G-*PLIP*: Knowledge graph neural network for structure-free protein-ligand affinity prediction

Simon J. Crouzet[1,2,3], Anja Maria Lieberherr[1,4], Kenneth Atz[1,5], Tobias Nilsson[1], Lisa Sach-Peltason[1], Alex T. Müller[1], Matteo Dal Peraro[2,3], and Jitao David Zhang[*1]

[1]Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, CH-4070 Basel, Switzerland.
[2]Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 1015, Switzerland.
[3]Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland.
[4]ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland.
[5]ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
[*]To whom correspondence should be addressed: *jitao_david.zhang@roche.com*

Todo list

Abstract

Protein-ligand interaction (PLI) shapes efficacy and safety profiles of small molecule drugs. Most existing methods rely on resource-intensive computation to predict PLI based on structural information. Here we show that a light-weight graph neural network (GNN), trained with quantitative PLIs of a small number of proteins and ligands, is able to predict the strength of unseen PLIs. The model has no direct access to structural information of protein-ligand complexes. Instead, the predictive power is provided by encoded knowledge of proteins and ligands, including primary protein sequence, gene expression, protein-protein interaction network, and structural similarities between ligands. Our novel model performs competitively with or better than structure-aware models. Our results suggest that existing PLI-prediction methods may be further improved by using *omics* data and representation learning techniques that embed biological and chemical knowledge.

Introduction

Most drug-like small molecules act as ligands binding to proteins [1, 2]. Protein-ligand interaction (PLI) with primary and secondary pharmacological targets critically determine the efficacy and safety profile of a drug [3, 4]. Prediction, quantification, and interpretation of PLI is, therefore an integrative component of pre-clinical drug discovery [5, 6].

A plethora of experiments has been developed for PLI quantification. Experimental methods include well-established low-throughput biophysical and biochemical binding assays, as well as high-dimensional omics-technology-based methods, such as chemoproteomics [7]. Low-dimensional assays are fast and accurate for a few protein targets that are of interest for the drug molecule [8]. However, our ability to probe proteome-wide PLIs with such assays is severely limited by the availability of large quantities of purified proteins of interest [9]. In addition, quantification of proteome-wide PLIs is labour-intensive, time-consuming, and often results in moderate sensitivity [10].

Computational methods predicting PLIs complement experimental approaches and are indispensable in drug discovery pipelines [11–13]. Some methods simulate molecular docking [14, 15]. Other methods rely on structure-based representations of the protein-compound complex, taking solved structures as input to machine-learning algorithms [16–19]. Accumulation of high-quality data derived from biochemical and biophysical screening [20], besides novel algorithms, empower such methods, although the underlying limitation remains, namely only a limited number of complex structures is solved. Nevertheless, screening procedures are progressively being complemented or replaced by modelling procedures [21, 22].

Besides the limitation of structures, most existing computational methods do not consider several biological factors that may contribute to PLI. For instance, most of them treat proteins as isolated entities, ignoring protein-protein interactions that are ubiquitous under physiological conditions [23]. Localization and expression of proteins are usually not considered, either. We reason that new prediction algorithms that are both able to circumvent existing limitations and able to embed additional knowledge into the prediction process may be valuable, especially if they lead to either a better performance, or a better understanding of the relative importance of diverse factors contributing to PLI, or both.

During the last years, applications of deep learning techniques have greatly expanded in drug discovery. [24] In particular, geometric deep learning and graph neural networks (GNNs) have been widely used for molecular representations [25], structure-based drug design [26–29], molecular property prediction [30–32], and protein-ligand binding predictions [33]. Besides its good performance in many tasks, two reasons make GNNs particularly appealing to us in the context of PLI prediction. First, in contrast to traditional deep learning, geometric deep learning can operate in non-Euclidean spaces, thus able to capture diverse information that is not easily embedded into Euclidean representation, such as graphs that consist of nodes and edges. [34]. Thus, GNNs are able to capture complex and interdependent relationship between biological entities in multiple scales. The benefit is particularly prominent in protein-relevant tasks, such as protein representation [**zhang2022protein**, **ingraham2019generative**, 29], protein folding [35], protein design [36, 37], and protein-protein interaction [38–43]. Second, various research groups have developed new techniques to improve the interpretability of GNNs [24, 44, 45]. Bearing these promises, challenges [16], and concerns [46] in mind, we consider GNN a natural candidate for the task of PLI prediction.

Here, we report a novel GNN model, termed G-*PLIP* (*Graph for Protein Ligand Interaction Prediction*), which predicts PLI quantitatively using non-structural features of proteins and ligands. We demonstrate that G-*PLIP* performs comparably with or better than state-of-the-art structure-aware PLI-prediction models, while the computational cost remains low thanks to its simple model architecture, and to the encapsulation of chemical sub-space into a single graph. We analyzed the contribution of individual feature types to the model's performances and demonstrated that primary sequences of proteins and molecular information of ligands contain rich information which can be exploited for PLI prediction.

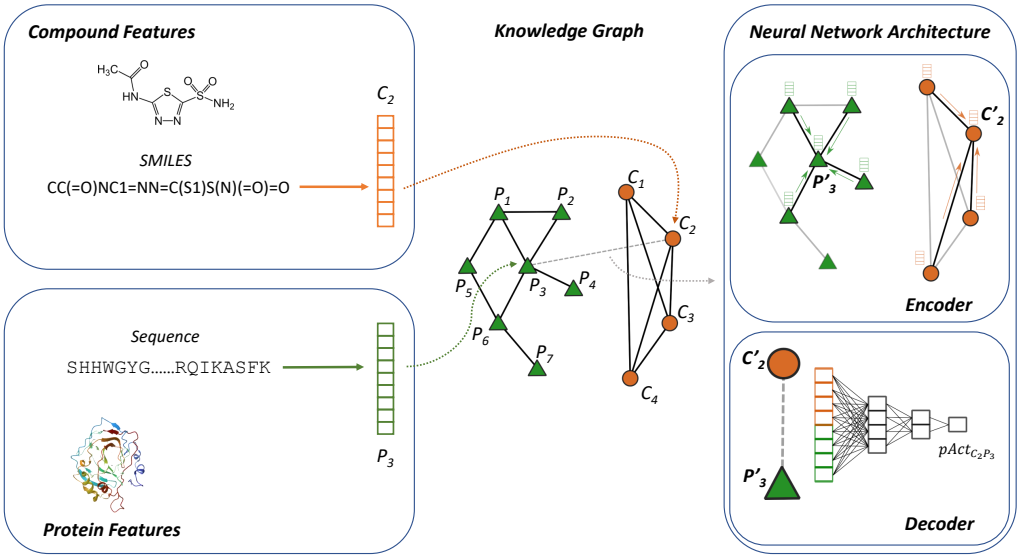

Fig. 1. Overview of a knowledge graph derived from protein-protein interactome and *in-vitro* pharmacology data. The heterogeneous graph consists of binary protein–protein interactions and drug–drug similarities. Information is integrated into the model in the form of protein and drug node feature vectors. The highlighted network encodes information of node P_3 and C_2 from their neighbours and decodes their embedding to predict the P_3-C_2 bioactivity ($pAct$). G-*PLIP* predicts drug-target activities, aiming to identify off-target effects and evaluate bioactivities of novel compounds. This example displays the prediction of drug-target bioactivity between *acetazolamide*, an anticonvulsant used to treat oedema, glaucoma or certain epilepsy, and the human protein carbonic anhydrase II encoded by the CA2 gene, which is responsible for the catalysis of the reversible hydration of carbon dioxide.

