J. Biol. Macromol., **19** (1), 5-10 (2019)

Review

Development of Practical Artificial Intelligence System for Drug Discovery and Its Application to Activity Prediction of Small Molecule Protein-Protein Interaction Modulators

Hirotsugu Komatsu*, Ken Ikeda, Takeshi Tanaka and Takao Matsuzaki

R&D and BD Division, Interprotein Corporation, SIC-2-205, 5-4-30 Nishi-hashimoto, Midori-ku, Sagamihara, 252-0131, Japan

Received May 13, 2019; Accepted May 23, 2019

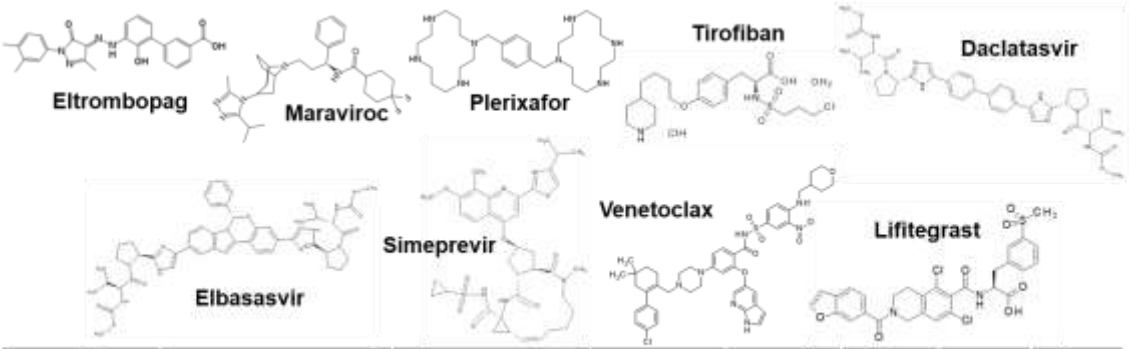
Degree of attention to protein-protein interaction (PPI) are increasing under circumstances of the exhaustion of conventional drug targets such as enzymes, receptors and channels whose ligands are non-protein. However, identification of potent small molecule PPI modulators are still challenging for pharmaceutical industries because in-house chemical libraries of pharmaceutical companies do not include strong and specific binders to PPI targets, and it is difficult to identify active compounds by high-throughput screening (HTS). On the other hand, to improve low productivity of drug research and developments (R&D), which is one of the most important long-standing issues for pharmaceutical industries, application of artificial intelligence (AI) to many stages of drug R&D is being tried by large number of pharmaceutical companies, academic institutions and biotech companies. However, a part of researchers of pharmaceutical companies also have opinions with a skeptic tone for the contribution of AI to the improvement of low productivity of drug R&D. Based on these situations, we review current approaches of application of AI to drug discovery researches and introduce Interprotein's approach to identify potent small molecule PPI modulators with an AI-based activity prediction system.

Key words: protein-protein interaction, *in silico* screening, artificial intelligence, activity prediction, ubiquitin-proteasome system

Introduction

In a couple of past decades, many challenging struggles have revealed that small molecules can act as effective protein-protein interaction (PPI) modulators although, prior to that, PPI modulation by small molecules had been considered to be substantially impossible. A part of potent and specific small molecule PPI modulators was evaluated for clinical availability and approved by regulatory agencies such as US Food and Drug Administration (FDA), European Medicines Agency (EMEA), and Pharmaceuticals and Medical Devices Agency (PMDA) of Japan. **Fig. 1** shows representative approved PPI modulators which exclude conventional PPI stabilizers such as FK506, rapamycin and paclitaxel. As an index for identification of potentially orally active small molecules with good pharmacokinetic profiles, the Lipinski's rule of five is widely used by many medicinal

*Correspondence author: Hirotsugu Komatsu
E-mail: komatsu@interprotein.com



Comp.	Target	Indication	Route	M. W.	LogP	HBD	HBA
Eltrombopag	TPO-R agonist	Chronic ITP	Oral	442.5	6.42	3	7
Maraviroc	CCR5 antagonist	HIV infection	Oral	513.7	3.48	1	7
Plerixafor	CXCR4 antagonist	Lymphoma/MM	S.C.	502.8	0.62	6	8
Tirofiban	gpIIb/IIIa antagonist	CV event prev.	I.V.	495.1	2.15	5	8
Daclatasvir	NS5A complex inhi.	HCV infection	Oral	738.9	5.33	4	10
Elbasvir	NS5A complex inhi.	HCV infection	Oral	882.0	7.14	4	12
Simeprevir	NS3/NS4 inhibitor	HCV infection	Oral	749.9	1.16	2	9
Venetoclax	Bcl2 inhibitor	CLL	Oral	866.4	6.74	3	11
Lifitegrast	LFA1 inhibitor	Dry eye	Eye drop	615.5	4.23	2	7

