zscale (delta in RMSE and PCC of 0.02; p = 0.02 and p = 0.04 respectively) baseline models on proprietary data. For the random ligand split, PCM-Zscale shows the best ligand generalisability potential. GraphDTA is statistically outperformed by PCM-Zscale (p = 0.002 for PCC) while performing statistically similar to the multitask model (p = 0.064 for PCC). The best kinase generalisability is achieved by GraphDTA but its performance is not statistically better than that of the KNN and PCM-Zscale models. Other models such as the multitask and the PCM-OHE could not be evaluated on the kinase split as their design does not enable the generalisation to new kinases.

Overall, these findings are similar to recent literature where KNN models were shown to have a comparable performance to QSAR models

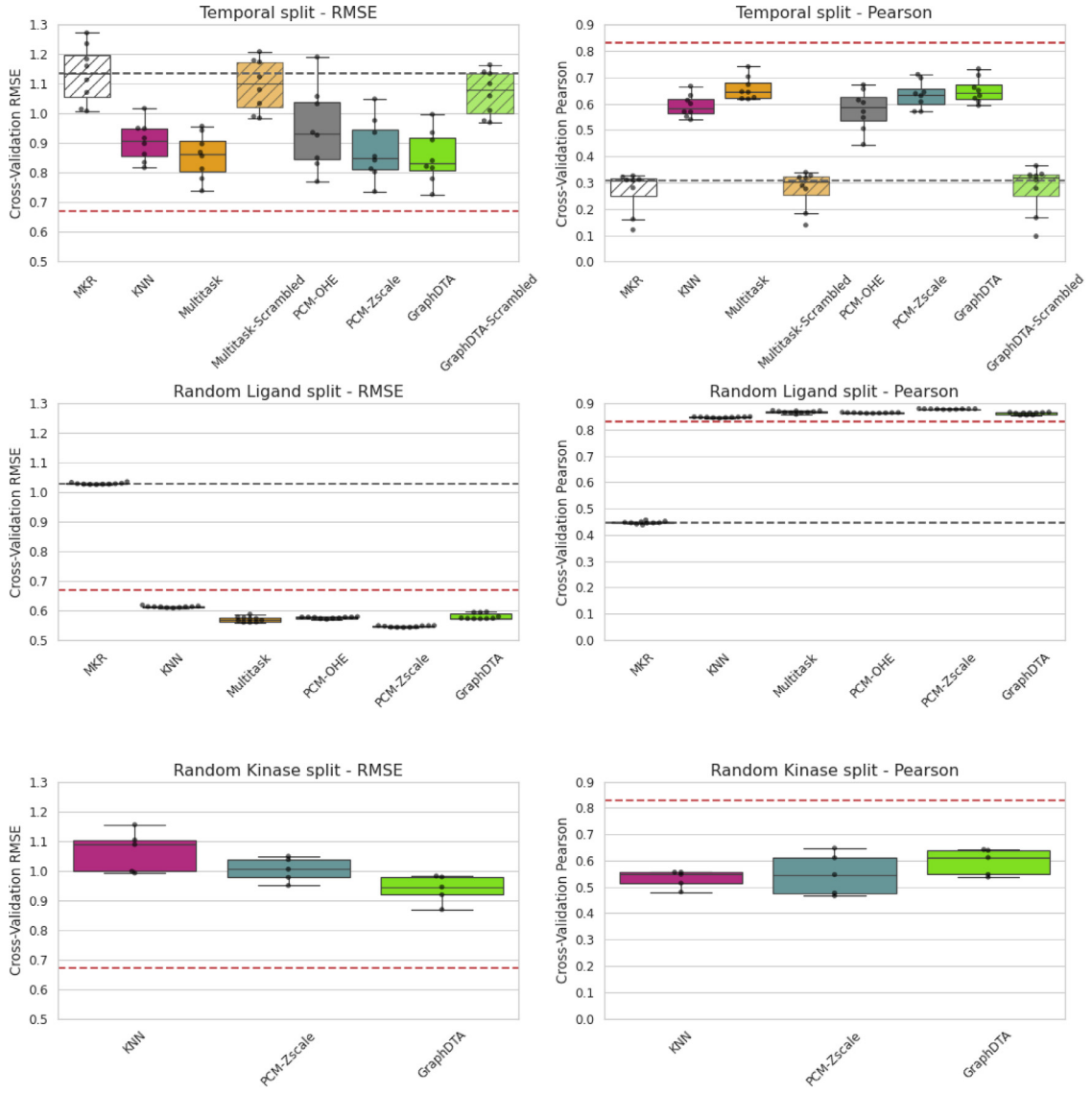

Fig. 3. Distribution of cross-validation performance results across folds. Performance is reported using two metrics, PCC (right) and RMSE (left), and across three splitting strategies: the temporal (upper), random ligand (middle), and kinase (lower) splitting strategies. Across the different spliting strategies, different relative model performance can be observed: in the random ligand split PCM-Zscale is the best performing model while GraphDTA is the best performing model in the temporal and kinase split. MAP is represented by the red dotted lines while the median performance achieved with the MKR is shown by the black dotted lines. The random prediction models for GraphDTA and the multitask model are indicated by the hashed boxes. All models perform statistically better than assigning the median potency per kinase, which highlights their predictive abilities beyond the learning of the pIC_{50} bias per kinase. The exact numerical scores can be found in Tables S1–S3. A summary of key statistical analysis on these results is provided in Table S4.

