ELSEVIER

DRUG DISCOVERY TODAY TECHNOLOGIES

Drug Discovery Today: Technologies

Vol. 32, No. 2019

Editors-in-Chief
Kelvin Lam – Simplex Pharma Advisors, Inc., Boston, MA, USA
Henk Timmerman – Vrije Universiteit, The Netherlands

Proteochemometrics – recent developments in bioactivity and selectivity modeling

Brandon J. Bongers, Adriaan. P. IJzerman, Gerard J.P. Van Westen*

Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA, Leiden, The Netherlands

Proteochemometrics is a machine learning based modeling approach relying on a combination of ligand and protein descriptors. With ongoing developments in machine learning and increases in public data the technique is more frequently applied in early drug discovery, typically in ligand–target binding prediction. Common applications include improvements to single target quantitative structure-activity relationship models, protein selectivity and promiscuity modeling, and large-scale deep learning approaches. The increase in predictive power using proteochemometrics is observed in multi-target bioactivity modeling, opening the door to more extensive studies covering whole protein families. On top of that, with deep learning fueling more complex and larger scale models, proteochemometrics allows faster and higher quality computational models supporting the design, make, test cycle.

Introduction

Proteochemometrics (PCM) is a statistical modeling technique used in bioactivity prediction [1,2]. PCM typically comprises the modeling of a biologically relevant endpoint using a concatenation of explicit ligand derived and explicit target derived descriptors via supervised machine learning [3,4]. These descriptors describe the properties and characteristics of the ligands and targets in a way that is compatible with machine learning. Frequently PCM uses non-linear machine learning techniques such as random forests (RF) or support vector machines (SVM), but it has also been applied using partial least squares (PLS) [1,3]. In the case of PLS, or other linear methods, cross-terms are needed for proper modeling. Cross terms are an additional set of descriptors which rely on mathematically combining ligand and targets descriptor blocks (e.g. via multiplication) [1]. PCM is based on the same principles as machine learning-based Quantitative Structure-Activity Relationship (QSAR) modeling and hence also uses similar validation methods (Fig. 1). In good practice, PCM models are validated using both (nested) cross-validation and external validation [5]. As with QSAR, PCM can either be classification or regression depending on the data set modeled. Hence model quality can be estimated using typical measures such as correlation coefficient (R^2) and root mean square error (RMSE) for regression-based models. Conversely, for classification-based models, use can be made of receiver operating characteristic (ROC) curves, sensitivity, specificity, and the Matthews correlation coefficient.

Descriptors

We consider PCM as any method that combines the interactions of multiple targets (typically proteins) with a group of

*Corresponding author.: Gerard J.P. Van Westen (gerard@lacdr.leidenuniv.nl)

 https://doi.org/10.1016/j.ddtec.2020.08.003

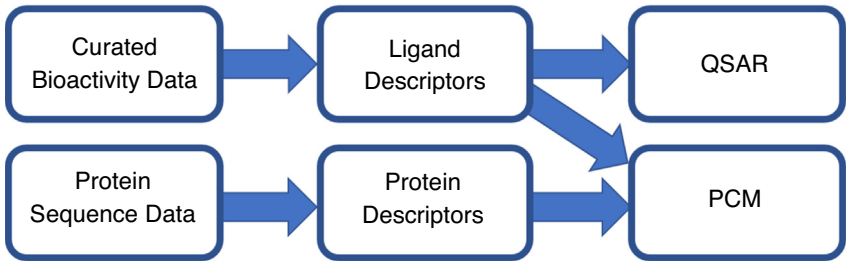

Fig. 1. A schematic overview of the difference between quantitative structure-activity relationship and proteochemometric modeling. QSAR uses only the ligand descriptors calculated from the ligands in the data and is typically single target. PCM on the other hand uses both ligand and protein descriptors to build its model and is typically multi-target. Consequently, a PCM model built on a single target is identical to a QSAR as there is no variation in target space.

ligands (typically small molecules) via supervised machine learning using an explicit target descriptor. The main reason of applying PCM over QSAR is that PCM allows pooling of data from related targets to increase the data available for modeling and allow use cases which are not possible with QSAR. Following our definition, targets can be a single protein with multiple binding sites (e.g. allosterism) or multiple related proteins. But also more complex targets such as cells can be converted to descriptors compatible with PCM modeling [6]. By using descriptors extracted from both targets and ligands, the ligand–target interaction space is modeled allowing for extrapolation in target space which is not possible with ligand only data [7]. To use sequence-based target descriptors, aligning these sequences is typically a requirement for best performance, but usage of alignment-independent descriptors has also been described [8]. Sequence based descriptors can be based on the whole protein sequence, limited to the known binding site, or limited to the varying amino acids (Table 1). Moreover, target descriptors do not necessarily need to be protein sequence based, other ways to describe targets can also be used (Table 1). To describe ligands, use can be made of well-established ligand descriptors such as physicochemical properties, circular fingerprints, or 3D descriptors.

Summarizing, PCM exploits the combination of the variation in multiple ligands and multiple targets. The resulting data set is larger and covers more information, which has been observed to lead to more robust models and allows insights and applications not available to ligand-only models (see below) [9].

Model interpretation and scope

After training and validating a PCM model, this model is then commonly interpreted to understand which descriptors correlate with the observed bioactivity. As PCM covers target and ligand information, interpretation can provide additional information beyond QSAR and lead to the identification of important ligand features (e.g. substructures) or target features (e.g. relevant amino acids) [10,11]. One thing to note is that a single target PCM is essentially a QSAR model as the variation in target space is then absent.

