

KNIME-based Analysis of Off-Target Effect of Drugs Related to The Molecular 2D Fingerprint

Nihayatul Karimah¹ and Gijs Schaftenaar²

¹Center for Pharmaceutical and Medical Technology, Agency for the Assessment and Application of Technology (BPPT), Serpong, Banten, Indonesia; ²Center for Molecular and Biomolecular Informatics, Radboud UMC, Nijmegen, The Netherlands

Corresponding author: Nihayatul Karimah, Center for Pharmaceutical and Medical Technology, Agency for the Assessment and Application of Technology (BPPT), LAPTIAB I, PUSPITEK, Serpong, Banten 15314, Indonesia; email: nihayatul.karimah@bppt.go.id

Received, February 26, 2021; Revised, May 3, 2021; Accepted, May 5, 2021; Published, May 14, 2021

ABSTRACT -- Purpose: Structurally similar molecules are likely to have similar biological activity. In this study, similarity searching based on molecular 2D fingerprint was performed to analyze off-target effects of drugs. The purpose of this study is to determine the correlation between the adverse effects and drug off-targets. **Methods:** A workflow was built using KNIME to run dataset preparation of twenty-nine targets from ChEMBL, generate molecular 2D fingerprints of the ligands, calculate the similarity between ligand sets, and compute the statistical significance using similarity ensemble approach (SEA). Tanimoto coefficients (Tc) are used as a measure of chemical similarity in which the values between 0.2 and 0.4 are the most common for the majority of ligand pairs and considered to be insignificant similar. **Result:** The majority of ligand sets are unrelated, as is evidenced by the intrinsic chemical differences and the classification of statistical significance based on expectation value. The rank-ordered expectation value of inter-target similarity showed a correlation with off-target effects of the known drugs. **Conclusion:** Similarity-searching using molecular 2D fingerprint can be applied to predict off-targets and correlate them to the adverse effects of the drugs. KNIME as an open-source data analytic platform is applicable to build a workflow for data mining of ChEMBL database and generating SEA statistical model.

INTRODUCTION

Adverse Drug Reaction (ADR) is an unwanted reaction after the administration of drug on usual dose and thus require the alteration of the dosage regimen or withdrawal of the drug [1]. It has become a large problem and deep concern in the public health system because it causes hospital admission and in-hospital morbidity. Lazarou et al. estimated that 6.7% of hospitalized patients have a serious adverse drug reaction with a fatality rate of 0.32%, which means there are more than 2,216,000 serious ADRs in hospitalized patients, causing over 106,000 deaths annually [2]. Several causes contribute to ADR that include modulation of the primary target of a drug [3], nonspecific interactions of reactive metabolites [4], and unintended activity at off-targets [5]. Some ADRs are commonly detected in pre-clinical phase before the new drugs are submitted and approved to be marketed. Accordingly, target identification and validation are important steps in drug development. Drug promiscuity is the reason for drug interaction with off-target receptors in which it is responsible for undesired side effects. A drug usually shows polypharmacology that causes unintended off-

target effects if it binds to more than one target with distinct disease pathways.

Cheminformatics has become a useful low-cost computational approach for predicting off-target hits of drugs by comparing their 2D fingerprints. A molecular 2D fingerprint represents molecules as bit strings to indicate the presence and the absence of particular features. It is a computationally efficient and effective method in many comparative studies of ligand chemistry such as similarity-based virtual screening, molecular diversity, and clustering chemical databases [6]. The similarity between two ligands characterized by molecular 2D fingerprints is usually quantified using the Tanimoto Coefficient that calculates the number of similar fragments between two molecules [7]. Based on the assumption that chemically similar ligands often have biologically similar receptors, Keiser et al. established a statistics-based cheminformatics technique, called Similarity Ensemble Approach (SEA), to relate protein pharmacology by 2D ligand structures. SEA is capable in analyzing 2D similarity of ligands that bind to the same target and predict the ligand-target interaction [8]. The probability of the raw similarity score being achieved by random chance alone, given the z-score, was converted to an expectation value (E-value). Furthermore, SEA

was successfully applied for predicting and testing the drug activity on a side effect confirmed by in vitro experiment. For the example, SEA predicted three targets for Chlorotrianisene which have significant E-value and two of them were experimentally confirmed [5]. It is also confirmed using SEA technique that Delavirdine, an HIV-1 reverse transcriptase inhibitor, can bind to Histamine H4 receptor and cause painful skin rash as side effect [8]. The aim of this work is to determine the correlation between the adverse effects and drug off-targets in twenty-nine receptors annotated in ChEMBL, a chemical database of bioactive molecules, based on ligands similarity using molecular 2D fingerprint and by applying SEA technique. KNIME (Konstanz Information Miner) was utilized as the main tool to perform SEA techniques. KNIME is an open-source platform that integrates numerous nodes for machine learning and data mining [9]. This tool is useful to facilitate works in the area of bioinformatics and cheminformatics.

METHOD

A KNIME workflow was built to run dataset preparation, generate molecular 2D fingerprints, calculate the ligands similarity, and compute the statistical significance (Supplementary Figure 1) [9].