Results

Prediction of drug-protein activity (pAct)

We formulate the task of protein-ligand interaction prediction as a multirelational link prediction problem in a heterogeneous graph. The formulation is inspired by the observations that heterogeneous multimodal graphs depict various entities and their interactions in a real-world system, and are therefore ideal for coarse-grained modelling of biochemical processes [38, 47]. We subsequently trained a GNN model to predict bioactivities of unseen protein-ligand pairs (Figure 1).

Table 1. G-*PLIP* benchmark on different settings, depicting the difference of performances between the default stratified split, the scaffold split, and two control conditions (edge labels randomly assigned, and edge labels shuffled).

Model	*RMSE*	*RBO*	R_p	R^2
G-*PLIP* - Baseline	0.862	0.680	0.525	0.246
G-*PLIP* - Scaffold Split	0.886	0.660	0.499	0.208
Random Edges	1.982	0.111	0.036	-0.015
Shuffled Edges	1.064	0.731	0.051	-0.033

We first assessed G-*PLIP* on a 80%-20% split by stratifying the split by proteins. Each of the proteins will have the same proportion of samples dispatched between the train and test set, for an equal representation across the sets to guarantee a fair evaluation. Proteins possessing only one sample were excluded from the split and retained for a specific use case. To reflect the fact that novel small-molecule screening compounds are often derived from a base molecule, known as the scaffold, which possesses properties of interest, and that multiple analogues are subsequently synthesized and tested, we also created train and test sets from the original scaffolds of our compounds to evaluate the model. G-*PLIP* produced a good accuracy (Table 1), particularly noteworthy with respect to its performances when subjected to randomized data.

Our model is able to learn from and predict protein-ligand interactions despite the fact that it does not use protein-ligand complex structures. We observe intriguing characteristics of the model. First, it displays a propensity to underestimate bioactivities with high values, while overestimating those with lower averages (as illustrated by Figure 2, panel a and b). The tendency lead to discrepancy between the performance judged by different metrics, with R_p and R^2 scores indicating a moderate performance and *RMSE* indicating a good

performance. Second, the model is also able to predict *pAct* values for missing links, i.e. cases where there is no quantitative information about the interaction between a protein and a ligand. Interestingly, the median predicted value of missing links is lower than the predicted values for non-missing links (Supplementary Figure S1). Note that in the context of protein-ligand interaction, a missing link does not necessarily mean a lack of interaction; it may arise due to a lack of measurement, a reporting bias, or due to the sensitivity of the assay. Finally, our model achieves a comparable performance for a wide range of targets classified by their molecular functions (Figure 2c).

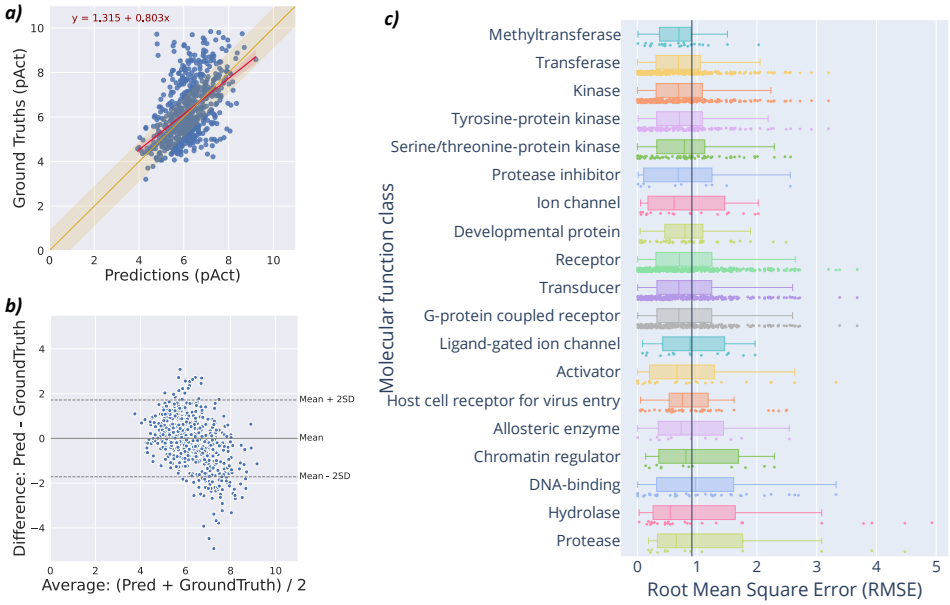

Fig. 2. a) Scatter plot of ground-truth values (*y*-axis) against predictions (*x*-axis). In yellow is highlighted the area of values below the mean $RMSE$. A fitted linear regression model prognosticating ground truths from predictions is sketched in red, with 95% confidence intervals indicated in shades. b) The MA plot portraying the differences between predicted values and ground truths (*y*-axis) against the average of the two (*x*-axis). Mean and two units of standard deviation of the differences are indicated in dashed lines. c) Model's performance on the test set across different molecular function classes of targets. Target classes are sorted by median $RMSE$ values. The black vertical line is the overall mean value across all molecular function classes. Molecular functions possessing less than 10 targets are not displayed.

Comparison with state-of-the-art pipelines

Having established a baseline performance, we next subjected G-*PLIP* to a comprehensive and rigorous evaluation. We applied the model to *PDBBind*, an independent dataset, and compared its outcomes with those of cutting-edge models leveraging structural information [16–19]. Despite the availability of structural data in the *PDBBind* dataset, our model only leveraged the amino-acid sequence of proteins and the SMILES representations of

compounds, in order to assess the merit of the structure-free methodology. Since our approach is grounded in human tissue gene expression data and tailored for applications in drug discovery, we confined the comparison to PLI involving human proteins.
Employing protocols established by prior studies [16–19], we ran our comparative study with two distinct setups. First, we assessed G-*PLIP* on the 2016 CASF Core Set (n=141), a subset of the *PDBBind* data, with the remaining database entries constituting the training set. Subsequently, we evaluated G-*PLIP* by a temporal splitting of the *PDBBind* dataset (denoted as the 2019 Hold-out Set), where the 2016 version serves as a training set, while the samples introduced in the 2019 and 2020 versions were utilized to test the model (n=667). We furthermore partitioned our training data into both a training and validation subset, with the latter aiding in the selection of the most optimal model across epochs.

G-*PLIP* performed significantly better than these peers, with an improvement of 23.2% compared to Yuel [48], 19.3% compared to K_{DEEP}, 21.7% compared to Fusion, 28.3% compared to Pafnucy [17] and 32.2% compared to the complete MPNN model (with merged protein, ligand and interaction graph) [16], judged by RMSE on the external 2016 CASF Core Set. To ensure the consistency of the model's performance, we conducted three distinct training runs and also reported the standard deviation (SD), confirming its reliability. Pearson correlation coefficients are reported to be slightly lower for G-*PLIP* on the core set, however it has been shown this metric can be illusory to evaluate predictive performances on regression tasks [49] and can be sensitive to the number of predicted samples. Satisfying Pearson correlation coefficients can be reached with models unable to predict relevant values. This hypothesis was corroborated by training a model using the aforementioned correlation coefficient as a loss function. While this yielded a Pearson correlation coefficient of 0.79 while proposing a RMSE of 6.12, and thus does not offer any practical predictive power (Supplementary Figure S3).