The property of PCM that it couples multiple targets into one large model lends itself for deep learning techniques rather well while it can also be applied using shallow learning techniques such as RF, SVM, and PLS [1,8]. Recently, deep learning techniques are more focused on making bioactivity predictions on very large datasets of targets and ligands, while shallow learning leans toward extracting expert knowledge about the system. The diverse applications make PCM an excellent tool for early drug design; multiple sources of knowledge can be integrated into a single model, and predictions can be made on which ligands will work on which targets.

Related approaches

However, the concept of combining data on multiple targets into one large bioactivity space has been explored in other applications as well. A detailed overview of all these other techniques was deemed out of scope of the current review, but we would like to highlight two specific examples. The first, matrix factorization, is a form of collaborative filtering that has been shown to work well on bioactivity data. The main approach in matrix factorization is approximating the weights in a ligand–target interaction matrix and their respective weighted similarity matrices by using the inner products of feature vectors. Essentially, the similarities between ligands and the similarities between targets are compared to create the combined ligand–target interaction space. Contrary to PCM these models cannot be readily interpreted to identify features leading to selectivity or promiscuity in descriptor space, however the approach has been shown to work well and we would like to refer the reader to interesting literature in the field on this [12–14].

Secondly, the creation of (chemogenomic) multi-task models that are built on the ligand (chemical) information of multiple targets, without explicit protein information. Multi-task models co-learn multiple tasks and can profit from shared knowledge, but cannot explicitly link this to features

Table 1. Overview of reviewed literature.

Reference	Section	Protein family/database	Main model quality parameter	Molecular descriptors	Protein descriptors
Giblin *et al.* [9]	Applications	Bromodomain	MCC = 0.83	Circular (Morgan)	Binding-site Physicochemical
Nazarshodeh *et al.* [24]	Applications	Carbonic Anhydrases	R^2 = 0.78, RMSE = 0.55	Circular and 3D	Alignment independent/ physicochemical
Shar *et al.* [25]	Applications	Psychoactive Drug Screen Program version 17	CV-CoD = 0.61 (SVM), 0.63 (RF)	Autocorrelation/ physicochemical	Autocorrelation/ physicochemical
Yordanov *et al.* [26]	Applications	Human Leukocyte Antigens	MCC = 0.86, 86% hit rate in top 10%	Physicochemical	Physicochemical
Cortés-Ciriano *et al.* [29]	Applications	Poly(ADP-ribose) polymerases	R^2 = 0.65, RMSE = 0.95	Circular (Morgan)	Alignment/binding site based
Rasti and Shahangian [30]	Applications	Thymidylate synthases	R^2 = 0.92	Grind	Grind
Rasti and Shahangian [31]	Applications	Phosphodiesterases	R^2 = 0.97,	Grind	Interaction-based (Grind) and Physicochemical
Hariri *et al.* [32]	Applications	Dihydrofolate reductase	R^2 = 0.87–0.89	Grind	Physicochemical (non-conserved binding site)
Hariri *et al.* [33]	Applications	Phosphoinositide 3-kinases	R^2 = 0.77	Grind	Binding-site Physicochemical
Merget and Sorgenfrei *et al.* [34]	Selectivity/ promiscuity	Kinases/ChEMBL version 22	AUC = 0.81, Kappa = 0.27	Circular (Morgan)	Binding site Physicochemical
Christmann-Franch *et al.* [35]	Selectivity/ promiscuity	Collected literature	R^2 = 0.74, RMSE = 0.41	Circular (Morgan)	Physicochemical
Lenselink *et al.* [8]	Selectivity/ promiscuity	ChEMBL version 20	MCC = 0.33,	Circular (Morgan)	Physicochemical
Rasti *et al.* [36]	Selectivity/ promiscuity	Carbonic anhydrases	R^2 = 0.77–0.88	Grind	Binding site/ physicochemical
Manoharan *et al.* [37]	Selectivity/ promiscuity	BACE1	R^2 = 0.77, CVpred = 0.73	Interaction fingerprints/ Glide XP score	Interaction Fingerprints/ Glide XP score
Tresadaern *et al.* [38]	Selectivity/ promiscuity	Metabotropic glutamate 7	ROC = 0.97, MCC = 0.58	Circular/physicochemical	Binding site Physicochemical
Qiu *et al.* [39]	Selectivity/ promiscuity	Nuclear receptors/ ONRLDB	AUC_int = 0.87, AUC_ext = 0.75	Physicochemical	Structure based/ physicochemical
Simeon *et al.* [40]	Selectivity/ promiscuity	Aromatase	R^2 = 0.90, RMSE = 0.42	Circular (Morgan)	Varying positions Physicochemical
Paricharak *et al.* [41]	Virtual screening	Dihydrofolate reductases	R^2 = 0.78, RMSE = 0.59	Circular/physicochemical	Binding site Physicochemical
Cortés-Ciriano *et al.* [11]	Virtual screening	Cyclooxygenases	R^2 = 0.65, RMSE = 0.71	Circular/physicochemical	Binding site Physicochemical
Burggraaf *et al.* [42]	Virtual screening	Sodium-dependent glucose co-transporters 1/2	MCC = 0.49, Hitrate = 39%	Circular/physicochemical	Full sequence alignment Physicochemical
Shaikh *et al.* [43]	Virtual screening	Zinc Database	MCC_int = 0.80–0.95, MCC_ext = 0.71–0.82	Circular (Morgan)	Alignment independent Physicochemical/3D structural
Zakharov *et al.* [45]	Deep learning	ChEMBL/PubChem	R^2 = 0.53, RMSE = 0.66 (best)	Circular (Morgan)/ Path-based (Avalon)	Alignment independent Physicochemical
Reker *et al.* [46]	Deep learning	ChEMBL SARfari/GLASS	MCC = 0.61 (best)	Circular (Morgan)	Alignment independent Physicochemical
Kim *et al.* [47]	Deep learning	ChEMBL version 20	MCC = 0.35, BEDROC = 0.84	Learned Vector-based (CDDD)	Learned sequence based (UniRep)
Jaeger *et al.* [48]	Deep learning	Kinases	CVpred = 0.86, MSE = 0.62	Learned Vector-based (Mol2Vec)	Learned Vector-based (ProtVec)
Jiménez *et al.* [49]	Deep learning	PDBind	R^2 = 0.82, RMSE = 1.27	Voxelized 3D-representation	Voxelized 3D representation
Playe and Stoven [50]	Deep learning	DrugBank subset	ROCAUC = 50–73, AUPR = 17–38	Learned graphNeural networks/circular	One-hot encoded CNN/ Physicochemical