Dataset preparation

As many as 29 targets of proteins from Homo sapiens were enlisted and extracted from online ChEMBL Database using a python source node and ChEMBL Web Resource Client package [10]. All the bioactivity data from each target were collected using ChEMBLdb Connector Input node which then the node also simultaneously extracted all the ChEMBL ID compounds for the ligands of the targets. The compounds from all target that have an IC50 value were selected and all

Equation to calculate z-score where

Eq. 1;

$$\begin{aligned}rs(S1, S2) &= \text{raw score of set } S1 \text{ vs set } S2 \\n(S1, S2) &= \text{size}(S1) \times \text{size}(S2) \\ \mu(x) &\approx (4.24 \times 10^{-4})x \quad [\text{Expected raw score mean}] \\ \sigma(x) &\approx (4.49 \times 10^{-3})x^{0.665} \quad [\text{Expected raw score std.dev}] \\ z &= (rs(S1, S2) - \mu(n(S1, S2))) / \sigma(n(S1, S2))\end{aligned}$$

Equation to calculate E-value:

Eq. 2:

$$P(Z > z) = 1 - \exp(-e^{-z\pi/\sqrt{6}} - \Gamma'(1)), \text{ where } \Gamma'(1) \text{ is the Euler Mascheroni constant } (\approx 577215665)$$

$$E(z) = P(z)N_{db}, \text{ where } N_{db} \text{ is the number of set comparison made in the database}$$

compounds with an unspecified IC50 value were excluded from the dataset. The remaining compounds were grouped on ingredient compound ChEMBL ID, target name, and target ChEMBL ID. Compounds with the same compound ChEMBL ID and target ChEMBL ID were obviously duplicated and removed. This made up a total of 11,714 unique ligands with a median and mean of 206 and 404 per target respectively. The "Join compound info" node was utilized to extract SMILES, a line notation for encoding molecular structures, of each compound from ChEMBL.

Set comparison

The conversion of a one-dimensional SMILES string to a 2D molecule and the calculation of molecular 2D fingerprints were accomplished by Indigo nodes for KNIME provided by GGA Software Services LLC [11]. The molecular 2D fingerprint of every ligand was computed with parameter size of "similarity" part of a fingerprint in 8-byte blocks. All pairs of ligands between two sets of targets were compared by a pair-wise similarity metrics with the widely used Tanimoto coefficient (Tc) as the similarity criterion. For set comparisons, all pair-wise Tcs between ligands from set 1 and ligands from set 2 above a threshold were summed up to calculate the raw set comparison score.

Statistical Model

The statistical model referred to SEA was applied to compute the statistical significance of the similarity result from set comparison. The z-scores of set comparison were calculated as a function of raw score, expected raw score, and standard deviation. The calculation of expectation value which signifies the probability of score occurrence by random chance alone is derived from the z-score. The cut off Tc value as a threshold is 0.57 and is used in SEA which considers the best fit z-score distribution to extreme value distribution [8].

RESULTS AND DISCUSSION

Patterns of Similarity

The aim of building a statistical model is to assess significance of a set similarity against a random background [12]. From the total of 29 receptors containing 11,714 unique ligands, the comparison between each ligand in each set and each ligand in every other set was made, resulting 137,217,796 total ligand pairs. Tanimoto coefficients (Tc) from each pair of ligands as a measure of chemical similarity were calculated. Tc values between 0.2 and 0.4 are the most common for the majority of ligand pairs and are considered insignificant similarity (figure 1). The intra-set similarity of 97 histamine H2 ligands obeyed this observation. As

many as 66.4% ligand pairs had Tc below 0.4, 22.1% ligand pairs had more significant Tc in the range of 0.6-1.0, and only 1.2% had Tc 1.0 (109 pairs of ligand are identical). The receptors extracted from ChEMBL contain many ligands, which have almost similar structures with only one atom difference in the side chain, resulting in the same fingerprint. The same pattern appeared in the comparison of histamine H2 ligands set versus α -2a adrenergic ligands set in which 90.0% ligand pairs had Tc < 0.4 and only 0.9% ligand pairs had significant Tc. When comparing the histamine H2 ligand set versus the bradykinin B1 ligands set, 98.8% ligand pairs had insignificant Tc < 0.4 and no identical ligand revealed. The raw score of histamine H2 ligand set against itself was 1694.46, while the raw score between histamine H2 and α -2a adrenergic ligand sets was 162.62 which is consistent with the low similarity for most of all ligand pairs. Considerably, histamine H2 and α -2a adrenergic ligand sets contain the antidepressant and antipsychotic drugs which have very weak biological activity on the histamine H2 receptor and only an indirect effect on the α -2a receptor. Antidepressants are very weak antagonist for H2 Histamine receptor [36] while psychotropic drugs have a little antihistamine H2 activity at high nontherapeutic concentrations [38]. Many set pairs had no ligand pairs with similarity scores above the threshold, such as α -1a adrenergic vs. bradykinin-B2 ligand sets, β -1 adrenergic vs. GABA receptor α -1 subunit ligand sets, and dopamine D1 vs. pregnane-X ligand sets. Structurally similar molecules do not always have similar biological activity.

