SCIENTIFIC REPORTS

OPEN

Received: 9 March 2017
Accepted: 7 August 2017
Published online: 20 September 2017

An iterative compound screening contest method for identifying target protein inhibitors using the tyrosine-protein kinase Yes

Shuntaro Chiba[1,2], Takashi Ishida[1,2,3], Kazuyoshi Ikeda[4], Masahiro Mochizuki[5], Reiji Teramoto[6], Y-h. Taguchi[7], Mitsuo Iwadate[8], Hideaki Umeyama[8], Chandrasekaran Ramakrishnan[9], A. Mary Thangakani[10], D. Velmurugan[10], M. Michael Gromiha[9], Tatsuya Okuno[11], Koya Kato[12], Shintaro Minami[13], George Chikenji[12], Shogo D. Suzuki[3], Keisuke Yanagisawa[3], Woong-Hee Shin[14], Daisuke Kihara[14,15], Kazuki Z. Yamamoto[16], Yoshitaka Moriwaki[17], Nobuaki Yasuo[3], Ryunosuke Yoshino[17,18], Sergey Zozulya[19,20], Petro Borysko[19,20], Roman Stavniichuk[19], Teruki Honma[1,3,21], Takatsugu Hirokawa[22,23,24], Yutaka Akiyama[1,2,3,22,24] & Masakazu Sekijima[1,2,3,18,24]

We propose a new iterative screening contest method to identify target protein inhibitors. After conducting a compound screening contest in 2014, we report results acquired from a contest held in 2015 in this study. Our aims were to identify target enzyme inhibitors and to benchmark a variety of computer-aided drug discovery methods under identical experimental conditions. In both contests, we employed the tyrosine-protein kinase Yes as an example target protein. Participating groups virtually screened possible inhibitors from a library containing 2.4 million compounds. Compounds were ranked based on functional scores obtained using their respective methods, and the top 181 compounds from

[1]Advanced Computational Drug Discovery Unit, Institute of Innovative Research, Tokyo Institute of Technology, J3-23 4259 Nagatsuta-cho, Midori-ku, Yokohama, 226-8501, Japan. [2]Education Academy of Computational Life Sciences, Tokyo Institute of Technology, J3-141 4259 Nagatsuta-cho, Midori-ku, Yokohama, 226-8501, Japan. [3]Department of Computer Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550, Japan. [4]Level Five Co. Ltd., Shiodome Shibarikyu Bldg., 1-2-3 Kaigan, Minato-ku, Tokyo, 105-0022, Japan. [5]IMSBIO Co., Ltd., Level 6 OWL TOWER, 4-21-1 Higashi-Ikebukuro, Toshima-ku, Tokyo, 170-0013, Japan. [6]Forerunner Pharma Research, Co., Ltd., Yokohama Bio Industry Center, 1-6 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan. [7]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan. [8]Department of Biological Sciences, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan. [9]Department of Biotechnology, Bhupat Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, 600036, Tamilnadu, India. [10]CAS in Crystallography and Biophysics and Bioinformatics Facility, University of Madras, Chennai, 600025, Tamilnadu, India. [11]Division of Neurogenetics, Nagoya University Graduate School of Medicine, 65 Tsurumai, Showa-ku, Nagoya, 466-8550, Japan. [12]Department of Computational Science and Engineering, Nagoya University, Furocho, Chikusa, Nagoya, 464-8603, Japan. [13]Department of Complex Systems Science, Graduate School of Information Science, Nagoya University, Furocho, Chikusa, Nagoya, 464-8601, Japan. [14]Department of Biological Sciences, Purdue University, Indiana, 47907, USA. [15]Department of Computer Science, Purdue University, Indiana, 47907, USA. [16]Isotope Science Center, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-0032, Japan. [17]Department of Biotechnology, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo, 113-8657, Japan. [18]Global Scientific Information and Computing Center, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550, Japan. [19]Bienta/Enamine Ltd., 78 Chervonotkatska Street, Kyiv, 02660, Ukraine. [20]National Taras Shevchenko University of Kyiv, 64/13 Volodymyrska Street, Kyiv, 01601, Ukraine. [21]Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan. [22]Molecular Profiling Research Center for Drug Discovery, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan. [23]Division of Biomedical Science, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki, 305-8575, Japan. [24]Initiative for Parallel Bioinformatics, Level 14 Hibiya Central Building, 1-2-9 Nishi-Shimbashi Minato-Ku, Tokyo, 105-0003, Japan. Correspondence and requests for materials should be addressed to M.S. (email: sekijima@c.titech.ac.jp)

each group were selected. Our results from the 2015 contest show an improved hit rate when compared to results from the 2014 contest. In addition, we have successfully identified a statistically-warranted method for identifying target inhibitors. Quantitative analysis of the most successful method gave additional insights into important characteristics of the method used.

Introducing a new drug to a market has become an enormous undertaking because of expanding research and development costs, which are estimated at over one billion USD[1–4]. With a view to reducing these costs, computational technology-driven approaches have been proven to be useful and have begun to be applied at various stages of the drug discovery campaign, including from target identification to clinical phases[3, 5]. For these stages, including the hit-compound identification for a target molecule, many computational methods have been devised to find compounds that are active from a compound library without resorting to high-throughput screening.

These computational methods use various approaches and experimental information; however, they are often divided into two categories: structure-based (SB) and ligand-based (LB). SB methods use an atomic-level structure of a target molecule. Most typical SB methods are molecular docking approaches that search the complex structure of a ligand, included in a compound library, and a target-molecule structure based on a scoring function. A ranking of docked compounds is calculated using these scores[6]. In contrast, LB methods use information of known active and/or inactive compounds related to a target molecule. LB methods generally calculate a ranking of compounds in a library using techniques such as a similarity search and machine learning[7]. Currently, various methods based on both SB and LB algorithms have been proposed for identifying hit compounds[6–8].

Although these methods are reasonably designed and seem to have the ability to enrich potent compounds toward higher ranks from a compound library, there are no set standards because the performance of a method often depends on the target molecule[9]. Hence, we cannot choose a method suitable for a specific target molecule before conducting experimental assessments. Thus, designating all resources for one method is risky. However, this risk may be reduced by collecting data from various computational methods. In addition, after conducting experimental assays, we can obtain information regarding a suitable method for the target.

To evaluate various methods for a target molecule, we held a compound-screening contest in 2014 to find inhibitors of the tyrosine-protein kinase Yes as an example target from a 2.2-million-compound library[10]. Ten groups participated in the contest and, in total, 600 compound-inhibition rates for enzymatic activity were assayed. We showed that the connected diversity of compounds proposed from all participant groups was larger than that proposed by any single group. This enabled the diversified screening of the compound library with reasonable methods. As a result, two compounds were identified as hit compounds. We had speculated that we could find methods that were significantly more likely to provide hit compounds than others based on the contest's results. However, this was not possible with a statistically significant measure because of the shortage in number of assayed compounds. In the previous contest, the most successful group found 2 hit compounds from 55 compounds assayed. Provided that an average hit rate was 2/600, the p-value calculated by the binomial test for the group was 0.015. Taking the problem of multiple comparisons into account by the Bonferroni correction, there were no methods that outperformed others. In addition, the experiment may fail to detect other good methods, because, even if a method has a 3% hit-rate potential, 18.7% of the trials would return 0 hit compounds with 55 assays. Thus, many more assays are required for reliable evaluation.

To evaluate our approach for collecting various methods to reduce the risk of allocating all resources towards one method, and for obtaining useful information regarding promising methods, we conducted another contest in this study. We increased the number of compounds to be assayed for each group to more than 180. We chose the same target molecule as in the previous contest, i.e., the tyrosine-protein kinase Yes, because participants could use protein structural information as well as active and inactive compound information for this target, as well as related kinases in the same family. While the structure of Yes has not been reported, many homologous protein structures are deposited in the Protein Data Bank (PDB)[11] (e.g., 1Y57 (Unphosphorylated state of the tyrosine-protein kinase Src. Positives to Yes=92%)[12], 2SRC (Phosphorylated state of the tyrosine-protein kinase Src, positives=92%)[13], 1OPK (the tyrosine-protein kinase Abl, positives=63%))[14]. Experimental information from active and inactive compounds for the target are deposited in open databases, such as BindingDB[15, 16], ChEMBL[17], DrugBank[18], and PubChem[19].