Table 1 (*Continued*)

Reference	Section	Protein family/database	Main model quality parameter	Molecular descriptors	Protein descriptors
Öztürk *et al.* [51]	Deep learning	Kinase datasets Davis and KIBA	DAVIS: R^2 = 0.63, AUPR = 0.71; KIBA: R^2 = 0.67, AUPR = 0.79	Learned smiles embedding	Learned sequence embedding
Lopez-del Rio *et al.* [52]	Deep learning	Collected literature	AUC = 0.85, BEDROC = 0.68, Kappa = 0.46 (recommended CV)	Learned smiles embedding/circular	Learned sequence embedding

Shown are protein (family) studied or data source, the numeric values of their primary parameter for model quality, and descriptor types used. When present external validation values are shown, and when present temporal validation is preferred over random split based validation. See main text for further details.

in the targets (whereas PCM can do this by using explicit descriptors). Similar to PCM, multi-task learning is especially useful in cases where there is little information for an individual target. Multi-task models have also been shown to perform better than individual QSAR models and sometimes on par with PCM. The reader is invited to read some of the recent work in this area [8,15].

Scope of the review

PCM has been utilized in very diverse applications to explore the ligand–target interaction space as we have reviewed before. Since then, in the last 5 years, usage of the technique has become more widespread (Fig. 2). More specifically, a shift can be observed from PCM applications as proof of concept in bioactivity modeling toward it becoming a more established technique with (diverse) applications, undergoing method refinement, and being applied by more groups. Many different protein families were investigated (e.g. kinases [16] and GPCRs [17]), and PCM was also actively explored in deep learning approaches [8,18]. Accompanied by a steady increase in computational power, the complexity of these PCM models increased as well. In this review, results from several recent key publications involving PCM developments are highlighted; for earlier work the reader is referred to our previous reviews from 2011 and 2015 [2,19]. In Table 1, an overview of all literature in this review is listed, alongside which protein (family) the authors studied, validation parameter values, and descriptor types used.

Applications of PCM

In the early 2000s, the use of machine learning in chemometrics became increasingly more popular in the form of Support Vector Machines (SVM) and Random Forests (RF) [20,21]. These applications were using the descriptions of compounds (features or independent variables) and linked them to biological activity (dependent variables) for individual targets. After the introduction of PCM, the use of protein descriptors in bioactivity modeling has become more widespread in the literature as also shown in Fig. 2 [22,23].

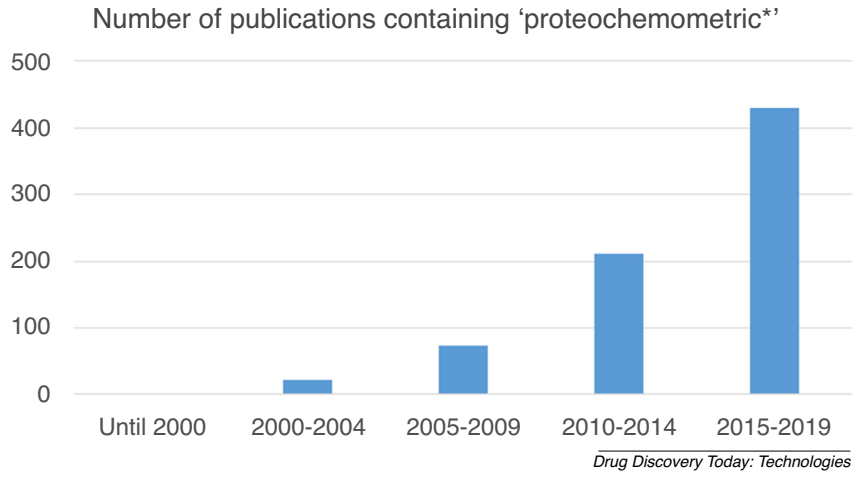

Fig. 2. Overview of published literature mentioning the term proteochemometrics on Google Scholar. The keyword was retrieved in abstract, title or full text. Totals were subsequently binned per period of 5 years. Compared to the early 2000s, an almost 20-fold increase is found in literature referencing proteochemometrics (data was obtained from Google Scholar in July 2020).

One of the main application areas was and is the modeling of several or all members of a protein family at once. As information about several proteins is contained in a PCM model, differences in bioactivity profiles between protein family members is captured. This is reflected in several studies that use PCM to model a full (sub)family of proteins. *Giblin et al.* started from a simple QSAR model using only descriptors based on the small molecules to generate the first set of models, and then created PCM models by adding protein descriptors for the bromodomain containing proteins (BRDs) [9]. With a Matthews Correlation Coefficient (MCC) of 0.83 (versus 0.50 from their QSAR model) they created a high performing predictor for this set of proteins. The authors also showed the use of peptide array data as protein descriptors, with similar performance (MCC of 0.80).

In a 2017 study from *Nazarshodeh et al.* carbonic anhydrases were the focus [24]. The authors here applied feature selection to optimize their PCM model, and showed an increase in performance when using k-nearest neighbor regression over their non-optimized models ($R^2 = 0.78$, RMSE = 0.55). Similarly, feature selection was applied in a broad study performed by *Shar et al.*, in which the authors investigated drug target interactions [25]. Both SVM and RF were applied on a wide selection of ligand–protein pairs and highlighted was how to select the most important compound descriptors from these models. A different approach was shown by *Yordanov et al.*, creating a PCM model that predicted binding to the seven most common human leukocyte antigens based on physicochemical property type descriptors [26–28].