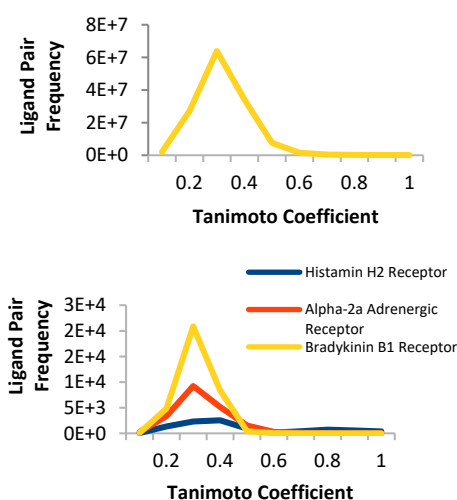


Figure 1. (Top) Insignificant similarities with Tc 0.2 - 0.4 have the highest frequency for all ligand pairs. (bottom) The distribution of ligand-ligand similarity for different ligand sets: Histamine H2 ligand set compared to themselves (blue), Histamine H2 ligand set compared to α -2a adrenergic ligand set (red), and Histamine H2 ligand set compared to Bradykinin B1 ligand set (yellow).

The patterns of similarity in the majority of ligand sets resembled that of histamine H2 – α -2a adrenergic ligand sets in which 9.04% of 841 set pairs had a raw score of zero. By calculating the mean and standard deviation of the raw score, it was revealed that as the product of set size increased, the number of ligand pairs which have Tc ≥ 0.57 also increased. It means that there is linear correlation between the size of ligand sets and the raw scores (Figure 2). The statistical model to quantitatively compare the significance of raw scores in different set sizes was represented by z-score and expectation value (E-value) in the context of random set similarity [8]. The z-score is the distance of a Tc from the mean measured in the number of standard deviation while E-value for Tc is the number of hits above the Tc that one can expect to find by chance alone when searching for a database of a particular size [12]. The smaller the E-value, the stronger the relationship between two receptors. Measuring E-value in random ligand similarity is beneficial to quantitatively identify pharmacological links among receptors based on similarity of the ligands that bind them and thus rank many potential targets [13]. Furthermore, it is also useful for large-scale test for drug repurposing. As many as 56.27% set comparisons had E-value > 1 which means there are more than 1 hits above Tc at the given ligand sets that one can expect to find at random. A heat map was built to classify expectation value of the ligand sets (supplementary Figure 2). GABA receptor α -1 subunit ligand set had the highest number of zero raw score which is why no E-value could be generated and makes the GABA ligand set the most dissimilar of all sets.

The comparison between histamine H2 ligand set against itself had z-score 857.74 and E-value zero, suggesting very high similarity, whereas histamine H2 and α -2a adrenergic ligand sets had z-score 47.40 and E-value 1.87×10^{-24} . Histamine H2 against bradykinin B1 ligand sets showed little similarity as was reflected by z-score -1.51 and E-value 5.22×10^3 . The ligand set comparison with a raw similarity score of zero could not be ranked.

On average, any given receptor was similar to seven other receptors with E-value < 10^{-10} . For some targets, the E-value declined rather steeply in the rank-ordered list (Table 1). For example, the set of type-1 angiotensin II receptor was similar to the histamine H1 ligand set and the ghrelin ligand set with E-values of 6.09×10^{-28} and 2.34×10^{-15} respectively. The third most significant ligand set was bradykinin B1 with E-value 9.20×10^{-5} . The monoamine oxidase A ligand set was similar to muscarinic acetylcholine M1, androgen, and progesterone ligand sets with E-value 4.38×10^{-22} , 3.29×10^{-21} , and 1.80×10^{-20} respectively. The next most significant ligand set was histamine H2

ligand set with E-value 7.38×10^{-5} . Some other ligand sets such as those of the targets melanocortin receptor 4, ghrelin receptor, neurotensin receptor 1, and voltage-gated N-type calcium channel α -1b subunit (N-type VGCC), were relatively promiscuous having more than six targets with E-value $< 10^{-40}$.

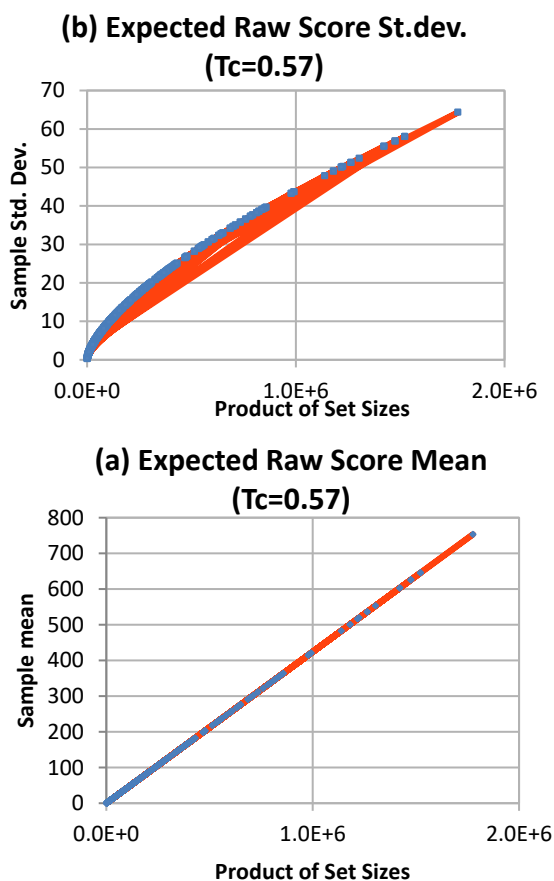


Figure 2. Plot and fit for expected mean (a) and expected standard deviation (b) of the random background of statistical model (N=765). The higher the set sizes, the higher the raw similarity scores of ligand pairs, which means the relation among the targets also depends on the set sizes.