Limited contribution of low-quality data to improve learning abilities

Previous studies remained inconclusive about the question of whether extending datasets by integrating low-quality affinity data would enhance the performance of machine-learning models. While some studies demonstrated improvements in this context [50], recent investigations have posited that such integration exerts minimal [19] or even misleading influence on model outcomes [51, 52]. To address this discrepancy, we conducted experiments using distinct subsets from the *PDBBind* collection: the *complete*, *refined*, and

Table 2. Performances of G-*PLIP* on several publicly available datasets over three independent trainings (reported as $mean \pm sd$), and comparison with state-of-the-art models. Results for Yuel [48] were manually compiled from a model retrained locally using the parameters proposed by the authors on the same train/test splits. Results for K_{DEEP} are reported according to the performances claimed in the original paper [19] (for the 2016 Core Set) and manually investigated by a later study [18] (for the 2019 Hold-out Set). Results for Fusion are reported according to the best performances reached during the study, namely by the *midlevel fusion* [18]. Performances for Pafnucy [17] and for MPNN-PLI were obtained from the original MPNN-PLI paper [16]

Comparison on PDB Bind

Model	2016 CASF Core Set		2019 Hold-out Set	
	RMSE	R_p	*RMSE*	R_p
G-*PLIP*	**1.024 ± 0.013**	0.781 ± 0.003	0.979 ± 0.011	0.679 ± 0.008
Yuel	1.334	0.782	**0.928**	**0.857**
MPNN-PLI	1.511	0.813	1.481	0.652
Pafnucy	1.429	0.773	1.642	0.456
K_{DEEP}	1.270	**0.820**	1.424	0.487
Fusion	1.308	0.810	1.338	0.545

other sets. Specifically, the *complete* set represents the entirety of available data from the *PDBBind* dataset, the *refined* set encompasses the curated high-quality data, and the *other* set comprises the data not present in the *refined* subset.

Each of those three subsets was employed for training and subsequently evaluated on both the 2016 CASF Core Set and the 2019 Hold-Out Set. The first setup provides a consistent benchmark using reliable test samples carefully selected and segregated from the three subsets (141 bioactivity values). In the second setup, we aimed to gauge the model's ability to generalize within the given subset. Here, the test data's composition directly mirrors the training subset's characteristics, enabling us to observe the model's behaviour when trained and evaluated on data of similar quality.

Table 3. Performances of G-*PLIP* on the full set and two subsets of the PDB Bind dataset. The 2016 CASF Core Set is external and thus constant over the subsets, while the 2019 Holdout Set is done by temporal splitting over the selected subset and thus dependent of its content.

Model	2016 Core Set		2019 Hold-out Set	
	RMSE	R_p	*RMSE*	R_p
Complete	1.024	0.613	0.968	0.689
Refined	1.137	0.677	1.116	0.748
Other	1.504	0.660	0.981	0.730

Results show that augmenting the dataset with lower-quality samples does not confer benefits to the model's performance (Table 3), and the model appears to struggle to capitalize on the size of the enlarged dataset. Indeed, when the model is evaluated in an

external reference set (CASF-2016 Core), comparable results are achieved between the *complete* and the *refined* subsets, and the model trained on the *other* subset fails to reach their baseline performances, despite demonstrating favourable metrics on the hold-out set. The model's performance on the temporal split appears influenced by other factors, possibly including the narrower distribution of bioactivity values.

Informativeness of graph features

Subsequently, we delve into comprehending the scope of G-*PLIP*'s capabilities and scrutinize the informativeness of its different features. We aim to examine the relative contributions of individual feature types for prediction, to elucidate its limitations, and to critically assess its relevance compared to standard techniques. We especially investigated the questions raised by Janela & Bajorath concerning the relevance of intricate deep learning methodologies [46, 53].

With the objective of unravelling how G-*PLIP* succeeds in capturing the signal for effective predictive power, we assessed partial versions of the model, utilizing combinatorial subsets of features. We started from the null model, a configuration encompassing protein-protein interactions and drug-drug similarities but without any specific node features. We then gradually added sets of protein or compound features to the null model. In each distinct configuration, we documented the optimal performance achieved, thereby facilitating an exploration of the informativeness of individual features and their contributions to enhance predictive capacity.

According to the results (Table 5) *k*mer counts is the most pivotal feature, yielding a substantial improvement of RMSE by 14.08% compared to the null model. Other two protein features, gene expression and sub-cellular localization do not significantly contribute to the prediction beyond the capability of *k*mers.

We made the interesting observation that protein node features emerge as more informative than their compound counterparts. Molecular fingerprints, as we had suspected, improved the prediction. We also anticipated the observation that neither the features derived from Lipinski rules, nor other chemical features commonly used to gauge physico-chemical properties of the molecules roughly, exhibited prominent informativeness since they reflect rather on the pharmacokinetic properties and drug-likeliness of the molecules rather than bioactivity.

Considering that *k*mer was the most prominent feature that we identified, we attempted to extend or substitute *k*mer counts with ESM2 [54], a state-of-the-art embedding of

proteins derived from evolutionary modelling. Yet we observed no discernible performance improvements S2. The simplicity and the good performance recommend *k*mer as an informative feature for any PLI prediction algorithm.

Another intriguing finding is that the model shows remarkably good performance even in the absence of protein-protein or compound-compound connectivity (Table 6). It's essential to clarify that in such models, despite the absence of connections, the graph structure is used implicitly. This is because encoder remains a key component of the model, and it is responsible for embedding the initial set of node features into a lower dimensional space. In addition, leaving out protein-protein and compound-compound connectivity leads to a small reduction in the performance. Nevertheless, we consider the observation interesting and worth reporting, because it underlies the power of GNNs to dissect relative contributions from node and edge features.

Table 4. G-*PLIP*'s performances compared over possible models.

Models	*RMSE*	*RBO*	R_p	R^2
Compounds-focused	0.956	0.661	0.377	0.122
Proteins-focused	0.914	0.598	0.444	0.161
only *k*Mer & Fingerprints	0.898	0.672	0.493	0.172
Complete	0.858	0.676	0.525	0.251

Discussions

Learning abilities of the model

Our structure-free approach compares favourably against advanced cutting-edge structure-aware models like MPNN-PLI, Pafnucy, K_{DEEP} or Fusion, as well as with novel structure-free models such as Yuel, when assessed in terms of RMSE.

Our evaluation highlights the importance of choosing an appropriate metric for benchmarking studies for the task of PLI prediction. As previously mentioned, the Pearson correlation coefficient can be misleading in evaluating regressive models, and we established that satisfactory scores can be achieved even by models incapable of predicting relevant values (Supplementary Figure S3). G-*PLIP*'s tendency to underestimate strong bioactivities likely explains its lower Pearson score in comparison to its peers.

