scientific reports

Check for updates

OPEN

A novel in silico scaffold-hopping method for drug repositioning in rare and intractable diseases

Mao Tanabe[1], Ryuichi Sakate[1], Jun Nakabayashi[2], Kyosuke Tsumura[2], Shino Ohira[2], Kaoru Iwato[2] & Tomonori Kimura[3,4,5]✉

In the field of rare and intractable diseases, new drug development is difficult and drug repositioning (DR) is a key method to improve this situation. In this study, we present a new method for finding DR candidates utilizing virtual screening, which integrates amino acid interaction mapping into scaffold-hopping (AI-AAM). At first, we used a spleen associated tyrosine kinase inhibitor as a reference to evaluate the technique, and succeeded in scaffold-hopping maintaining the pharmacological activity. Then we applied this method to five drugs and obtained 144 compounds with diverse structures. Among these, 31 compounds were known to target the same proteins as their reference compounds and 113 compounds were known to target different proteins. We found that AI-AAM dominantly selected functionally similar compounds; thus, these selected compounds may represent improved alternatives to their reference compounds. Moreover, the latter compounds were presumed to bind to the targets of their references as well. This new "compound-target" information provided DR candidates that could be utilized for future drug development.

Approximately 7000 rare and intractable diseases (RIDs) have been defined to date, affecting an estimated 300 million people worldwide[1]. These diseases largely reduce patients' quality of life throughout their lives. Though the unmet medical needs are very high in this field, new drug development for RIDs is difficult. The reason is that the number of affected patients is too small for pharmaceutical companies to invest in targeting these diseases[2,3] and the mechanism of onset of many RIDs still remains to be elucidated. Therefore, attention is focused on drug-repositioning (DR) methods, finding a candidate drug previously developed for other diseases[4–6]. When relatively little information is available for the disease, phenotypic screening and target-based method are selected from existing DR methods[5]. We considered that a drug which was rather effective in a RID and known to target some protein, even when its action mechanism was not thoroughly known, could be replaced by improved alternative drug using target-based method. Target-based methods include in vitro and in vivo high-throughput screening of drugs and in silico (computational) screening of drugs from libraries[7,8]. Computational methods, virtual screening (VS) techniques are becoming increasingly popular as these are continuously being developed, improved, and made available. The examples of the state-of-the art reviews are those of Gimeno[9], Guputa[10] and so on. VS techniques are generally classified into two major categories: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS). SBVS encompasses methods that exploit the three-dimensional (3D) structure of the target and molecular docking[11,12]. LBVS can be employed when the target structure of sufficient quality for docking simulations is not available and some binders for the target binding pocket are already known[13,14]. LBVS mainly includes methods based on similarity, in which the relationships between compounds in a given library and known binding molecules for the target are examined by similarity measurements using suitable molecular descriptors[15]. To find compounds that are structurally diverse but share some biological activity, scaffold-hopping, which is a LBVS approach, has been widely attempted[16]. However, because only a few active ligands are available to be used as references, the hit compounds found by LBVS lack novelty[17,18]. As a wide diversity of hit structures is important for improved properties, some hybrid strategies that integrate both SBVS and LBVS techniques have been proposed to overcome the weakness of LBVS[18,19]. In this study, we developed a new methodology called

[1]Laboratory of Rare Disease Information and Resource Library, Center for Intractable Diseases and ImmunoGenomics Research, National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Ibaraki, Osaka, Japan. [2]Analysis Technology Center, FUJIFILM Corporation, 210 Nakanuma, Minami-ashigara, Kanagawa, Japan. [3]Reverse Translational Research Project, National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Ibaraki-City, Osaka, Japan. [4]KAGAMI Project, National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Ibaraki, Osaka, Japan. [5]Department of Nephrology, Osaka University Graduate School of Medicine, Suita, Osaka, Japan. ✉email: t-kimura@kid.med.osaka-u.ac.jp

AI-AAM, which enabled obtaining candidates with a wide variety of structures by using only the ligand-based virtual scaffold-hopping. Our hypothesis was that the interactions between a ligand and the set of amino acids could represent the interaction between a ligand and its target protein. By introducing Amino Acid Mapping (AAM), the descriptor of the interactions of a compound with amino acids, to the scaffold-hopping technique, we aimed to discover the compounds that have preserved interactions with their targets.

In this report, we aimed to examine the possibility of DR using AI-AAM with 6 compounds as reference. An overview of this study is shown in Fig. 1. At first, we selected a SYK inhibitor as reference to evaluate the technique experimentally, focusing on the pharmacological activity of a hit with the different scaffold from the reference. Then we applied this method to other five reference compounds selected from DDrare, a database of Drug Development for Rare Diseases, and examined whether the hit compounds were structurally diverse and target the same proteins as reference compounds. Moreover, on the basis of the target information of hit compounds, we investigated the pharmacological functions of hits or inferred the new compound-target connection. Last, we discuss the possibility of DR using AI-AAM, based on the results.

Results

The search for compounds using the AI-AAM technique

AI-AAM was applied to the drugs in the field of RIDs or in clinical trials as reference compounds. Information regarding the drugs was obtained from DDrare, a database of drugs and their target information used in clinical trials of rare diseases (https://ddrare.nibiohn.go.jp/) (Fig. 1a). For the experimental validation of the technique, we selected a known SYK inhibitor candidate BIIB-057 as reference on the basis of target information in DDrare. This screening also had the meaning "prospective study", because BIIB-057 had not been approved and we aimed to examine whether alternatives could be obtained by scaffold-hopping. Then, for the detailed analysis of the characteristics of the screening, we chose 5 compounds (i.e., aldosterone, testosterone, sildenafil, sunitinib and celecoxib) from the compounds registered with both DDrare and Directory of Useful Decoys, Enhanced (DUD-E), as reference (Fig. 1a, b). DUD-E is a database of useful decoys designed to help benchmark molecular docking programs[20]. This analysis was regarded as "retrospective study", as the reference compounds were approved drugs. Based on these 6 compounds, a search for DR candidate compounds was performed using AI-AAM (Fig. 1a). In the chemical library, 44,503 compounds were preprocessed successfully and subjected to screening by AI-AAM, among which 1251 compounds (neutral compounds, 808; monovalent cations, 443) had target information in DrugBank (https://go.drugbank.com/) and were used for comparative analysis of target (Fig. 1c). For both the reference and candidate compounds, AAM descriptors, which describe the set of interactions between amino acids and the compound, were calculated and the compounds with similar AAM descriptors were screened from compound libraries and identified as hits (Fig. 1d, see "Calculations of AAM descriptors" in the "Methods" for more details). The hits were analyzed and validated in terms of comprehensiveness and specificity.

Experimental validation for a hit compound identified with AI-AAM

As mentioned above, we selected a known SYK inhibitor candidate BIIB-057 as reference to discover other lead compounds. SYK is a non-receptor tyrosine kinase associated with many RIDs and is considered to be a worthy drug target. With BIIB-057 as the reference, 18 compounds with similar AAM descriptors were identified, and one of them, XC608, which had the highest AAM similarity score to BIIB-075 and a scaffold that differed from the reference, was selected to evaluate the levels of inhibitory activity for the target, SYK (Fig. 2a). The IC50 values for BIIB-057 and XC608 were 3.9 nM and 3.3 nM, respectively. These values are close to each other, suggesting that XC608 inhibits SYK activity as effectively as BIIB-057. The purity of BIIB-057 and XC608 measured by High Performance Liquid Chromatography (HPLC) was 100% and 96%, respectively (see Supplementary Fig. S1 online). These results indicate that the compounds predicted to have similar pharmacological activity to their reference compound by AI-AAM have also experimentally obtained IC50 values, which represent the potency of the inhibitor, close to the reference.

Next, kinase profiling of BIIB-057 and XC608 was performed to examine their selectivity. Among 24 kinases examined, 2 and 14 were inhibited at least 50% by BIIB-057 and XC608, respectively (Fig. 2b). These results suggest that BIIB-057 inhibits SYK and PAK5 selectively, while XC608 inhibits many kinases including SYK with a lower selectivity.

These results showed that the screening setting the threshold value of AAM similarity scores for 0.7 provided the compounds with different structure and selectivity than the reference, maintaining the pharmacological activity.

Comprehensiveness of screening using AI-AAM

A total of 1275 compounds comprising 25–488 new compounds per reference compound, were obtained as hits. When the compounds were limited to those registered in DrugBank, there were 144 compounds identified in total, with 2–70 based on each reference (Table 1. For chemical structures of representative compounds, see Supplementary S2–S6 on line). The compounds registered in DrugBank have target information, and Table 1 shows the result of aggregation according to whether their known targets are the same as those of the reference compounds.