[46]. Despite these reports and results, KNN, RF and multitask models are not often considered as baselines when assessing DL PCM models, and hence the performances obtained here provide strong evidence that such simple controls should be more routinely included in PCM model evaluations.

In this section we have provided empirical evidence that the temporal splitting strategy provides a more realistic assessment of true prospective performance as it simultaneously considers a shift in the chemical and protein space. We have also demonstrated that with a temporal split the state-of-the-art PCM model performs similarly to a multitask model on a proprietary dataset. We also highlight the importance of considering other modelling approaches, such as KNN, RF and multitask models, as benchmarks for the evaluation of DL PCM models. We call attention to the remaining gap to MAP. This suggests that there is a hypothetical window to be exploited by better performing models in the future.

Performance over time

In this section we sought to explore the performance of the different modelling approaches for the temporal splitting strategy on a communal test set. This evaluation highlights how the performance of the models varies due to progressive accumulation of training data and the expansion of the chemical space explored in training. Results from this analysis are presented in Fig. 4.

On average, the test performance of all models improves when trained on progressively more data, with overall cross-method PCC-means increasing from 0.28 ± 0.08 (<2005) to 0.41 ± 0.1 (<2017). During testing, all models except for the KNN offer some predictive ability across all time points, with a performance greater that the MKR. Both GraphDTA and the multitask model outperform their respective random prediction models across all folds. The relative performance across models thus varies between testing and validation. During val-