GNNs offer a new approach to structure-free prediction of protein-ligand interaction. The knowledge graph structure inherent to our methodology allows the encapsulation

Table 5. G-*PLIP*'s performances on the Roche dataset with several sets of features activated. 'Null' uses the graph topology but neither protein nor compound features. Indicated are the sets of features used in addition to the null model: Protein features: (K) kMer counts; (GE) gene expression and (SL) for subcellular localization. Compound features: (F) fingerprints (*i.e.*, extended connectivity fingerprint [ECFP]), (L) for Lipinski feature,s and (Ch) for other chemical features. $-\Delta_{RMSE}$ represents the change in RMSE compared to the null model.

(a) Features informativeness

Model	*RMSE*	$-\Delta_{RMSE}$	*RBO*	R_p	R^2
Null	1.065	(*ref*)	0.596	0.061	-0.034
(K)	0.915	*+14.1%*	0.596	0.438	0.141
(GE)	0.948	*+11.0%*	0.595	0.398	0.125
(SL)	0.991	*+6.9%*	0.595	0.309	0.062
(K + GE)	0.911	*+14.5%*	0.596	0.448	0.151
(K + SL)	0.914	*+14.2%*	0.596	0.444	0.157
(GE + SL)	0.935	*+12.2%*	0.595	0.419	0.139
(K + GE + SL)	0.914	*+14.2%*	0.596	0.443	0.150
(F)	0.991	*+6.9%*	0.650	0.320	0.027
(L)	1.091	*-2.4%*	0.617	0.042	-0.177
(Ch)	1.028	*+3.5%*	0.648	0.235	0.025
(F + L)	1.017	*+4.5%*	0.650	0.300	0.039
(F + Ch)	1.004	*+5.7%*	0.654	0.308	0.027
(L + Ch)	1.022	*+4.0%*	0.648	0.257	0.052
(F + L + Ch)	1.011	*+5.1%*	0.650	0.301	0.045
Complete	0.858	*19.4%*	0.676	0.525	0.251

of informative protein and compound embeddings, obviating the need for an excessively complex model architecture. In this work, we only included a basic set of features in order to achieve a minimal and parsimonious model structure. Further work can elaborate based on our model to include more information thanks to the open and adaptive structure provided by GNNs.

The minimalistic architecture of G-*PLIP*, with only 8,449 trainable parameters, facilitates training on non-professional computing units without any high-end GPU and is robust to overfitting. As a reference, Yuel [48], the sole structure-free method tackling the same task we are aware of, comprises 1,411,523 trainable parameters.

Inference of the model

Remarkably, the model exhibits commendable learning capabilities even in the absence of protein-protein or compound-compound connectivity, with the encoder layers being used to encapsulate the original features within a low-dimensional space without any convolutional mechanisms or aggregation. Regardless of the PPI database employed,

Table 6. a)G-*PLIP*'performances with different graph topologies. 'w/o PPI' (*i.e.*, protein-protein interaction) doesn't take into account the protein-protein connectivity but uses the other available features and connections. 'w/o CompSim' (*i.e.*, compound similarity) doesn't use the compound-compound similarities but still integrates compound-compound connectivity (without weight on the edges). 'w/o CompLink' doesn't connect compounds at all but use normally protein-protein connections. Similarly, 'w/o PPI & CompLink' uses neither protein-protein nor compound-compound connectivities. Each setup normally retains all other features and connections, and $-\Delta_{RMSE}$ is displayed compared to the complete model. 'w/ DTIs' (*i.e.*, drug-target interaction) draws the drug-target interactions from the train set above every other connection mentioned, meaning the encoder won't be limited to the protein or compound subgraph anymore. b)G-*PLIP*'performances only using graph topologies. 'only PPI' utilizes only protein-protein connectivity. 'only CompSim' integrates compound-compound similarities. 'only PPI & CompSim' normally connects compounds and proteins. In all of these cases, other features are excluded, and $-\Delta_{RMSE}$ is reported in comparison to the null model.

(a) Graph informativeness, from the complete model

Model	*RMSE*	$-\Delta_{RMSE}$	*RBO*	R_p	R^2
Complete	0.858	(*ref*)	0.676	0.525	0.251
w/o PPI	0.857	*+0.1%*	0.686	0.534	0.262
w/o CompSim	0.883	*-2.9%*	0.677	0.509	0.209
w/o CompLink	0.870	*-1.4%*	0.679	0.521	0.223
w/o PPI & CompLink	0.883	*-2.9%*	0.685	0.517	0.212
w/ DTIs	0.854	*+0.5%*	0.693	0.541	0.269

(b) Graph informativeness, from the null model

Model	*RMSE*	$-\Delta_{RMSE}$	*RBO*	R_p	R^2
Null	1.065	(*ref*)	0.596	0.061	-0.034
only PPI	1.066	*-0.1%*	0.596	0.061	-0.042
only CompSim	1.065	*0.0%*	0.596	0.061	-0.034
only PPI & CompSim	1.065	*0.0%*	0.596	0.061	-0.034

whether the PPI depicts physically interacting pairs only or a mix of functional and physical interactions, the protein-protein connectivity seems to make a limited contribution to the prediction. These observations, coupled with measures of node feature informativeness, suggest that the information aggregated from neighbouring nodes is currently not being effectively utilized by the machine. One may wonder whether the current convolutional layers are suited to aggregate information from a biological knowledge graph. After all, previous reports and studies indicate that deep learning neural networks can encounter challenges when attempting to harness the learning potential of complex patterns they possess [46]. We think that the protein-protein interaction is only one facet within a broader spectrum of relationships that a graph can depict. Other elements such as structural similarities, folding resemblance, or evolutionary distances could be explored with the architecture that we proposed.

Moreover, our GNN approach demonstrates its suitability for the task compared to standard machine learning models (Supp Table S1c). We therefore establish the graph architecture

as an expressive framework capable of embedding a wide array of biological information. In another word, our work demonstrates that GNN is uniquely empowered to identify parsimonious models for a prediction task at hand. In line with this, we found that a restricted model using only *k*Mer and molecular fingerprints exhibits similar performances with, only 4.7% worse than, the complete model. While the trials with ESM2 did not yield improvements, standard protein and compound representations could still potentially be substituted by more advanced representation models (Table S2) in regard to embeddings' performances on molecule properties prediction [55].

Considering that *k*mer was the most prominent feature that we identified, we attempted to extend or substitute *k*mer counts with ESM2 [54], a state-of-the-art embedding of proteins derived from evolutionary modelling. Yet we observed no discernible performance improvements S2. The simplicity and the good performance recommend *k*mer as an informative feature for any PLI prediction algorithm.

Our observation regarding the irrelevance of an extension of the dataset using lower-quality data confirms what has been previously seen in previous studies [19, 51]. Integrating data from biased or untidy biochemical assays carries a risk, even when performances on the dataset split initially appear promising. This underscores the necessity of systematic cross-validation against an external reference test set (in our case, the CASF 2016 Core set, remaining constant over the experiments). Evaluations on traditional random or temporal split can extend the biases of the original set.