We estimated how comprehensively AI-AAM identified the compounds whose known targets were the same proteins as the reference compounds. After all compounds that were known to target the same proteins as the references were counted, we examined how many of these compounds were obtained as hits using AI-AAM (Table 1 "Extraction rate"). For aldosterone, testosterone, and sildenafil, the extraction rate was greater than 60%. For sunitinib, the extraction rate was 33.3% when the target was limited to KIT only, whereas it was 50% when it was defined as the percentage of the hits to any of the 8 targets including KIT. The reason for the calculation limited

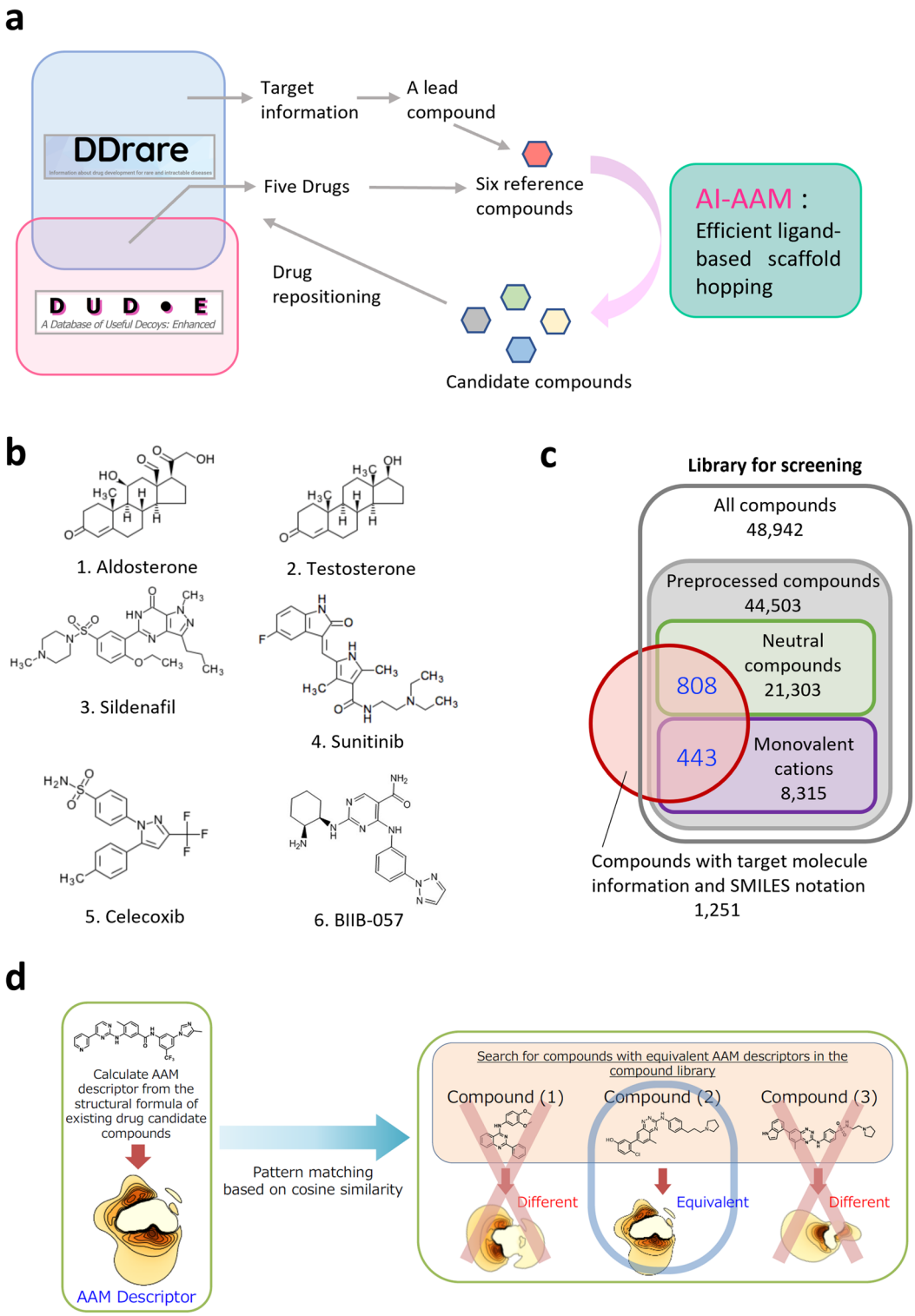

Figure 1. Schematic representation of this study. (**a**) A lead compound, based on the target information of DDrare, and five drugs contained in both Directory of Useful Decoys, Enhanced (DUD-E) and DDrare were selected as reference compounds. Chemical libraries were then explored by scaffold-hopping using AI-AAM. Identified compounds are candidates for drug repositioning for rare and intractable diseases. (**b**) The structural formulae of six reference compounds. (**c**) The range of the search for novel hit compounds. The neutral compounds and the monovalent cations including the five reference compounds constitute a part of the Namiki compound library set for repositioning (Namiki Shoji Co., Ltd.), which has 48,942 types in total. The number of the compounds preprocessed successfully was 44,503. Among these compounds, those that are also registered in the DrugBank and written in SMILES notation were selected. The number of these compounds is 1251 that consisted of 808 neutral and 443 monovalent cation compounds. Within the limits of these compounds, the search for novel hits was performed. (**d**) Schematic representation of virtual screening by AI-AAM.

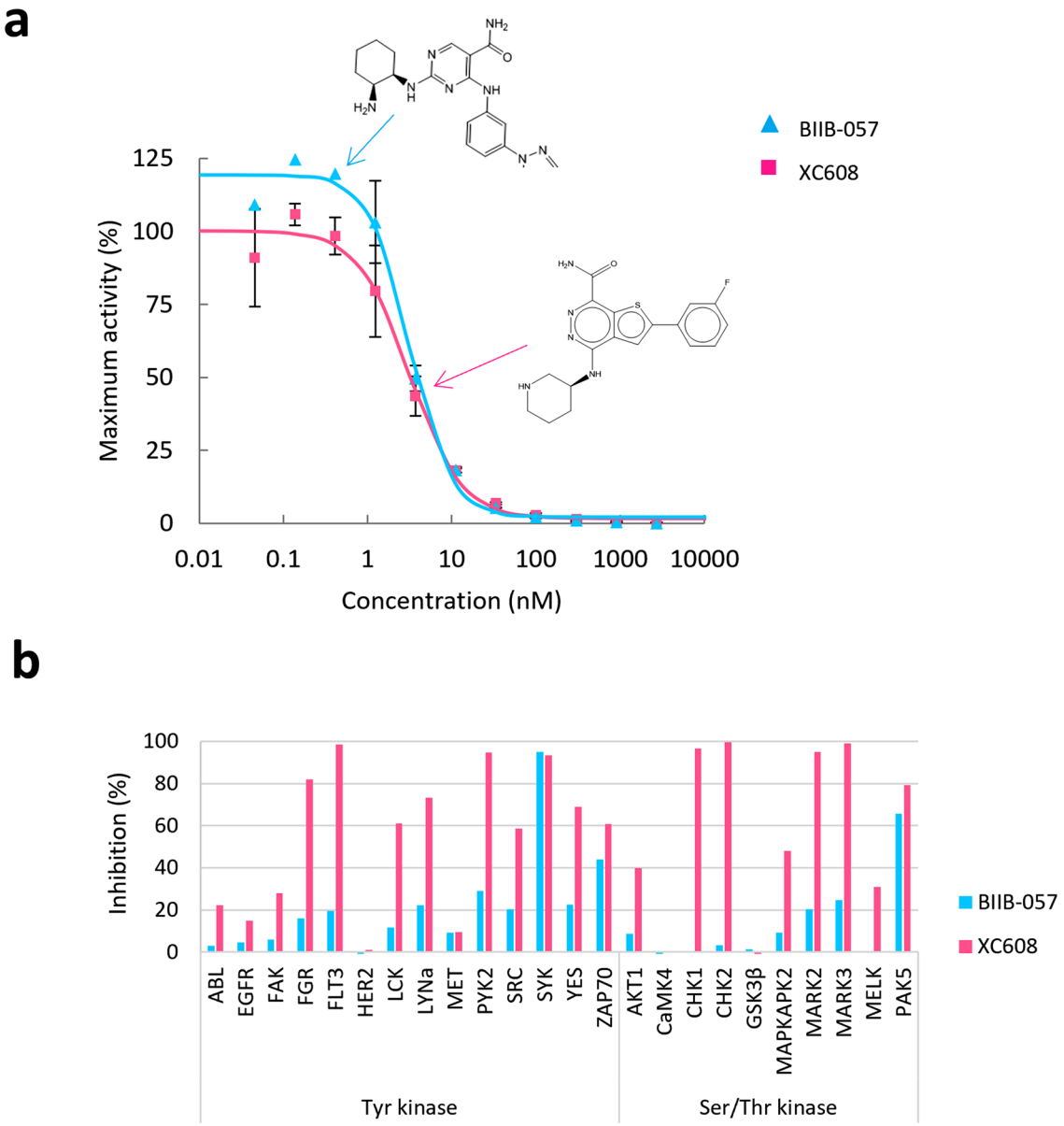

Figure 2. Comparison between BIIB-057, the reference, and XC608, the hit, in inhibitory activity for kinases. (**a**) SYK kinase activity dose–response curves using BIIB-057 and XC608. The kinase activity IC50 values for BIIB-057 and XC608 are 3.9 nM and 3.3 nM, respectively. (**b**) Kinase profiling of BIIB-057 and XC608 using mobility shift assay (ATP concentration: Km value of each kinase, Concentration of each compound: 50 nM).