The majority of ligand sets are unrelated because most ligand sets are strongly related to only a few others, as is evidenced by the intrinsic chemical differences and the classification of statistical significance between significant (E-value < 1.0) and insignificant (E-value > 1.0). The number of the resembling ligand sets may depend on the number of ligand set in database (Ndb). Sodium channel protein type V α -subunit strongly related to only three other receptors: Voltage-gated N-type calcium channel α -1B subunit, gastric inhibitory polypeptide receptor, and ghrelin receptor, with the E-value ranging from zero to 8.58×10^{-13} . The four highest ranking receptors against the histamine H2 receptor were dopamine D1 receptor, histamine H1 receptor, α -2a adrenergic receptor and muscarinic acetylcholine

receptor M1. They have relatively similar drugs (Tc ranging from 0.6 to 1.0), amongst which are known drugs classified into antihistamine, anticholinergic, and antidepressant.

Off-targets Effects Analysis

The drugs that are structurally quite similar to others do not necessarily exhibit the same activity because there are many other factors involved in drug-bioactivity relationship, such as molecular properties. The predicted off-target receptors in this work cannot replace direct experimentation such as in vitro and in vivo because of some issues, e.g. drug disposition, but at least it can beneficially prioritize off-targets for consideration in preclinical or clinical testing. ChEMBL database provides bioactivity information of more than 285,000 ligands, which modulate more than 1,500 different human targets with affinities better than $30 \mu\text{M}$ [5]. The prediction of off-target effect (Table 2) is focused on known single agents annotated in ChEMBL. The selection of off-target receptor should be guided by significant E-value of the ligand sets or the number of similar ligands of the related receptor. Most ligand pairs that have Tc 1.0 were not included in the analysis because they are the same ligands that exhibit poor bioactivity in the other targets. However, if the ligand pairs that have Tc 1 have good IC₅₀ to the targets, they are included in the analysis because it will bring information about the side effects. ChEMBL provides large-scale bioactivity data for more than 1 million compounds. ChEMBL users who retrieve the data for specific purpose still need to filter the data and do data mining.

Relcovaptan as selective antagonist of vasopressin V1a receptor is predicted to bind bradykinin B1 receptor, ghrelin receptor, melanocortin receptor 4, β -1 adrenergic receptor, and muscarinic acetylcholine receptor M1. Although vasopressin V1a receptor against bradykinin B1 receptor and β -1 adrenergic receptor has E-values of 6.38×10^{-3} and 7.93 respectively and therefore are considered less significant. However, Relcovaptan had a higher similarity with more than ten ligands of bradykinin B1 receptor and β -1 adrenergic receptor with Tc ranging from 0.57 to 0.71. On the other hand, the drug was only similar to one ligand from the N-type VGCC set while the two sets showed a high similarity with E-value of 1.40×10^{-41} . So, single drugs with a high similarity score are not always representative for the calculated E-value between ligand sets. The most common side effect of Relcovaptan is thirst, pollakiuria (increased daytime urination), and dry mouth [14]. Other study reported that the most frequent adverse effects were thirst, nausea, hypotension, constipation, dizziness, and dry mouth [15]. The dry mouth side effect of

Relcovaptan may relate to antagonist activity to β -1 adrenergic and Muscarinic M1 receptor because the drug resembles the other drugs that antagonize both receptors. Beta-adrenoceptor antagonist affects salivary composition and significantly decreases amylase activity [16]. Muscarinic M1 and muscarinic M3 receptors play an important role in mediating salivary fluid secretion and muscarinic M4 receptors regulate salivary protein secretion [17]. Although the hypotension side

effect of Relcovaptan was reported infrequently, it is interesting to investigate the effect of the drug to β -1 adrenoceptor since the drug is very similar to drugs on the market for the β -1 adrenoceptor.

Daltroban as antagonist of thromboxane A2 receptor and Samixogrel as antiplatelet were similar to ligands of the glucocorticoid, progesterone, β -1 adrenergic and bradykinin B1 receptors, with E-value better than 2.06×10^{-14} and Tc ranging from 0.61 to 0.75.