The compound screening contest was organized by the Initiative for Parallel Bioinformatics (IPAB). It started on January 15, 2015 and ended on March 20, 2015. Eleven groups participated in the contest. The participants were asked to propose a prioritized set of 400 compounds. We selected approximately top 180 compounds from the prioritized list from each group and, in total, 1,991 unique compounds were assayed. Ten potent compounds with half-maximal inhibitory concentrations (IC_{50}) less than 10 μmol L^{-1} were identified. Overview of the procedure is shown in Fig. 1. Among the 11 methods, a successful method was identified for this target in terms of hit rate, and the salient features of this method are discussed.

Methods

Preparation of compound library. A compound library was originally provided by Enamine Ltd. and contained 2,382,017 of the available compounds in their inventory. We searched the known inhibitors of Src-family kinases shown in Table S1 (Supporting Information) that met with certain criteria from ChEMBL (version 19)[20] and BindingDB[15, 16] to eliminate them from the original library. These criteria included compounds with IC_{50} <10 μmol L^{-1}, K_i <10 μmol L^{-1}, K_d <10 μmol L^{-1}, and inhibition rates >30%, where we did not take experimental conditions into consideration. We found 3,528 unique compounds, hereafter referred to as the known inhibitors of the contest, among which 24 compounds were identified and eliminated from the original compound library. We also excluded compounds interacting with a number of proteins. We searched compounds that inhibit

Group	Modeling of Yes structure		Ligand preparation	Processing method of compound library		
	3D structure prediction methods/tools	Template PDB ID		Filter class	Actives	Inactive
1	—	—	—	LB: 1D and 2D PaDEL descriptor[49]	The top 7 compounds in PubChem (AID 686947)[22]	The rest of compounds
2	—	—	—	LB: Morgan descriptor[50]	80% of PubChem (AID 686947)[22] (The rest was used to validate the model built.)	
3	—	—	—	LB: Morgan2[50] and atom pairs descriptors[51] Protein: ProtFP[52] and Z-scales[53] Experimental conditions	Eliminated.sdf.zip,[a] PubChem (see details in Table S6 of the Supporting Info.) & IPAB2014[b]	
4	Homology modeling (*FAMS*)[54]	1Y57[12]	*Open Babel*[55]	SB: *ChooseLD*[25]		
5	Homology modeling (*MODELLER*)[27]	1OPK[14], 1IEP[26]	*LigPrep*[56]	Hybrid (LB & SB): *Glide*[28, 57–59] and pharmacophore-based screening[29, 30]	IPAB2014[b]	IPAB2014[b]
6	Homology modeling (*MODELLER*)[27]	2SRC[13]	*OMEGA*[60]	Hybrid (LB & SB): *VS-APPLE*[31]	Eliminated.sdf.zip[a]	DUD-E[9]
7	—	—	—	LB: Physicochemical properties and topological descriptors complied in *Canvas*[46, 47]	Eliminated.sdf.zip[a]	IPAB2014[b]
8	Homology modeling (*GalaxyTBM*)[61]	2H8H[62], 1KSW[63], 1FMK[64]	*OMEGA*[60]	SB: *PL-PatchSurfer2* (primitive version)[65, 66]	—	—
9	Homologous protein structure themselves were used.	1YI6[67], 3G5D[68],	*OMEGA*[60]	Hybrid (LB → SB) LB: Drug-like filtering (*SYBYL-X 2.0*), SB: *OEDocking*[69–71]		
10	Homology modeling (*MODELLER* in *HHpred*[72], *PSIPRED*)[73], followed by MD simulation[74–76] (*GROMACS*)	2H8H[62]	*OMEGA*[60]	Hybrid (LB → SB) LB: *ROCS* (ligand-shape-based method)[77] SB: *Molegro Virtual Docker*[78]	List 1–3 for Ligand-based filtering (See Section *Methods used by each group* in the Supporting Info.)	—
11	Homology modeling (*Prime*[1, 2])	1Y57[12]	*LigPrep*[56]	SB: *Glide*[57–59], followed by filtering based on conserved binding modes of docking poses	Actives (IC_{50} <1 μM) in ChEMBL, IPAB2014[b]	300 compounds from IPAB2014[b]

Table 1. Summary of methods used by participant groups. Software names are given in italics. [a]Known Src-kinase inhibitors distributed by IPAB (see Preparation of compound library section). [b]Inhibitory assay results of the previous contest[10], in which experimental conditions were the same as this study. PDB = protein data bank; LB = ligand-based; SB = structure-based; IPAB = Initiative for Parallel Bioinformatics; MD = molecular dynamics;

more than four proteins from ChEMBL, using the same inhibition criteria, and 5009 compounds were identified. This number was reduced to 245 compounds by filtering for drug-likeness, as defined in Table S2 (Supporting Information). From there, 54 compounds were identified and eliminated from the original library. Finally, the processed library contained 2,381,939 compounds, and it was distributed to participants of the contest. All compound IDs in this study correspond to Enamine Ltd. IDs.

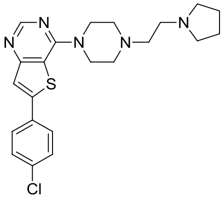
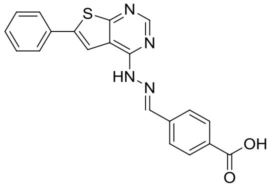
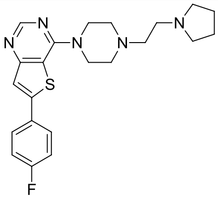
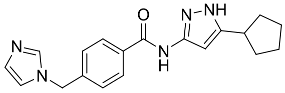
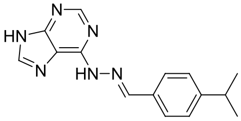
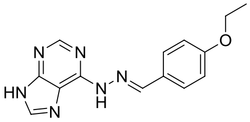
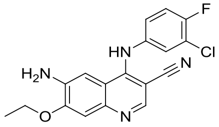
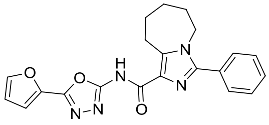
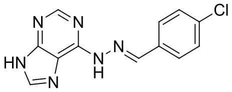
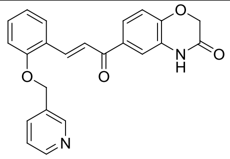
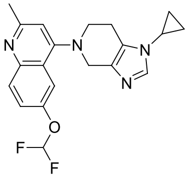
Methods participated. We accepted 11 groups, which are referred to as G1–G11 hereafter, which proposed various methods (shown in Table 1). A detailed description from groups and proposed compounds of SMILES in prioritized order are given in the section *Methods used by each group* of the supporting information and supplemental materials. Here, we briefly describe each method.

G1: A structure-activity-relationship (SAR) model was built employing balanced random forests[21]. Ligand descriptors of PubChem bioactive data[22] for Yes kinase were used as the training set, in which seven compounds with IC_{50} <1 nmol L^{-1} were selected as active compounds and the other 832 compounds were designated as inactive.

G2: An SAR model was built employing a deep neural network model, in which descriptors of randomly-chosen 80% of the PubChem bioactive data[22] were used as a training set and the other 20% comprised the test set, each of which contained active and non-active compounds. Promising compounds based on the SAR model were selected, followed by a filtering of drug-likeness and diverse selection.

G3: Compounds that were physicochemically similar to those of known inhibitors were filtered using a modified QED[23]. A randomized tree model[24] was built on the bases of the concatenated descriptors of known inhibitors, their target kinases, and experimental conditions (concentration of reagents) and was applied to filter compounds. Out-of-bag validation showed a good correlation between predicted and experimental values. The filtered compounds were re-ranked by three metrics: (1) the original ranking, (2) prioritized by ligand efficiency based on the number of heavy atoms, and (3) the novelty of compounds to the top 1,000 of the original ranking compared with Src-family inhibitors. The proposed compounds were rotationally picked up from the three ranks.