Similarly, improving on previous work, *Cortés-Ciriano et al.* investigated Poly(ADP ribose) polymerases (limiting themselves to the binding site) [29]. They showed an increased performance on both the interpolation, using their 181 compounds for training, and extrapolation to newer compounds. *Rasti and Shahangian* also extrapolate to novel compounds in their 2017 study, exploring thymidylate synthases [30] and phosphodiesterases [31]. *Hariri et al.* designed ligands that were identified as potent and highly selective toward dihydrofolate reductase [32] and toward phosphoinositide 3-kinases [33], respectively.

Whereas PCM is frequently shown to lead to better absolute performance, the addition of target information also allows applications not possible with QSAR such as extrapolation toward untested targets and selectivity prediction as will be highlighted below.

Selectivity and promiscuity prediction

While we have addressed the use in defined protein (sub) families above, PCM can also be applied to full protein superfamilies. *Merget and Sorgenfrei et al.* demonstrated this by using publicly available kinase profiling data to construct a kinome PCM model [34]. This extensive model was used to make kinome-wide predictions after being thoroughly cross-validated with a hold-out partition (AUC 0.81, kappa 0.27, sensitivity 0.63). In this way bioactivity of untested compounds can be predicted across the full kinome (provided their chemical structure is within the applicability domain). Prior to that, another kinome-wide approach was performed by *Christmann-Franck et al.* who applied regression instead of classification and reached an R^2 of 0.74 and RMSE of 0.41[35]. In addition, they were able to interpret their model to rationalize observed (in)activities of inhibitors on specific kinase by 3D structural follow up.

Building on this work, *Lenselink et al.* extended the concept and applied PCM to all high-quality bioactivity data in ChEMBL (version 20) crossing protein subfamily boundaries [8] They were able to improve over individual QSAR models, leading to a PCM with an MCC of 0.33 versus 0.22 for QSAR. Note that this result was obtained with a deep learning application of PCM (see below for further studies using deep learning). These broad studies demonstrate the potential of PCM to also be applied to diverse protein families.

The opposite of these protein superfamily wide PCM models can be found in selectivity approaches within protein families. *Rasti et al.* opted to investigate the selective power of their carbonic anhydrase inhibitors models [36]. They showed that a robust model can predict this selectivity, based on the selective interactions between individual protein family members and compounds by including them as separate descriptors. This also provides a way to estimate selectivity in the design of new inhibitors. A similar approach was performed by *Manoharan et al.* who focused on the BACE1 protein [37]. The authors used structural information, both interaction fingerprints and docking scores, as their main tool to create descriptors and subsequently a statistical model was trained on this data. This led to a predictive BACE1 inhibitor model, where the docking model provided insight into water mediated interactions.

Making the link to different species, *Tresadern et al.* investigated allosteric modulators for the metabotropic glutamate 7 receptor by inclusion of the entire protein family including a rat homologue [38]. The authors showed that following this PCM approach allowed for higher quality hits in their validation as compared to regular QSAR approaches (mainly by a reduction of the false positive rates as confirmed by their prospective validation by a secondary assay).

These examples highlight one of the strengths of creating protein family specific models, as it also allows bioactivity predictions between family members lacking affinity measurements. As such, PCM models lend themselves as powerful tools to compare the effects of slight alterations between protein family members on the bioactivity. PCM to explore selectivity can also be applied to investigate both ligand and protein specific features. A study performed by *Qiu et al.* sought out to analyze the bioactivity of compounds toward

nuclear receptors [39]. They used this data to identify scaffolds overrepresented in both active and inactive compound-protein interactions. A similar approach was performed to aromatase inhibitors by *Simeon et al.*, who analyzed feature importance in their high-performing PCM models (R^2 of 0.90 and RMSE of 0.71) [40]. Combining structural fingerprints as descriptors with the protein structure for aromatase, individual amino acid interactions were identified as important. Noteworthy is that they only focused on the amino acid positions that varied among the studied variants, effectively a form of feature selection.

Applications of PCM in virtual screening

As shown previously, main applications of PCM are typically in improving QSAR predictions and creating prospective models. However, PCM can also be incorporated in the design of novel molecules. *Paricharak et al.* used target prediction and PCM models applied to dihydrofolate reductase (DHFR) inhibitors to combine predicting poly-pharmacology and affinity [41]. For the former, a Bayesian target prediction approach (Recall: 79%) assessed whether a compound was active or inactive as a DHFR inhibitor (and estimate activity on other targets). Affinity on DHFR was predicted using a PCM model (R^2: 0.79; RMSE: 0.59). This combination allowed for a modeling approach that could identify novel ligands for the DFHR.

In order to increase prediction reliability in virtual screening, a common way of combining multiple models is ensemble modeling. In this way multiple models vote and together form one prediction. *Cortés-Ciriano et al.* used this ensemble modeling on the cyclooxygenases (COX) family to improve their base PCM models [11]. An R^2 of 0.59 was found for the COX selectivity of 1086 compounds, which outperformed any of the older singular PCM models. Combining several PCM models can thus be worth investigating as proper implementation leads to an increase in predictive power. One point of criticism can be that *Paracharak* and *Cortés-Ciriano* did not experimentally validate the models and only performed virtual screening.