Reference compound		Hit compounds by AI-AAM (n)		Reported compounds targeting the same protein (n) (A)	Extraction rate (B/A) (%)	EF value
Name	Target	Same target (B)	Different target			
Aldosterone	NR3C2	7	37	11	63.6	86.5
Testosterone	AR	18	52	24	75.0	36.3
Sildenafil	PDE5A	2	11	3	66.7	33.3
Sunitinib	KIT	2	13	6	33.3	
	Any target of sunitinib[a]	10	5	20	50.0	8.3
Celecoxib	PTGS2	2	0	18	11.1	95.2

Table 1. The reference compounds and the hit compounds, with the extraction rate and the enrichment factor (EF) value of the hit compounds known to target the same protein as its reference. [a]CSF1R, FLT3, PDGFR, RET, FLT1, KDR, FLT4, KIT.

to KIT was that the binding conformation of sunitinib was based on the complex with KIT (See "Preparation of compound conformation" in the "Methods"). As a whole, screening by using AI-AAM on the basis of 5 references detected 11–75% of known compounds whose targets were the same as those of the reference compounds.

Then, we calculated the enrichment factor (EF) (see "The enrichment factor" in the "Methods" for more details). The random hit rates were approximately 0.01–0.25%, while the predicted hit rates were 1–8%. The predicted hit rates were apparently lower than the "extraction rates" shown in Table 1 because the denominators contained not only compounds that target the same proteins as references, but also those that target different proteins or have no target-molecule information. As shown in Table 1, the hit rate was improved by approximately 10–100 times (see Supplementary Fig. S7 online).

Binding free energies between compounds and target proteins

For both the reference compounds and the hit compounds, the free energy of compound binding to their target was compared. For each reference compound, we chose a compound with the highest AAM similarity among the hits whose known targets were the same as or different from those of the reference, and designated them "cmpd. (same target)" or "cmpd. (different target)" respectively. Then we calculated the compound-target binding free energy by Fragment Molecular Orbital (FMO) method. As a result, the compound-target binding energies of "cmpd. (same target)" were almost equal to those of "cmpd. (different target)" for all systems. On the other hand, there was a 10–20 kcal/mol energy difference between the reference compounds and "cmpd. (same target)" (see Supplementary Fig. S8 and Table S1 online). One of the reasons for the difference might be that the structures of the target proteins were determined using reference compound—target cocrystal structures and it was optimized for the reference compound. Or, it might be simply because "cmpds. (same target)" had lower activity than the reference compounds. In any case, these findings suggest that the binding energies of hits whose known targets were different from those of the reference were not higher than those of the hits targeting the same proteins as the reference.

Structure of the compounds identified using AI-AAM

The structural similarity of hits to the reference compounds was examined. In a graph representing AAM similarity on the vertical axis and Tanimoto coefficient on the horizontal axis, the hits were plotted (Fig. 3). For the compounds identified using aldosterone as a reference, Tanimoto coefficients were within the range of 0.1–0.6 (Fig. 3a). These values are rather low, which means that the hits include those with low structural similarity to the reference. Even for compounds whose known targets are the same as those of the reference compounds, Tanimoto coefficients are not always large. Moreover, most of the hits identified using sunitinib as a reference had a Tanimoto coefficient of less than 0.2 (Fig. 3b). In this way, the hits with low similarity could generally be identified, although there were some differences in the range of the Tanimoto coefficient depending on the reference compounds.

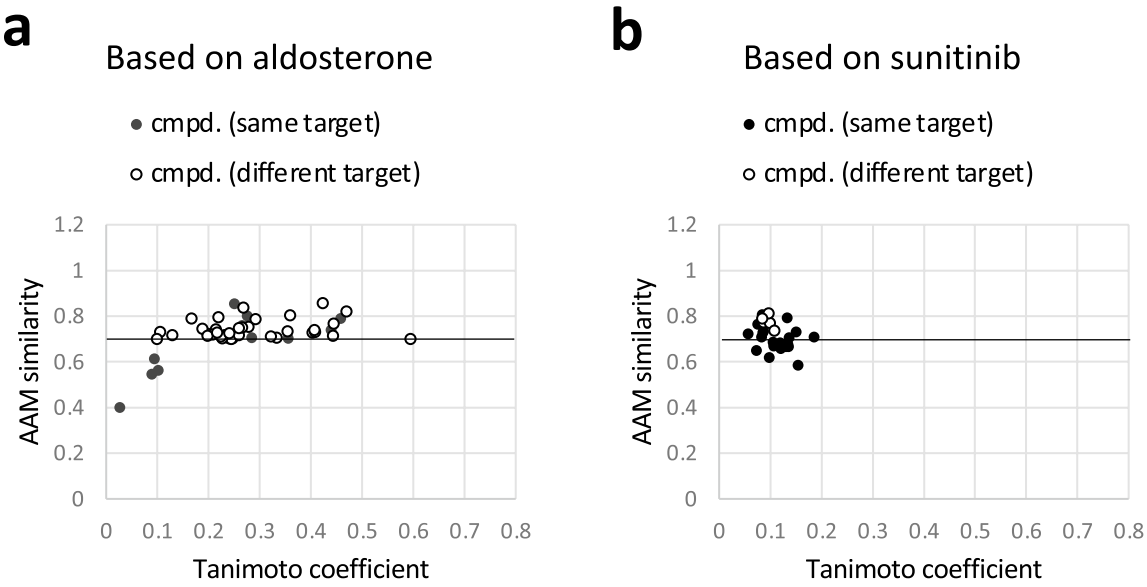

Figure 3. Tanimoto coefficients of hits and non-hits. Hits and non-hits based on aldosterone (**a** target: NR3C2) or sunitinib (**b** target: KIT). The threshold value of AAM similarity as the boundary of hits and non-hits is 0.7. The hits include many compounds that target different proteins than the reference as well as those which target the same proteins as the reference. Although the non-hits whose targets are the same protein as the references are also included, those known to target different proteins than the references are not shown.

Specificity of screening using AI-AAM

As shown in Table 1, there were compounds whose known targets were the same as those of the reference compound among the hits, but more compounds known to target proteins different from the reference were identified as hits. Moreover, the compound-target binding energies of the compounds of the former and the latter, with the highest AAM similarity among the hits respectively, were approximately equal to each other for all systems, suggesting that the latter also bind to the targets of their reference compounds (Fig. 4a). For example, among 44 hits identified with aldosterone as the reference, only 7 compounds were known to target NR3C2, the same protein targeted by the reference (Fig. 4b). In regards to the compounds screened on the basis of sunitinib, only 10 out of 15 hits were known to target any of the targets of sunitinib (Fig. 4b).

An analysis to investigate the reason why the known targets of many hit compounds were different from those of the reference compounds was conducted. Among the compounds known to bind to the same targets as the reference compounds, some have the same biological function as the references, while others function in a different manner. We assumed that, as AI-AAM identified the compounds with the same function as the references, the comprehensiveness was underestimated. To verify this hypothesis, the functions of hits and non-hits were analyzed (Fig. 5, Supplementary Table S2 online).

As shown in Fig. 5a[21], aldosterone is an agonist for nuclear receptor subfamily 3 group C member 2 (NR3C2, also known as MR). Eleven compounds are known to bind to NR3C2 in addition to aldosterone: 3 agonists and 8 antagonists. Seven hits were identified with aldosterone as the reference: 3 agonists and 4 antagonists. By using AI-AAM, agonists with the same function as the reference were obtained at a rate of 100%, while antagonists were identified at a rate of 50%.

Testosterone is an agonist for androgen receptor (AR) (Fig. 5b)[22]. In addition to testosterone, 24 compounds are known to bind to AR: 14 agonists, 9 antagonists and a modulator. The number of hits identified applying AI-AAM based on testosterone was 18, 14 of which were agonists. The agonists, which have the same function as the reference, were obtained at a rate of 100%, while antagonists were identified at a rate of 44.4%. All 6 compounds that were not identified by using AI-AAM, even though these were known to target AR, were antagonists and a modulator. Chi-square test of independence shows a statistically significant relationship ($p = 0.00162^{**}$) between the pharmacological action type (agonist or antagonist) and the hit rate (see Supplementary Table S3 online).