Table 1. The rank-ordered relation of four-example ligand sets to the others

Rank	Target Query	Similar Target	Size	E-value	Tc1	Max Tc
1		Type-1 angiotensin II receptor	734	0	902	1
2	Type-1	Histamine H1 receptor	266	6.09×10^{-28}	0	0.78
3	angiotensin II	Ghrelin receptor	643	2.34×10^{-15}	1	1
4	receptor	Bradykinin B1 receptor	357	9.20×10^{-05}	0	0.7
5		Bradykinin B2 receptor	227	2.00×10^3	0	0.7
1		Sodium channel type V α -subunit	246	0	272	1
2	Sodium channel	N-type VGCC	323	1.12×10^{-36}	0	0.69
3	protein type V α -	Gastric inhibitory polypeptide receptor	110	4.33×10^{-18}	0	0.8
4	subunit	Ghrelin receptor	643	8.58×10^{-13}	1	1
5		L-type VGCC	30	8.03×10^{-6}	4	1
1		Monoamine oxidase A	1067	0	1995	1
2		Muscarinic acetylcholine receptor M1	570	4.38×10^{-22}	0	0.85
3		Androgen Receptor	1107	3.29×10^{-21}	7	1
4	Monoamine	Progesterone receptor	1142	1.80×10^{-20}	0	0.86
5	oxidase A	Histamine H2 receptor	97	7.38×10^{-5}	0	0.8
6		Dopamine D1 receptor	108	1.84×10^{-4}	3	1
7		α -2a adrenergic receptor	206	1.68×10^{-1}	7	1
1		Histamine H2 receptor	97	0	109	1
2		Dopamine D1 receptor	108	4.91×10^{-39}	36	1
3	Histamine H2	Histamine H1 receptor	266	2.50×10^{-35}	37	1
4	receptor	α -2a adrenergic receptor	206	1.87×10^{-24}	36	1
5		Melanocortin receptor 4	744	2.53×10^{-22}	6	1
6		Muscarinic acetylcholine receptor M1	570	4.92×10^{-22}	34	1
7		Monoamine oxidase A	1067	7.38×10^{-5}	0	0.8

Table 2. Predicted off-target of drug and its correlation to reported adverse effect

Known Drug	Pharmacological Action	Related Target	E-value	Reported adverse effect
Relcovaptan	Vasopressin V1a antagonist	Muscarinic M1 Receptor	6.11×10^{-10}	Dry mouth
Flecainide	Sodium Channel Blocker	Ghrelin Receptor	8.58×10^{-13}	Loss of appetite
Lidocaine	Sodium Channel Blocker	Ghrelin Receptor	8.58×10^{-14}	Loss of appetite
Norethindrone	Progesterone Receptor Agonist	Vasopressin V1a receptor	7.65×10^{-31}	Fluid retention
Flunarizine	N-type VGCC Blocker	Melanocortin Receptor 4	2.47×10^{-118}	Weight gain
Verapamil	L-type VGCC Blocker	Dopamine D1 Receptor	1.32×10^{-17}	Antipsychotic effect
Saralasin	Partial agonist of angiotensin II receptor	Cholecystokinin receptor	4.51×10^3	Amelioration of acute pancreatitis
Valsartan	Antiangiotensin II	Ghrelin Receptor	2.34×10^{-15}	Loss of appetite
		Melanocortin Receptor 4	1.30×10^{-1}	Weight gain
Diphenidol	Antimuscarinic M1	N-type VGCC	1.56×10^{-157}	Account for antiemetic and antivertigo effect

Despite Samixogrel is not an antagonist of thromboxane A2 receptor, it has low inhibitory concentration against thromboxane A2 receptor with IC₅₀ 19 nM [18]. The adverse effects of Daltroban and Samixogrel were only sparsely reported. For future antithrombin drug development, it is better to consider the drug promiscuity towards glucocorticoid, progesterone, and bradykinin B1 ligands as well as the mechanism of action to the related receptors. Sodium channel blockers, Flecainide and Lidocaine, have the adverse effect of decreasing appetite as reported by FDA [19], which is presumably related to ghrelin receptor because ghrelin is responsible for stimulating hunger. While the adverse effect of drugs having affinity with ghrelin receptor may reduce appetite, both Flecainide and Lidocaine have a similarity to ghrelin antagonists with Tc 0.57-0.58 and significant E-value of 8.58×10^{-13} . Medroxyprogesterone Acetate (MPA) is an agonist of progesterone, androgen, and glucocorticoid receptor with high affinity of binding [20]. Mifepristone, an antiprogesterone and antiglucocorticoid, has weak antagonism towards androgen receptor [21]. Onapristone is antiprogesterone only and it was reported to have low affinity to glucocorticoid and androgen receptor. It is in line with the result that these drugs as antiprogesterone exhibited strong relation to androgen and glucocorticoid ligands with diverse Tc ranging from 0.57 to 1.0 and E-value of 3.57×10^{-110} for androgen ligand set and 6.92×10^{-249} for glucocorticoid ligand set. The drug informer reported 48 cases of fluid retention side effects with the drug Norethindrone [19]. Vasopressin, as antidiuretic hormone, plays an important role in fluid-electrolyte balance, control of vascular tone, and cardiovascular contractility [22]. Although Norethindrone did not show significant similarity to vasopressin V1a ligands (no Tc ≥ 0.57), it is intriguing that progesterone ligand set has highly significant relation to vasopressin V1a ligand with E value 7.65×10^{-31} .