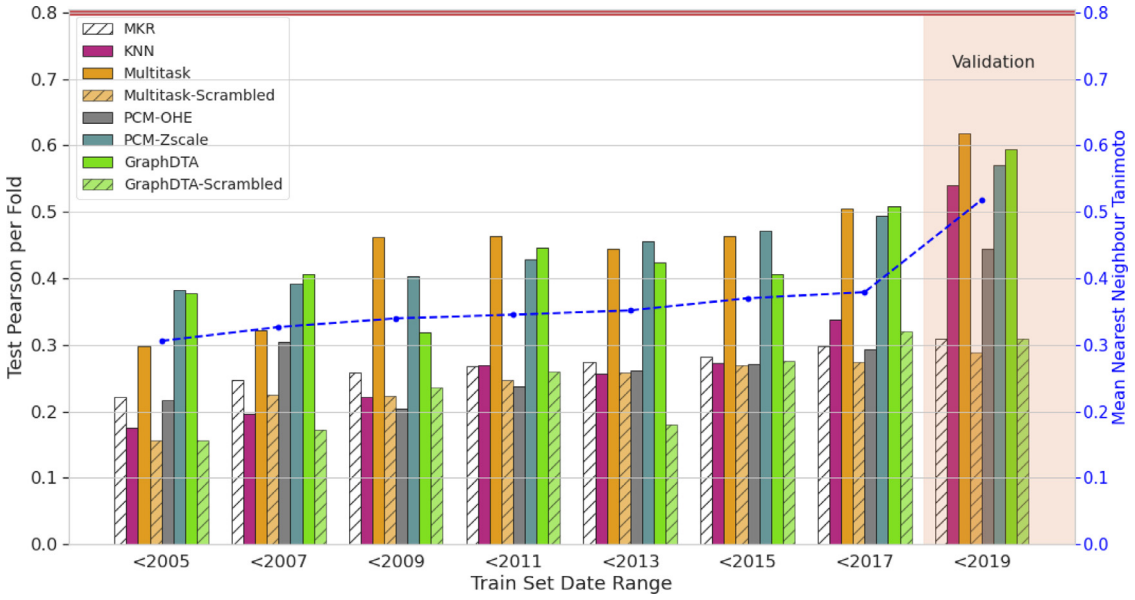

Fig. 4. Temporal split test results indicate time-dependent performance behavior toward more recent data. The performance of models improves when trained on progressively more data (from left to right on the Train Set Date Range), expanding the chemical space explored in training, leading to higher Tanimoto similarity between train and test set compounds. The mean Tanimoto similarity across all folds is indicated by the blue line while the MAP is indicated by the red line. The random prediction models for GraphDTA and the multitask model are indicated by the hashed bars.

idation, when trained on all data available, GraphDTA and the multitask model achieved comparatively the highest performance, outperforming the KNN and the RF PCM models by a relatively small margin of 0.05. However, during testing, the KNN achieves differing performance to GraphDTA, where the aforementioned outperforms the KNN model by a larger margin of 0.20 on the first fold ("<2005″). This fold is trained on a much smaller subsection of the data (more than 2 times smaller) and the chemical space explored is further away from the test set (average Tanimoto similarity of 0.31 ± 0.08 versus 0.52 ± 0.17 for validation). This highlights that the KNN model has limited predictivity when trained on smaller datasets with larger shifts in chemical space between training and testing. On the other contrary, GraphDTA and the multitask model achieve some predictive performance beyond the MKR in conditions when the tested molecules are further away from training chemical space. The superior performance in these conditions suggests they are better equipped at handling larger shifts in chemical space.

To summarise, despite the comparable overall performances between the KNN, the multitask and GraphDTA highlighted during cross-validation, differences in performance can be observed when assessing the effect of progressive accumulation of training data and the expansion of the chemical space explored in training. GraphDTA and the multitask model outperform the KNN in conditions where there is a more pronounced shift in the chemical space explored between training and testing, suggesting have a wider domain of applicability.

Performance on understudied kinases

In this section we sought to assess the ability of the models to predict ligand affinity for understudied kinases as described in the method section. This aspect is of vital importance in industry to enable the design of competitive inhibitors for novel targets that can either be wild type proteins that remain mainly unexplored or mutated proteins in the context of drug resistance. In this vein, Fig. 5 highlights the cross-validation performance of the models on understudied kinases. For all models (KNN, GraphDTA, Multitask, PCM-OHE and PCM-Zscale), the performance is considerably lower when only understudied kinases are considered but some predictive ability remains, with PCCs significantly above that of the median regressor (MKR) for all models except for the multitask and PCM-OHE models. Both, GraphDTA and the multitask model significantly outperform their respective random prediction models further highlighting that both models retain some predictive ability. Results of the statistical analysis are provided in Table S6. While the PCM-Zscale and GraphDTA models marginally outperform the multitask and KNN models, GraphDTA offers a significant improvement in predictive abilities for understudied kinases over the KNN ($p = 0.008$) but no significant improvement in predictive abilities were found versus the multitask model ($p = 0.148$) or the PCM-Zscale model ($p = 0.078$).

