論文 / 著書情報
Article / Book Information

Title	Quantitative estimate of protein-protein interaction targeting drug-likeness
Authors	Takatsugu Kosugi, Masahito Ohue
Citation	In Proceedings of The 18th IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2021), , , pp. 1-8
Pub. date	2021, 10
Copyright	(c) 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
DOI	http://dx.doi.org/10.1109/CIBCB49929.2021.9562931
Note	This file is author (final) version.

Quantitative Estimate of Protein-Protein Interaction Targeting Drug-likeness

Takatsugu Kosugi
Department of Computer Science
School of Computing
Tokyo Institute of Technology
Kanagawa, Japan
kosugi@li.c.titech.ac.jp

Masahito Ohue
Department of Computer Science
School of Computing
Tokyo Institute of Technology
Kanagawa, Japan
ohue@c.titech.ac.jp

***Abstract*—The quantification of drug-likeness is very useful for screening drug candidates. The quantitative estimate of drug-likeness (QED) is the most commonly used quantitative drug efficacy assessment method proposed by Bickerton et al. However, QED is not considered suitable for screening compounds that target protein-protein interactions (PPI), which have garnered significant interest in recent years. Therefore, we developed a method called the quantitative estimate of protein-protein interaction targeting drug-likeness (QEPPI), specifically for early-stage screening of PPI-targeting compounds. QEPPI is an extension of the QED method for PPI-targeting drugs and developed using the QED concept, involving modeling physicochemical properties based on the information available on the drug. QEPPI models the physicochemical properties of compounds that have been reported in the literature to act on PPIs. Compounds in iPPI-DB, which comprises PPI inhibitors and stabilizers, and FDA-approved drugs were evaluated using QEPPI. The results showed that QEPPI is more suitable for the early screening of PPI-targeting compounds than QED. QEPPI was also considered an extended concept of "Rule-of-Four" (RO4), a PPI inhibitor index proposed by Morelli et al. We have been able to turn a discrete value indicator into a continuous value indicator. To compare the discriminatory performance of QEPPI and RO4, we evaluated their discriminatory performance using the datasets of PPI-target compounds and FDA-approved drugs using F-score and other indices. Results of the F-score of RO4 and QEPPI were 0.446 and 0.499, respectively. QEPPI demonstrated better performance and enabled quantification of drug-likeness for early-stage PPI drug discovery. Hence, it could be used as an initial filter for efficient screening of PPI-targeting compounds, which has been difficult in the past.**

***Index Terms*—Drug discovery, protein-protein interaction (PPI), drug-likeness filter, QED, QEPPI, PPI inhibitor**

I. Introduction

Protein-protein interactions (PPIs) have been garnering interest as drug targets since the early 2000s [1]–[5]. However, it is difficult to design drugs for PPIs based on conventional rules, such as Lipinski's rule of five (RO5) [6], [7] because their physicochemical characteristics are very different from those of conventional drug targets [8], [9]. In fact, only a few PPI inhibitors have been approved to date, and a few PPI-targeting drug candidates have advanced in clinical trials to subsequent phases [10]. Owing to this, it would be beneficial to develop an index that can be used to computationally select compounds that are likely to target PPIs.

Quantitative estimate of drug-likeness (QED), proposed in 2012 [11], is an index of drug-likeness modeled using the information available on marketed drugs and is widely used in current small-molecule drug discovery for computational methods [12], [13] and evaluation of drug-like properties [14]. QED index models drug-like properties using data available from 771 orally administered drugs already approved by the U.S. Food and Drug Administration (FDA). However, it is not an appropriate measure for PPI-targeting compounds, which require relatively large surface area of protein to interact with. Therefore, the development of new measures would be advantageous for PPI-targeting drugs [15].

QEX [16] and QEPT [17] are examples of QED remodeling. These methods are based on the concept of QED involving modeling physicochemical properties. In the case of QEX, the target compounds act on each target protein. In the case of QEPT, the target compounds are organic chemicals obtained from plant roots. These compounds represent quintessential successful models of their physicochemical properties.

The idea is to remodel a PPI-targeting drug based on an already approved PPI-targeting drug. However, many molecular optimizations need to be performed before approval. Even PPI-targeting compounds were optimized to have general characteristics of drugs, such as RO5 (low molecular weight, water solubility, etc.).

However, indices such as QED are mainly used in the early-stages of drug discovery, that is, seed compound discovery. The metrics modeled from PPI-targeting drugs already available in the market are idealistic and unsuitable for the early stage.

Therefore, in this study, we developed a method called QEPPI (**Q**uantitative **E**stimate of **P**rotein-**P**rotein **I**nteraction targeting drug-likeness), which is useful for early-stage PPI-targeting drug discovery, based on data from compounds that have undergone extensive PPI inhibition or stabilization experiments, rather than data from marketed PPI-targeting drugs.