Limitations

Biologically, G-*PLIP*performed comparably well across molecular function classes (Figure 2). The findings partially reflect the historical focus and current practice of small-molecule drug discovery. The relatively better performance for kinases, ion channels, and G protein-coupled receptors (GPCRs), may derive from the facts that many small molecules are discovered with such proteins as targets [56], and that they also comprise a large body of secondary pharmacological targets [57]. On the other side, some transcription factors (DNA-binding) are known to be difficult targets for small molecules[58]. The relatively variable performance of our model for multiple drug classes such as transcription factors, human hydrolases, and human proteases calls for alternative modelling strategies.

Methodologically, despite clearly improved performances compared with recent models, G-*PLIP* persistently demonstrates a propensity to underestimate potent bioactivities (Figure 2), which remains critical for the identification of putative interactions to be explored by drug discovery projects.

Computationally, modelling PLI with a knowledge graph bears a significant memory cost. It may be mitigated by refactoring the data structure or by fine-tuning the compound similarity edges. These adjustments may improve memory efficiency and scalability.

Conclusion

Predicting protein-ligand interactions quantitatively represents an unsolved challenge at the core of drug discovery. Here we present a structure-free approach that exhibits enhanced predictive accuracy compared to existing methods. Further, we demonstrate that information from the biochemical space can be encapsulated and leveraged by a deep learning procedure. It only requires an amino-acid sequence of proteins and SMILES representation of ligands, which makes it easily applicable. Despite its simplicity, G-*PLIP* exhibits remarkable learning and predictive capabilities.We demonstrate that G-*PLIP* can efficiently be trained from a relatively small qualitative dataset and yet proficiently extrapolate on unseen interactions from learned parameters. Composed of only a few layers, the model is swiftly trainable on non-specialized platforms and its computational performances position it as a suitable candidate for large-scale retraining or deployment.

Methods

Architecture of a heterogeneous GNN

We constructed a graph structure featuring two distinct node types: ligands encompass drugs or drug-like molecules, and proteins encompass target proteins. The graph incorporates three types of links: protein-protein links embodying physical binary interactions between proteins, ligand-ligand links reflecting chemical similarity between compounds, and protein-ligand links representing bioactivities between ligands and protein targets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1a}$$

$$T_s(A,B) = \frac{\sum (X_i \bigwedge Y_i)}{\sum (X_i \bigvee Y_i)} \tag{1b}$$

1. The Jaccard and Tanimoto scores.

$$\mathbf{x}'_i = \mathbf{W}_1 \mathbf{x}_i + \mathbf{W}_2 \cdot \text{mean}_{j \in \mathcal{N}(i)} \mathbf{x}_j \tag{2}$$

2. The GraphSAGE operator, from Hamilton et al. [59]

Protein features

The encoder leverages three distinctive types of protein features, each contributing to a comprehensive understanding of protein characteristics:

- K-mer count: K-mer counting from protein sequences empowers G-*PLIP* with an understanding of protein characteristics. This entails tallying the occurrences of every substring of length k within a given sequence, yielding a commonly employed distance measure. [60] GNNs, with their message passing, facilitate the learning of similarities and traits among target proteins within the network. Given the relatively modest size of drug-like molecules, their binding sites should be narrowed and inspecting *3*-mers will give us a comprehensive overview of the catalytic site and its immediate vicinity. Here, we then use 3-mers as features, calculated from the protein sequence available on UniProt [61].

- Subcellular location: Subcellular location is complex to capture but still crucial as it sheds light on how the proteins interact with their surroundings. Proteins' roles within the cellular machinery can often differ based on their location, thereby influencing their properties. We use information pertaining to the mature protein's location and topology in the cell, extracted from UniProt. Subcellular locations are grouped into 31 main classes, and we employ one-hot encoding for representation.

- Gene expression: Abundance of proteins, or as a proxy, its encoding mRNAs which can be measured by diverse sequencing techniques such as RNA-seq. Gene expression patterns have been correlated with drug effects [62], and can be adapted depending on the cell or tissue type of interest. For each protein, we incorporated normalized Transcripts Per Million (nTPM) from RNA-seq data, averaged across the corresponding expressed genes. We used gene expression data from the Human Protein Atlas [63].

Compound features

Similarly, the encoder harnesses three distinct types of compound features, each providing a comprehensive grasp of molecular attributes:

- Lipinski Rule of Five: Aiming to evaluate druglikeness, this rule is widely used in pharmacology, assessing whether a compound possesses attributes making it suitable as an oral drug. Lipinski's Rule of Five [64] states that, in general, an orally active drug has:

  1. Not more than 5 hydrogen bond donors (OH and NH groups)
  2. Not more than 10 hydrogen bond acceptors (notably N and O)
  3. A molecular weight under 500 g/mol
  4. An octanol-water partition coefficient (log P) that does not exceed 5

  Rather than to record violations of the rule, which happen increasingly more for certain modern drug classes [65, 66] and often disregarded nowadays, we chose to input the numerical values of the descriptors proposed above, allowing the model to set its own customized rules form the Rule of Five parameters.

- Chemical features: In order to capture the main characteristics of a compound, we also extracted the following chemical features:

  1. Topological Polar Surface Area ($Å^2$)
  2. Number of rotatable bonds (#)
  3. Molecular Refractivity (MR) [67]
  4. Number of atoms (#)
  5. Number of aromatic rings (#)
  6. Number of aliphatic rings (#)
  7. Percentage of atoms belonging to aromatic rings (%)

- Fingerprints: We generated Morgan fingerprints [68] from the SMILES representation of compounds and folded them into a bit vector. We proposed by default extended-connectivity fingerprints (ECFP) with a diameter of 4, *i.e.*, ECFP4. ECFPs are circular topological fingerprints designed for molecular characterization or similarity searching, being widely employed as molecular descriptors for machine learning models [11]. Those identifiers represent substructures present within the molecule, by assigning integer identifiers to each non-hydrogen atom of the input molecule and iteratively updating them to capture larger and larger circular neighbourhoods. Those final integer identifiers can be converted to fixed-length bit strings, with the risk of collision and therefore loss of information increased as the bit string is shorter.

As neural networks can be quite sensitive to background noise, we encoded our fingerprints in a 256 bits vector whereas chemoinformaticians usually use encodings ranging from 1024 to 4096 bits [69, 70]. Nevertheless, we let this setting open to modifications, alongside the choice between ECFP and FCFP (Functional-Class Fingerprints [68]).

Interaction networks

By default, we extracted 15,749 protein-protein interactions (PPIs) between 939 human proteins from the PPI database (compiled by [47]). We linked all the ligands together to form a full graph, where ligand-ligand connections symbolize compound similarities quantified by Tanimoto scores (Equation 1b), a measure resembling the Jaccard Index (Equation 1a). The Jaccard index measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. Likewise, the Tanimoto score (also referred as Tanimoto similarity ratio) expresses a similarity ratio over Morgan fingerprint bitmaps, defined as the number of common bits divided by the bit length. Similarity values range from 0 to 1, where 1 indicates identical molecules (accounting for potential information loss during bit encoding). Our pipeline permits customization of ligand connectivity to consider only connections surpassing a specified Tanimoto score threshold. By default, we incorporated all compound-compound similarities (keeping the threshold to 0.0), giving a complete subgraph of compounds. It results in a graph with $C + P$ nodes and $\frac{C(C-1)}{2} + P \cdot \frac{k}{2}$ edges, with P the number of proteins, C the number of ligands and k the average degree of connectivity of proteins, defined as $\frac{2nb_{PPI}}{P}$ (since the graph is undirected) depending on the PPI database used. In the default configuration, the graph has two distinct connected components, thus we attempted to modelise the protein-ligand edges in one of the experiments (Table 6, reported as 'w/ DTIs').