To assess how the performance on understudied kinases varies due to progressive accumulation of training data and the expansion of the chemical space explored in training, the performance for the 18 understudied kinases in the communal temporal test set is evaluated and the results are presented in Fig. 6. In early folds (trained on data before 2015), the models exhibit limited predictivity with PCCs close to that of the MKR and GraphDTA and the multitask showing similar performances to their random prediction models. It is only in the last testing fold (<2017), when the ligand space explored during training more closely resembles the ligands in the test set (with an average Tanimoto similarity of 0.34 ± 0.13 in Fold 7, "<2017″ versus a score of 0.26 ± 0.11 in Fold 1, "2005″), that the performance increases, with all models except for PCM-OHE showing very moderate predictive abilities.

Taken together, this suggests that for these understudied kinases, all models except for PCM-OHE only offer limited predictive abilities that are dependent on similar ligands being explored in training. Despite the hypothesis that PCM models can better leverage cross-target information and thus generalise better to novel targets, GraphDTA does not significantly outperform a KNN and shows comparable performance to multitask models in the prediction of bioactivity for understudied kinases. Since the overall predictive performance of all models remains limited, the PCM models showed the best overall predictive performance, and there is a distinct difference in performance between a RF model with little protein information (PCM-OHE) and some protein information (PCM-Zscale), there is potentially an opportunity for future PCM models to further optimise cross-target learning with better protein descriptors to improve performance for understudied kinases.

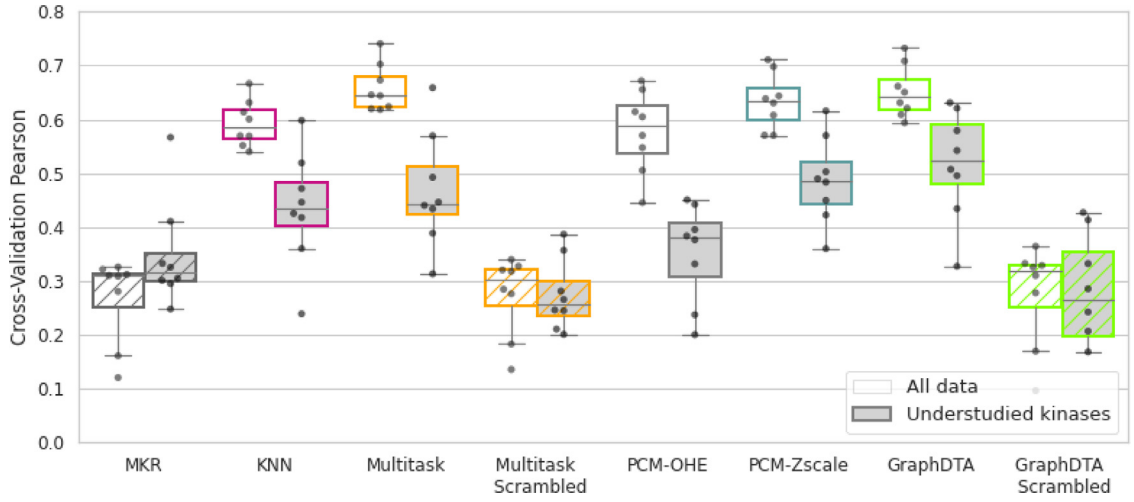

Fig. 5. Decreased performance for understudied kinases in temporal cross-validation. The performance of GraphDTA is statistically comparable to the multitask model when considering all targets and only understudied kinases. All methods except for the multitask and PCM-OHE models are statistically superior to the MKR for understudied kinases. GraphDTA and the multitask model significantly outperformed their respective random prediction models (indicated by the hashed boxes), highlighting that some predictive ability is retained for understudied kinases.