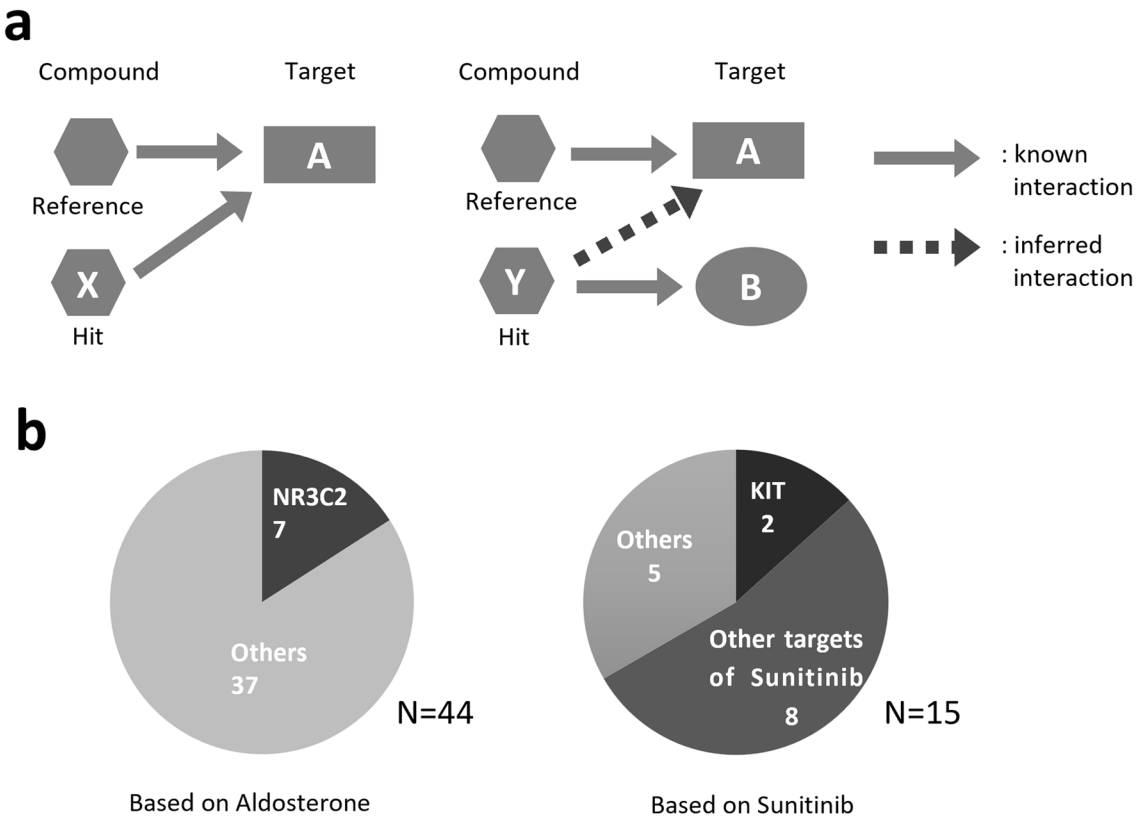

Figure 4. The targets of reference and hit compounds. (**a**) The reference and hit compounds and their target proteins. There are two types of hit-target interaction cases; cases where the known targets of the hit compounds are the same as those of the reference compounds and cases where the known targets of the reference and hit compounds differ. In the latter cases the targets of the reference compounds are presumed to be the unknown targets of hit compounds. (**b**) (Left) Hit compounds screened on the basis of aldosterone. Classification is based on the known targets: NR3C2 and others. (Right) Hit compounds screened on the basis of sunitinib. Classification is based on the known targets: KIT, other targets of sunitinib, and others.

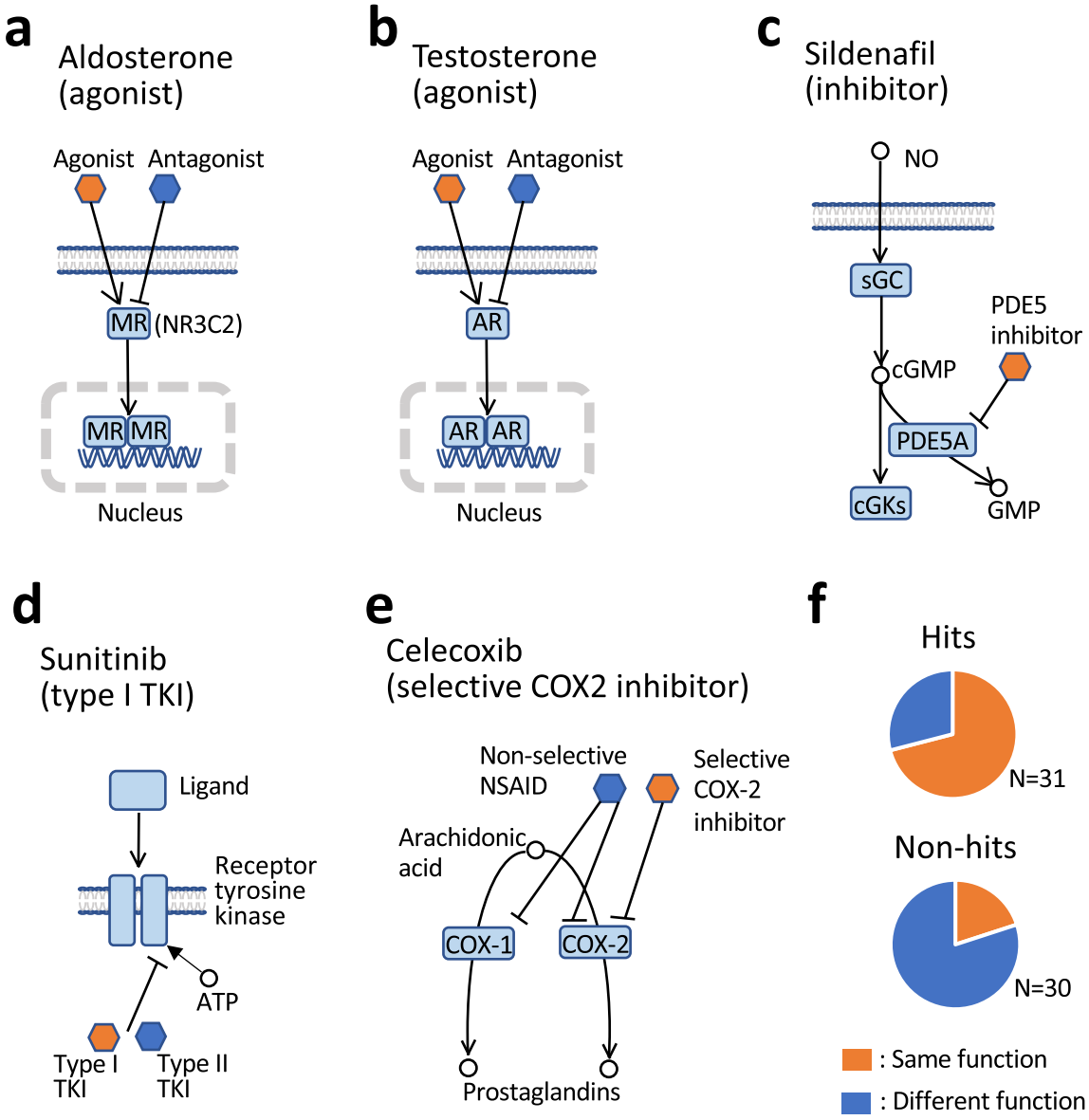

Figure 5. Pharmacological actions of reference compounds and the classification of hits and non-hits based on whether or not their function is the same as the reference. (**a**) Aldosterone targets NR3C2 as an agonist. NR3C2 is a ligand-activated nuclear transcription factor. NR3C2s in the inactive state reside primarily in the cytoplasm and are then transported to the nucleus, dimerize and form a transcription complex with DNA hormone response elements to initiate transcription of genes. The NR3C2 has affinity for both its primary physiological agonists and antagonists developed to treat diseases. (**b**) Testosterone targets the AR as an agonist. As with NR3C2, AR is nuclear transcription factor. Upon binding to endogenous ligands such as testosterone, AR translocates to the nucleus and regulate the transcription of genes. Apart from the endogenous ligands (agonists), exogenous ligands such as environmental chemicals and pharmaceuticals can also interact with AR as agonists or antagonists. (**c**) Sildenafi targets PDE5A. NO mediates its biological effects by activating sGC and increasing cGMP synthesis. As cGMP is degraded by PDE5A, its levels are maintained by inhibition of PDE5A by, for example, sildenafil. (**d**) Sunitinib is a type I tyrosine kinase inhibitor (TKI). Several kinases responsible for cell growth and proliferation are hyperactivated in various tumors. TKIs are the largest group of kinase-inhibiting small molecules. Most of the compounds, including sunitinib, act by blocking the ATP-binding site of the target molecule and the binding modes are classified as type I and type II depending on whether the compounds bind competitively with ATP using the 'DFG-in' (type I) conformation or the 'DFG-out' (type II) conformation. (**e**) Celecoxib targets COX-2 (gene symbol: PTGS2). NSAIDs (non-steroidal anti-inflammatory drugs) inhibit the enzyme cyclooxygenase (COX), which mediates the conversion of arachidonic acid to inflammatory prostaglandins. COX enzyme can exist in two forms: COX-1, the constitutive isoform; or COX-2, the inducible isoform. Selective COX-2 inhibitors are a subclass of NSAIDs that have a much greater affinity for the COX-2 enzyme, whereas non-selective NSAIDs inhibit both COX-1 and COX-2. (**f**) The ratios of the compounds with functions that are same as or different from reference compounds to hits or non-hits, which is based on the summary of the compounds identified using each of 4 compounds (i.e., aldosterone, testosterone, sunitinib and celecoxib).

Although phosphodiesterase 5A (PDE5A) inhibitors (Fig. 5c)[23] identified using sildenafil as a reference were not classified according to their functions, their hit rate was originally high. Tyrosine kinase inhibitors (Fig. 5d)[24] such as sunitinib could be classified into type I and type II[25,26]. Although type I inhibitors, with the same function as sunitinib, were identified more often than type II, this was not statistically significant.