PCM can also be a powerful tool in the absence of structural data. In a study by *Burggraaff et al.*, the sodium-dependent glucose co-transporters 1 and 2 (SGLT1/SGLT2) were addressed [42]. As there was no structural information available for SGLT, the authors created a SGLT PCM model derived from publicly available and in-house data for SGLT1 and SGLT2. Virtual screening led to an experimentally validated hit-rate of 39%, underscoring the relevance of the followed approach. *Shaikh et al.* concluded in their study on drug-target interactions that using sequence-based descriptors performed just as well as structure-based ones [43]. They assessed the sc-PDB database, which contains annotated druggable binding sites. The authors suggest that 'non-interactive' datapoints (i.e. untested experimentally) in the dataset should not be classified as negative by default, as this skews the model to underpredicting affinities. A better approach is leaving these datapoints out or predicting their values with a statistical model trained on the available data, in a way a form of active learning and interpolation. Their other major finding, contrary to structure-based descriptors, sequence-based descriptors are more widely available, allow for broader coverage of data, and with that a better possibility of filling in these 'non-interactive' datapoints.

Combining the above findings, it appears that structure-based descriptors are more predictive in protein family specific instances (with high levels of detail), while sequence-based descriptors are better suited for wide PCM models (where such high level is lacking). A careful consideration must be made on which descriptors to use, and is mainly dependent on the amount and type of data available.

Deep learning applications of PCM

As the field of computational drug discovery moved forward in the last decade, Deep Learning (DL) became increasingly more relevant in both activity modeling and de novo structure generation [44]. The power of deep learning lies in its ability to model high dimensional data sets with a high accuracy and to learn directly on data instead of using engineered features or descriptors. The downside of this method is the low interpretability and very high hyperparameter dimensionality. This low interpretability is found in the (multiple) hidden layers of neural networks, a black box where information is transformed and weighted by the algorithms without any clear feedback. We will highlight some key DL applications here.

As databases become increasingly larger, so do the data set size and training time used for PCM models to model bioactivity. DL can be sped up using graphical processing units (GPUs). This was demonstrated by *Lenselink et al.* who were to apply a deep learning PCM model to the whole high-quality benchmark set they extracted from ChEMBL [8]. Similarly, *Zakharov et al.* introduced DL to improve the performance of these very large PCM models by using GPUs and a novel architecture [45]. In their work they integrate both multitask deep learning models and consensus models and observed an increase in accuracy of the predictions. Conversely, *Reker et al.* showed with DL methods that specific, high performing models can be constructed from large bioactivity datasets that do not encompass a full matrix [46]. Moreover, the researchers noted issues with combinatorial screening: predicting the activity for every possible combination of target/ligand is time consuming and expensive. Using their developed method, they extracted the most informative target/ligand interactions and achieved similar performance with 25% of the data.

Studies applying DL in bioactivity modeling frequently focus on improving information encoding or learn data

representations before applying supervised machine learning. *Kim et al.* extracted ligand and target descriptors based on a comprehensive benchmark set of compound-protein interactions [47]. These descriptors were then incorporated into a PCM model that performed better than previously examined models based on engineered descriptors. This study shows a significant improvement in descriptor creation and data preprocessing, cutting down on the time needed to do this manually while improving accuracy.

In descriptor representation, *Jaeger et al.* created deep neural networks to further automate the integration of structural information [48]. Compound–target pair information was extracted from text sources and a learned vector based representation was created for both (Mol2Vec and ProtVec) to function as descriptor. After supervised learning, this approach showed an increase in predictive power, and could overcome drawbacks such as sparseness in the original dataset. *Jiménez et al.* investigated 3D-convolutional neural networks and found that using 3D generated structural descriptors gave similar results to established methods, but improved on the speed and performance of the method itself [49]. Both these methods show increased performance for PCM models in their respective areas, and allow for better predictions through the use of DL in descriptor creation and model application.

Beyond improving activity modeling and descriptors, other areas of improvement have been reported. *Playe and Stoven* compared deep learning based descriptors (graph neural network and one-hot encoded convolutional neural network based respectively) with literature-based descriptors (i.e. shallow learning) [50]. They concluded that while on large datasets deep learning performed better than shallow learning, the opposite was true for the smaller datasets. This reinforces the idea that expert knowledge works well on specific protein families, and that PCM models created with that limitation in mind predict bioactivity well. For larger datasets, deep learning is preferred to identify relationships and interactions that are not directly found in experimental validations.

Öztürk et al. showed an effective manner for predicting affinities using sequence type data for both compounds and proteins from which an embedding was created. They apply this description form to PCM models on kinases [51]. Built on similar descriptor types, *Lopez-del Rio et al.* compared cross-validation approaches in a PCM deep learning environment [52]. The authors concluded that using a restrictive cross-validation method on a PCM model resulted in more reliable predictions. These studies reinforce the idea that targeted predictions (such as a small selected database, expert knowledge on a protein family, 3D structural information) are valuable to explore as they have more reliable and significantly powerful results compared to large dataset wide predictions.

What is common in all deep learning applications is that PCM bioactivity predictions are still in their infancy and