Link predictions

The model adopts a non-linear multiconvolutional neural network with an encoder-decoder architecture: the encoder generates node embeddings *via* convolutional layers, and the decoder predicts bioactivities. The proposed GNN consists of two primary components: (a) an encoder producing node embeddings from the original graph, and (b) a decoder using embedding features to forecast drug-target activities. The encoder processes an input graph $\mathcal{G}$ and generates a d-dimensional embedding for each node, aiming to incorporate original node information. This encoder contains two distinct components: the first

encapsulates proteins and binary protein interactions, while the second handles ligands and compound-compound similarities.

Both components of the encoder consist of a singular convolutional layer, updating the node features within a latent dimension using the GraphSAGE operator (Equation 2) designed for low-dimensional embeddings of nodes in large graphs [59]. The GraphSAGE operator serves as learnable function that generates embeddings by aggregating information from a node's local neighbourhood. The decoder is composed of two linear layers, which employ concatenated embeddings of the protein target and compound molecule to predict bioactivity as a continuous value. A dropout layer is interposed between the two linear layers to prevent overfitting. Given that bioactivities are commonly expressed in logarithmic scale (pAct, defined as -log10(Ka)), we used the Rectified Linear Unit (ReLU) activation function to rectify any negative values within the layers that wouldn't align with biological meaning.

Prediction of drug-protein activity (pAct)

Losses calculation

G-*PLIP* is predicting drug-target bioactivity values (pAct), therefore prompting the use of Root Mean Square Error (RMSE) as the loss function for model training and backward propagation. We opted for a regression task, primarily due to the inherent subjectivity associated with any threshold value used to categorize ground truth data, and secondly because of the goal to emphasize possible off-target effects by directly comparing bioactivities. The model has been trained using the Adam optimizer [71] with a learning rate of 0.001 and a batch size of 512 over 500 epochs. It possesses only 8449 trainable parameters, due to its relatively simple architecture.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i - y_i\right)^2}{n}} \tag{3}$$

$$\mathrm{R_p} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}} \tag{4}$$

$$\mathrm{R}^2 = \frac{N \sum xy - \sum x \sum y}{\sqrt{\left[N \sum x^2 - \left(\sum x\right)^2\right]\left[N \sum y^2 - \left(\sum y\right)^2\right]}} \tag{5}$$

RMSE stands as the prevalent loss metric for regression models and is conventionally used to assess model performances, whereas we also reported several other metrics. Commonly

used, the Pearson Correlation Coefficient (R_p) gauges the linear correlation between the ground truths and the predicted values. Nonetheless, its exclusive focus on correlation can yield diverse interpretations and has been a subject of concern in regression evaluation [49]. Consequently, we also reported the Coefficient of Determination R^2.

Furthermore, for accurate identification of off-target activities, the ranking of predicted bioactivities in relation to one another holds more significance than the individual predicted values. In essence, for drug discovery purposes, pinpointing the most significant off-target effects of a molecular compound surpasses the importance of predicting exact bioactivity values. Identifying the most likely off-target effects of a drug candidate is crucial to understand the molecular mechanisms underlying drug efficacy and safety. Indeed, when a prominent off-target effect is predicted, further analysis can elucidate how the target protein is implicated in critical pathways relevant to the subject's well-being, and if it might hijack the expected molecular mechanism (e.g. if the initially focused target and the off-target are both membrane proteins, the molecule is highly likely to be sidetracked and not interact with the expected target). Conversely, identifying off-target effects also provides new insight into existing drugs, unveiling their potential to treat diseases for which they were not initially designed, such as the repurposing of the immunosuppressant mTOR inhibitor rapamycin [72]. To meet such demands, we proposed the use of new metrics, comparing the rankings rather than ranked values. We, therefore, introduced the use of Rank-Biased Overlap (RBO) [73] to evaluate G-*PLIP*'s proficiency to highlight accurate off-target effects.

$$\begin{aligned} \text{RBO} = (1-p)\sum_{d=1}^{D} p^{d-1} \cdot \frac{|S_{:d} \cap T_{:d}|}{d}, \\ \text{with } S \text{ and } T \text{ two rankings of length } D, \\ \text{and } p \text{ the steepness parameter.} \end{aligned} \tag{6}$$

Designed to measure similarity between nonconjoint, indefinite or incomplete ranking lists, RBO is tuneably top-weighted (parameter p) and provides a consistent way to handle tied ranks and prefix rankings of varying lengths.

Datasets

Roche internal dataset

The development of the model has been possible using Pathway-Annotated Chemical Ensemble (PACE), an internal extension of the published Roche SPARK library. PACE

incorporates diverse compound sets from both internal and external sources, serving various general and specialized purposes [74]. We collected 2972 PLI measures for 2057 different compounds in PACE. Previous studies claimed contradictory results about the extension of datasets by integrating low-quality affinity data to improve the performance of machine learning scoring models, some observed improvements [50] whereas recent studies suggested that it has no significant impact [19] or even that it is misleading for the model [51, 52]. This internal dataset, characterized by uniformity due to consistent protocols and handlers within a single institution, facilitated the development of our method, from the feature design to model architecture building and parameter tuning.
The concept of chemical scaffolds is an abstraction of the actual chemical structure, focusing on the topological and functional framework of a molecule. Originally described by Bemis & Murcko in 1996 [75], the concept has found wide application in medicinal chemistry and to date, many different variants of defining molecular scaffolds exist (Hu et al [76]). It is nowadays a standard approach in small-molecule drug discovery, where a pharmacophore or a scaffold is first identified based on available data (HTS, phenotypic or target-based screening, in silico molecular modelling, etc.), and from where a library of derivative compounds is probed to find putative compound showing optimum potency and selectivity as well as favourable ADMET profile (hit-to-lead and lead optimization of compound series).
In our context, we utilize the original Bemis & Murcko (BM) scaffolds, which divide compounds into R-groups, linkers, and ring systems and derive the framework (a scaffold) by removing all R-groups while retaining all ring systems and linkers between them. We use scaffolds to identify molecules that belong to the same chemotype and split into train and test sets that do not overlap in these terms. This approach ensures maximal dissimilarity of train and test sets and reflects a realistic drug discovery scenario where models are applied on previously unseen scaffolds.

We further evaluated G-*PLIP* on a publicly available dataset, aiming to benchmark its performance against competitive models.