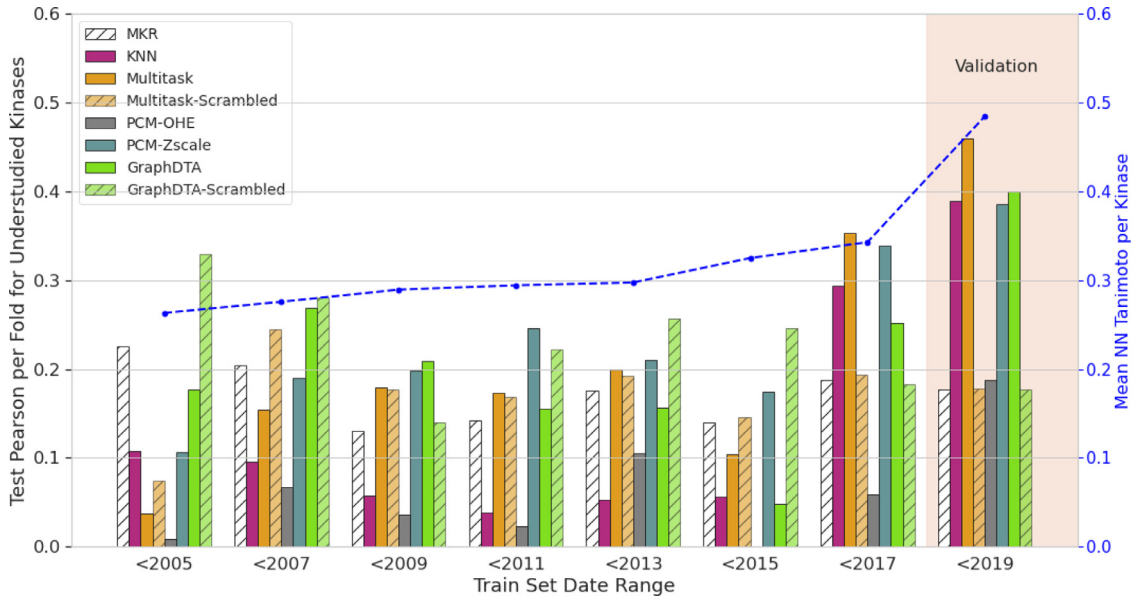

Fig. 6. Temporal performance in predicting affinity for understudied kinases. The performance is assessed for 18 understudied kinases in the test set. The mean nearest neighbour Tanimoto similarity per kinase (blue) is also presented across the folds. For the first six folds (<2015), poor model performance is observed with results comparable to that of the MKR (hatched white bar) and random prediction (hatched bars) highlighting a lack of predictive ability when no similar compounds are found in the training set.

Performance in the context of ligand promiscuity and selectivity

As discussed in the Introduction section, a key challenge in the development of effective kinase inhibitors is ligand selectivity. In this vein, we further sought to evaluate the models in the context of assessing kinase inhibitor selectivity. High-throughput virtual screening solutions for protein kinases should enable the identification of selective kinase inhibitors and therefore they should be sensitive and selective enough to predict differences in affinities per ligand across several targets. For ligands evaluated over at least 25 targets, the sensitivity and selectivity of the models was assessed as well as the model's ability to predict ligand affinity across targets using the concordance index, as described in the Method section. The results highlighted in Fig. 7, demonstrate that even though all models are statistically outperforming the MKR, the ability of the models to correctly reflect relative ligand affinity across targets remained limited. The multitask model, KNN and GraphDTA only achieved a CI 0.71 ± 0.1 compared to the MKR (CI 0.62 ± 0.1). GraphDTA significantly outperformed all models except for the multitask and PCM-Zscale models. While all models achieved good specificity, poor sensitivity was observed across models (mean true positive rates ranging between 0.30 and 0.54) and thus highlighting their limited ability to correctly identify instances where the ligands are active across specific targets. GraphDTA was the second-best performing model, it was significantly outperformed by the KNN and achieved statistically similar results to the multitask model. Combined, these results suggest that all models have limited predictive power to evaluate ligand promiscuity with a poor ability to identify actives across specific targets.