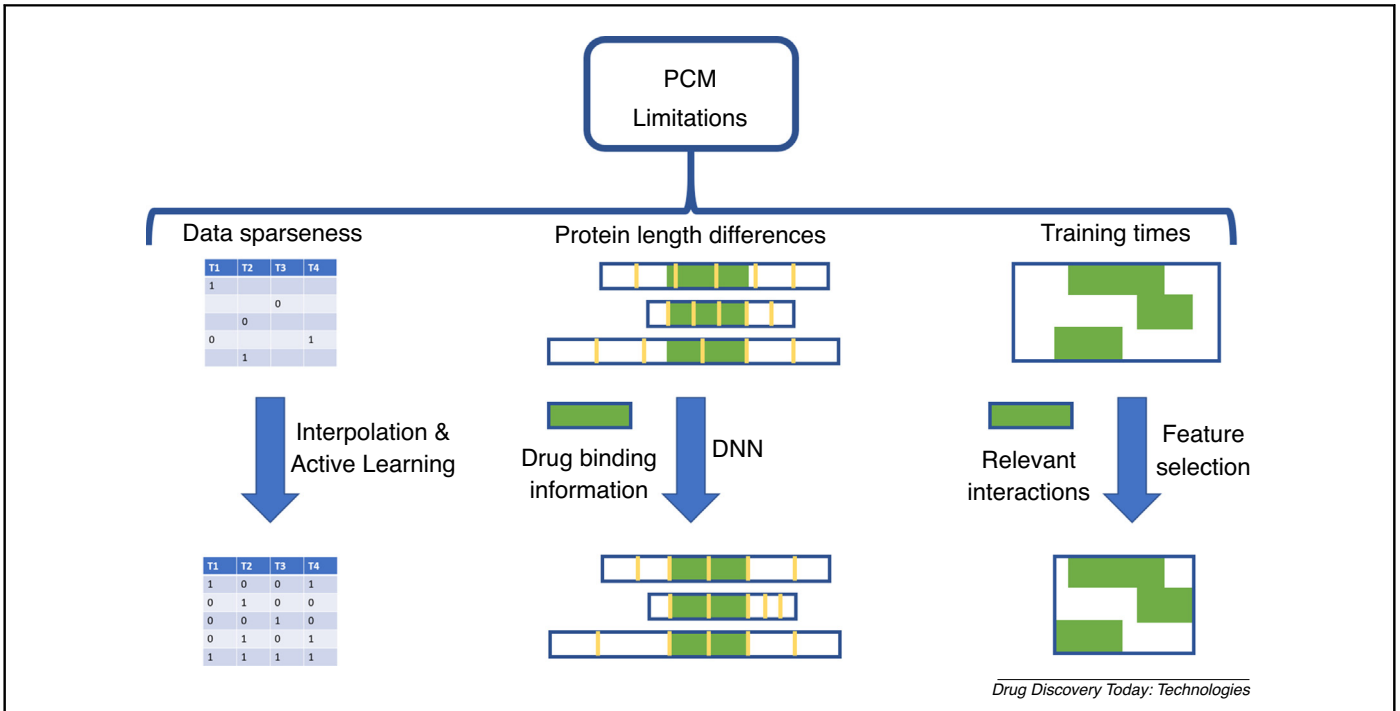

Fig. 3. A schematic overview of several limitations of proteochemometrics and possible mitigation methods. The left-hand side shows data sparseness, where interpolation and active learning can be used to infer activity in the empty cells and complete the matrix as efficiently as possible. The middle shows how differing protein lengths can hamper protein descriptor creation. Manually assigning relevant drug binding information or learning a protein embedding can mitigate this problem. The right-hand side shows increasing training times with increasing data. Each PCM has to be trained over a large matrix, which is time consuming. Feature selection can be applied to limit the data set to the relevant information cutting down training times.

intensive research is ongoing. As research progresses and algorithms improve, the quality of these predictions will increase and ultimately lead to a better hit-to-drug rate. However, there are also some limitations to PCM, three of which will be discussed below.

Limitations

There are several known PCM modeling limitations, here we will address data sparseness, inability to align protein sequences, and training times (Fig. 3). Sparse data leads to a reduction in model reliability. Predictive power increases when coverage of the ligand–protein interaction matrix is increased, however, these data are not always available. PCMs can be used to iteratively fill this matrix in an active learning framework, improving model quality [46]. It should be noted that expansion of the data ultimately requires experimental validation beyond only model predictions.

A second limitation is encountered when using PCMs on a diverse set of proteins that cannot always be easily aligned (in particular if different families are incorporated). Two potential mitigations can be considered in such cases, firstly a manual assignment of relevant amino acids that make up the ligand binding site. Secondly, a specific neural network (e.g., learned embeddings [53,54]) can be trained to extract protein descriptors from the sequence space.

A final limitation is that a massive increase in data can lead to significant increases in training times. In particular this is true for PCM, which uses massive ligand–target interaction matrices when focusing on multiple proteins at once (such as a superfamily). While GPUs speed up the training process, this may not be sufficient. Feature selection can then be used (e.g., on a representative subset of the full dataset) to decrease the size of the full dataset to a more manageable amount of data.

Conclusions

As the knowledge about compound-protein interactions expands, the power of statistical modeling and PCM models constructed on these data, increases. PCM models perform best for specific research questions, such as affinity and selectivity predictions of compounds acting on restricted protein families. For larger more diverse datasets the quality of specific predictions tends to be less precise. However, the application area of PCM models continues to expand, as their predictive power is shown in an increasing set of studies.

A typical application should be followed by experimentally validating novel predictions by these models. Concepts such as ensembling, scaffold-to-target investigation, as well as deep learning implementations, increase the validity and usefulness of PCM models. As neural network technology continues to develop, more powerful applications of PCM models in deep learning drug design are anticipated along with novel forms to embed chemical and target information.

Finally, as drug development is still a long and costly endeavor, proper usage of PCM models can decrease the cost and increase the success rate of the early design stages. The increase in (publicly) available drug interaction data goes hand in hand with deep learning methods able to handle this data. Therefore, a combination of both deep learning and PCM models shows a strong contender for early drug design and should provide new insights and eventually new drugs.

Conflict of interests

None declared.

References

[1] Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg JES. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim Biophys Acta - Gen Subj 2001;1525:180–90. http://dx.doi.org/10.1016/S0304-4165(00)00187-2.

[2] van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Medchemcomm 2011;2:16. http://dx.doi.org/10.1039/C0MD00165A.

[3] Van Westen GJP, Swier RF, Wegner JK, IJzerman AP, Van Vlijmen HWT, Bender A. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J Cheminform 2013;5:41. http://dx.doi.org/10.1186/1758-2946-5-41.