978-1-6654-0112-8/21/$31.00

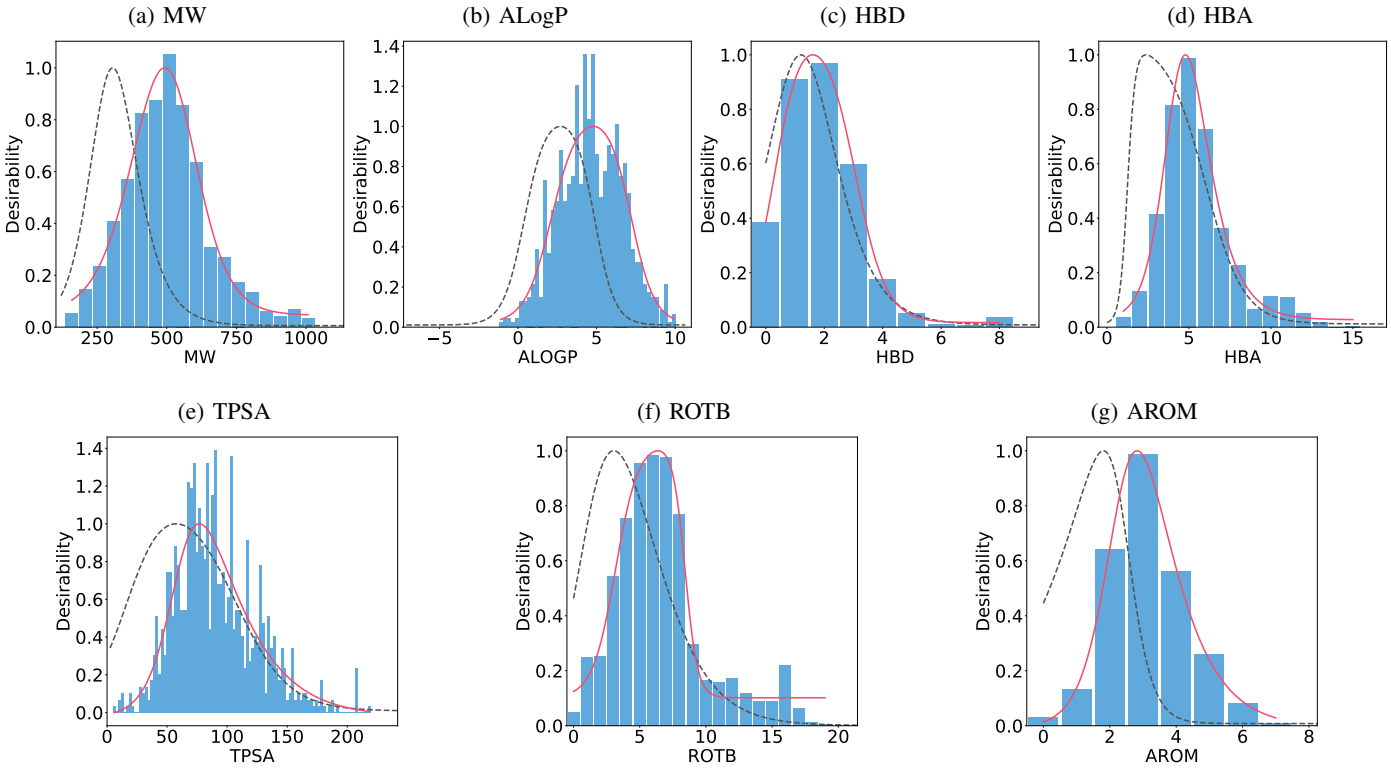

Fig. 1: Histograms of seven molecular physicochemical properties for a set of non-redundant compounds of iPPI-DB. (a)–(g), molecular weight (MW) (a), LogP value estimated by Ghose-Crippen method (ALogP) (b), number of hydrogen bond donors (HBD) (c), number of hydrogen bond acceptors (HBA) (d), Topological molecular polar surface area (TPSA) (e), number of rotatable bonds (ROTB) (f), and number of aromatic rings (AROM) (g). The solid red lines describe the ADS function (1) used to model the QEPPI histograms. The black dashed lines describe the ADS function used to model the QED histograms.

II. Results

A. Model building for QEPPI

QEPPI is an indicator for the early-stages of PPI drug discovery, and the prerequisites for the data set to model QEPPI are listed as follows:

- Not limited to those in the approval phase or marketed after approval, as various optimizations will be made during the approval phase.
- Not limited to PPI structures or complexes of protein and PPI-targeting compound with known structures.

The reason for these requirements is that a drug undergoes many molecular optimizations before it is approved. Therefore, if only the approved compounds are used as the data set for creating the model, too many ideal compounds will be determined unsuitable for this purpose. In addition, data on various candidate compounds are necessary for the initial stage of the search. According to a recent review, Shin *et al*. [15], if only compounds with known structures are selected for X-ray crystallography of proteins and ligands, the amount of data that can be handled will involve tens to hundreds of compounds, although more than the 720,000 human PPIs are known BioGrid [18] Current Build Statistics (4.3.196) - April 2021). Considering the number of known PPIs, we believe that the structural information on PPIs is still insufficient. We hypothesized that the information of target proteins and their ligand compounds would be sufficient without requiring three-dimensional structures.

Therefore, we used iPPI-DB [19], which was manually curated from the literature. In total, 2,361 PPI-targeting compounds are registered in this database (as of April 21, 2021), which are primarily derived from PPI inhibition or stabilization experiments. The number of compounds registered in DrugBank was 43, which is approximately 1.8% of the total. The quality and quantity of the data meet the requirements of the dataset for modeling QEPPI.