Classified as calcium channel blocker, Flunarizine is a non-selective drug that has antihistaminic effect. It is effective to be used in the prophylaxis of migraine, occlusive peripheral vascular disease, vertigo of central and peripheral origin, and as an adjuvant in the therapy of epilepsy [23]. Many side effects of the drug were reported such as weight gain, drowsiness, extrapyramidal effect, and depression [24, 25]. In a comparative trial of Flunarizine and Propanolol in the prevention of migraine, nine out of twenty-eight patients experienced weight gain [26]. Considering melanocortin receptor 4 (MCR4) is responsible in the regulation of appetite and energy expenditure [27], the weight gain side effect of Flunarizine is

predicted to be related to this receptor because the drug had high Tc values, ranging from 0.57 to 0.72 with more than twenty MCR4 ligands. The Voltage-gated N-type calcium channel against MCR4 ligand sets had highly significant E-value of 2.47×10^{-118} . Verapamil as another calcium channel blocker has significant similarity with dopamine D1 ligands. The Tc between them ranging from 0.58 to 0.67 and the E-value of the ligand sets between voltage-gated N-type calcium channel and dopamine D1 was 1.32×10^{-17} . This result was corroborated by a paper stating Verapamil having an antipsychotic effect [28]. Moreover, short-term Verapamil administration could increase dopamine level in striatal while long-term verapamil administration could decrease it [29]. Furthermore, dopamine is involved in several psychotic disorders such as schizophrenia and bipolar. Increasing dopamine levels are related to hallucination and delusion [30].

Saralasin, a partial agonist of angiotensin II receptor, was reported to attenuate oxidative stress and tissue injury in cerulein-induced acute pancreatitis [31]. In various animal models of acute pancreatitis, cholecystokinin-receptor antagonist has ameliorated the injury response [32]. Those reports correspond to the result in this study that showed the relation of Saralasin to cholecystokinin ligand set in which it resembled more than ten ligands of cholecystokinin receptor despite the E-value (4.51×10^3) was not significant. Saralasin is predicted to have antagonist activity to Cholecystokinin receptor, mediating amelioration of acute pancreatitis. Another report about inhibition of injury induced intimal hyperplasia by Saralasin mentioned that weight gain of rats was not reduced in the group treated with Saralasin [33]. This is backed up by a high similarity score of the angiotensin II ligand set to the ghrelin and melanocortin 4 ligand sets with E-value 2.34×10^{-15} and 1.30×10^{-1} respectively and diverse Tc ranging from 0.58 to 0.92. Saralasin is also similar to many ligands of the neurotensin receptor 1 with a significant E-value of 2.18×10^{-3} , however no literature could be found to support this side effect. Valsartan, an angiotensin II receptor antagonist, has less common side effect of weight gain and loss of appetite [34]. As mentioned above, weight gain is correlated to melanocortin receptor 4 while loss of appetite is correlated to ghrelin receptor. The result showed that Valsartan was similar to many ligands of both receptors and strongly related to both receptors with E-value 1.30×10^{-1} and 2.34×10^{-15} respectively.

Cyclizine, an antihistamine and antimuscarinic, had significant similarity to more than twenty ligands of melanocortin 4 with Tc ranging from 0.57 to 0.74 and highly significant E-value of 1.30×10^{-46} . However, the side effect

pertaining to melanocortin receptor 4 such as weight alteration has not been reported. Diphenidol is antiemetic and antivertigo that has anticholinergic activity by muscarinic receptors blockade. The result in this study showed very strong relation of Diphenidol as antimuscarinic with voltage-gated N type calcium channel (VGCC) which has E-value 1.56×10^{-157} and Tc ranging from 0.57 to 0.80. One report confirmed that Diphenidol inhibited VGCC which has an inhibiting effect on neuronal excitability, that accounts for the antiemetic and antivertigo effects of this drug [35].

The whole correlation between adverse effects and off-target receptors of all drugs within the ligand sets cannot be elucidated because of four main reasons. First, only twenty-nine targets were used in this study. Second, the drugs such as famotidine and atropine are only similar to the ligand set of their receptor. No predicted side effects were found for famotidine and atropine within these limited ligand sets. Third, the drugs such as Chlorpromazine and Promethazine resembled diverse ligand sets with significant E-value. However, reported side effects did not correspond to the side effects predicted from significant E-values, which means the predicted side effects are possibly not severe. Fourth, some drugs (e.g. Amitriptyline, Metergoline, Selegiline) existed in the ligand set of another target instead of their primary target (figure 3). For instance, Amytriptiline is prescribed as antidepressant to act on serotonin transporter, but it existed in the histamine H2 ligand set with high IC50 value. Whereas, serotonin transporter was not included in the list of ligand sets. It causes ambiguous side-effect prediction whether the predicted side effects are the real side effects or even the main therapeutic effect.

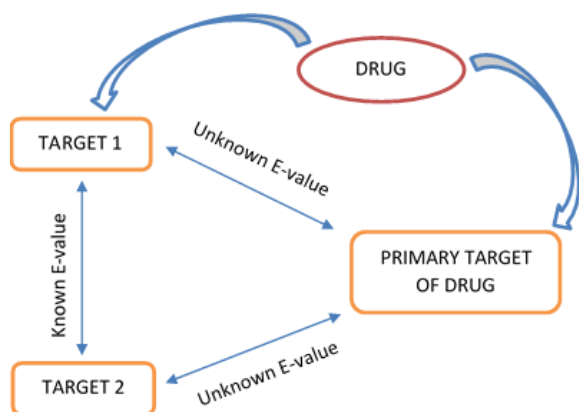


Figure 3. Some drugs existed in the ligand set of another target instead of their primary target. The E-value between the primary target and the other targets cannot be determined causing it is ambiguous to predict side effect in this case.