In the next step, the models' ability to predict differences in affinities per ligand across similar targets was assessed by concentrating on four common selectivity examples: selectivity between (1) CDK1 and

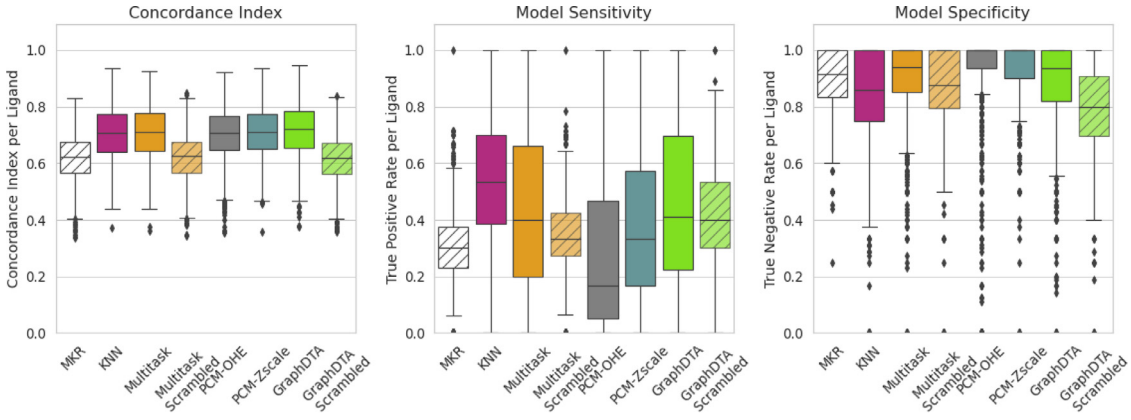

Fig. 7. Performance on the prediction of ligand promiscuity. The ability of models to assess promiscuity is assess by evaluating the Concordance index and the models' sensitivity and specificity on a per ligand basis (N = 927). All models show a moderate performance and that significantly outperforms a simple MKR.

Table 1
Performance in predicting ligand specificity. The performance is evaluated with the PCC between the experimental and predicted delta in affinities for four pairs of similar targets.

PCC between predicted and experimental pIC_{50} delta	Multitask	PCM-OHE	PCM-Zscale	GraphDTA
Δ(CDK1-CDK2)	0.44	0.43	0.46	0.41
Δ(SIK2-SIK3)	0.47	0.33	0.37	0.48
Δ(IGF1R-FGFR1)	0.55	0.55	0.60	0.48
Δ(CLK2-CLK3)	0.50	0.41	0.49	0.54

CDK2, (2) SIK2 and SIK3, (3) IGF1R and FGFR1, (4) CLK2 and CLK3. To achieve this, we sought to relate the PCC between the experimental and predicted delta in affinities for each pair of targets, which is shown in Table 1. The multitask model outperformed GraphDTA in two out of the four examples, while the opposite was true for the other two examples. The PCM-Zscale was the best performing model in two examples whilst it was amongst the two worst performing models for the other two examples. The variability in relative performance indicates that GraphDTA does not offer superior predictivity over a multitask model for the assessment of ligand selectivity for the two related targets analysed here, and highlights the importance of benchmarking or trialling different algorithms for different kinase projects. For all four examples, GraphDTA achieved a PCC of 0.5 between the true and predicted deltas in pIC_{50}, which highlights a relatively limited performance in predicting relative delta affinities between two kinases.

The relatively poor predictive abilities of PCM models in the context of kinase promiscuity and selectivity highlighted in this study, indicate that this modelling technique provides only a limited benefit in an industrial setting where designing selective kinase inhibitors remains a key challenge. Since the PCM-Zscale model outperformed the PCM-OHE model in all four examples, we believe that better protein descriptors could facilitate the learning of the interactions between protein and ligand features and should thus in principle improve these predictive abilities.