We built the QEPPI model using the data selected after clustering for non-redundancy (see Methods for details.). The histograms of the distributions of seven molecular physicochemical properties, namely, molecular weight (MW), LogP value estimated by Ghose-Crippen method (ALogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), topological molecular polar surface area (TPSA), number of rotatable bonds (ROTB), and number of aromatic rings (AROM) are shown in Fig. 1. The distribution peaks of each physicochemical property are demonstrated in Table I.

Fig. 1 and Table I show that oral drugs and PPI-targeting compounds have very different properties. Table I shows that

TABLE I: Distribution peaks of each molecular physicochemical property

	MW	ALogP	HBD	HBA	TPSA	ROTB	AROM
QED	305.8	2.70	1.20	2.38	57.5	3.04	1.8
QEPPI	492.7	4.78	1.61	4.79	76.9	6.37	2.8

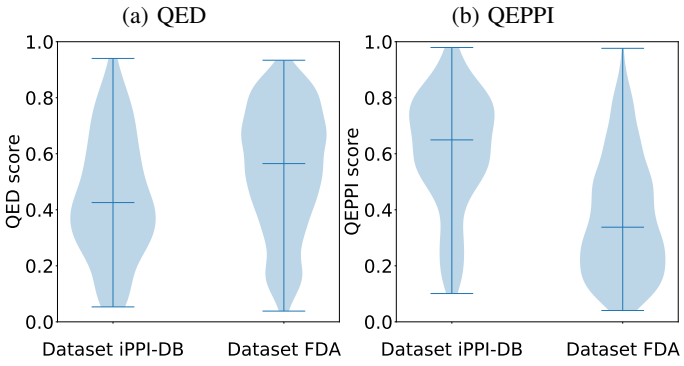

Fig. 2: Distribution of QED and QEPPI in PPI-targeting compounds Dataset and FDA-approved drug Dataset. Each filled area extends to represent the entire data range, with optional lines at the median. The QED score was calculated for both datasets (a). The QEPPI score was calculated for both datasets (b).

the peak values of all properties were higher for QEPPI than those for QED.

In particular, the major difference between QEPPI and QED is the peak value of ALogP (QEPPI: 4.78, QED: 2.70), suggesting that low lipophilicity and high hydrophilicity are important for oral drugs in terms of oral absorption. This suggests that QEPPI can capture PPI-targeting drug-like properties compared to QED and can play a different role in the seed compound discovery process, which is the early-stage of drug discovery.

B. Evaluation of QEPPI

To evaluate whether QEPPI, developed in this study, is a more useful indicator for early-stage PPI drug discovery than QED, we obtained data on 321 PPI-targeting compounds from the iPPI-DB that were not used for model building (Dataset iPPI-DB). In addition, we obtained data on 1,609 FDA-approved drugs, excluding duplicates (Dataset FDA). First, the QED score was calculated using these data, and the distribution of these values is shown in Fig. 2(a). Similarly, the QEPPI score was calculated, and the distribution of the values is shown in Fig. 2(b).

Fig. 2(a) shows that PPI-targeting compounds exhibit a lower distribution of QED scores than conventional drugs, suggesting that QED is not an appropriate measure for PPI-targeting compounds as it typically represents oral drug-like properties rather than drug-likeness. Fig. 2(b) shows that PPI-targeting compounds exhibit a higher distribution of QEPPI scores than conventional drugs, and a QEPPI threshold of 0.5 is sufficient to identify approximately 75% of PPI-targeting

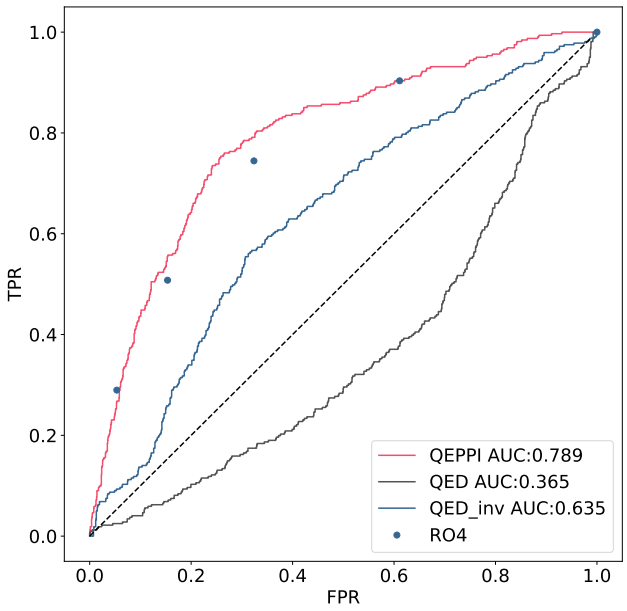

Fig. 3: Comparison of QEPPI with other measures of drug-like properties in ROC curves. All ROC curves show that the true positive rate against the false positive rate describes the differences in performance for classifying compounds as PPI-targeting compounds. The red, black, and blue lines represent the ROC curves for QEPPI, QED, and $1 - \mathrm{QED}$ (QED_inv), respectively. The five blue dots are plotted as points that allowed 0 to 4 violations of RO4. The dashed black line represents a random prediction of the dataset.