The PDBBind dataset

The PDBBind database, curated by Wang *et al.*[77, 78], presents a comprehensive collection of qualitative binding affinity data for a comprehensive list of biomolecular complexes, enriched with detailed structural information. PDBBind serves as one of the gold standards used to assess structure-based PLI models, such as for the popular Comparative

Assessment of Scoring Functions (CASF) benchmark. PDBBind has originally been updated on an annual base to follow the growth of the Protein Data Bank (PDB), the pace of the updates has notwithstanding been slowed down the past years. The general set of PDBBind 2019 includes 17652 protein-ligand complexes, with 4852 of them being in the refined set which contains complexes with high-quality structural data. Within this database, the PDBBind core set aims at providing a relatively small set of high-quality protein-ligand complexes for evaluation and comparison of docking/scoring methods, and is notably used as the test set for the Comparative Assessment of Scoring Functions (CASF) benchmark. The last version of this core set, used for CASF-2016, contains 285 protein-ligands complexes from 57 target proteins.

Recurrence of samples between the Roche internal dataset and the PDBBind dataset has been evaluated both for the compounds and the compound-target bioactivities. The datasets are highly divergent, with a Jaccard index (Equation 1a) of 0.02 between available compounds (207 compounds) and of 0.004 between known compound-target bioactivities (41 samples).

Model inference

We iteratively compared multiple instances of G-*PLIP* investigate the impact of different features. Given the vast number of possible feature combinations, we restricted our experimentation on (a) informativeness of features in a built graph; (b) informativeness of subgraph connectivity with built features; (c) informativeness of subgraph connectivity without any features; (d) usage of known drug-target interactions within the graph itself; (e) usage of different interactomes; (f) usage of different gene expressions; (g) exclusive focus on proteins or on compounds; (h) relevance of a simplest representation. We reported the four metrics (RMSE, RBO, R_p and R^2) alongside the improvement, defined by the formula 7 expressing the improvement from the null to the complete model.

$$-\Delta_{RMSE}(model) = \frac{RMSE_{ref} - RMSE_{model}}{RMSE_{ref}} \tag{7}$$

To distinguish the learned biochemical insights of the model from any artificial knowledge obtained solely from graph topology, we defined a null model trained on randomly labelled edges, with bioactivities drawn from a uniform distribution between 3 and 11 (pAct values considered as extremes). To guard our lessons learned against any coincidence, we repeated this experience 20 times with different random seeds and reported the averaged

results as being representative of the null model.

We then tested G-*PLIP* on a dataset of randomly connected edges, with bioactivities shuffled alongside the known drug-target pairs from the existing data, and performed this experiment with the seed used for further tests relative to feature informativeness (Table 1). This procedure aimed to distinguish potential bias stemming from the dataset curation protocol from the model's actual learning ability.
Finally, we compared the results of G-*PLIP*'s architecture to a state-of-the-art machine learning model (LightGBM, one of the most recent gradient boosting decision tree frameworks [79]), both using the same input features to evaluate the gain of a graph neural network architecture because of concerns raised by previous studies ([16, 52, 80]) about the learning abilities of deep learning models for inferring binding affinities.

Code availability

A reference implementation of G-*PLIP* platform, based on PyTorch [81] and PyTorch Geometric [82], will be available at `https://github.com/simoncrouzet/g-plip` after acceptance of the manuscript for publication. With 16 cores of a Xeon-Gold CPU running at 2.1 GHz, the model takes approximately 2,670 minutes to train on the PDBBind *complete* dataset, leading to a carbon footprint of 1.84 kg.eCO^2.

References

1. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **9,** 646–652 (2002).
2. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. *Journal of Medicinal Chemistry* **53,** 5061–5084 (2010).
3. Nicolaou, C. A. & Brown, N. Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies* **10,** e427–e435 (2013).
4. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432,** 855–861 (2004).
5. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2,** 573–584 (2020).
6. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **19,** 353–364 (2020).
7. Algar, W. R., Hildebrandt, N., Vogel, S. S. & Medintz, I. L. FRET as a biomolecular research tool - understanding its potential while avoiding pitfalls. *Nature Methods* **16,** 815–829 (2019).
8. Huang, L. & Zhang, C. Microscale Thermophoresis (MST) to Detect the Interaction Between Purified Protein and Small Molecule. *Methods in Molecular Biology* **2213,** 187–193 (2021).
9. Kaur, U. *et al.* Proteome-Wide Structural Biology: An Emerging Field for the Structural Analysis of Proteins on the Proteomic Scale. *Journal of Proteome Research* **17,** 3614–3627 (2018).

10. Backus, K. M. *et al.* Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534,** 570–574 (2016).

11. Carracedo-Reboredo, P. *et al.* A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal* **19,** 4538–4558 (2021).

12. Qin, T., Zhu, Z., Wang, X. S., Xia, J. & Wu, S. Computational representations of protein–ligand interfaces for structure-based virtual screening. *Expert Opinion on Drug Discovery* **16,** 1175–1192 (2021).

13. Romano, J. D. & Tatonetti, N. P. Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives. *Frontiers in Genetics* **10** (2019).

14. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology. Computation Resources for Molecular Biology* **428,** 720–725 (2016).

15. Honorato, R. V. *et al.* Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Frontiers in Molecular Biosciences* **8** (2021).

16. Volkov, M. *et al.* On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry* **65,** 7946–7958 (2022).

17. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34,** 3666–3674 (2018).

18. Jones, D. *et al.* Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *Journal of Chemical Information and Modeling* **61,** 1583–1592 (2021).

19. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **58,** 287–296 (2018).

20. Sutherland, J. J., Yonchev, D., Fekete, A. & Urban, L. A preclinical secondary pharmacology resource illuminates target-adverse drug reaction associations of marketed drugs. *Nature Communications* **14,** 4323 (2023).

21. Di Lascio, E., Gerebtzoff, G. & Rodríguez-Pérez, R. Systematic Evaluation of Local and Global Machine Learning Models for the Prediction of ADME Properties. *Molecular Pharmaceutics* **20,** 1758–1767 (2023).

22. Melnikov, F., Anger, L. T. & Hasselgren, C. Toward Quantitative Models in Safety Assessment: A Case Study to Show Impact of Dose–Response Inference on hERG Inhibition Models. *International Journal of Molecular Sciences* **24,** 635 (2023).

23. Zhang, J. D. & Wiemann, S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25,** 1470–1471 (2009).

24. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* **16,** 949–959 (2021).

25. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nature Machine Intelligence* **3,** 1023–1032 (2021).

26. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology* **79,** 102548 (2023).

27. Schneuing, A. *et al. Structure-based Drug Design with Equivariant Diffusion Models* 2022.

28. Luo, S., Guan, J., Ma, J. & Peng, J. *A 3D Generative Model for Structure-Based Drug Design* in *35th Conference on Neural Information Processing Systems (NeurIPS)* (2021).

29. Peng, X. *et al. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets* in *Proceedings of the 39th International Conference on Machine Learning (ICML)* **162** (2022), 17644–17655.

30. Atz, K., Isert, C., Böcker, M. N., Jiménez-Luna, J. & Schneider, G. Δ-Quantum machine-learning for medicinal chemistry. *Physical Chemistry Chemical Physics* **24,** 10775–10783 (2022).

31. Isert, C., Kromann, J. C., Stiefl, N., Schneider, G. & Lewis, R. A. Machine Learning for Fast, Quantum Mechanics-Based Approximation of Drug Lipophilicity. *ACS Omega* **8,** 2046–2056 (2023).