Non-steroidal anti-inflammatory drugs (NSAIDs), such as celecoxib, inhibit cyclooxygenase (COX) and are classified into selective COX-2 inhibitors and non-selective NSAIDs (Fig. 5e)[27,28]. In addition to celecoxib, there are 7 compounds known to be selective for COX-2 (approved nomenclature for gene symbol: PTGS2). The number of non-selective NSAIDs was 10. Two hits identified using AI-AAM and celecoxib as the reference were selective inhibitors, and no non-selective NSAIDs were screened. Non-hit compounds included a cyclooxygenase-inhibiting nitric-oxide (NO) donator (CINOD) in addition to selective COX-2 inhibitors and non-selective NSAIDs. Moreover, there was a significant difference between the averages of AAM similarity of selective COX-2 inhibitors and non-selective NSAIDs regardless of hit or non-hit status (t-test, $p = 0.0137$*, see Supplementary Fig. S9 online).

The summary of hit and non-hit compounds identified using each of these 4 compounds (i.e., aldosterone, testosterone, sunitinib and celecoxib) as reference is shown in Table 2 and Fig. 5f. All compounds in this table target the same proteins as their reference compounds. Each hit and non-hit was further classified based on whether or not it had the same pharmacological mechanism of action as the reference. Chi-square test of independence showed a statistically significant relationship between two categorical variables: the result of screening (i.e., hits or non-hits) and function ($p = 0.000065$**). The screening using AI-AAM was highly selective of function.

Possibility of novel drug-target interactions suggested by AI-AAM

As mentioned above, there were compounds among the hits whose known targets were the same as those of the reference compound, while many compounds without information regarding whether they bind to the same protein as the references were also identified as hits (Table 1). We examined what kind of molecules were the targets when hits were known to target different proteins than the reference compound. As shown in Fig. 6, hits were classified on the basis of the biological functions of their known targets (i.e., nuclear receptor, enzyme, GPCR, and ion channel). The target information was obtained from DrugBank, and the target proteins were classified according to KEGG BRITE[29].

The number of hits identified using AI-AAM with aldosterone as the reference was 44, with 37 known to target proteins other than NR3C2 (DrugBank, KEGG) (Fig. 6a). However, many of the known targets of the 37 compounds belonged to the same nuclear receptor family as NR3C2.

The number of hits screened with testosterone as the reference was 70, 52 of which were known to target different proteins other than AR (Fig. 6b). Most of these targets belonged to the same nuclear receptor family as AR, the targets of testosterone.

Sildenafil is known to target a hydrolase, PDE5A. Although 61% of the 13 hits obtained with sildenafil as the reference were known to target enzymes including hydrolases, the known targets of the remaining 31% and 8% of hits were GPCRs and ion channels, respectively (Fig. 6c). The known targets of sunitinib are some tyrosine kinases, and 67% of the 15 hits identified with sunitinib as reference were known to target the same tyrosine kinases as sunitinib. However, 20% and 13% of the hits were known to target other tyrosine kinases and GPCRs, respectively (Fig. 6d).

These results, together with those of other studies, indicate that many compounds identified through hits target multiple proteins of both the same and different families (see Discussion for more details).

Discussion

In drug discovery, it is needed to obtain the improved compounds with excellent function and reduced side effects. To promote drug development for RIDs, we developed a new method of virtual screening, integrating AAM into scaffold-hopping, a LBVS technique. By applying this method to 5 compounds in DDrare, many hit compounds with diverse structures and the same affinity for a given target were obtained. As the EF values are equal to or greater than those of many SBVS techniques[30–32], AI-AAM can be considered as the LBVS method to find the various compounds that target the same protein as the references with equal efficiency to SBVS methods. Moreover, our results show that XC608 identified with BIIB-057 as the reference has pharmacological activity equal to the reference. As well, for the compounds screened based on 5 compounds in DDrare, those with the same function as the reference tended to dominate the hits. This tendency was statistically significant regarding the compounds based on testosterone (target: AR) and celecoxib (target: PTGS2) (see Supplementary Table S3 and Fig. S9 online). Singam et al.[33] reported that agonists and antagonists of AR exhibit distinct binding modes: agonists form an H-bond with either Thr877 or Asn705, while this interaction is absent for antagonists. Other studies[28,34] have reported that three amino acid differences between the COX-2 and COX-1 (gene symbol: PTGS1) active sites have major implications for the selectivity profile of inhibitors. In this way, pharmacological action is

	Same function	Different function	SUM
Hits	22	9	31
Non-hits	6	24	30
SUM	28	33	61

Table 2. Contingency table for function data with row and column totals.

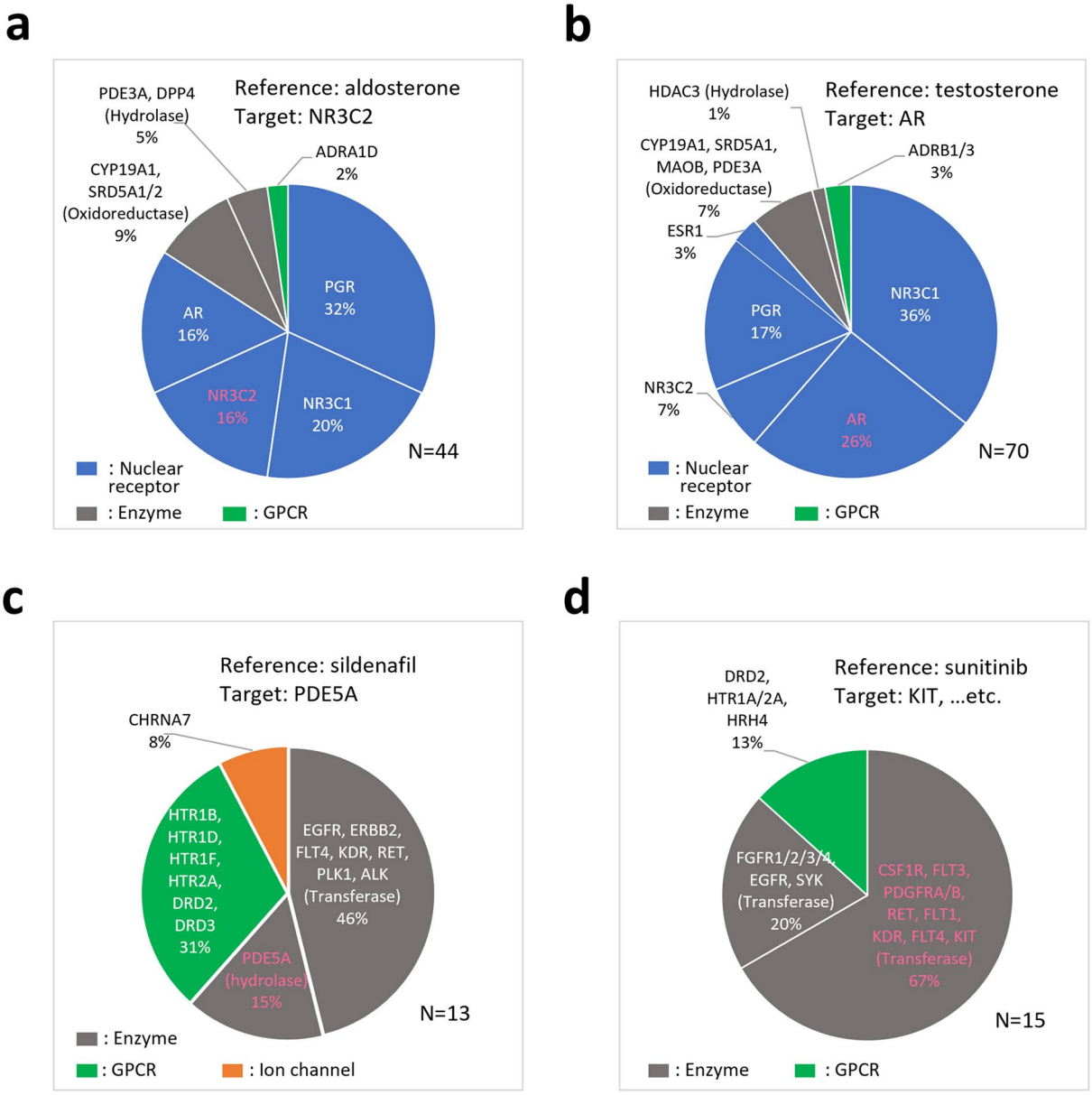

Figure 6. Hit compounds classified on the basis of the biological functions of their known targets. Hit compounds screened on the basis of each reference compound (**a** aldosterone, **b** testosterone, **c** sildenafil, **d** sunitinib). Classification is based on the biological functions of their known targets. Pink letters represent the same targets as the reference compounds.

presumed to be closely related to the binding site of the target, especially some amino acids, and the differences in the function of the drugs are likely related to AAM similarity, as AI-AAM mainly considers the interaction between the candidate compound and each of the 20 amino acids. In other words, our results suggest that AI-AAM describes the interaction accurately, focusing on the essential qualities of interaction. In this way, if their known targets are the same as those of the reference compounds, the compounds likely are similar in function to the reference compounds, unlike those searched only on the basis of the target information in the databases. Moreover, even for such compounds, the structures are not always similar to the reference compounds, contributing to the expanded pharmacological space and the possibility of drug improvement by "hopping" from one scaffold to another[16].