compounds. Furthermore, there are few PPI-targeted compounds in the FDA dataset, so the smaller QEPPI values in the FDA dataset compared to the iPPI dataset are consistent. To evaluate the quantitative performance of QEPPI and QED in identifying PPI-targeting compounds, we calculated the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). The true positive rate (TPR) and false-positive rate (FPR) were calculated to plot the ROC curve. Fig. 3 shows the ROC curves obtained from QEPPI, QED, and the value of $1 - \mathrm{QED}$. For QED, the AUC was less than 0.5 (0.365). This result was worse than that of randomly selected compounds. This is consistent with the results of Fig. 2(a), which shows that the AUC of QEPPI (0.789) is higher than that of $1 - \mathrm{QED}$ (0.635), indicating that QEPPI clearly performs better than QED ($1 - \mathrm{QED}$) in identifying whether a compound is likely to be a PPI-targeting compound.

C. QEPPI extends the Rule-of-Four

Morelli *et al*. proposed the "Rule-of-Four" (RO4) to evaluate PPI inhibitors [8]. This proposal was based on a statistical analysis of 39 PPI inhibitors in 2P2Idb [21] (currently, 2P2Idb is not accessible. Thus, we cannot use data from the 2P2Idb). They calculated the general characteristics of the chemical space in which PPI inhibitors differ from FDA-approved drugs.

TABLE II: Confusion matrix based on RO4 with one violation.

	passed	failed
positive	163	158
negative	247	1,362

TABLE III: Confusion matrix based on QEPPI scores with the threshold value of 0.5196.

	passed	failed
positive	236	85
negative	389	1,220

As a result, RO4 consisted of the following four criteria for physicochemical properties:

- MW must be higher than 400
- ALogP should be higher than 4.
- HBA should be higher than 4.
- the number of rings (RING) should be higher than 4.

Fig. 3 shows that the ROC curve of QEPPI and each point of RO4 are very close to each other. The result suggested that QEPPI is a general extension of the RO4 concept.

The threshold value of QEPPI can be adjusted. We calculated the threshold value of QEPPI (QEPPI score higher than 0.5196) such that the F-score was maximized. We then used the Dataset iPPI-DB as a positive sample and the Dataset FDA as a negative sample to compare the discriminative performance of RO4 allowed one violation and QEPPI. The confusion matrix and F-score results for RO4 and QEPPI are described in Tables II, III, and IV.

Table IV shows that the F-score of QEPPI is 0.499 and the F-score of RO4 is 0.446. Indicating that QEPPI performs better than RO4.

Finally, in order to compare the classification performance of two different metrics, namely, RO4 (rule-based) and QEPPI (threshold-based), we compared the value of recall between the same value of precision and the value of precision between the same value of recall. The Precision-Recall curve is shown in Fig. 4. Since RO4 is rule-based, we plotted the curves for all violations from one to four. As a result, each point of RO4, although not all RO4 points, is plotted on the lower side of the Precision-Recall curve of QEPPI.

III. Discussion

A. The advantage of QEPPI

Theoretically, we represent the ideal values for each physicochemical property that is characteristic of that dataset. This is because the frequency of compounds with that property was the highest in the dataset. Therefore, these properties are expected to reflect the nature of the target proteins. Furthermore, since QED is modeled using FDA-approved oral drugs, it is expected to reflect absorption, distribution, metabolism, excretion, and toxicity (ADMET). In contrast, the dataset used for QEPPI involves many PPI-targeting compounds and does not involve any optimizations. Hence, the peak values for all physicochemical properties were higher for QEPPI than those for QED.

TABLE IV: Precision, Recall, and F-score values for one violation of RO4 and QEPPI score with the threshold value of 0.5196.

	Precision	Recall	F-score
RO4	0.398	0.508	0.446
QEPPI	0.378	0.735	0.499

TABLE V: RO4 violations in the dataset used for QEPPI modeling.

	MW	ALOGP	HBA	RING
violation	243	378	347	532
no violation	764	629	660	475
violation rate	0.241	0.375	0.345	0.528

The advantage of QEPPI is that it allows model building using only target data. It does not require appropriate negative samples. The performance of machine learning classifiers is poor in problem settings where positive and negative samples are imbalanced [20]. Therefore, QEPPI may be more effective than machine learning models in conditions where appropriate negative samples are difficult to obtain from public databases.

RO4 is rule-based; it is basically impossible to adjust certain threshold values. However, the threshold values of QEPPI developed in this study can be adjusted such that the desired sensitivity and specificity are achieved.