LogP, partition coefficient (a value indicating hydrophobicity of whole molecule); HBD, hydrogen bond donor; HBA, hydrogen bond acceptor; Underlined letters: in violation of the Lipinski's rule of five.

Fig. 1. Representative approved small molecule PPI modulators.

chemists. In this rule, it is recommended that molecular weight should be less than 500. Interestingly, however, molecular weights of 7 compounds out of 9 molecules listed in **Fig. 1** were more than 500, which suggests that small molecule PPI modulators have a need for higher molecular weight than conventional orally active small molecule drugs**.** Thus, the fact that many approved small molecule PPI modulators are in violation of the Lipinski's rule of five might indicate that regulation of PPIs by small molecules is a still challenging approach. On the other hand, to improve the efficiency and provability of drug discovery of small molecule PPI modulators, application of artificial intelligence (AI)-based technologies is expected. However, current AI-based technologies including activity prediction are mainly used for identification of non-PPI-targeted compounds. AI-based drug discovery technologies for activity prediction of small molecules are divided into structure-based and ligand-based approaches, and most of all AI-based drug discovery biotech companies seem to conduct the activity prediction by ligand-based approaches. In addition, many current ligand-based approaches are roughly categorized into an analysis of advanced quantitative structure-activity relationship (QSAR), but a part of researches involved in drug discovery have an impression that QSAR-based approaches do not well function in many cases for PPIs. Based on these situations, we tried to develop a practical AI system for drug discovery and apply it to identification of potent small molecule PPI modulators.

Outline of Interprotein's approach for AI-based drug discovery

One of the major problems for pharmaceutical companies is low productivity of drug research and development (R&D), and the purpose of application of AI to drug R&D is to improve the

Table 1. Differences between main current AI-based drug discovery technologies and AI-guided INTENDD from the aspect of activity prediction of small molecules.

Compared point	Main current technologies	AI-guided INTENDD
Main aim	Comprehensive support of selection of new drug candidates including toxicological and pharmacokinetic aspects	Focused on selection of highly active compounds for lead optimization
Approach for activity prediction	Ligand-based or mixture of ligand- and structure-based	Purely structure-based
Training data for deep learning	Partial structure information of ligands at 2D- or 3D-level	Atomic information at 3D-level (including unique hyper parameters produced by proprietary data preprocessing method)
Predicted activity classification	2 Classes	8 Classes
Main target	Non-PPIs	PPIs

productivity. The major cause of the low productivity is mainly failures in phase II studies, which is often called phase II attrition **[1]**, and the biggest reason for the failure is lack of efficacy **[2]**. These facts suggest that the most important factors in drug R&D are "validity of concept (including selection of drug target)" and "selection of the compound that shows efficacy in human at a high probability". Almost all biotech companies conducting AI-based drug discovery are comprehensively supporting selection of new drug candidates including toxicological and pharmacokinetic aspects. On the other hand, Interprotein is focusing on "prediction of activity for human proteins" based on the above-mentioned view point. On the basis of this concept, Interprotein established a new AI-implemented activity prediction system and named it AI-guided INTerprotein's Engine for New Drug Design (INTENDD). AI-guided INTENDD was constructed based on the accumulation of the many experiences and successful results obtained from the examination with INTENDD as a basal technology, which is a binding mechanism-based proprietary *in silico* screening system discriminated from conventional binding energy-based docking methodologies. **Table 1** shows the differences between main current AI-based drug discovery technologies and AI-guided INTENDD from the aspect of activity prediction of small molecules. Approach of AI-guided INTENDD is purely structure-based, which enables to newly identify active compounds without pre-reported active ligand information. Unique training data represent 3D- and atomic-level (but not partial structure-level) interactions between target proteins and compounds with known bioactivities. Such training data are produced by proprietary data-