However, for the hits identified using 5 compounds as the reference, more compounds had known targets that differed from the targets of the reference compounds than those that targeted the same protein as the reference compounds. This suggests that the former also binds to the targets of reference compounds. The results of the calculation of the compound-target free energy also underscore the inference. Even if the known targets of the hits are different from those of the reference compounds, in many cases these belong to the same gene family. For example, all NR3C1 (GR), NR3C2, NR3C3 (PGR), and AR belong to the nuclear receptor family and have the same composition of functional domains, one of which, the steroid-binding domain (LBD), has sequence

conservation to a certain degree[35]. Therefore, there may be some compounds active against multiple proteins of this family. However, the compounds that were inferred to interact with proteins of a different family were also not negligible. For example, the results showed that 2 of 15 hit compounds identified with sunitinib (target: receptor tyrosine kinases) as the reference were known to target aminergic GPCRs (i.e., DRD2, HTR2A, HTR1A, and HRH4); for 9 of 17 hits with BIIB-057 (target: SYK) as the reference, the known targets were GPCRs including aminergic GPCRs (i.e., HRH1, HTR1D, HTR1B, DRD2) (see Supplementary Fig. S10 online)[36]. This corresponds to Paolini's report that a quarter of all of the compounds with multitarget activity (known as promiscuous compounds) are active across different gene families and aminergic GPCRs and protein kinases exhibit the greatest intra- as well as inter-gene family promiscuity[37]. Taken together, these results indicate the possibility of novel interaction between compounds and proteins, leading to multi-target information. The percentage of promiscuous compounds to the whole is reported to be approximately 20%[38,39]. However, we believe that there are still many unknown interactions, as many ChEMBL data regarding the activity of compounds against targets are not yet reflected in the target information of databases such as DrugBank[40,41]. The kinase assay in our study also showed that XC608 targets many kinases other than SYK, although the target of XC608, identified with BIIB-057 as the reference, was inferred to be merely SYK using AI-AAM, and it was validated by the experiment. It is probably another example to show the successful scaffold-hopping that the compounds with various sets of targets were screened on the basis of each reference compound.

In our previous study[42], we invented a score R_{gene} for disease pairs sharing drug targets in RIDs, which represent a common mechanism of drug action underling drug repositionability. If a hit becomes known to share a target with the reference compound, the value of R_{gene} will rise. This means that the degree of drug repositionability between the indications of a hit and its reference are higher than that before the application of AI-AAM. The known target of a hit, prednisolone, screened with testosterone (target: AR) as reference is not AR, but NR3C1 (GR). However, one of the indications of prednisolone is muscular dystrophy, which is also known to be that of testosterone. Although AR is expressed at high levels in muscle[43,44], in the reports about corticosteroids for the treatment of Duchenne muscular dystrophy (DMD), this was not mentioned and the authors reported that the precise mechanism by which corticosteroids increase strength in DMD is not known[45,46]. To confirm the possibility that prednisolone has pharmacological effects via the target of the reference compound, there is a need for further studies, but this may be an example of retrospective validation of this technique for DR. Recent approaches linked to network biology, the so-called 'Network Pharmacology" are moving away from the current 'one disease-one target-one drug paradigm of drug discovery' that is becoming increasingly inefficient[47]. These approaches simultaneously target two or more proteins within disease-associated protein-networks[48–50]. The multi-target information that can be accumulated by our method will serve to clarify the mechanisms of action of drugs and, consequently, the disease mechanisms.

Methods

Reference compounds

For experimental validation of the technique

DDrare, a database of Drug Development for Rare diseases, was searched to identify proteins that are related to systemic lupus erythematosus (SLE), and SYK was finally selected as the target protein of the study. SLE is known as a RID, and there is a high need for new drugs to treat it. SYK is a non-receptor tyrosine kinase and was found to be related to many RIDs including SLE. Starting from a known SYK inhibitor candidate BIIB-057[51], we aimed to discover other lead compounds.

For detailed analysis

DUD-E database, a database of 22,886 active compounds and their affinities against 102 targets[52], was searched to find compounds that are also included in DDrare, and nine compounds were found. Four of them were removed because they were not covered by AI-AAM, such as a nucleic acid analogue (interaction with nucleic acids is important rather than with amino acids) and unsaturated fatty acid (too flexible). The remaining five compounds were used here as the reference compounds: aldosterone, testosterone, sildenafil, sunitinib and celecoxib.

Chemical library for virtual screenings

For both the prospective and the retrospective studies, we used a commercial library provided by NAMIKI SHOJI Co., Ltd. (https://www.namiki-s.co.jp/), which is composed of biologically active compounds including those in the clinical trial phase and approved drugs. All compounds in the library were preprocessed (desalted and normalized) by the MolStandardize module of RDKit v.2018.09.1, and 44,503 compounds preprocessed successfully were subjected to screening by AI-AAM.

Calculation of AAM descriptors

AAM descriptors of each compound were calculated as distribution of centers of mass of amino-acid "probes" around the compound by using molecular dynamics (MD) simulation. However, its high computational cost makes it impractical for use in large-scale virtual screenings. Here we employed deep learning techniques to accelerate those calculations. The computational details are as follows:

Preparation of compound conformations

Experimentally determined binding conformations were used for calculating AAM descriptors of reference compounds except for BIIB-057. Cocrystal structures were downloaded from RCSB[53] in PDB format under the PDB ID 2AA2 (aldosterone), 2AM9 (testosterone), 1UDT (sildenafil), 3G0E (sunitinib) and 3LN1 (celecoxib) and 3D structures of the ligands were extracted from the PDB files. Binding conformations of BIIB-057 and

library compounds were unknown, and thus we considered 100 conformations generated by Discovery Studio 2020 (BIOVIA). Charge states at pH = 7.0 of all compounds including the reference compounds were also predicted by Discovery Studio.

Amino acid probes
In this study, not all 20 natural amino acids were considered but some were selected depending on the charge states of the reference compounds to reduce computational costs: asparagine, cysteine (deprotonated), phenylalanine, and threonine for monocationic reference compounds, and histidine (having protons both on the epsilon and delta nitrogen) was added for neutral reference compounds. The detailed method is mentioned in "Calculation of AAM similarity scores" section. To clearly highlight differences in AAM descriptors among amino acids, we removed their backbone chains and employed only side chains as the amino-acid probes.

Force fields
The generalized AMBER force field 2 (GAFF2)[54] was employed to describe atomic interactions of reference and library compounds. Partial atomic charges were obtained by the restrained electrostatic potential (RESP) method[55]. The antechamber[56] in the AmberTools19 package was used both to assign atom and bond types and to calculate the charges. For the RESP fit, electrostatic potentials of all the compounds were obtained from single-point quantum calculations at the HF/6-31G* level with the conformations optimized by the PM6 method. All of the quantum calculations were carried out by using Gaussian 16[57]. For amino acid probes, the ff14SB force field was employed. As a water model, TIP3P was used.

Calculations of AAM descriptors
In the beginning, we calculated AAM descriptors of various compounds as training data for deep learning by MD calculation. All MD calculations were performed using the gromacs-5.1.5 package[58]. As initial structures for the simulations, we considered pair structures of a compound and an amino acid probe which was placed randomly in the vicinity of the compound. A water box was created by tleap in AmberTools19 package where a minimum distance between any atom in the two molecules and an edge of a periodic box was set to 8.0 Å. A total of 100 pair structures were generated per compound, and all the structures were optimized and equilibrated with Berendsen thermostat and barostat. Here the coordinates of the compound were fixed, temperature T = 300 K, and pressure $P = 1$ atm. After a 100 ns production run with Nosé-Hoover thermostat and Parrinello-Rahman barostat, the AAM descriptor was calculated as an average distribution of a center of mass of the amino acid probe over 100 pair structures by using in-house software.

To learn AAM descriptors calculated by MD simulation, we adopted the pix2pix-type generative adversarial network model (https://doi.org/10.48550/arXiv.1611.07004). The original pix2pix was for 2D image data, and thus we extended it for 3D distribution data such as AAM descriptors. As explanatory variables, intermolecular potential energy surfaces (PES) between each probe atom and compounds were calculated. Note that PES calculations can be performed analytically and are therefore very fast. We confirmed that the obtained predictor can predict AAM descriptors of various compounds with sufficient accuracy. Here the number of training data (the number of compounds used for generating training data) was 100.