QEPPI indices are primarily intended to be used in the early-stage of PPI drug discovery, the seed compound discovery stage. Hence, better discrimination performance is desirable. Fig. 4 shows that QEPPI has a higher Precision at the same Recall and a higher Recall at the same Precision than RO4 with one violation, two, and four violations of RO4. Comparing Precision Recall AUC, QEPPI was 0.422, QED, a measure of oral drug-like properties, was 0.134, and QED_inv, a measure of 1-QED, was 0.238, indicating that among these measures, QEPPI could identify PPI target compounds most accurately. The rules of RO4 are based on only 39 PPI inhibitors, and as with RO5, the hard cutoff for each physicochemical property is debatable. For example, a molecular weight of 401 is a pass, whereas 399 is a violation. In fact, Table I shows that the peak value for MW is approximately 500 and the peak values for ALogP and HBA are slightly higher than 4. This means that many compounds would demonstrate violation of the RO4 criteria. For the 1007 PPI target compounds used in the QEPPI model, the result of calculating whether each physicochemical property used in RO4 violates the four criteria is shown in Table V. Table V shows that violation percentage of WM, ALOP, and HBA are 24.1%, 37.5%, and 34.5%, respectively. Especially for RING, which is a physicochemical property were used only for RO4, more than 50% compounds demonstrated violation of this property.

The aforementioned results suggest that QEPPI is more useful and suitable than the conventional drug discovery indices QED and RO4, a proposal for the index of PPI-targeting compounds, in designing a useful index for the early detection of PPI drug.

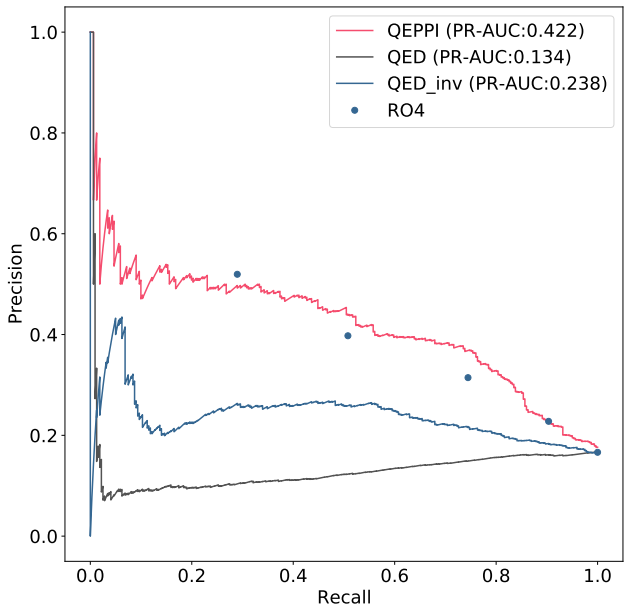

Fig. 4: Comparison of QEPPI with RO4 in Precision-Recall curve. The Precision-Recall curve shows that precision against recall value, which describes the differences in performance for classifying compounds as PPI-target compounds. The red, black, and blue lines represent the ROC curves for QEPPI, QED, and $1 - \mathrm{QED}$ (QED_inv), respectively. The five blue dots represent the points that allowed 0 to 4 violations of RO4, respectively.

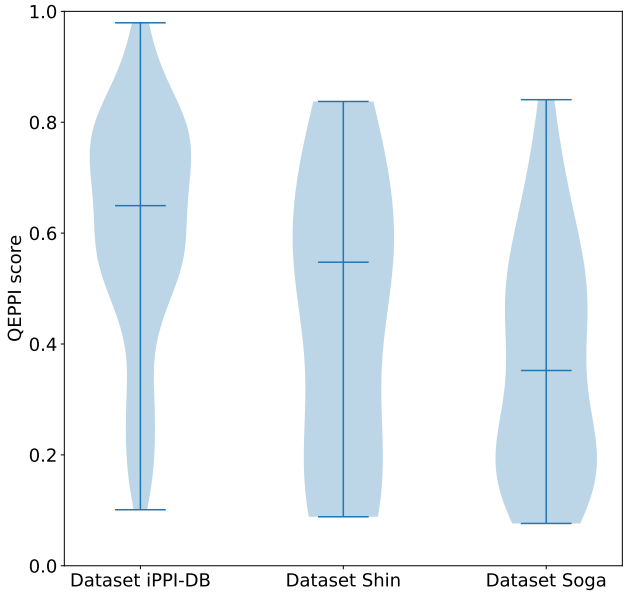

Fig. 5: Distribution of QEPPI with respect to marketed drugs or compounds in the clinical phase PPI-targeting compounds Dataset (Dataset Shin), non-redundant small molecule Dataset (Dataset Soga) and Dataset iPPI-DB for control. Each filled area extends to represent the entire data range, with optional lines at the median. The QEPPI score was calculated for all the datasets.

B. Application of QEPPI to PPI-targeting compounds and other small molecule drugs in clinical trials

In 2020, Shin *et al*. reported a review of PPI-targeting drug designs. We applied QEPPI to two datasets in this review [15]. The datasets are simply described as the PPI-targeting compounds dataset with X-ray crystallography results (Dataset Shin) and the non-PPI dataset used in the review (Dataset Soga) (See Methods for details) [22].