Conclusion

In this study, we have provided a comprehensive assessment of the predictive abilities of a state-of-the-art PCM model on proprietary data to assess the potential of PCM modelling as a virtual screening tool in an industry setting. The performance of the state-of-the-art GraphDTA PCM model on proprietary data appears more limited compared to the performance suggested in the original publication, with a mean PCC of 0.65. We partially attributed the overly optimistic performance to the very lenient random ligand splitting strategy employed in the original publication which did not reflect the shift in chemical space observed over time in an industrial setting. We therefore recommend the use of stricter ligand splits for the evaluation of PCM models. The relative performance of GraphDTA to other modelling approaches was also limited. GraphDTA performed comparably to a multitask model and only marginally outperformed a KNN and a RF PCM model with a Z-scale description of the proteins. This highlights the importance of considering other modelling approaches as benchmarks for the evaluation of PCM models. By considering a more comprehensive performance evaluation, we showed that, despite a reasonable overall performance, a relatively poor performance was still observed in key areas vital for the design of competitive kinase inhibitors: predicting the affinity for understudied kinases and predicting ligand selectivity and promiscuity. Despite the hypothesis that PCM models can better leverage cross-target information and thus generalise better to novel targets, GraphDTA does not show improved predictive abilities in these areas compared to multitask modelling. Nevertheless, the RF PCM model (with Z-score description) exhibits improved predictive abilities compared to the RF PCM model with one-hot-encoding of the proteins. We feel that the evaluation of PCM models in the context of understudied targets and ligand selectivity and promiscuity is an important yet often unexplored evaluation criteria for new DL PCM models, especially for regression models. The moderate overall performance of the PCM model and relatively poor performance in predicting ligand promiscuity and selectivity limit the application of PCM models as a virtual screening tool for kinase inhibitors in an industry setting. We call attention to the remaining gap to MAP which suggests an opportunity for the development of better performing PCM models in the future. An interesting next step that falls out of the scope of this project, would be to evaluate if more chemically relevant descriptors and other state-of-the-art modelling techniques such as attention mechanisms, improve the learning of protein-ligand interactions and thus performance in an industrial setting on the key metrics highlighted in this paper.

Availability

The source code for the adapted graphDTA model is available in the following GitHub repository: https://github.com/AstraZeneca/AdaptedGraphDTA.

Acknowledgements

A.S. is a fellow of the AstraZeneca Postdoctoral Research Programme.

Declaration of Competing Interest

The authors declare the following interest(s): A.S. received funding from AstraZeneca; L.H.M., O.E., and G.R.R. are employees of AstraZeneca.

Data availability

The data that has been used is confidential.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100079.