G4: The Yes protein structure was built using BLAST search with the Yes sequence. Homologous proteins having a ligand of the Yes sequence were searched and the bound ligands were remapped to the built protein, which was used for the docking[25] of known inhibitors considering remapped ligands. Based on the ability to pick up inhibitors, Yes and ligand pairs were selected. These structures were used for the docking of library compounds.

			95% CI μM[a]		
Compound ID	**Chemical Structure**	**IC_{50} μM**	**lower**	**upper**	**Group**
Z64663950		0.26	0.22	0.31	3
Z49895016[d]		0.30	0.23	0.38	3
Z64663944		0.35	0.13	0.99	3
Z1229984790		0.71	0.24	2.10	10
Z57745314[d]		1.16	0.51	2.62	3
Z57745304[d]		1.9	1.5	2.4	3
Z199512484		3.0	1.9	4.7	3
Z410927360		3.4	3.2	3.6	10
Z295464022[d]		5.0	3.5	7.3	3
Z449737600[d]		7.0	5.2	9.3	11
Z1252403274[b]		20.0	15.6	25.6	11
Continued					

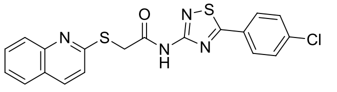
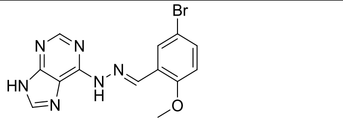
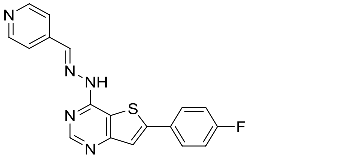
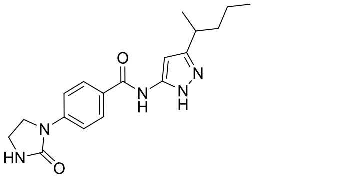
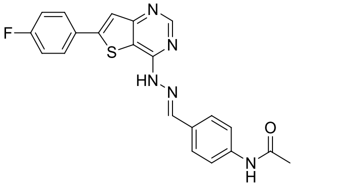
Compound ID	Chemical Structure	IC_{50} µM	95% CI µM[a]		Group
			lower	upper	
Z275023406[b]		37.4	16.9	82.5	5
Z57745307[cd]		—	—	—	3
Z50080378[cd]		—	—	—	3
Z1283491630[c]		—	—	—	5
Z50080181[cd]		—	—	—	3

Table 2. IC_{50} values of compounds that passed the validation assay (the 2nd screening). Inhibition rates from the first and second screenings are shown in Tables S1 and S2 of the Supporting Information along with the canonical SMILES. The final reagent concentrations were 5.5-nmol L^{-1} Yes, 0.013-mmol L^{-1} ATP, and 0.2-mg mL^{-1} substrate (poly Glu-Tyr peptides, Glu:Tyr = 4:1). (*a*) 95% confidence interval. Some compounds are not a hit because of insufficient potency (*b*) or a bad dose-dependence relationship (*c*). (*d*) These compounds are hydrazones or a potential Michel acceptor (see sections "Experimental procedure and screening of potential inhibitors" and "Comparison of ligand-based and structure-based methods"). IC_{50} = inhibitory concentrations; CI = confidence interval.

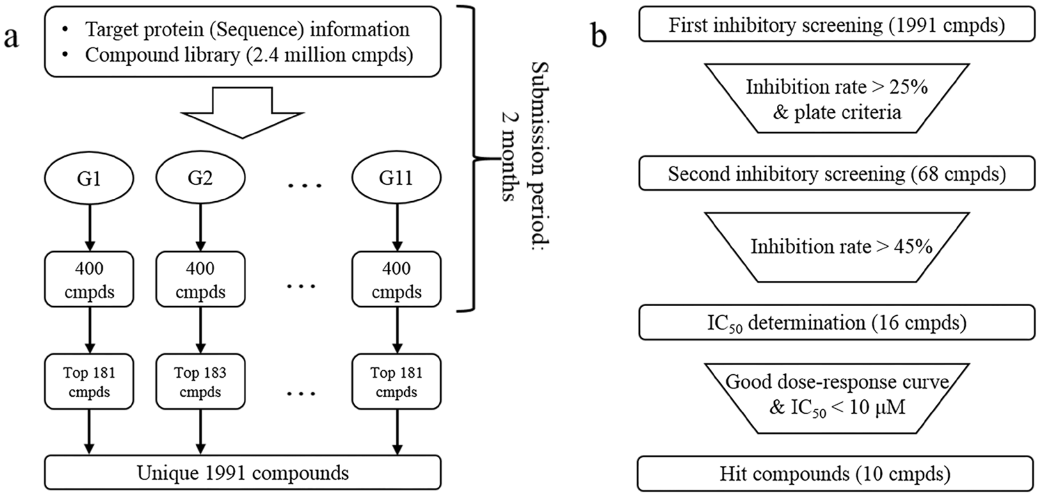

a
• Target protein (Sequence) information
• Compound library (2.4 million cmpds)
Submission period: 2 months
G1
G2
. . .
G11
400 cmpds
400 cmpds
400 cmpds
Top 181 cmpds
Top 183 cmpds
Top 181 cmpds
Unique 1991 compounds
b
First inhibitory screening (1991 cmpds)
Inhibition rate > 25% & plate criteria
Second inhibitory screening (68 cmpds)
Inhibition rate > 45%
IC50 determination (16 cmpds)
Good dose-response curve & IC50 < 10 μM
Hit compounds (10 cmpds)

Figure 1. (**a**) The flowchart of the contest. The participated groups (G1–G11) proposed 400 compounds (cmpds) with a prioritized rank from compound library using their own methods. The proposed compounds that were not stocked-out were selected until the number of compounds reached 181 for each group. If there is a duplication in the proposed compounds from different groups, such group attained additional compounds to be assayed. This is the reason why there are differences among the number of selected compounds of each group. Finally, the selected compounds were assayed. (**b**) The screening flow of the compounds in the experimental assay. The filtering criteria are shown in a trapezium.

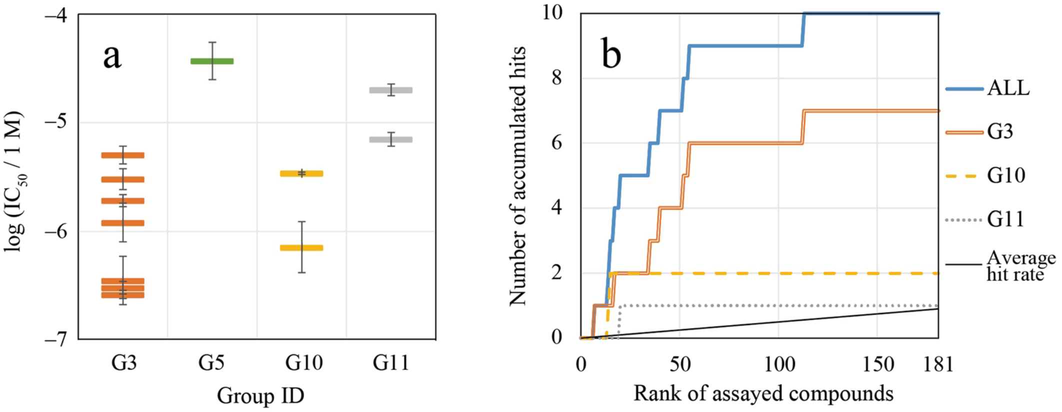


Figure 2. (**a**) The IC_{50} of compounds from each group, where results of those groups whose compounds did not proceed to the IC_{50} analysis are omitted. The compounds of log (IC_{50}/1 M) less than −5 are hit compounds. The error bars represent a 68% confidence interval estimated from the IC_{50} assay. (**b**) The number of hit compounds included within a prioritized rank of compounds that were proposed from each group.