We also applied QEPPI to the data set. The distribution of the QEPPI is shown in Fig. 5. Fig. 5 shows that the distribution of QEPPI is higher for the Dataset Shin than the Dataset Soga. However, the Dataset Shin has a relatively lower distribution of QEPPIs compared to the Dataset iPPI-DB, a dataset of PPI-targeting compounds. This may be since the Dataset Shin is comprises marketed drugs or compounds in the clinical phase and has already been subjected to various optimizations. The results suggest that QEPPI can function effectively as a filter in the early-stages of drug discovery, which was its original purpose. However, it is less effective as a restriction for compounds in the market or clinical stages.

In this study, some PPI-targeting compounds with low QEPPI values showed small molecular weights compared to the peak. A previous study showed that the size and complexity of the binding interface of PPIs varied depending on the target. If the interface was relatively less complex and small, some PPI-targeting compounds with relatively small molecular weights can sufficiently block the binding interface. When the binding interface is relatively more complex, the binding interface tends to be wide, and only a PPI-targeting compound with a large molecular weight can sufficiently block the binding interface [23]. Evaluating the average PPI-targeting compounds using the iPPI-DB as a dataset of various types of PPI-targeting compounds would be advantageous. In future studies, we may design indices that are more specific to PPI-targeting compounds, such as the size of the binding interface or the PPI family. This is similar to the proposal of QEX. The approach will become feasible as more data are deposited in the database.

IV. Conclusion

QEPPI is based on the concept of QED, which models the physicochemical properties of a target compound and can quantify the PPI-targeting drug-likeness of interest compounds by using the PPI inhibitors and stabilizers as the target compound. The metric is proposed to be used in the early detection of PPI drugs.

RO4 was proposed as a rule-based approach with respect to a statistical analysis of the physicochemical characteristics of actual PPI inhibitors. QED is also based on the distribution data of the physicochemical properties of oral drugs and has garnered significant interest in early-stage drug discovery. However, it is not suitable for early-stage screening of PPI-

TABLE VI: The RDKit functions used to calculate the molecular properties used in QEPPI and RO4

property	RDKit function
MW	`Chem.rdMolDescriptors.CalcExactMolWt`
ALogP	`Chem.Crippen.MolLogP`
HBD	`Chem.rdMolDescriptors.CalcNumHBD`
HBA	`Chem.rdMolDescriptors.CalcNumHBA`
TPSA	`Chem.rdMolDescriptors.CalcTPSA`
ROTB	`Chem.rdMolDescriptors.CalcNumRotatableBonds`
AROM	`Chem.rdMolDescriptors.CalcNumAromaticRings`
RING	`Chem.rdMolDescriptors.CalcNumRings`

targeting compounds because the physicochemical properties of PPI-targeting compounds differ significantly from those of oral drugs. In addition, compared to the rule-based approach of RO4, QEPPI is based on the basic distribution data of physicochemical properties of more PPI-targeting compounds. Unlike rule-based indices, when many parameters of physicochemical properties are ideal, certain unfavorable parameters of properties may still be acceptable, making it an extremely useful indicator specifically for early-stage screening of compounds targeting PPIs.

We expect that QEPPI will lead to the development of PPI-based drugs along with consequent improvements in the accuracy of QEPPI as more PPI-targeting compounds are registered in the database.

V. Materials and Methods

A. Calculation of QEPPI

QEPPI was calculated using essentially the same procedure as that of the original QED, except that it was modeled using compounds curated in the iPPI-DB. We did not use 'ALERTS' among the physicochemical properties. The algorithms used are described below: In the first modeling step, RDKit (2020.09.1) was used to calculate seven molecular physicochemical properties: molecular weight (MW), LogP value estimated by the Wildman-Crippen method [24] (ALogP), the number of hydrogen bond donors (HBD), the number of hydrogen bond acceptors (HBA), topological molecular polar surface area (TPSA), the number of rotatable bonds (ROTB), and the number of aromatic rings (AROM). Table VI lists the RDKit functions used to calculate these properties.

Then, a histogram of each property was created and fitted to the asymmetric double sigmoid (ADS) function $Q(x)$ shown in (1) by implementing the Levenberg-Marquardt algorithm in SciPy (version 1.6.1).

$$Q(x) = a + \frac{b}{1+\exp\left(-\frac{x-c+\frac{d}{2}}{e}\right)}\left[1 - \frac{b}{1+\exp\left(-\frac{x-c-\frac{d}{2}}{f}\right)}\right] \tag{1}$$

All fitting functions ($Q_{\mathrm{MW}}(x)$, $Q_{\mathrm{ALogP}}(x)$, $Q_{\mathrm{HBD}}(x)$, $Q_{\mathrm{HBA}}(x)$, $Q_{\mathrm{TPSA}}(x)$, $Q_{\mathrm{ROTB}}(x)$, and $Q_{\mathrm{AROM}}(x)$) were divided by the maximum value and normalized to a maximum value of 1. The normalized function $\tilde{Q}_i(x)(i \in \{\mathrm{MW}, \mathrm{ALogP}, \mathrm{HBD}, \mathrm{HBA}, \mathrm{TPSA}, \mathrm{ROTB}, \mathrm{AROM}\})$ was used as the desirability function. Finally, the QEPPI score of compound k was assigned as the weighted geometric mean of all desirability functions [25], as shown in (2).