Calculations of AAM similarity scores

AAM similarity scores were defined as the cosine similarities between AAM descriptors of the reference compounds and library compounds, and were calculated by the following process. We first defined a tensor of a reference compound for the AAM similarity score calculation as $g_a(\boldsymbol{r})\theta(\boldsymbol{r})$, where $g_a(\boldsymbol{r})$ is an AAM descriptor of an amino acid probe a ($a = 1 \sim N_a$ is an index of amino acids, and $N_a = 4$ or 5) for a reference compound, $\boldsymbol{r}$ is a coordinate where the center of mass of the reference compound is taken as the origin, and

$$\theta(r) = \begin{cases} 1\left(|\boldsymbol{r}| \le 10\ [\text{Å}]\right) \\ 0 \quad \left(|\boldsymbol{r}| 10\ [\text{Å}]\right) \end{cases} \tag{1}$$

Now, the cosine similarity between AAM descriptors of the reference compound and a library compound can be defined as follows:

$$\cos(R,\boldsymbol{s}) \equiv \frac{1}{N_a} = \sum_{a=1}^{N_a} \frac{\iiint d^3\boldsymbol{r}\left[G_a(\boldsymbol{r}) - G_a^{(1)}(\boldsymbol{s})\right]\left[g_a(R(\boldsymbol{r}+\boldsymbol{s})) - g_a^{(1)}(\boldsymbol{s})\right]\theta(\boldsymbol{r}+\boldsymbol{s})}{\sqrt{G_a^{(2)}(\boldsymbol{s}) - G_a^{(1)}(\boldsymbol{s})G_a^{(1)}(\boldsymbol{s})}\sqrt{g_a^{(2)} - g_a^{(1)}g_a^{(1)}}\iiint d^3\boldsymbol{r}\theta(\boldsymbol{r}+\boldsymbol{s})} \tag{2}$$

where $G_a(\boldsymbol{r})$ is an AAM descriptor of an amino acid a for the library compound, $\boldsymbol{s} = (X, Y, Z)$ is a translation vector, $R = R(\alpha, \beta, \gamma)$ is a rotation matrix and $\alpha \sim \gamma$ are Euler angles, and

$$G_a^{(1)}(\boldsymbol{s}) \equiv \frac{\iiint d^3\boldsymbol{r}G_a(\boldsymbol{r})\theta(\boldsymbol{r}+\boldsymbol{s})}{\iiint d^3\boldsymbol{r}\theta(\boldsymbol{r}+\boldsymbol{s})} \tag{3}$$

$$G_a^{(2)}(\boldsymbol{s}) \equiv \frac{\iiint d^3\boldsymbol{r}G_a(\boldsymbol{r})G_a(\boldsymbol{r})\theta(\boldsymbol{r}+\boldsymbol{s})}{\iiint d^3\boldsymbol{r}\theta(\boldsymbol{r}+\boldsymbol{s})} \tag{4}$$

$$g_a^{(1)} \equiv \frac{\iiint d^3\boldsymbol{r}g_a(\boldsymbol{r})\theta(\boldsymbol{r})}{\iiint d^3\boldsymbol{r}\theta(\boldsymbol{r})} \tag{5}$$

$$g_a^{(2)} \equiv \frac{\iiint d^3\boldsymbol{r} g_a(\boldsymbol{r}) g_a(\boldsymbol{r}) \theta(\boldsymbol{r})}{\iiint d^3\boldsymbol{r} \theta(\boldsymbol{r})} \tag{6}$$

$\boldsymbol{s}$ and R were determined to satisfy the following condition by grid search with a step size of 1 Å for X, Y, and Z and 10 degrees for α, β and γ:

$$\left(\text{AAM similarity score}\right) = \max_{(R,\boldsymbol{s})} \{\cos(R, \boldsymbol{s})\} \tag{7}$$

However, this requires high computational cost because the integration over $\boldsymbol{r}$ must be done $N_a \times N_R \times N_{\boldsymbol{s}}$ times (N_R and $N_{\boldsymbol{s}}$ are the numbers of grids for R and $\boldsymbol{s}$). In a practical calculation, we reduced it by applying the singular value decomposition (SVD) method. Substituting

$$g_a(R(\boldsymbol{r}+\boldsymbol{s})) - g_a^{(1)}(\boldsymbol{s}) = \sum_{\mu=1}^{N_\mu} V_{a\mu}(R) \Xi_{a\mu} U_{a\mu}(\boldsymbol{r}+\boldsymbol{s}) \tag{8}$$

into Eq. (2), the following expressions can be obtained:

$$\cos(R, \boldsymbol{s}) = \frac{1}{N_a} \sum_{a=1}^{N_a} \sum_{\mu=1}^{N_\mu} V_{a\mu}(R) \Xi_{a\mu} \lambda_{a\mu}(\boldsymbol{s}) \tag{9}$$

$$\lambda_{a\mu}(\boldsymbol{s}) \equiv \frac{\iiint d^3\boldsymbol{r} \left[G_a(\boldsymbol{r}) - G_a^{(1)}(\boldsymbol{s})\right] U_{a\mu}(\boldsymbol{r}+\boldsymbol{s}) \theta(\boldsymbol{r}+\boldsymbol{s})}{\sqrt{G_a^{(2)}(\boldsymbol{s}) - G_a^{(1)}(\boldsymbol{s}) G_a^{(1)}(\boldsymbol{s})} \sqrt{g_a^{(2)} - g_a^{(1)} g_a^{(1)}} \iiint d^3\boldsymbol{r} \theta(\boldsymbol{r}+\boldsymbol{s})} \tag{10}$$

where a cutoff N_μ was determined so that a summation of explained variance ratios exceeded 0.99. It is clear from these expressions that the number of integrations over $\boldsymbol{r}$ was decreased from $N_a \times N_R \times N_{\boldsymbol{s}}$ to $N_a \times N_\mu \times N_{\boldsymbol{s}}$.

We calculated AAM similarity scores of all pairs of the reference compounds and library compounds, and extracted compounds whose AAM similarity scores were > 0.7 as candidate compounds.

By the way, AAM descriptors and AAM similarity scores were also calculated to select some amino acids as probes described in "Amino acid probes" section. At the beginning of this operation, we calculated AAM descriptors for each reference compound, using each of 20 amino acid probes that carried different electric charges. Then, AAM similarity scores of all the pairs of two amino acid probes were calculated. With respect to each pair of two amino acid probes, the AAM similarity scores for all the reference compounds were added and averaged. With this distance matrix, hierarchical clustering was performed. As a result, we got some clusters of amino acid probes, and calculated the average of AAM descriptors of the amino acid probes for each cluster. Then, from each cluster, we chose an amino acid probe whose AAM descriptor was proximate to the average: asparagine, cysteine (deprotonated), phenylalanine, and threonine for monocationic reference compounds, and histidine (having protons both on the epsilon and delta nitrogen) was added for neutral reference compounds.

Calculations of Tanimoto coefficients

Morgan fingerprint with radius 2 and 2048 bits was used for calculations of Tanimoto coefficients between library and reference compounds. All the calculations were done by using RDKit v.2018.09.1.

The enrichment factor (EF)

EF values are commonly used in virtual screening evaluation as accuracy metrics. The EF value is defined as the ratio between the predicted hit rate and the random hit rate[30,31]. At first, the hit rate of the compounds known to target the same proteins as their reference was estimated when AI-AAM was not applied. To calculate the hit rate, the number of compounds that were known to target the same proteins as the reference and were contained in the subset of the NAMIKI library, which consists of the compounds having the same electrical charge as the reference, was counted. This was then divided by the total number of the compounds contained in the library ("random hit rate"). Subsequently, the hit rate of the compounds whose known targets were the same as those of the reference compound was estimated when AI-AAM was applied. The value was the number of the hit compounds known to target the same proteins as the reference divided by the total number of hits ("predicted hit rate"). EF values were calculated by dividing the "predicted hit rate" by the "random hit rate" for each reference compound.

In vitro SYK kinase assay

The experiments to evaluate the inhibitory activity levels of compounds were carried out according to the manufacturer's instructions on the SYK Kinase Enzyme System (Promega, Wisconsin, USA) and ADP-Glo™ Kinase Assay (Promega). By measuring luciferase activity using the Ensight Multimode Plate Reader (PerkinElmer), the inhibition rate was calculated by comparing the OD value to the negative and positive control wells. BIIB-057 and XC608 were analyzed using an Shimazu Prominence HPLC system equipped with Shiseido capcell pak C18 UG120 column (5 µm, 4.6 mm × 150 mm), reversed phase HPLC column eluting with a solvent gradient A:B; where A = 0.1% formic acid and 10 mM ammonium acetate in H_2O and B = 0.1% formic acid and 10 mM ammonium acetate in CH_3CN/H_2O = 95/5 (Gradient: as follows, Detection: UV 254 nm) at a flow rate 1 mL/min. The column temperature was kept at 40 ℃. The gradient was from 5 to 100% (20 min gradient % of B and 5 min isocratic hold).

Mobility shift assay

Kinase profiling of BIIB-057 and XC608 against a panel of 24 kinases was performed to examine the selectivity by using mobility shift assay at Carna Biosciences, Inc. (Kobe, Japan) (https://www.carnabio.com/english/product/msa.html). Drug concentrations were both set to 50 nM, and ATP concentrations were approximately equal to the K_m value for each kinase.

Calculation of binding free energy

The fragment molecular orbital (FMO) method was employed to calculate binding energies between compounds and target proteins. As the complex structures of reference compounds and their target proteins, cocrystal structures (PDB ID: 2AA2, 2AM9, 1UDT, 3G0E, and 3LN1) were used. The complex structures of library compounds were created as follows: first, the compounds were translated and rotated with R and $\boldsymbol{s}$ obtained from the processes of calculating AAM similarity scores (see "Calculations of AAM similarity scores" in the "Methods"); second, the obtained structures were combined with protein structures extracted from the cocrystal structures of the corresponding reference compounds. All complex structures were improved by optimizing conformations and positions of the compounds using the DFTB method (200 steps) with GAMESS v2020.2[59]. Here the structures of the target proteins were kept fixed. FMO calculations (MP2/6-31G* level of theory) were then performed based on the improved complex structures. Here the target proteins were fragmented into amino acids, and solvent effects (water) were included using the PCM method. The binding energies were calculated by subtracting the sum of energies of the compound alone and the target protein alone from the energy of the complex structure.