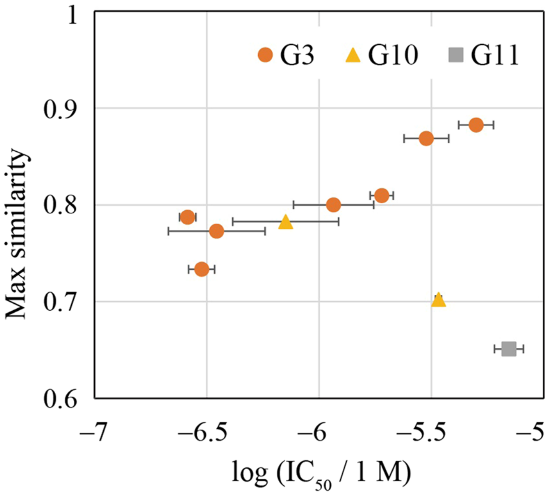


Figure 3. Similarity of each hit compound to known Src-family kinase inhibitors (see Section Preparation of compound library) is plotted against experimental inhibition activity. The error bars represent 68% confidence intervals estimated from IC_{50} assays. The similarity in these figures was calculated with the Tanimoto coefficient of the MACCS descriptor[41]. A chemical structure of the most similar compound of each hit is shown in Table S5 of the Supporting Information with its ChEMBL ID and literature.

G5: The crystal structures of Abl kinase, available for both IN and OUT conformations[14, 26], were taken as templates and respective structures were built for Yes kinase[27]. PD166326, a type I inhibitor (IN), and imatinib, a type II inhibitor (OUT) were co-crystallized with Abl kinase and docked with the IN and OUT models built for Yes kinase. On the basis of physicochemical properties, the initial compound library was filtered. Actives and decoys[10] were added to the filtered compounds and subjected to docking[28] combined with pharmacophore-based virtual screening[29, 30]. The same set of actives and decoys were included to validate the screening results. Finally, the top hit compounds from the pharmacophore-based virtual screening of DFG-IN and DFG-OUT conformations were applied.

G6: A virtual screening method[31] was applied to the compound library that performed 3D structural comparison based on a multiple-ligand template built from known multiple inhibitors using a geometric hashing technique. If a steric clash between a compound and the target protein was found, a score for a given ligand pose was penalized. Twenty complex structures of homologous proteins of Yes and its ligands deposited in the PDB were selected on the basis of the ability to discriminate actives from decoys through docking. The selected 20 proteins and their bound ligands were superimposed by the protein structure alignment program MICAN[32, 33] to the Yes structure model built based on the closest homology of Yes[27].

G7: A deep neural network was trained based on physicochemical and topological descriptors of active and inactive ligands. Hyperparameters of the deep neural network (e.g., a number of hidden layers) were also optimized using a random search[34] based on receiver-operating-characteristic (ROC) curves calculated using 5-fold

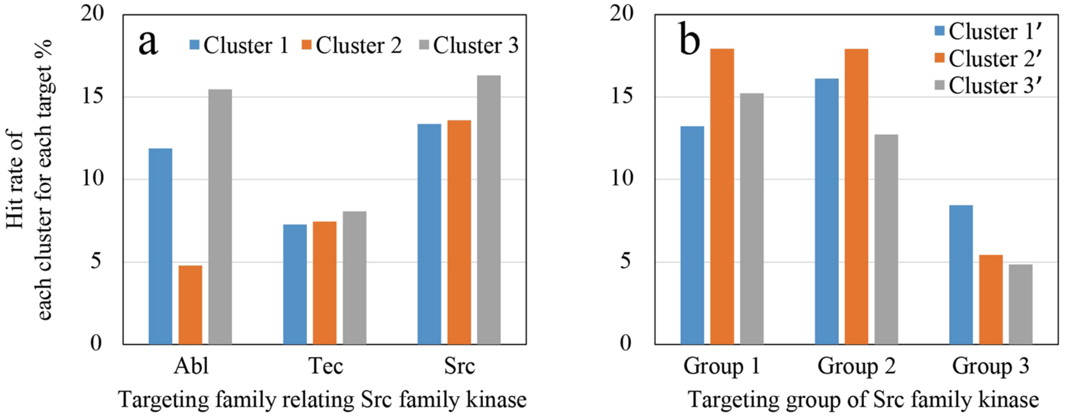


Figure 4. (**a**) Hit rate of compounds in each cluster with respect to the three kinase families. The hit rate was calculated by dividing the number of hit compounds by the number of compounds with inhibition rates that were measured to the family. (**b**) Hit rate of compounds in each cluster with respect to the three groups of Src-family kinases. The 11 kinases defined by the kinome were classified into three groups: Group 1: Src, Fyn, Yes, Fgr; Group 2: Blk, Hck, Lck Lyn; and Group 3: Frk Srm, Brk based on the kinome[42]. The clustering was calculated with Canvas[46, 47] based on the k-means algorithm[48] of the MACCS descriptor[41]. The clusters in Fig. 4a do not correspond to those in Fig. 4b.

	Reagent concentration (μmol L^{-1})			
Feature	**Compound**	**ATP**	**Mg^{2+b}**	**pH**
Average	4.6	91	5500	7.3
Minimum	4.6×10^{-4}	10	0	7.5
Maximum	670	200	2×10^4	7.0
Standard deviation	20	27	3900	0.2

Table 3. Range of experimental conditions used for training a machine learning technique[a]. [a]In addition to these features, dummy parameters that distinguish sources of experimental studies were combined with the training set. [b]This range was calculated based on the actual training set used, which included trivial mistakes in retrieving experimental parameters. As Mg^{2+} usually coexists adequately in assay samples, it would not affect inhibition rates. G3 confirmed that removing the concentration of Mg^{2+} from the training set did not affect the result after participation in the contest.

cross-validation procedures in terms of known ligands. The model that gave the best ROC curve was applied to filter the compound library.

G8: The target protein structure was built from homologous proteins and its binding pocket was converted into three-dimensional Zernike descriptors (3DZD). Ligand structures from the compound library were also converted to the 3DZD and the compatibility of each ligand to the pocket was used to select a potential inhibitor.

G9: Homologous proteins of Yes were downloaded from the PDB and docking pockets that were distant from the ATP/substrate-binding pockets were searched to find allosteric sites. Among the prepared candidate structures, two structures that showed higher docking scores from a relatively small number of compounds were chosen for the production run. Docked compounds were prepared by filtering similar known inhibitors (85% similarity) from the compound library. Visual inspection was applied to eliminate compounds that did not have drug-likeness.

G10: First, known potent compounds were used to filter the compound library to be used for subsequent docking. The Yes protein and ligand complex structure were built by homology modeling, followed by a molecular dynamics (MD) simulation of the complex to relax the structure. The 40-ns structure of the complex was used for docking.

G11: Protein ligand complex structures were built from three homologous proteins. Docking of active and inactive compounds was applied to each structure and the ability to separate active from inactive compounds was evaluated. Those displaying reasonable ability were used for docking of the compound library. Docking poses of high-ranked compounds were re-ranked using scores that considered the similarity and dissimilarity of docking poses among active and inactive compounds.