$$\mathrm{QEPPI}_k = \exp\left(\frac{\sum_i w_i \ln\left(\tilde{Q}_i\right)}{\sum_i w_i}\right) \tag{2}$$

The seven weights were thoroughly tested from 0 to 1 in increments of 0.25, and the average of the 1,000 combinations of weights that resulted in the highest Shannon entropy was adopted. The Shannon entropy of the model was calculated as shown in (3). Where n represents the number of compounds used in the modeling.

$$\mathrm{entropy} = -\sum_{k=1}^{n} \mathrm{QEPPI}_k \log_2 \mathrm{QEPPI}_k \tag{3}$$

B. Calculation of QED and RO4

To evaluate the filtering performance of QEPPI, QED and RO4 were calculated and used as a comparison. The QED score was calculated using the `Chem.QED.qed` method of RDKit. RO4 is calculated from four properties, MW, ALogP, HBA, and the number of ring structures (RING). MW, ALogP, and HBA were calculated using the same methods as QEPPI (Table V). RING was calculated using the `Chem.rdMolDescriptors.CalcNumRings` method of RDKit.

C. Dataset

To create a non-redundant dataset for the QEPPI model, we downloaded 2,361 SMILES and other data of compounds registered in iPPI-DB, and 1,007 compounds were selected from all clusters one by one with the best activities determined by clustering with Bemis-Murcko atomic frameworks [26].

As a dataset for the evaluation of QEPPI, 321 compounds were selected from all clusters one by one with the best activities from all clusters of compounds that were not used for model building (Dataset iPPI-DB).

As a dataset for small molecule compounds, we obtained SMILES and other data of compounds called "DrugBank FDA only" compounds from the catalog of ZINC [27] and removed duplicates by InChI, resulting in 1,609 compounds (Dataset FDA).

As a dataset for PPI-targeting compounds in the approval or clinical trial stages, 14 PDB IDs and other data pertaining to small-molecule compounds were obtained from [15] (Dataset Shin).

As a dataset of non-PPI ligands, which involved known non-redundant protein-ligand complexes evaluated via X-ray crystallography, we obtained 40 PDB IDs of single-molecule ligands obtained from [22] (Dataset Soga).

For datasets for which only PDB IDs were available, the IDs were converted to SMILES using PDB's GraphQL-based API [28].

D. Performance measures

With the "Dataset iPPI-DB" as the positive dataset and the "Dataset FDA" as the negative dataset, samples scored above a certain threshold by QEPPI or QED scoring were predicted to be positive, and samples scored below the threshold were predicted to be negative. The performance measures used are shown in (4) and (5).

$$\mathrm{TPR} = \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5}$$

where TP, FP, FN, TN, TPR, and FPR are the number of true positives, false positives, false negatives, true negatives, and true-positive and false-positive ratios, respectively.

Furthermore, the F-score shown in (6) was used to evaluate the discrimination performance, and the Precision shown in (7) was used for the Precision-Recall curve (Recall already shown in (4)).

$$\text{F-score} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

To evaluate the ROC and Precision-Recall curves, the QEPPI threshold was calculated based on all the QEPPI values in the data set, TP, FP, FN, and TN. In Table III, the threshold was calculated, where the F-score was maximized.

ACKNOWLEDGMENT

This work was partially supported by KAKENHI (20H04280) from the Japan Society for the Promotion of Science (JSPS), ACT-X (JPMJAX20A3) from the Japan Science and Technology Agency (JST), and Mizuho Foundation for the Promotion of Sciences. The authors thank Editage (www.editage.com) for English language editing.

REFERENCES

[1] P. L. Toogood. "Inhibition of protein–protein association by small molecules: Approaches and progress," *J. Med. Chem.*, vol. 45, no. 8, pp. 1543–1558, March 2002. doi:10.1021/jm010468s.

[2] M. R. Arkin, and J. A. Wells. "Small-molecule inhibitors of protein–protein interactions: progressing towards the dream," *Nat. Rev. Drug Discov.*, vol. 3, no. 4, pp. 301–317, April 2004. doi:10.1038/nrd1343.

[3] K. K. Dev. "Making protein interactions druggable: targeting PDZ domains," *Nat. Rev. Drug Discov.*, vol. 3, no. 12, pp. 1047–1056, December 2004. doi:10.1038/nrd1578.

[4] L. Jin, W. Wang, and G. Fang. "Targeting protein–protein interaction by small molecules," *Annu. Rev. Pharmacol. Toxicol.*, vol. 54, no. 1, pp. 435–456, October 2014. doi:10.1146/annurev-pharmtox-011613-140028.

[5] A. A. Ivanov, F. R. Khuri, and H. Fu. "Targeting protein–protein interactions as an anticancer strategy," *Trends Pharmacol. Sci.*, vol. 34, no. 7, pp. 393–400, July 2013. doi:10.1016/j.tips.2013.04.007.

[6] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.*, vol. 23, issues 1–3, pp. 3–25, January 1997. doi:10.1016/S0169-409X(96)00423-1.

[7] C. A. Lipinski. "Lead- and drug-like compounds: the rule-of-five revolution," *Drug Discov. Today Technol.*, vol. 1, no. 4, pp. 337–341, December 2004. doi:10.1016/j.ddtec.2004.11.007.