Statistical analysis

All statistical analyses were performed with Excel. To identify the relationship between two categorical variables, Chi-square test was performed. To calculate the probability of significant difference between two groups, two-tailed t-test (unpaired) was performed. $^*P < 0.05$ were considered as statistically significant. Before the t-test, F-test was performed to determine if the two samples had equal variance.

Data availability

The datasets and the chemical structures written in SMILES notation analyzed during the current study are available from the corresponding author on reasonable request.

Received: 16 June 2023; Accepted: 3 November 2023
Published online: 08 November 2023

References

1. Nguengang, W. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
2. Blin, O., Lefebvre, M.-N., Rascol, O. & Micallef, J. Orphan drug clinical development. *Therapie* **75**, 141–147 (2020).
3. Cremers, S. & Aronson, J. K. Drugs for rare disorders. *Br. J. Clin. Pharmacol.* **83**, 1607–1613 (2017).
4. Pushpakom, S. *et al.* Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
5. Jin, G. & Wong, S. T. C. Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines. *Drug Discov. Today* **19**, 637–644 (2014).
6. Delavan, B. *et al.* Computational drug repositioning for rare diseases in the era of precision medicine. *Drug Discov. Today* **23**, 382–394 (2018).
7. Swamidass, S. J. Mining small-molecule screens to repurpose drugs. *Brief Bioinform.* **12**, 327–335 (2011).
8. Ekins, S., Mestres, J. & Testa, B. In silico pharmacology for drug discovery: Methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* **152**, 9–20 (2007).
9. Gimeno, A. *et al.* The light and dark sides of virtual screening: What is there to know?. *Int. J. Mol. Sci.* **20**, 1375 (2019).
10. Gupta, R. *et al.* Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. *Mol. Divers.* **25**, 1315–1360 (2021).
11. 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).
12. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: A review. *Biophys. Rev.* **9**, 91–102 (2017).
13. Jorgensen, W. L. Rusting of the lock and key model for protein-ligand binding. *Science* **254**, 954–955 (1991).
14. Glaab, E. Building a virtual ligand screening pipeline using free software: A survey. *Brief Bioinform.* **17**, 352–366 (2016).
15. Rica, E., Alvarez, S. & Serratosa, F. Ligand-based virtual screening based on the graph edit distance. *Int. J. Mol. Sci.* **22**, 12751 (2021).
16. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. "Scaffold-Hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed. Engl.* **38**, 2894–2896 (1999).
17. Wang, W. *et al.* Combined strategies in structure-based virtual screening. *Phys. Chem. Chem. Phys.* **22**, 3149–3159 (2020).
18. Kumar, A. & Zhang, K. Y. J. Hierarchical virtual screening approaches in small molecule drug discovery. *Methods* **71**, 26–37 (2015).
19. Vazquez, J., Lopez, M., Gibert, E., Herrero, E. & Luque, F. J. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules* **25**, 4723 (2020).
20. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
21. Gomez-Sanchez, E. P. Third-generation mineralocorticoid receptor antagonists: Why do we need a fourth?. *J. Cardiovasc. Pharmacol.* **67**, 26–38 (2016).
22. Grossmann, M. E., Huang, H. & Tindall, D. J. Androgen receptor signaling in androgen-refractory prostate cancer. *J. Natl. Cancer Inst.* **93**, 1687–1697 (2001).
23. Andersonn, K.-E. PDE5 inhibitors—pharmacology and clinical applications 20 years after sildenafil discovery. *Br. J. Pharmacol.* **175**, 2554–2565 (2018).
24. Lacouture, M. E. *et al.* Evolving strategies for the management of hand-foot skin reaction associated with the multitargeted kinase inhibitors sorafenib and sunitinib. *Oncologist* **13**, 1001–1011 (2008).
25. Wang, B. *et al.* An overview of kinase downregulators and recent advances in discovery approaches. *Signal Transduct. Target Ther.* **6**, 423 (2021).

26. Zhao, Z. *et al.* Exploration of type II binding mode: A privileged approach for kinase inhibitor focused drug discovery?. *ACS Chem. Biol.* **9**, 1230–1241 (2014).
27. Allaj, V., Guo, C. & Nie, D. Non-steroid anti-inflammatory drugs, prostaglandins, and cancer. *Cell Biosci.* **3**, 8 (2013).
28. Zarghi, A. & Arfaei, S. Selective COX-2 inhibitors: A review of their structure-activity relationships. *Iran J. Pharm. Res.* **10**, 655–683 (2011).
29. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
30. Li, J., Liu, W., Song, Y. & Xia, J. Improved method of structure-based virtual screening based on ensemble learning. *RSC Adv.* **10**, 7609–7618 (2020).
31. Yasuo, N. & Sekijima, M. Improved method of structure-based virtual screening via interaction-energy-based learning. *J. Chem. Inf. Model.* **59**, 1050–1061 (2019).
32. Xu, L. *et al.* Molecular modeling of the 3D structure of 5-HT(1A)R: Discovery of novel 5-HT(1A)R agonists via dynamic pharmacophore-based virtual screening. *J. Chem. Inf. Model.* **53**, 3202–3211 (2013).
33. Singam, E. R. A., Tachachartvanich, P., La Merrill, M. A., Martyn, T. S. & Durkin, K. A. Structural dynamics of agonist and antagonist binding to the androgen receptor. *J. Phys. Chem. B.* **123**, 7657–7666 (2019).
34. Kurumbail, R. G. *et al.* Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature* **384**, 644–648 (1996).
35. Baker, M. E. & Katsu, Y. 30 years of the mineralocorticoid receptor: Evolution of the mineralocorticoid receptor: Sequence, structure and function. *J. Endocrinol.* **234**, T1–T16 (2017).
36. Chan, H. C. S., Li, Y., Dahoun, T., Vogel, H. & Yuan, S. New binding sites, new opportunities for GPCR drug discovery. *Trends Biochem. Sci.* **44**, 312–330 (2019).
37. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815 (2006).
38. Feldmann, C., Miljković, F., Yonchev, D. & Bajorath, J. Identifying promiscuous compounds with activity against different target classes. *Molecules* **24**, 4185 (2019).
39. Ramsey, R. R., Popovic-Nikolic, M. R., Nikolic, K., Uliassi, E. & Bolognesi, M. L. A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* **7**, 3 (2018).
40. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
41. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
42. Sakate, R. & Kimura, T. Drug target gene-based analyses of drug repositionability in rare and intractable diseases. *Sci. Rep.* **11**, 12338 (2021).
43. Ruizeveld de Winter, J. A. *et al.* Androgen receptor expression in human tissues: An immunohistochemical study. *J. Histochem. Cytochem.* **39**, 927–936 (1991).
44. Bookout, A. L. *et al.* Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell* **126**, 789–799 (2006).
45. Matthews, E., Brassington, R., Kuntzer, T., Jichi, F. & Manzur, A. Y. Corticosteroids for the treatment of Duchenne muscular dystrophy. *Cochrane Database Syst. Rev.* **2016**, CD003725 (2016).
46. Tanoury, Z. A. *et al.* Prednisolone rescues Duchenne muscular dystrophy phenotypes in human pluripotent stem cell-derived skeletal muscle in vitro. *Proc. Natl. Acad. Sci. U S A* **118**, e2022960118 (2021).
47. Casas, A. I. *et al.* From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc. Natl. Acad. Sci. U S A* **116**, 7129–7136 (2019).
48. Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **4**, 682–690 (2008).
49. Nogales, C. *et al.* Network pharmacology: Curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.* **43**, 136–150 (2022).
50. Poornima, P., Kumar, J. D., Zhao, Q., Blunder, M. & Efferth, T. Network pharmacology of cancer: From understanding of complex interactomes to the design of multi-target specific therapeutics from nature. *Pharmacol. Res.* **111**, 290–302 (2016).
51. Gebhard, T. *et al.* Discovery and profiling of a selective and efficacious syk inhibitor. *J. Med. Chem.* **58**(4), 1950–1963 (2015).
52. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801 (2006).
53. wwPDB Consortium. Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
54. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **25**, 1157–1173 (2004).
55. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).
56. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Antechamber: An accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc.* **222**, U403 (2001).
57. Frisch, M. J. *et al.* Gaussian 16, Revision C.01. Gaussian, Inc., Wallingford CT (2016).
58. Pronk, S. *et al.* GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
59. Barca, G. M. J. *et al.* Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **152**, 154102 (2020).

Acknowledgements

We thank Yuko Usami for supporting the original data construction. This study was supported in part by a Grant from the Japan Society for the Promotion of Science (JSPS, Grant Number 20K12056). The images of chemical compounds were obtained from KEGG with permission.

Author contributions

Conceptualization and design: M.T., R.S., T.K.; Acquisition and construction of data: J.N., K.T., S.O., K.I.; Analysis and interpretation of data: M.T., R.S., J.N., K.T., S.O., K.I., T.K.; Supervision: T.K.; Writing manuscript-draft: M.T., R.S., T.K.; Writing revised manuscript: M.T., T.K. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-46648-1.

Correspondence and requests for materials should be addressed to T.K.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023