Screening of Compounds

Experimental procedure and screening of potential inhibitors. All inhibitory assays of the phosphorylation activity of Yes were performed in accordance with the Promega Technical Manual for the ADP-Glo™ Kinase Assay (Fitchburg, WI, USA. Catalog number: V9102). The human recombinant Yes [a.a. 2–543 (end)] was purchased from BPS Bioscience (catalog number: 40488). The details of the assay protocol and reagent

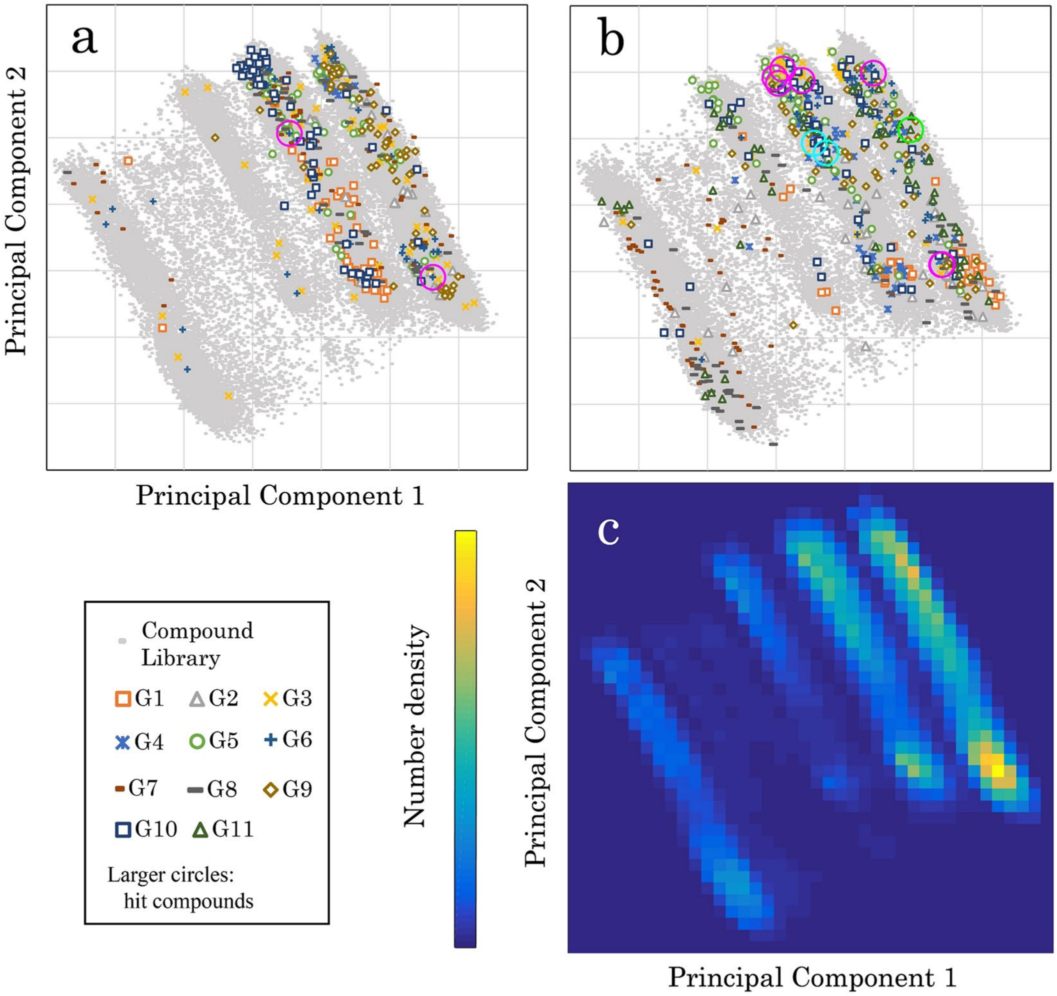


Figure 5. Diversified screening by collecting various computational methods. Principal component analysis of the library compounds in this study was applied, in which the MACCS fingerprint was used. The cumulative variance of the principal component (PC) 1 and 2 are 26% and 49%. (**a**) Compounds proposed from groups participating in the previous contest are projected to the PC1 and PC2. Two hit compounds confirmed based on IC_{50} determination are plotted. To avoid the complication of symbols, the top 60 compounds in the proposed list are shown. As for the compound library, a randomly chosen 2.5% of all the compounds are shown. (**b**) The same analysis as (**a**) is conducted using data from this study. Ten hit compounds (magenta for G3, cyan for G10, green for G11) are plotted. (**c**) Number density in the PC1 and PC2 of all the compounds are shown.

information are given elsewhere[10]. Here, we briefly describe the screening of the compounds based on inhibition activity and results.

The screening was conducted in three steps consisting of the first screening, the second screening, and the IC_{50} determination, as can be illustrated in Fig. 1b. First, we determined the inhibition rates of the 1,991 compounds. Each compound was placed in 4 wells of a 384-well plate. In total, 80 compounds were assayed on one plate and the other wells were used for positive and negative controls. The compounds were randomly placed on plates so that compounds proposed by one group were not placed on a plate together. A mean of four inhibition rates for each compound was compared to criteria for the first screening. These criteria included that the inhibition rate was greater than 25% and the inhibition rate was greater than the mean plus three-fold of the standard deviation of the plate on which the compound was assayed, where observed inhibition rates of positive and negative controls were not taken into consideration. As a results, 68 compounds passed the screening. Information of the dropped compounds is given in Table S3 of the Supporting Information. Second, the inhibition rates of the screened compounds were determined on one plate using the same procedure of the first screening, where compounds were dissolved from fresh powder. As a result, 16 compounds showed inhibition rates greater than the threshold of the second screening (i.e., approximately 50%). Information for screened and dropped compounds is given in Table S4 of the Supporting Information. Screened compounds were then evaluated for their IC_{50} values. The chemical structure and assay results of these compounds are given in Table 2.

Among the 16 compounds, 10 compounds showed an IC_{50} less than our hit criterion, which was an IC_{50} less than 10 μmol L^{-1}, as shown in Table 2. These compounds showed a clear dose-response relationship (DRR) as can be seen in Figure S1 of the Supporting Information. As for Z1252403274 and Z275023406, which showed a good DRR, they were not defined as hit compounds because of insufficient potency. The other four compounds, Z50080378, Z57745307, Z50080181, and Z1283491630, did not reveal a DRR, having "inhibition activity" around 50% in the whole range of concentrations, which may be due to their non-specific interactions with the target (promiscuous protein binding, protein aggregation) or solubility-related issues. For these reasons, these compounds were excluded from consideration. Note that we confirmed that the threshold used for the second screening was reasonable as can be seen in Figure S2 of the Supporting Information.

The 10 hit compounds were compared to the pan-assay interference compounds (PAINS) filters, filters A, B, and C described in the literature[35], which suggests potential functional groups of frequent hitters extracted from HTS assays. We found that the 10 compounds do not have these potential functional groups. This means that all the hit compounds are promising for further investigation. It should be noted that some hit compounds have "questionable" chemotypes from a medicinal chemistry point of view, i.e., hydrazones (Z49895016, Z57745314, Z57745304, Z295464022) and a potential Michael acceptor (Z449737600), which may present a reactivity/toxicity liability. In the present study, we did not exclude them because only the biochemical assays, not cell-based, were used for screening in this study. This strongly decreases the chance of getting false positives with these compounds during the primary screen. The emphasis was also placed on avoiding the potential loss of any active scaffolds identified by the competing computational groups, rather than on the early elimination of less desirable chemical series. Substituting hydrazones with their non-reactive isosteres during hit-to-lead optimization is a feasible medicinal chemistry endeavor as is illustrated by some research publications[36–38]. We will discuss these hydrazone-containing compounds in more details in the following section. The Michael acceptor could be substituted by an amide group between an acid and a cyclical secondary amine, to retain molecular rigidity.

Discussion

Hit rate of assayed compounds.

The total number of hit compounds was 10 (Table 2), seven of which were proposed by G3, two by G10, and one by G11. G3 outperformed the other groups in terms of the number of hits and potencies of the compounds and was followed by G10 and G11 as can be clearly seen in Fig. 2a. Performances of these methods in terms of a hit rate, compared to an average rate of all the methods, can be evaluated by the binomial test while eliminating the problem of multiple comparisons by applying the Bonferroni correction. Assuming the hit rate of all the compounds is 10/1991, the p-values for G3, G10, and G11 were 4×10^{-5}, 0.2, and 0.6, respectively. Hence, we confirmed that the method of G3 was statistically warranted.

We can also evaluate these methods by how they enriched active compounds in their prioritized ranks. The hit compounds of these methods were enriched toward higher ranks as can be seen in Fig. 2b. These methods showed better enrichment compared to the average rate.

These results suggest that the methods employed by these groups could reasonably distinguish active compounds. In the present study, we will mainly focus on these three methods.