CONCLUSION

This study has shown that similarity-searching using molecular 2D fingerprint can be applied to predict off-targets and correlate them to the adverse effects of the drugs. KNIME as an open-source data analytic platform is applicable to build a workflow for data mining of ChEMBL database and generating SEA statistical model. One of the limitations of this study is the under-population of the ligand sets that were used to extract adverse effects prediction due to the limited amount of computer memory available. However, one can explore the promiscuity of certain drug or substance with this simple task and thus speed up the effort of developing the new drug. The drug-targets association generated from this study still cannot replace direct experimentation, e.g., in vitro, and in vivo, but it can prioritize the off-targets for preclinical and clinical testing. Some predictions of off-targets and their correlations to adverse effect in this study still need subsequent investigation.

ACKNOWLEDGEMENTS

The authors thank Dr. Celia van Gelder and Prof. Dr. Gert Vriend for giving the permission to use the resources at Centre of Molecular and Biomolecular Informatics (CMBI), Radboud University Nijmegen, to complete this project. This work was supported by Centre of Molecular and Biomolecular Informatics (CMBI), Radboud University Nijmegen, and Indonesia Endowment Fund for Education (LPDP) Scholarship.

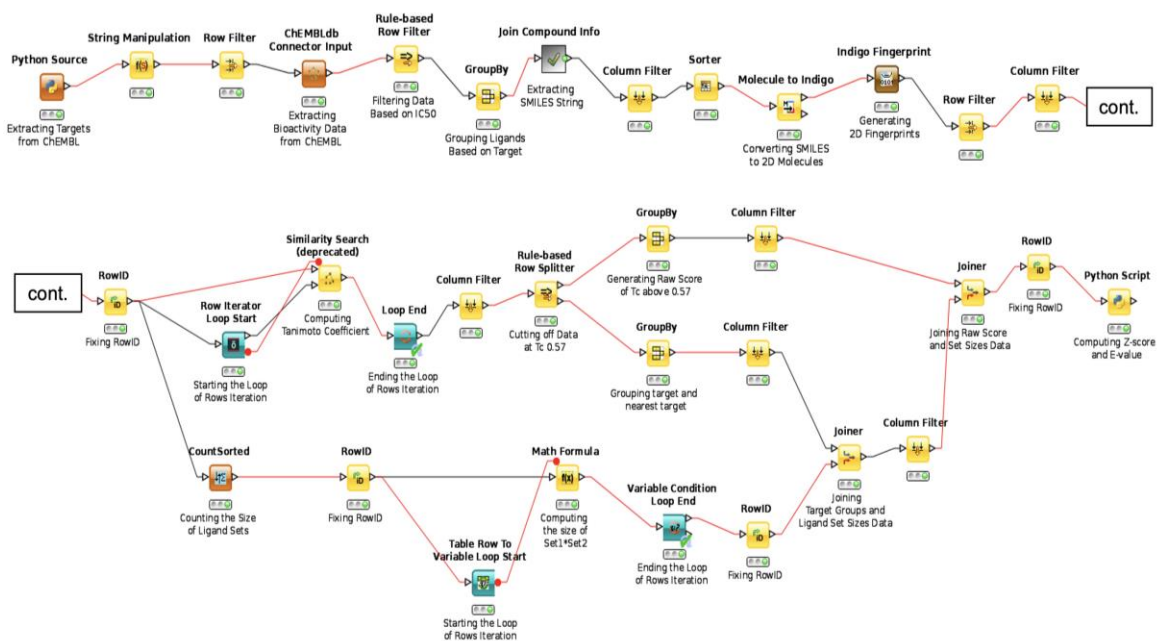
REFERENCES

1. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *Lancet*. 356:1255–1259 (2000).
2. Lazarou J, Pomeranz B, Corey PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA*. 279:1200–1205 (1998).
3. Wong, D., Wang, M., Cheng, Y. & Fitzgerald, G. A. Cardiovascular hazard and non-steroidal anti-inflammatory drugs. *Curr. Opin. Pharmacol.* 5: 204–210 (2005)
4. Antunes, A. M. M. *et al.* Protein adducts as prospective biomarkers of nevirapine toxicity. *Chem. Res. Toxicol.* 23: 1714–1725 (2010).
5. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK,