[4] Van Westen GJP, Swier RF, Cortes-Ciriano I, Wegner JK, Overington JP, IJzerman AP, et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. J Cheminform 2013;5:42. http://dx.doi.org/10.1186/1758-2946-5-42.

[5] Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J Cheminform 2014;6:1–19.

[6] Corté S-Ciriano I, et al. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics 2016;32:85–95. http://dx.doi.org/10.1093/bioinformatics/btv529.

[7] Lapinsh M, Prusis P, Uhlé S, Wikberg JES. Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. Bioinformatics 2005;21:4289–96. http://dx.doi.org/10.1093/bioinformatics/bti703.

[8] Lenselink EB, ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform 2017;9:45. http://dx.doi.org/10.1186/s13321-017-0232-0.

[9] Giblin KA, Hughes SJ, Boyd H, Hansson P, Bender A. Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins. J Chem Inf Model 2018;58:1870–88. http://dx.doi.org/10.1021/acs.jcim.8b00400.

[10] Subramanian V, Ain QU, Henno H, Pietilä LO, Fuchs JE, Prusis P, et al. 3D proteochemometrics: using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases. Medchemcomm 2017;8:1037–45. http://dx.doi.org/10.1039/c6md00701e.

[11] Cortes-Ciriano I, Murrell DS, Van Westen GJ, Bender A, Malliavin TE. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. J Cheminform 2015;7:1–18. http://dx.doi.org/10.1186/s13321-014-0049-z.

[12] Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. Proc ACM SIGKDD Int Conf Knowl Discov Data Min vol Part F1288 2013;1025–33. http://dx.doi.org/10.1145/2487575.2487670.

[13] Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, et al. Repurposing high-throughput image assays enables biological activity