Comparison of ligand-based and structure-based methods.

The proposed methods were classified into LB and SB approaches. The LB approaches were defined as those methods that used active and/or inactive ligand information for relevant kinases regardless of the incorporation of protein structure. The SB approaches included methods that used protein structure and did not use ligand information for filtering of the compound library.

Hit rate. The groups that found hit compounds, G3, G10, and G11, can be classified into LB, LB→SB, and SB methods as tabulated in Table 1. G3 and G10 used ligand information in a direct way to filter the compound library; only G11 used ligand information in an indirect way, i.e., the selection of a protein structure used for docking. In this sense, it was only a single compound that was proposed by an SB method. Hence, compared to LB methods, it was very difficult to find a hit compound using an SB method in this study.

The proposed compounds from G3 were selected based on the three prioritized ranks (see the explanation of G3 in the section Methods participated). Four and three of the hit compounds of G3 were found by the original rank (1) and ligand-efficiency-based rank (2), respectively. No compounds were found from the novelty-based rank, which may indicate that finding novel compounds using an LB approach is difficult.

Novelty. It is of great importance to obtain a number of novel hit compounds in drug discovery[39]. We compared which hit compounds from LB or SB gave novel compounds in this study. First, we calculated similarities between each hit compounds and known Src-family inhibitors defined in the Preparation of compound library section. Among the similarities calculated for each of the compounds, the maximum value was assigned to the compound as the max similarity. The most novel compound was proposed by G11 (SB), which used docking for the selection of compounds, as can be shown in Fig. 3. The second was proposed by G10 (LB→SB), which used known inhibitors to filter the compound library followed by docking. Almost all the other compounds were proposed by G3 (LB), which used known active and inactive compounds to build a compound filter. Among seven hits from G3, four compounds were hydrazone (Z49895016, Z57745314, Z57745304, Z295464022) and had a similar scaffold as can be shown by their structures. This was because the training set G3 used contained 65 known hydrazone-containing compounds, in which 58 compounds had inhibition rate greater than 50%, in the total number of compounds used of 2040. Among the compounds used, 56 compounds of the hydrazone-containing compounds were derived from Published Kinase Inhibitor Set (PKIS), which collected results of kinase panel experiments of 367 kinase inhibitors and was released from GlaxoSmithKline. A similar scaffold was reported by a clustering analysis of PKIS[40]. This shows a clear dependency of the LB method on training data set used. Hence, we could say that an LB method is more likely to give similar hit compounds to known inhibitors in our contest. Conversely, one can resort to a method that uses an SB approach to obtain novel hit compounds. We also confirmed that hit compounds that were proposed by different groups were not similar to each other (see Figure S3 of the Supporting Information).

Characteristics of the most successful method.

Among all the groups, the hit rate of compounds proposed by G3 was statistically confirmed to be higher than the others. We summarize the salient characteristics of the method here. G3 employed a machine learning technique based on a training set that combined three kinds of data for known active and inactive compounds of the Src and relevant kinase families. These data included compound descriptors, experimental conditions when the inhibition rate was measured, and target protein information. Inhibition rates were used for training the model instead of inhibition constants or IC_{50} values because inhibition rates for compounds were relatively abundant. In some cases, G3 used inhibition rates that were measured for the determination of IC_{50} values. Among the LB methods, experimental conditions and protein information were not used except for G3. Here we focus on these two characteristics and investigate the significance of incorporating these features.

Incorporation of experimental conditions for machine learning.

G3 included several experimental conditions, as compiled in Table 3, when training the machine learning model using inhibition rates. This was based on the fact that an inhibition rate of a compound depends on experimental conditions (e.g., concentrations of compounds, enzyme, and ATP) and that experimental conditions can differ in different studies. Hence, incorporating these conditions in the training data sounds reasonable. Experimental conditions that accompanied the known compound information that G3 used were diverse, as seen in Table 3. The range of concentrations was broad, indicating that it is dangerous to build an SAR model based only on inhibition rates or IC_{50} values from different experimental studies.

To test the significance of incorporating experimental conditions, G3 conducted an OOB validation with and without experimental conditions. Excluding the experimental condition made prediction accuracy (R^2) decrease from 0.82 to 0.44. We believe that, especially in the case of building an SAR model as G3 conducted, considering experimental conditions would be crucially important if data sets are based on several experimental conditions. As which experimental conditions were significantly important was not clear in this study, further investigation and validation of the insights we obtained are needed. Further, incorporating substrate concentration, which was not used by G3, may help improve prediction accuracy.

Incorporation of protein information for machine learning.

G3 used compound information for Src, Tec, and Abl kinase families, which are closely related[42] (The ChEMBL IDs and references used are tabulated in Table S6 of the Supporting Information). While some compounds interact with a broad range of kinases, others have selectivity to a specific kinase[43, 44]. To evaluate the selectivity of compounds that G3 used, we clustered the compounds into three groups and calculated the hit rates of compounds in each cluster with respect to each kinase group, in which the hit criterion was defined to be 50% inhibition. As can be seen in Fig. 4a, each cluster did not interact with each kinase family equally. This means that there is some selectivity of compounds in the three kinase families. To evaluate the selectivity of compounds within the Src family, we clustered the compounds that had experimental information available into three groups. As can be seen in Fig. 4b, the selectivity persists in these groups. The selectivity of Group 3 of Src-family kinases was different from the other two groups. This may be consistent with the fact that Group 3 is distantly related to the other groups[45]. Hence, incorporating protein information to compound descriptors and experimental conditions may improve prediction accuracy.

An OOB validation showed that excluding protein descriptors made the prediction accuracy worse, i.e., the R^2 decreased from 0.82 to 0.73, indicating that distinguishing protein targets was meaningful in this study. As the trained model can provide a potency of a compound for each kinase used in the training set, we could obtain a selective compound for a specific kinase. Interestingly, only combining a compound's descriptor and protein information did not improve the prediction accuracy compared to using compound descriptors simply as a training set, i.e., the R^2 was only improved from 0.43 to 0.44 by introducing protein information. This means that protein information becomes useful when it is used with experimental conditions for a training set.

Comparison to the previous contest.

Comparing this study with the previous contest would give useful information. As we noted about the previous contest[10], collecting various computational methods enables diversified screening in the chemical space of the contest library compared with a single method, as can be seen in Fig. 5 and Figure S4 of the Supporting Information. This reflects the diversity of hit compounds, as can be seen in Fig. 5b and Figure S3 of the Supporting Information. The contest-based approach can provide diverse hit compounds than a single method can do. In addition, comparing the chemical diversity of hit compounds of this study (Fig. 5b) to the previous contest (Fig. 5a), hit compounds obtained in this study had broader diversity.

The total hit rate improved from 2/600 to 10/1991 (hit compounds/assayed compounds). This improvement is remarkable considering that we eliminated known inhibitors of the Src-family from the contest library this time. In the previous contest, we eliminated known inhibitors of Yes, but all the hit compounds in the previous contest were known inhibitors of other Src-family kinases.

As we have discussed in "Experimental procedure and screening of potential inhibitors" and "Comparison of ligand-based and structure-based methods" sections, we decided not to exclude the hydrzones (Z49895016, Z57745314, Z57745304, Z295464022) and the potential Michael acceptor (Z449737600) from the hit list. However, it would be worth comparing this study to the previous contest with eliminating them from the list, because regarding them as possible compounds for lead optimization remains a matter of debate. The total hit rate decreases from 10/1991 to 5/1991, which is comparable to the previous hit rate of 2/600. Even though the questionable compounds were eliminated, considering the absence of known Src-family inhibitors in the compound library used, improvement of the second contest is warranted.

We speculate that iterative participation provides the opportunity for improvement in each method because the three groups that proposed hit compounds participated in both contests. Note that 92% of the compounds in the compound library in this study were included in the previous contest library and the ten hit compounds were also included in the compound library of the previous contest.