- Urban L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* (2012).
6. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*. 11:1046–53 (2006).
 7. Willett, P. et al. Implementation of nearest-neighbour searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* 26, 36–41 (1986)
 8. Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nature Biotechnol.* 25: 197–206 (2007).
 9. Berthold, M.R. *et al.* KNIME: the Konstanz Information Miner. in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* 319–326 (Springer, 2007).
 10. PyPI, ChEMBL Web Resource Client. <https://pypi.python.org/simple/chembl-webresource-client/> retrieved on August 15th 2020.
 11. Indigo Cheminformatics Toolkit. <https://www.knime.com/community/indigo> retrieved on August 17th 2020.
 12. Baldi P, Nasr R. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values. *Journal of Chemical Information and Modeling* (2010).
 13. Liu, X., al. In silico target fishing: addressing a "Big Data" problem by ligand-based similarity rankings with data fusion. *J. Cheminform.* 6: 33 (2014).
 14. Morphy, R., Rankovic, Z. Designed multiple ligands. An emerging drug discovery paradigm. *Journal of Medicinal Chemistry*. 48: 6523–6543. 4 (2005).
 15. Gines P., Guevara M. Hyponatremia in cirrhosis: Pathogenesis, clinical significance and management. *Hepatology*. 48:1002-10 (2008).
 16. Konstam MA, Gheorghide M, Burnett JC, Grinfeld L, Maggioni AP, Swedberg K, *et al.* Effects of oral Tolvaptan in patients hospitalized for worsening heart failure. The Everest outcome trial. *JAMA*. 297:1319-31 (2007).
 17. Nederfors T. Xerostomia: prevalence and pharmacotherapy. With special reference to beta-adrenoceptor antagonists. *Swed. Dent. J. Suppl.* 116: 1–70 (1996).
 18. Tobin G, Ryberg AT, Gentle S, and Edwards AV. Distribution and function of muscarinic receptor subtypes in the ovine submandibular gland. *J. Appl. Physiol.* 100: 1215–1223 (2006).
 19. Drug Informer: side effect search. <https://www.druginformer.com> retrieved on September 18th 2020.
 20. Schindler AE, Campagnoli C, Druckmann R, Huber J, Pasqualini JR, Schweppe KW, Thijssen JH. "Classification and pharmacology of progestins". *Maturitas*. 61 (1-2): 171–80 (2008).
 21. Heikinheimo O, Kekkonen R, Lahtenmaki P. "The pharmacokinetics of mifepristone in humans reveal insights into differential mechanisms of antiprogestin action". *Contraception*. 68 (6): 421–6 (2003).
 22. C.R. Lee, M.L. Watkins, J.H. Patterson, *et al.* Vasopressin: a new target for the treatment of heart failure. *Am. Heart J.* 146: 9–18 (2003).
 23. Flunarizine. MeSH <http://www.ncbi.nlm.nih.gov/mesh/68005444> retrieved on September 19th 2020.
 24. Holmes B, Brogden RN, Heel RC, Speight TM, Avery GS. *Flunarizine*. A review of its pharmacodynamic and pharmacokinetic properties and therapeutic use. *Drugs*. 27: 6–44 (1984).
 25. Mangone CA, Herskovits E. Extrapyramidal and depressive side reactions with flunarizine and cinarizine. *J. Neurol. Neurosurg. Psychiatry*. 52: 288–289 (1989).
 26. Shimell CJ, Fritz VU, Levien SL: A comparative trial of flunarizine and propranolol in the prevention of migraine. *S. Afr. Med. J.* 77:75–77 (1990).
 27. Fani L, Bak S, Delhanty P, van Rossum EF, van den Akker EL. The melanocortin-4 receptor as target for obesity treatment: a systematic review of emerging pharmacological therapeutic options. *Int. J. Obes.* 38: 163–169 (2014).
 28. Umukoro S, Bakre TO, Onwuchekwa C. Antipsychotic and sedative effect of calcium channel blockers in mice. *Afr. J. Med. Med. Sci.* 39 Suppl: 61-6 (2010).
 29. Sitges M, Guarneros A. *Chronic verapamil modifies striatal and frontal cortex dopamine levels.* *Eur Neuropsychopharmacol.* 8 (2): 105-11 (1998).
 30. van Os J, Kapur S. Schizophrenia. *The Lancet*. 374 (9690): 635-45 (2009).
 31. Ip, SP, Tsang, SW, Wong, TP, Che, CT, Leung, PS. Saralasin, a nonspecific angiotensin II receptor antagonist, attenuates oxidative stress and tissue injury in cerulein-induced acute pancreatitis. *Pancreas*. 26: 224-229 (2003).
 32. Nordback, I., Clemens, J., Cameron, J. The role of cholecystokinin in the pathogenesis

- of acute pancreatitis in the isolated pancreas preparation. *Surgery*. 109: 301–306 (1991).
33. Pan XM, Nelken N, Colyvas N, Rapp JH. Inhibition of injury induced intimal hyperplasia by saralasin in rats. *Journal Of Vascular Surgery*. 15(4): 693-698. (1992)
 34. Valsartan Side Effects. <http://www.drugs.com/> retrieved on September 21st 2020.
 35. Leung, Y.M.;Wong, K.L.;Cheng, K.S.;Kuo, C.S.;Su, T.H.;Chen, Y.W.;Cheng, T.H. "Inhibition of voltage-gated K⁺ channels and Ca²⁺ channels by diphenidol." *Pharmacological reports: PR*. 64, 3: 739-44 (2012).
 36. Watanabe T, Wada H. Histaminergic neurons: morphology and function. Boca Raton: CRC Press (1991).
 37. Alvarez FJ, Galindo A, Velasco A. Effect of psychotropic drugs on histamine H₂-receptors in rat isolated uterus. *Methods and findings in experimental and clinical pharmacology*. 14 (6): 431-5 (1992).

Supplementary figure 1. KNIME Workflow



Supplementary figure 2. Heat map of the expectation value between ligand sets. Red: $0 < E\text{-value} \leq 10^{-50}$; Light brown: $10^{-50} < E\text{-value} \leq 10^{-20}$; Light blue: $10^{-20} < E\text{-value} \leq 10^{-5}$; Dark blue: $E\text{-value} > 10^{-5}$