References

[1] Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. Mol Inform 2016;35(1):3–14.
[2] Lipinski CF, et al. Advances and perspectives in applying deep learning for drug design and discovery. Front Robot AI 2019;6:108.
[3] Askr H, et al. Deep learning in drug discovery: an integrative review and future challenges. Artif Intell Rev 2023;56(7):5975–6037.
[4] Kimber TB, Chen Y, Volkamer A. Deep Learning in Virtual Screening: recent Applications and Developments. Int J Mol Sci 2021;22(9):4435.
[5] Bongers BJ, Ijzerman AP, Van Westen GJP. Proteochemometrics – recent developments in bioactivity and selectivity modeling. Drug Discov Today: Technol 2019;32-33:89–98.
[6] Giblin KA, et al. Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins. J Chem Inf Model 2018;58(9):1870–88.
[7] Shar PA, et al. Pred-binding: large-scale protein–ligand binding affinity prediction. J Enzyme Inhib Med Chem 2016;31(6):1443–50.
[8] Cortés-Ciriano I, Bender A, Malliavin T. Prediction of PARP inhibition with proteochemometric modelling and conformal prediction. Mol Inf 2015;34(6–7):357–66.
[9] Rodríguez-Pérez R, Bajorath J. Multitask machine learning for classifying highly and weakly potent kinase inhibitors. ACS Omega 2019;4(2):4367–75.
[10] Sturm N, et al. Industry-scale application and evaluation of deep learning for drug target prediction. J Cheminform 2020;12(1):26.
[11] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34(17):i821–9.
[12] Nguyen T, et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 2020;37(8):1140–7.
[13] Jiang M, et al. Drug–target affinity prediction using graph neural network and contact maps. RSC Adv 2020;10(35):20701–12.
[14] Voitsitskyi T, et al. 3DProtDTA: the deep learning model for drug-target affinity prediction based on the residue-level protein graphs. RSC Adv 2023;13(15):10261–72.
[15] Wang S, et al. MCN-CPI: multiscale convolutional network for compound-protein interaction prediction. Biomolecules 2021;11(8).
[16] Torrisi M, et al. Improving the assessment of deep learning models in the context of drug-target interaction prediction. bioRxiv 2022;04:20.
[17] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model 2018;58(5):916–32.
[18] Yang K, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59(8):3370–88.
[19] Cichońska A, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun 2021;12(1):3307.
[20] Cohen P, Cross D, Jänne PA. Kinase drug discovery 20 years after imatinib: progress and future directions. Nat Rev Drug Discov 2021;20(7):551–69.
[21] Cohen P. Protein kinases–the major drug targets of the twenty-first century? Nat Rev Drug Discov 2002;1(4):309–15.
[22] Cohen P, Alessi DR. Kinase drug discovery – what's next in the field? ACS Chem. Biol. 2013;8(1):96–104.
[23] Lopez-del Rio A, Picart-Armada S, Perera-Lluna A. Balancing data on deep learning-based proteochemometric activity classification. J Chem Inf Model 2021;61(4):1657–69.
[24] Chen L, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PLoS ONE 2019;14(8):e0220113.
[25] Volkov M, et al. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. J Med Chem 2022;65(11):7946–58.
[26] Sundar V, Colwell L. The effect of debiasing protein ligand binding data on generalization. J Chem Inf Model 2020;60(1):56–62.
[27] Yang J, Shen C, Huang N. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. Front Pharmacol 2020;11:69.
[28] Papadatos G, et al. Activity, assay and target data curation and quality in the ChEMBL database. J Comput Aided Mol Des 2015;29(9):885–96.
[29] Tang J, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 2014;54(3):735–43.
[30] Bento AP, et al. The ChEMBL bioactivity database: an update. Nucl Acid Res 2014;42(Database issue):D1083–90.
[31] Gilson MK, et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucl Acid Res 2016;44(D1):D1045–53.
[32] Mervin LH, et al. Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. J Cheminform 2021;13(1):62.
[33] Ji Y., et al. DrugOOD: out-of-Distribution (OOD) Dataset Curator and Benchmark for A I-aided Drug Discovery A Focus on Affinity Prediction Problems with Noise Annotations. arXiv; 2022.
[34] Béquignon OJM, et al. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. J Cheminform 2023;15(1):3.
[35] Kramer C, et al. A comprehensive company database analysis of biological assay variability. Drug Discov Today 2016;21(8):1213–21.
[36] Davis MI, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 2011;29(11):1046–51.
[37] Subramanian V, et al. 3D proteochemometrics: using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases. Medchemcomm 2017;8(5):1037–45.
[38] Subramanian V, et al. Predictive proteochemometric models for kinases derived from 3D protein field-based descriptors. Medchemcomm 2016;7(5):1007–15.
[39] Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 2013;53(4):783–90.
[40] Pedregosa F, et al. Scikit-learn: machine Learning in Python. Journal of Machine Learning Research 2012:12.
[41] Born J, et al. Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3D effects in a 1D model. J Chem Inf Model 2022;62(2):240–57.
[42] van Linden OP, et al. KLIFS: a knowledge-based structural database to navigate kinase-ligand interaction space. J Med Chem 2014;57(2):249–77.
[43] Modi V, Dunbrack RL. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. Sci Rep 2019;9(1):19790.
[44] Sandberg M, et al. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. J Med Chem 1998;41(14):2481–91.
[45] *RDKit: open-source cheminformatics*. cited 2022; Available from: https://www.rdkit.org/.
[46] Janela T, Bajorath J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. Nat Mach Intell 2022.
[47] Conover WJ. On methods of handling ties in the wilcoxon signed-rank test. J Am Stat Assoc 1973;68(344):985–8.
[48] Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw 2019;4:1317.
[49] Sorgenfrei FA, Fulle S, Merget B. Kinome-wide profiling prediction of small molecules. ChemMedChem 2018;13(6):495–9.
[50] Cáceres EL, Mew NC, Keiser MJ. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. J Chem Inf Model 2020;60(12):5957–70.