We expected to distinguish promising methods by increasing the number of compounds assayed. However, even if the number of assayed compounds for each group was reduced to the approximate number of assayed compounds in the previous contest (55), almost all hit compounds can be found (see Fig. 2b). The p-values for G3, G10, and G11 improve to 6×10^{-7}, 0.04, and 0.26, respectively. Hence, the method of G3 is statistically warranted. Apparently, we could reduce the number of compounds to assay in this sense. However, a sufficient number of compounds to assay is necessary to detect a method with a modest hit rate. If a method has a hit rate of 3%, at least one hit compound can be found in 99.6% of the time in this experiment that assayed 180 compounds for each group. In this regard, the experiment did not miss promising methods with a significant hit rate.

Conclusion

The compound screening contest to predict potential inhibitors of the tyrosine-protein kinase Yes from the 2.4-million-compound library was held not only to identify potent inhibitors for the target and but also to benchmark various methods based on the same experimental conditions, in which 11 groups participated. Among 1,991 assayed compounds, ten hit compounds with IC_{50} values less than $10\,\mu mol\,L^{-1}$ were identified, which are not likely to be frequent hitters in terms of the fact that they passed PAINS filters. Comparing this study with the previous contest, which was held by the same organizer with the same target[10], the hit rate improved and the diversity of hit compounds grew broader.

The participating groups employed various approaches, which were classified as LB or SB approaches. Comparison of the LB and SB approaches by the three groups which proposed hits showed that the LB approach was more likely to give hit compounds, whereas the SB approach gives more novel hit compounds in our contest.

The characteristics of the most successful LB method, which identified seven hit compounds, were studied in terms of the training data set that the group used for a machine learning technique. We found that incorporation of experimental conditions, e.g., concentration of compounds under which inhibition rates were measured, significantly contributed to the prediction accuracy. In addition, the incorporation of protein descriptors to distinguish known compounds' target kinase was found partly to contribute to improved prediction accuracy.

We confirmed that a contest-based approach to identify potential inhibitors of a target protein can be successful in identifying promising hit compounds. Moreover, it can provide an initial benchmark of various methods and suggests promising approaches for the target system. Extensive exploitation and further investigation of these methods should lead to additional novel hit compounds in the drug discovery process.

References

1. Paul, S. M. *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203–214, doi:10.1038/nrd3078 (2010).
2. Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. & Greyson, D. The cost of drug development: a systematic review. *Health Policy* **100**, 4–17, doi:10.1016/j.healthpol.2010.12.002 (2011).
3. Loging, W. T. *Bioinformatics and Computational Biology in Drug Discovery and Development*. (Cambridge University Press, 2016).
4. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **47**, 20–33, doi:10.1016/j.jhealeco.2016.01.012 (2016).
5. Ou-Yang, S. S. *et al*. Computational drug discovery. *Acta Pharmacol. Sin.* **33**, 1131–1140, doi:10.1038/aps.2012.109 (2012).
6. Meng, X. Y., Zhang, H. X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **7**, 146–157 (2011).
7. Acharya, C., Coop, A., Polli, J. E. & Mackerell, A. D. Jr. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided Drug Des.* **7**, 10–22 (2011).
8. Lionta, E., Spyrou, G., Vassilatis, D. K. & Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* **14**, 1923–1938 (2014).
9. von Korff, M., Freyss, J. & Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **49**, 209–231, doi:10.1021/ci800303k (2009).
10. Chiba, S. *et al*. Identification of potential inhibitors based on compound proposal contest: Tyrosine-protein kinase Yes as a target. *Sci. Rep.* **5**, 17209, doi:10.1038/srep17209 (2015).
11. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242, doi:10.1093/nar/28.1.235 (2000).
12. Cowan-Jacob, S. W. *et al*. The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. *Structure* **13**, 861–871, doi:10.1016/j.str.2005.03.012 (2005).
13. Xu, W. Q., Doshi, A., Lei, M., Eck, M. J. & Harrison, S. C. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol. Cell* **3**, 629–638, doi:10.1016/S1097-2765(00)80356-1 (1999).
14. Nagar, B. *et al*. Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell* **112**, 859–871, doi:10.1016/S0092-8674(03)00194-6 (2003).
15. Liu, T. Q., Lin, Y. M., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201, doi:10.1093/nar/gkl999 (2007).
16. Chen, X., Lin, Y., Liu, M. & Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **18**, 130–139, doi:10.1093/bioinformatics/18.1.130 (2002).
17. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107, doi:10.1093/nar/gkr777 (2012).
18. Knox, C. *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **39**, D1035–1041, doi:10.1093/nar/gkq1126 (2011).
19. Li, Q., Cheng, T., Wang, Y. & Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discov Today* **15**, 1052–1057, doi:10.1016/j.drudis.2010.10.003 (2010).
20. Bento, A. P. *et al*. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090, doi:10.1093/nar/gkt1031 (2014).
21. Svetnik, V. *et al*. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958, doi:10.1021/ci034160g (2003).