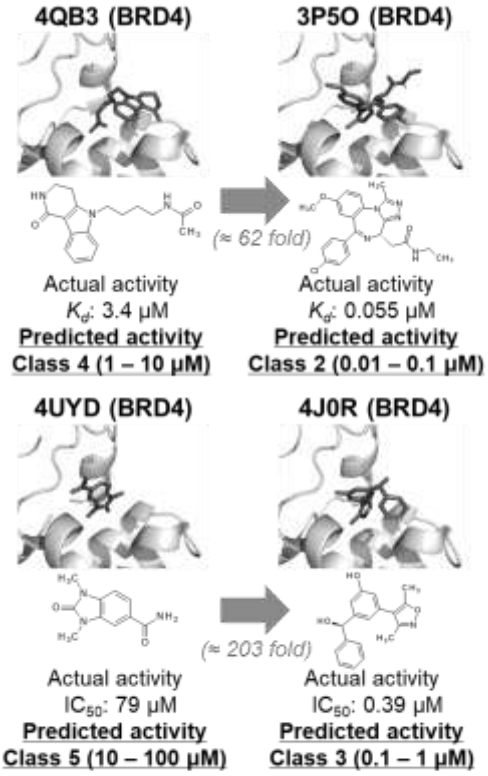


Fig. 2. Comparison between actual and predicted activities of BRD4 inhibitors.

preprocessing method. Predicted activity can be divided into 8 classes based on the order of active concentration. Furthermore, its specialty fields include challenging drug targets such as PPIs and ubiquitin-proteasome system-related proteins.

Application of AI-guided INTENDD to PPI targets

1) Confirmatory study with public information on co-crystal structure and activity

As an example of activity prediction of PPI inhibitors, we conducted a confirmatory study with public data on co-crystal structure (Protein Data Bank, PDB) and activity (PDBbind) of bromodomain-containing protein 4 (BRD4) inhibitors (**Fig. 2**). When we compared the activities of 4 compounds between actual PDBbind and AI-guided INTENDD prediction data, the both values showed a good consistency; e.g., while actual K_d value of the compound of 4QB3.pdb was 3.4 μmol/L, AI-guided INTENDD predicted that the activity was classified into "Class 4 (1 – 10 μmol/L)", indicating that the actual activity was included in the predicted class. Other 3 compounds also showed good correspondence between actual activities and predicted classes.

2) Activity prediction of compounds without co-crystal structure information

Although the results of the confirmative study described above were satisfactory, it is rare that co-crystal structure information has been clarified before activity prediction in the practical drug discovery researches. Therefore, we tried to predict activities of compounds without co-crystal structure information using small molecule inhibitors for Runx1, which is a transcription factor mainly related to pathophysiology of acute myelocytic leukemia (AML) and functions by association with CBFβ. When we classified activity of 142 compounds into 8 classes, a good consistence was observed between actual and predicted activities (**Fig. 3**). Actual inhibitory activities for the Runx1/CBFβ – DNA binding were assessed by a standard surface plasmon resonance (SPR) method and indicated on the horizontal axis, and predicted activities were indicated on the vertical axis. Dark grey-colored