32. Satorras, V. G., Hoogeboom, E. & Welling, M. E (n) equivariant graph neural networks. *Proceedings of the 38th International Conference on Machine Learning (ICML),* 9323–9332 (2021).

33. Liu, X., Luo, Y., Song, S. & Peng, J. Pre-training of Graph Neural Network for Modeling Effects of Mutations on Protein-Protein Binding Affinity. en. *PLOS Computational Biology* **17,** e1009284 (2021).

34. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* **34,** 18–42 (2017).

35. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. & Jaakkola, T. *EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction* in *Proceedings of the 39th International Conference on Machine Learning (ICML)* **162** (2022), 20503–20521.

36. Rudden, L. S. P., Hijazi, M. & Barth, P. Deep learning approaches for conformational flexibility and switching properties in protein design. *Frontiers in Molecular Biosciences* **9** (2022).

37. Krapp, L. F., Meireles, F. A., Abriata, L. A. & Dal Peraro, M. *Context-aware geometric deep learning for protein sequence design* 2023.

38. Mohseni Behbahani, Y., Crouzet, S., Laine, E. & Carbone, A. Deep Local Analysis evaluates protein docking conformations with locally oriented cubes. *Bioinformatics* **38,** 4505–4512 (2022).

39. Mohseni Behbahani, Y., Laine, E. & Carbone, A. Deep Local Analysis deconstructs protein–protein interfaces and accurately estimates binding affinity changes upon mutation. *Bioinformatics* **39,** i544–i552 (2023).

40. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17,** 184–192 (2020).

41. Zeng, M. *et al.* Protein-protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* **36,** 1114–1120 (2020).

42. Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **34,** i802–i810 (2018).

43. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications* **14,** 2175 (2023).

44. Tan, X. *et al.* Current Advances and Limitations of Deep Learning in Anticancer Drug Sensitivity Prediction. *Current Topics in Medicinal Chemistry* **20,** 1858–1867 (2020).

45. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **18,** 463–477 (2019).

46. Janela, T. & Bajorath, J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models. *Nature Machine Intelligence,* 1–10 (2022).

47. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34,** i457–i466 (2018).

48. Wang, J. & Dokholyan, N. V. Yuel: Improving the Generalizability of Structure-Free Compound–Protein Interaction Prediction. *Journal of Chemical Information and Modeling* **62,** 463–471 (2022).

49. Waldmann, P. On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction. *Frontiers in Genetics* **10** (2019).

50. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. J. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* **20,** 10947–10962 (2015).

51. Klarner, L., Reutlinger, M., Schindler, T., Deane, C. & Morris, G. *Bias in the Benchmark: Systematic experimental errors in bioactivity databases confound multi-task and meta-learning algorithms* in *2nd AI4Science Workshop at the 39th International Conference on Machine Learning (ICML),* (2022).

52. Yang, J., Shen, C. & Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Frontiers in Pharmacology* **11** (2020).

53. Isert, C., Atz, K., Riniker, S. & Schneider, G. *Exploring protein-ligand binding affinity prediction with electron density-based geometric deep learning* 2023.

54. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).

55. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4,** 1256–1264 (2022).

56. Dougall, I. G. & Unitt, J. in *The Practice of Medicinal Chemistry (Fourth Edition)* (eds Wermuth, C. G., Aldous, D., Raboisson, P. & Rognan, D.) 15–43 (2015).

57. Valentin, J.-P. *et al.* In vitro secondary pharmacological profiling: An IQ-DruSafe industry survey on current practices. *Journal of Pharmacological and Toxicological Methods* **93,** 7–14 (2018).

58. Samarasinghe, K. T. G. *et al.* Targeted degradation of transcription factors by TRAFTACs: TRAnscription Factor TArgeting Chimeras. *Cell Chemical Biology* **28,** 648–661.e5 (2021).

59. Hamilton, W. L., Ying, R. & Leskovec, J. *Inductive Representation Learning on Large Graphs* in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* (2017), 1025–1035.

60. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

61. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51,** D523–D531 (2023).

62. Evans, W. E. & Guy, R. K. Gene expression as a drug discovery tool. *Nature Genetics* **36,** 214–215 (2004).

63. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Science Advances* **7** (2021).

64. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **46,** 3–26 (2001).

65. Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *Journal of Medicinal Chemistry* **62,** 1701–1714 (2019).

66. Cui, Y. *et al.* A Bidirectional Permeability Assay for beyond Rule of 5 Compounds. *Pharmaceutics* **13,** 1146 (2021).

67. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **39,** 868–873 (1999).

68. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50,** 742–754 (2010).

69. Fassio, A. V. *et al.* Prioritizing Virtual Screening with Interpretable Interaction Fingerprints. *Journal of Chemical Information and Modeling* **62,** 4300–4318 (2022).

70. Wei, Y., Li, W., Du, T., Hong, Z. & Lin, J. Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method. *International Journal of Molecular Sciences* **20,** 3572 (2019).

71. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* (2014).

72. Palve, V., Liao, Y., Remsing Rix, L. L. & Rix, U. Turning liabilities into opportunities: Off-target based drug repurposing in cancer. *Seminars in Cancer Biology. Drug Repurposing for Cancer* **68,** 209–229 (2021).

73. Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* **28,** 20:1–20:38 (2010).

74. Roudnicky, F. *et al.* Inducers of the endothelial cell barrier identified through chemogenomic screening in genome-edited hPSC-endothelial cells. *Proceedings of the National Academy of Sciences* **117,** 19854–19865 (2020).

75. Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **39,** 2887–2893 (1996).

76. Hu, Y., Stumpfe, D. & Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *Journal of Medicinal Chemistry* **59,** 4062–4076 (2016).

77. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **47,** 2977–2980 (2004).

78. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry* **48,** 4111–4119 (2005).

79. Ke, G. *et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree* in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* (2017), 3149–3157.

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

81. Paszke, A. *et al.* in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019).

82. Fey, M. & Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric* in *Representation Learning on Graphs and Manifolds Workshop of the 7th International Conference on Learning Representations (ICLR)* (2019).

Acknowledgements

This work has been conducted under the sponsorship of the internship program "Think Tank in Innovation and Sustainability #IP2TIS" initiated by André Hoffman, vice-chairman of the board of Roche Holding Ltd. We thank Leo Klarner for valuable discussions, as well as to our colleagues from the Predictive Modeling and Data Analytics (PMDA) chapter of Pharmaceutical Sciences for their continuous support.

Author contributions

S.J.C. conceived the idea, wrote the code, performed the experiments and wrote the manuscript. A.M.L. helped with the code writing and performed some experiments. K.A. wrote the manuscript and helped the project guidance with insightful discussions. L.S-P. and A.T.M. helped with the project guidance and the preparation of the manuscript with insightful discussions. T.N. contributed to the data gathering and helped to review the manuscript. M.DP. participated in reviewing the manuscript. J.D.Z. originated the idea, guided the project and wrote the manuscript.

Conflicts of Interest

S.J.C., A.M.L., K.A., T.N., L.S-P., A.T.M and J.D.Z. have been or are employees of F. Hoffmann - La Roche Ltd.