22. Patel, P. R. *et al.* Identification of potent Yes1 kinase inhibitors using a library screening approach. *Bioorg. Med. Chem. Lett.* **23**, 4398–4403, doi:10.1016/j.bmcl.2013.05.072 (2013).
23. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98, doi:10.1038/nchem.1243 (2012).
24. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63**, 3–42, doi:10.1007/s10994-006-6226-1 (2006).
25. Takaya, D. *et al.* Bioinformatics based ligand-docking and in-silico screening. *Chem. Pharm. Bull.* **56**, 742–744, doi:10.1248/cpb.56.742 (2008).
26. Nagar, B. *et al.* Crystal Structures of the Kinase Domain of c-Abl in Complex with the Small Molecule Inhibitors PD173955 and Imatinib (STI-571). *Cancer Res.* **62**, 4236–4243 (2002).
27. Fiser, A. & Sali, A. MODELLER: Generation and refinement of homology-based protein structure models. *Method Enzymol* **374**, 461–491, doi:10.1016/S0076-6879(03)74020-8 (2003).
28. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **49**, 6177–6196, doi:10.1021/jm051256o (2006).
29. Salam, N. K., Nuti, R. & Sherman, W. Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis. *J. Chem. Inf. Model.* **49**, 2356–2368, doi:10.1021/ci900212v (2009).
30. Loving, K., Salam, N. K. & Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput. Aided Mol. Des.* **23**, 541–554, doi:10.1007/s10822-009-9268-1 (2009).
31. Okuno, T., Kato, K., Terada, T. P., Sasai, M. & Chikenji, G. VS-APPLE: A Virtual Screening Algorithm Using Promiscuous Protein-Ligand Complexes. *J. Chem. Inf. Model.* **55**, 1108–1119, doi:10.1021/acs.jcim.5b00134 (2015).
32. Minami, S., Sawada, K. & Chikenji, G. MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C(alpha) only models, Alternative alignments, and Non-sequential alignments. *BMC Bioinformatics* **14**, 24, doi:10.1186/1471-2105-14-24 (2013).
33. Minami, S., Sawada, K. & Chikenji, G. How a spatial arrangement of secondary structure elements is dispersed in the universe of protein folds. *Plos One* **9**, e107959, doi:10.1371/journal.pone.0107959 (2014).
34. Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization. *J Mach Learn Res* **13**, 281–305 (2012).
35. Baell, J. B. & Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **53**, 2719–2740, doi:10.1021/jm901137j (2010).
36. Mayer, N. *et al.* Structure-activity studies in the development of a hydrazone based inhibitor of adipose-triglyceride lipase (ATGL). *Bioorganic & medicinal chemistry* **23**, 2904–2916, doi:10.1016/j.bmc.2015.02.051 (2015).
37. Yogeeswari, P., Menon, N., Semwal, A., Arjun, M. & Sriram, D. Discovery of molecules for the treatment of neuropathic pain: synthesis, antiallodynic and antihyperalgesic activities of 5-(4-nitrophenyl)furoic-2-acid hydrazones. *Eur. J. Med. Chem.* **46**, 2964–2970, doi:10.1016/j.ejmech.2011.04.021 (2011).
38. Senger, M. R., Fraga, C. A., Dantas, R. F. & Silva, F. P. Jr. Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discov Today* **21**, 868–872, doi:10.1016/j.drudis.2016.02.004 (2016).
39. Owens, P. K. *et al.* A decade of innovation in pharmaceutical R&D: the Chorus model. *Nat. Rev. Drug Discov.* **14**, 17–28, doi:10.1038/nrd4497 (2015).
40. Dranchak, P. *et al.* Profile of the GSK published protein kinase inhibitor set across ATP-dependent and-independent luciferases: implications for reporter-gene assays. *Plos One* **8**, e57888, doi:10.1371/journal.pone.0057888 (2013).
41. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280, doi:10.1021/ci010132r (2002).
42. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934, doi:10.1126/science.1075762 (2002).
43. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045, doi:10.1038/nbt.2017 (2011).
44. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051, doi:10.1038/nbt.1990 (2011).
45. Roskoski, R. Jr. Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. *Pharmacol. Res.* **94**, 9–25, doi:10.1016/j.phrs.2015.01.003 (2015).
46. Canvas v. 2.8 (Schrödinger, LLC, New York, NY, 2016).
47. Duan, J. X., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **29**, 157–170, doi:10.1016/j.jmgm.2010.05.008 (2010).
48. Lloyd, S. P. Least-Squares Quantization in Pcm. *IEEE Trans. Inf. Theory* **28**, 129–137, doi:10.1109/Tit.1982.1056489 (1982).
49. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32**, 1466–1474, doi:10.1002/jcc.21707 (2011).
50. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754, doi:10.1021/ci100050t (2010).
51. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73, doi:10.1021/ci00046a002 (1985).
52. van Westen, G. J. *et al.* Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. *Plos One* **6**, e27518, doi:10.1371/journal.pone.0027518 (2011).
53. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. & Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **41**, 2481–2491, doi:10.1021/jm9700575 (1998).
54. Umeyama, H. & Iwadate, M. FAMS and FAMSBASE for protein structure. *Curr. Protoc. Bioinformatics* **Chapter 5**, Unit5 2, doi:10.1002/0471250953.bi0502s04 (2004).
55. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33, doi:10.1186/1758-2946-3-33 (2011).
56. LigPrep v. 3.2 (Schrödinger, LLC, New York, NY, 2014).
57. Glide v. 6.0 (Schrödinger, LLC, New York, NY, 2014).
58. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749, doi:10.1021/jm0306430 (2004).
59. Halgren, T. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47**, 1750–1759, doi:10.1021/jm030644s (2004).
60. Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **50**, 572–584, doi:10.1021/ci100031x (2010).
61. Ko, J., Park, H. & Seok, C. GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* **13**, 1–8, doi:10.1186/1471-2105-13-198 (2012).
62. Hennequin, L. F. *et al.* N-(5-Chloro-1,3-benzodioxol-4-yl)−7-[2-(4-methylpiperazin-1-yl)ethoxy]−5- (tetrahydro-2H-pyran-4-yloxy)quinazolin-4-amine, a Novel, Highly Selective, Orally Available, Dual-Specific c-Src/Abl Kinase Inhibitor. *J. Med. Chem.* **49**, 6465–6488, doi:10.1021/jm060434q (2006).
63. Witucki, L. A. *et al.* Mutant Tyrosine Kinases with Unnatural Nucleotide Specificity Retain the Structure and Phospho-Acceptor Specificity of the Wild-Type Enzyme. *Chemistry & Biology* **9**, 25–33, doi:10.1016/S1074-5521(02)00091-1 (2002).

64. Xu, W., Harrison, S. C. & Eck, M. J. Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385**, 595–602, doi:10.1038/385595a0 (1997).
65. Hu, B., Zhu, X., Monroe, L., Bures, M. G. & Kihara, D. PL-PatchSurfer: a novel molecular local surface-based method for exploring protein-ligand interactions. *Int. J. Mol. Sci.* **15**, 15122–15145, doi:10.3390/ijms150915122 (2014).
66. Shin, W. H., Christoffer, C. W., Wang, J. & Kihara, D. PL-PatchSurfer2: Improved Local Surface Matching-Based Virtual Screening Method That Is Tolerant to Target and Ligand Structure Variation. *J. Chem. Inf. Model.* **56**, 1676–1691, doi:10.1021/acs.jcim.6b00163 (2016).
67. Fleury, D., Sarubbi, E., Courjaud, A., Guitton, J. & Ducruix, A. Structure of the unphosphorylated c-terminal tail segment of the src kinase and its role in src activity regulation. *To be published*.
68. Bauerova-Hlinkova, V., Dvorsky, R., Perecko, D., Povazanec, F. & Sevcik, J. Structure of RNase Sa2 complexes with mononucleotides–new aspects of catalytic reaction and substrate recognition. *FEBS J* **276**, 4156–4168, doi:10.1111/j.1742-4658.2009.07125.x (2009).
69. OEDOCKING v. 3.2.0.2 (Santa Fe, NM).
70. McGann, M. F. R. E. D. and HYBRID docking performance on standardized datasets. *J. Comput. Aided Mol. Des.* **26**, 897–906, doi:10.1007/s10822-012-9584-8 (2012).
71. McGann, M. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **51**, 578–596, doi:10.1021/ci100436p (2011).
72. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–248, doi:10.1093/nar/gki408 (2005).
73. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–357, doi:10.1093/nar/gkt381 (2013).
74. Abraham, M. J. *et al*. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25, doi:10.1016/j.softx.2015.06.001 (2015).
75. Maier, J. A. *et al*. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713, doi:10.1021/acs.jctc.5b00255 (2015).
76. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174, doi:10.1002/jcc.20035 (2004).
77. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74–82, doi:10.1021/jm0603365 (2007).
78. Thomsen, R. & Christensen, M. H. MolDock: a new technique for high-accuracy molecular docking. *J. Med. Chem.* **49**, 3315–3321, doi:10.1021/jm051197e (2006).

Acknowledgements

We gratefully acknowledge the financial support of Research Organization for Information Science and Technology (RIST), The Japan Biological Informatics Consortium (JBIC), HPCTECH Corporation, Schrödinger K.K., IMSBIO Co., Ltd., Dassault Systemes Biovia K.K., DiscoveResource, Inc., DataDirect Networks Japan, Inc., DELL, Namiki Shoji Co., Ltd., NEC Corporation, MITSUI KNOWLEDGE INDUSTRY CO., LTD., Lisit, Co., Ltd., Leave a Nest Co., Ltd., Level Five Co., Ltd., which made it possible to complete our contest. We are deeply grateful to Tokyo Institute of Technology, Ministry of Economy, Trade and Industry (METI), New Energy and Industrial Technology Development Organization (NEDO), Japan Pharmaceutical Manufacturers Association (JPMA), Information Processing Society of Japan (IPSJ), Japanese Society of Bioinformatics (JSBi), Chem-Bio Informatics (CBI) Society, PC Cluster Consortium, The Science News Ltd., Nikkan Kogyo Shimbun Ltd. We would like to offer our special thanks to Mr. Toshiaki Miyaki, Dr. Kazuki Ohno, and Ms. Kanako Ozeki. This work is partially supported by the Research Complex Program "Wellbeing Research Campus: Creating new values through technological and social innovation" from Japan Science and Technology Agency, JST and the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research) from Japan Agency for Medical Research and Development, AMED.

Author Contributions

All authors made substantial contributions to this study and article. S.C., T.I., Y.A. and M.S. organized and operated the contest. S.Z., P.B., and R.S. carried out inhibitory assays. S.C., T.I., K.I., M.M., Te.H., and Ta.H. evaluated data. R.T., H.h.T., M.I., H.U., M.M., C.R., A.M.T., D.V., M.M.G., T.O., K.K., S.M., G.C., S.D.S., K.Y., W.H.S., D.K., K.Z.Y., Y.M., N.Y., and R.Y. participated the contest and predicted hit compound for target protein by their method. S.C., T.I., and M.S. wrote the main manuscript text. All authors approved this version to be published.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-10275-4

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.