[8] X. Morelli, R. Bourgeas, and P. Roche. "Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I)," *Curr. Opin. Chem. Biol.*, vol. 15, no. 4, pp. 475–481, August 2011. doi:10.1016/j.cbpa.2011.05.024.

[9] O. Sperandio, C. H. Reynès, A.-C. Camproux, and B. O. Villoutreix. "Rationalizing the chemical space of protein–protein interaction inhibitors," *Drug Discov. Today*, vol. 15, no. 5–6, pp. 220–229, March 2010. doi:10.1016/j.drudis.2009.11.007.

[10] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson. "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, January 2018. doi:10.1093/nar/gkx1037.

[11] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins. "Quantifying the chemical beauty of drugs," *Nat. Chem.*, vol. 4, no. 2, pp. 90–98, February 2012. doi:10.1038/nchem.1243.

[12] N. De Cao, and T. Kipf. "MolGAN: An implicit generative model for small molecular graphs," *ICML'18 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, pp. 1–11, 2018.

[13] E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, and A. Zhavoronkov. "Reinforced adversarial neural computer for *de novo* molecular design," *J. Chem. Inf. Model.*, vol. 58, no. 6, pp. 1194–1204, June 2018. doi:10.1021/acs.jcim.7b00690.

[14] K. D. Warner, C. E. Hajdin, and K. M. Weeks. "Principles for targeting RNA with drug-like small molecules," *Nat. Rev. Drug Discov.*, vol. 17, no. 8, pp. 547–558, August 2018. doi:10.1038/nrd.2018.93.

[15] W. H. Shin, K. Kumazawa, K. Imai, T. Hirokawa, and D. Kihara."Current challenges and opportunities in designing protein–protein interaction targeted drugs," *Adv. Appl. Bioinform. Chem.*, vol. 13, pp. 11–25, September 2020. doi:10.2147/AABC.S23554

[16] M. Mochizuki, S. D. Suzuki, K. Yanagisawa, M. Ohue, and Y. Akiyama. "QEX: target-specific druglikeness filter enhances ligand-based virtual screening," *Mol. Divers.*, vol. 23, no. 1, pp. 11–18, February 2019. doi:10.1007/s11030-018-9842-3.

[17] M. A. Limmer, and J. G. Burken. "Plant translocation of organic compounds: Molecular and physicochemical predictors," *Environ. Sci. Technol. Lett.*, vol. 1, no. 2, pp. 156–161, February 2014. doi:10.1021/ez400214q.

[18] R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, and M. Tyers. "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Sci.*, vol. 30, no. 1, pp. 187–200, January 2021. doi:10.1002/pro.3978.

[19] R. Torchet, K. Druart, L. C. Ruano, A. Moine-Franel, H. Borges, O. Doppelt-Azeroual, B. Brancotte, F. Mareuil, M. Nilges, H. Ménager, and O. Sperandio. "The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators," *Bioinformatics*, vol. 37, no. 1, pp. 89–96, January 2021. doi:10.1093/bioinformatics/btaa1091.

[20] C. Xu, and S. A. Jackson. "Machine learning and complex biological data," *Genome Biol.*, vol. 20, Article No. December 76, 2019. doi:10.1186/s13059-019-1689-0.

[21] M.-J. Basse, S., Betzi, X., Morelli, P., Roche. "2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions," *Database*, vol. 2016, no. baw007, March 2016. doi:10.1093/database/baw007.

[22] S. Soga, H. Shirai, M. Kobori, and N. Hirayama. "Use of amino acid composition to predict ligand-binding sites," *J. Chem. Inf. Model.*, vol. 47, no. 2, pp. 400–406, March 2007. doi:10.1021/ci6002202.

[23] M. R. Arkin, Y. Tang, and J. A. Wells. "Small-molecule inhibitors of protein-protein interactions: Progressing toward reality," *Chem. Biol.*, vol. 21, no. 9. pp. 1102–1114, September 2014. doi:10.1016/j.chembiol.2014.09.001.

[24] S. A. Wildman, and G. M. Crippen. "Prediction of physicochemical parameters by atomic contributions," *J. Chem. Inf. Comput. Sci.*, vol. 39, no. 5, pp. 868–873, August 1999. doi:10.1021/ci9903071.

[25] E. C. Harrington. "The desirability function," *Ind. Qual. Control*, vol. 21, no. 10, pp. 494–498, 1965.

[26] G. W. Bemis, and M. A. Murcko. "The properties of known drugs. 1. molecular frameworks," *J. Med. Chem.*, vol. 39, no. 15, pp. 2887–2893, January 1996. doi:10.1021/jm9602928.

[27] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, and R. A. Sayle. "ZINC20—A free ultralarge-scale chemical database for ligand discovery," *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 6065–6073, December 2020. doi:10.1021/acs.jcim.0c00675.

[28] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. Feng, S. Ganesan, D. S. Goodsell, S. Ghosh, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, C. L. Lawson, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Persikova, C. Randle, A. Rose, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, Y.-P. Tao, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, and M. Zhuravleva. "RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D437–D451, January 2021. doi:10.1093/nar/gkaa1038.