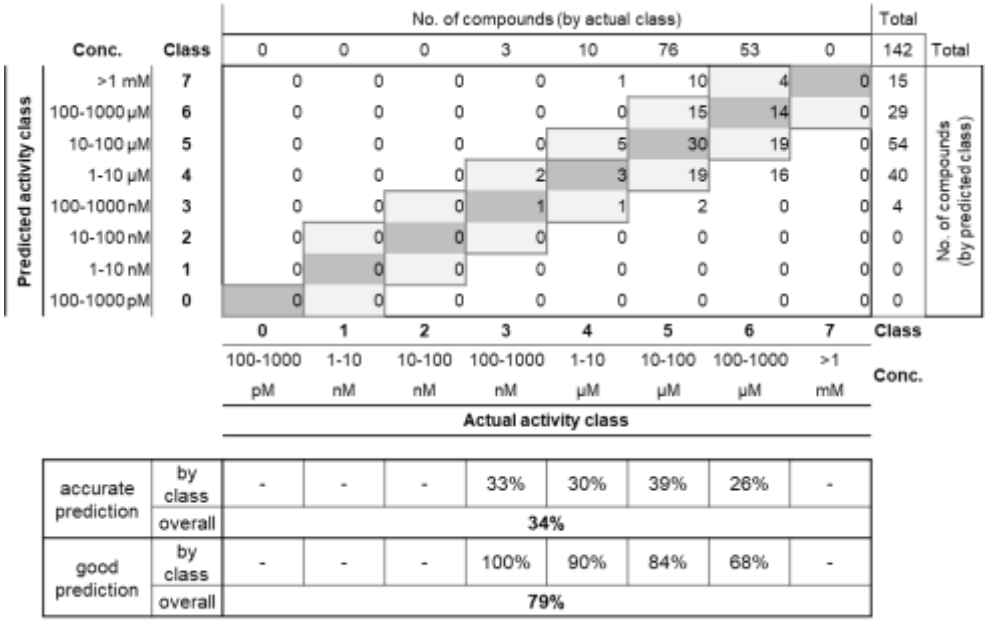


Fig. 3. Activity prediction of Runx1/CBFβ inhibitors by AI-guided INTENDD.

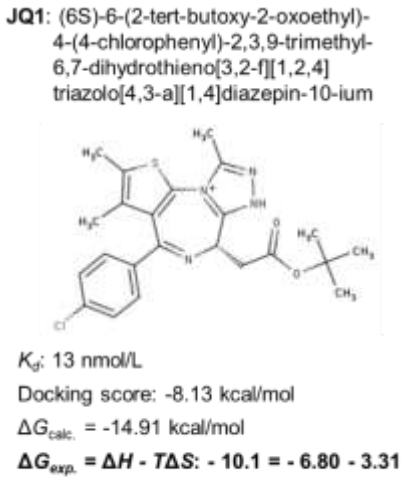


Fig. 4. Structure and property of JQ1.

cells show that actual activity is included in the predicted single activity class (defined as "Accurate Prediction") and flames consisting of dark gray- and light grey-colored cells (the range consisting of 3 classes of the predicted single activity class, one order more potent class and one order less potent class) were defined as "Good Prediction". The result showed a rising positive slope from bottom left to top right demonstrating an overall correspondence between actual and predicted activities. Overall rates of "Accurate Prediction" and "Good Prediction" were 34% and 79%, respectively. Importantly, AI-guided INTENDD gave high Good Prediction rates of more than 90% for relatively potent (Class 3 and 4) compounds.

3) Activity prediction of the compound with a good balance of enthalpy- and entropy-driven binding energy

One of the recent strategies for lead optimization is identification of the compound with a good balance of enthalpy and entropy components for binding free energy, and many drug discovery researchers consider that the compound with a certain ratio of enthalpy and enthalpy gains (but not losses) is favorable **[3]**. Based on this strategy, we are currently

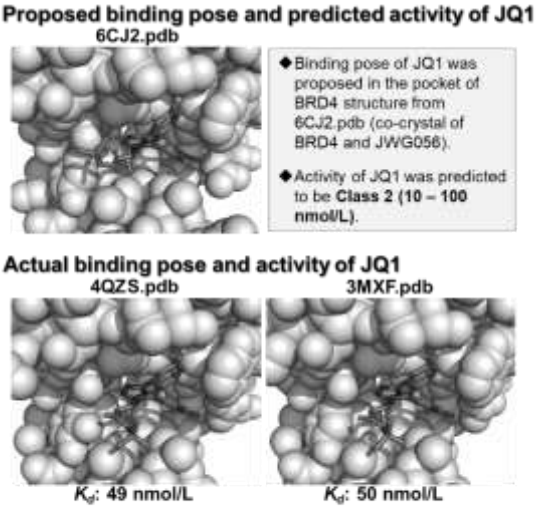


Fig. 5. Comparison of proposed binding pose and predicted activity with actual binding pose and activity of JQ1.

conducting many studies for hit identification with INTENDD, which is a basal *in silico* screening technology of AI-guided INTENDD and enables proposal of hit candidates with a good balance of enthalpic and entropic binding energy. Because input parameters of INTENDD are partly overlapping with those of AI-guided INTENDD, it is expected that AI-guided INTENDD can predict activity of the compound with a favorable balance of enthalpy and entropy.