prediction for drug discovery. Cell Chem Biol 2018;25. http://dx.doi.org/10.1016/j.chembiol.2018.01.015. 611-618.e3.
[14] Zhang J, Li C, Lin Y, Shao Y, Li S. Computational drug repositioning using collaborative filtering via multi-source fusion. Expert Syst Appl 2017;84:281–9. http://dx.doi.org/10.1016/j.eswa.2017.05.004.
[15] Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Doğan T. DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 2020;11:2531–57. http://dx.doi.org/10.1039/c9sc03414e.
[16] Lapins M, Wikberg JES. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. BMC Bioinformatics 2010;11:1–15. http://dx.doi.org/10.1186/1471-2105-11-339.
[17] Gao J, Huang Q, Wu D, Zhang Q, Zhang Y, Chen T, et al. Study on human GPCR-inhibitor interactions by proteochemometric modeling. Gene 2013;518:124–31. http://dx.doi.org/10.1016/j.gene.2012.11.061.
[18] Lee M, Kim H, Joe H, Kim HG. Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery. J Cheminform 2019;11:1–16. http://dx.doi.org/10.1186/s13321-019-0368-1.
[19] Cortés-Ciriano I, Ain QU, Subramanian V, Lenselink EB, Endez-Lucio OM, Ijzerman AP, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects; 2015. http://dx.doi.org/10.1039/c4md00216d.
[20] Demiriz A, Demiriz A, Bennett KP, Breneman CM, Embrechts MJ. Support vector machine regression in chemometrics. Comput Sci Stat Proc 33RD Symp Interface 2001.
[21] Scott IM, Lin W, Liakata M, Wood JE, Vermeer CP, Allaway D, et al. Merits of random forests emerge in evaluation of chemometric classifiers by external validation. Anal Chim Acta 2013;801:22–33. http://dx.doi.org/10.1016/j.aca.2013.09.027.
[22] Liu L, He D, Yang S, Xu Y. Applying chemometrics approaches to model and predict the binding affinities between the human amphiphysin SH3 domain and its peptide ligands (suplementary material). Protein Pept Lett 2010;17:246–53. http://dx.doi.org/10.2174/092986610790226085.
[23] Lindström A, Pettersson F, Linusson A. Quantitative protein descriptors for secondary structure characterization and protein classification. Chemom Intell Lab Syst 2009;95:74–85. http://dx.doi.org/10.1016/j.chemolab.2008.08.006.
[24] Nazarshodeh E, Sheikhpour R, Gharaghani S, Sarram MA. A novel proteochemometrics model for predicting the inhibition of nine carbonic anhydrase isoforms based on supervised Laplacian score and k-nearest neighbour regression. SAR QSAR Environ Res 2018;29:419–37. http://dx.doi.org/10.1080/1062936X.2018.1447995.
[25] Shar PA, Tao W, Gao S, Huang C, Li B, Zhang W, et al. Pred-binding: large-scale protein–ligand binding affinity prediction. J Enzyme Inhib Med Chem 2016;31:1443–50. http://dx.doi.org/10.3109/14756366.2016.1144594.
[26] Yordanov V, Dimitrov I, Doytchinova I. Proteochemometrics-based prediction of peptide binding to HLA-DP proteins. J Chem Inf Model 2018;58:297–304. http://dx.doi.org/10.1021/acs.jcim.7b00026.
[27] Dimitrov I, Doytchinova I. Peptide binding prediction to five most frequent HLA-DQ proteins – a proteochemometric approach. Mol Inform 2015;34:467–76. http://dx.doi.org/10.1002/minf.201400150.
[28] Yordanov V, Dimitrov I, Doytchinova I. Proteochemometric analysis of peptides binding to human leukocyte antigen (Hla) proteins from locus Dp, vol. 64. 2017.
[29] Cortés-Ciriano I, Bender A, Malliavin T. Prediction of PARP inhibition with proteochemometric modelling and conformal prediction. Mol Inform 2015;34:357–66. http://dx.doi.org/10.1002/minf.201400165.
[30] Rasti B, Shahangian SS. Proteochemometric modeling of the origin of thymidylate synthase inhibition. Chem Biol Drug Des 2018;91:1007–16. http://dx.doi.org/10.1111/cbdd.13163.
[31] Rasti B, Schaduangrat N, Shahangian SS, Nantasenamat C. Exploring the origin of phosphodiesterase inhibition: via proteochemometric modeling. RSC Adv 2017;7:28056–68. http://dx.doi.org/10.1039/c7ra02332d.
[32] Hariri S, Ghasemi JB, Shirini F, Rasti B. Probing the origin of dihydrofolate reductase inhibition via proteochemometric modeling. J Chemom 2019;33:e3090. http://dx.doi.org/10.1002/cem.3090.
[33] Hariri S, Rasti B, Mirpour M, Vaghar-Lahijani G, Attar F, Shiri F. Structural insights into the origin of phosphoinositide 3-kinase inhibition. Struct Chem 2020;1–18. http://dx.doi.org/10.1007/s11224-020-01510-2.
[34] Sorgenfrei FA, Fulle S, Merget B. Kinome-wide profiling prediction of small molecules. ChemMedChem 2018;13:495–9. http://dx.doi.org/10.1002/cmdc.201700180.
[35] Christmann-Franck S, Van Westen GJP, Papadatos G, Beltran Escudie F, Roberts A, Overington JP, et al. Unprecedently large-scale kinase inhibitor set enabling the accurate prediction of compound-kinase activities: a way toward selective promiscuity by design? J Chem Inf Model 2016;56:1654–75. http://dx.doi.org/10.1021/acs.jcim.6b00122.
[36] Rasti B, Karimi-Jafari MH, Ghasemi JB. Quantitative characterization of the interaction space of the mammalian carbonic anhydrase isoforms I, II, VII, IX, XII, and XIV and their inhibitors, using the proteochemometric approach. Chem Biol Drug Des 2016;88:341–53. http://dx.doi.org/10.1111/cbdd.12759.
[37] Manoharan P, Chennoju K, Ghoshal N. Target specific proteochemometric model development for BACE1 – protein flexibility and structural water are critical in virtual screening. Mol Biosyst 2015;11:1955–72. http://dx.doi.org/10.1039/c5mb00088b.
[38] Tresadern G, Trabanco AA, Pérez-Benito L, Overington JP, Van Vlijmen HWT, Van Westen GJP. Identification of allosteric modulators of metabotropic glutamate 7 receptor using proteochemometric modeling. J Chem Inf Model 2017;57:2976–85. http://dx.doi.org/10.1021/acs.jcim.7b00338.
[39] Qiu T, Wu D, Qiu J, Cao Z. Finding the molecular scaffold of nuclear receptor inhibitors through high-throughput screening based on proteochemometric modelling. J Cheminform 2018;10:21. http://dx.doi.org/10.1186/s13321-018-0275-x.
[40] Simeon S, Spjuth O, Lapins M, Nabu S, Anuwongcharoen N, Prachayasittikul V, et al. Origin of aromatase inhibitory activity via proteochemometric modeling. PeerJ 2016;2016:e1979. http://dx.doi.org/10.7717/peerj.1979.
[41] Paricharak S, Cortés-Ciriano I, IJzerman AP, Malliavin TE, Bender A. Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules. J Cheminform 2015;7:15. http://dx.doi.org/10.1186/s13321-015-0063-9.
[42] Burggraaff L, Oranje P, Gouka R, Van Der Pijl P, Geldof M, Van Vlijmen HWT, et al. Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling. J Cheminform 2019;11:1–10. http://dx.doi.org/10.1186/s13321-019-0337-8.
[43] Shaikh N, Sharma M, Garg P. An improved approach for predicting drug–target interaction: proteochemometrics to molecular docking. Mol Biosyst 2016;12:1006–14. http://dx.doi.org/10.1039/c5mb00650c.
[44] Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharm 2016;13:2524–30. http://dx.doi.org/10.1021/acs.molpharmaceut.6b00248.
[45] Zakharov AV, Zhao T, Nguyen DT, Peryea T, Sheils T, Yasgar A, et al. Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. J Chem Inf Model 2019;59:4613–24. http://dx.doi.org/10.1021/acs.jcim.9b00526.
[46] Reker D, Schneider P, Schneider G, Brown J. Active learning for computational chemogenomics. Future Med Chem 2017;9:381–402. http://dx.doi.org/10.4155/fmc-2016-0197.
[47] Kim P, Winter R, Clevert D-A. Deep protein-ligand binding prediction using unsupervised learned representations; 2020. http://dx.doi.org/10.26434/CHEMRXIV.11523117.V1.
[48] Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model 2018;58:27–35. http://dx.doi.org/10.1021/acs.jcim.7b00616.
[49] Jiménez J, kaki M, Martínez-Rosell G, De Fabritiis G. KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural

networks. J Chem Inf Model 2018;58:287–96. http://dx.doi.org/10.1021/acs.jcim.7b00650.

[50] Playe B, Stoven V. Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity. J Cheminform 2020;12:1–18. http://dx.doi.org/10.1186/s13321-020-0413-0.

[51] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 2018;vol. 34:i821–9. http://dx.doi.org/10.1093/bioinformatics/bty593.

[52] Lopez-Del Rio A, Nonell-Canals A, Vidal D, Perera-Lluna A. Evaluation of cross-validation strategies in sequence-based binding prediction using deep learning. J Chem Inf Model 2019;59:1645–57. http://dx.doi.org/10.1021/acs.jcim.8b00663.

[53] Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics 2018;34:2642–8. http://dx.doi.org/10.1093/bioinformatics/bty178.

[54] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16:1315–22. http://dx.doi.org/10.1038/s41592-019-0598-1.