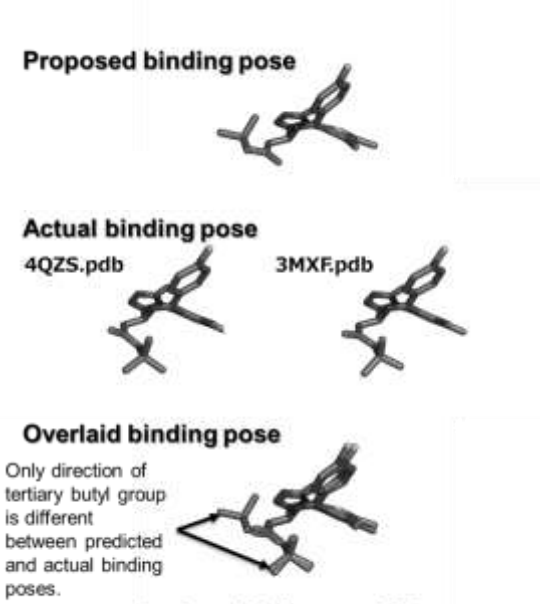


Fig. 6. Proposed and actual binding poses of JQ1 (overlaid structures).

Thus, we confirmed whether AI-guided INTENDD can give a precise activity prediction for a well characterized BRD4 inhibitor, JQ1, whose ΔH and $-T\Delta S$ values are -6.80 and -3.31, respectively **[4]**, and considered to be almost ideal (**Fig. 4**). Although co-crystal structures of JQ1 are reported as 4QZs.pdb and 3MXF.pdb, a binding pose of JQ1 was proposed in the pocket of other BRD4 structure from 6CJ2.pdb (co-crystal of BRD4 and JWG056).in a purposeful manner, and predicted activity was Class 2 (10 – 100 nmol/L) (**Fig. 5**). On the other hand, actual activities (K_d values) corresponding to 4QZS.pdb and 3MXF.pdb were 49 **[5]** and 50 nmol/L **[6]**, respectively. In addition, when we compared the proposed binding pose (on 6CJ2.pdb) with actual binding pose (4QZS.pdb and 3MXF.pdb), both binding poses showed a good consistence except only direction of tertiary butyl group is different between predicted and actual binding poses (**Fig. 6**). These results strongly suggest that AI-guided INTENDD enables activity prediction of the compounds with a favorable balance of enthalpic and entropic binding energy and contributes to efficient identification of optimized small molecule PPI modulators.

Future vision

Interprotein is currently conducting *in silico* screening from around 10 million compounds that we can readily purchase in Japan, and based on our experience, the potency of hit compounds seems to be limited to 100 nmol/L order in the best case. For promotion of the process for lead optimization, it is preferable to obtain more potent compounds as early as possible, but such compounds are rarely included in the commercially available real compound libraries. To solve this problem, we are currently trying to establish an AI system which proposes novel compounds by *de novo* drug design independently of structure and activity of existing compounds. Moreover, as another approach, advantage of virtual compound libraries is being taken. Reymond reported that the number of virtual compounds with selected existing partial structures was estimated to be 166.4 billion when the number of molecular-composing atoms was increased up to 17 types **[7]**. Interprotein's proprietary data-preprocessing method and resultant unique hyperparameters are expected to contribute to activity prediction of huge number of compounds and improvement of productivity of drug R&D.

References

[1] Patel, D. D., Antoni, C., Freedmen, S. J., Antoni, C., Freedman, S. J., Levesque, M. C., and Sundy, J. S. (2017) *J. Allergy Clin. Immunol.*, **140**, 685-687.

[2] Harrison, R. K. (2016) *Nat. Rev. Drug Discov.*, **15**, 817-818.

[3] Klebe, G. (2015) *Nat. Rev. Drug Discov.*, **14**, 95-110.

[4] Tanaka, M., Roberts, J. M., Seo, H.-S., Souza, A., Paulk., J., Scott, T. G., DeAngelo, S. L., Dhe-Paganon, S., and Bradner, J. E. (2016) *Nat. Chem. Biol.,* **12**, 1089–1096.

[5] Mishra, N. K., Urick, A. K., Ember, S. W. J., Schönbrunn, E., and Pomerantz, W. C. (2014) *ACS Chem. Biol.*, **9**, 2755-2760.

[6] Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W. B., Fedorov, O., Morse, E. M., Keates, T., Hickman, T. T., Felletar, I., Philpott, M., Munro, S., McKeown, M. R., Wang, Y., Christie, A. L., West, N., Cameron, M. J., Schwartz, B., Heightman, T. D., La Thangue, N., French, C. A., Wiest, O., Kung, A. L., Knapp, S., and Bradner, J. E. (2010) *Nature*, **468**, 1067-1073.

[7] Reymond, J.-L. (2015) *Acc. Chem. Res.*, **48**, 722−730.

Communicated by Sumio Ishijima