

# Assessment of a Rule-Based Virtual Screening Technology (INDDEx) on a Benchmark Data Set

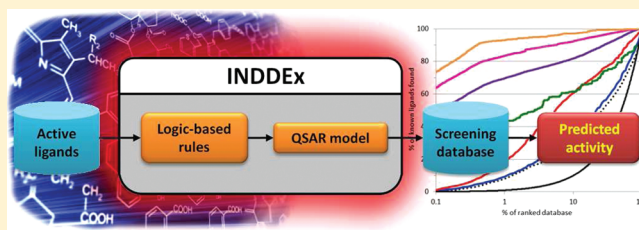
Christopher R. Reynolds,<sup>\*,†,§</sup> Ata C. Amini,<sup>§</sup> Stephen H. Muggleton,<sup>‡,§</sup> and Michael J. E. Sternberg<sup>†,§</sup>

<sup>†</sup>Department of Life Science, Imperial College London, London, SW7 2AZ United Kingdom

<sup>‡</sup>Department of Computing, Imperial College London, London, SW7 2BZ United Kingdom

<sup>§</sup>Equinox Pharma Ltd., Incubator, Bessemer Building, Prince Consort Road, London, SW7 2BP United Kingdom

**ABSTRACT:** The Investigational Novel Drug Discovery by Example (INDDEx) package has been developed to find active compounds by linking activity to chemical substructure and to guide the process of further drug development. INDDEx is a machine-learning technique, based on forming qualitative logical rules about substructural features of active molecules, weighting the rules to form a quantitative model, and then using the model to screen a molecular database. INDDEx is shown to be able to learn from multiple active compounds and to be useful for scaffold-hopping when performing virtual screening, giving high retrieval rates even when learning from a small number of compounds. Across the data sets tested, at 1% of the data, INDDEx was found to have average enrichment factors of 69.2, 82.7, and 90.4 when learning from 2, 4, and 8 active ligands, respectively. At 0.1% of the data, INDDEx had average enrichment factors of 492, 631, and 707 when learning from 2, 4, and 8 active ligands, respectively. Excluding all ligands with more than 0.5 Tanimoto Maximum Common Substructure, INDDEx had average enrichment factors at 1% of 52.3, 63.6, and 66.9 when learning from 2, 4, and 8 active ligands, respectively. The performance of INDDEx is compared with that of eHiTS LASSO, PharmaGist, and DOCK.



## INTRODUCTION

Virtual screening is a computational technique used for searching a large library of compounds to identify those that are likely to be biologically active. It is used when an experimental high-throughput screening would be unfeasible because of cost, time, or the number of compounds involved, but it can also be used to complement high-throughput screening by identifying a list of candidate compounds for testing. Many different virtual screening methods have been developed, as reviewed in, for example, Schneider.<sup>1</sup> The two main strategies are structure-based and ligand-based. Structure-based requires the known structure of the protein target; ligand-based derives its models from the molecular structure of known active ligands. This paper reports a ligand-based virtual screening method implemented in the program INDDEx (Investigational Novel Drug Discovery by Example), which is a drug discovery system that uses a patented combination of machine-learning processes.<sup>2</sup> INDDEx is compared to two other methods of ligand-based virtual screening and one of structure-based docking.

It is often desirable that virtual screening returns compounds that are not only highly active but also structurally diverse from each other and from the training data. Retrieving a compound that has the same mode of action as, but significant topological differences to, the input compound is known as “scaffold-hopping”. Ligand-based methods learn from existing molecules and so can tend toward finding molecules similar to ones used as training data. Having a wide range of diversity gives greater

choice of structure to develop, allowing structures to be chosen to avoid side-effects. Wider diversity also gives more chance of finding greater activity, because searching through structural space can get stuck in activity local minima. Compounds with sufficient topological differences from currently used molecules have the opportunity to be patented.

Structure-based screening uses the 3D structure of the target to model the docking of potential compounds to the target, and applying a scoring function to rank the potential activity of each compound.<sup>3–5</sup> The structure-based method used for comparison in this paper was DOCK,<sup>6,7</sup> but other commonly used structure-based screening programs include FlexX, FRED, GLIDE, GOLD, SLIDE, SURFLEX, and QXP, as described and compared in Kellenberger et al.<sup>8</sup> Scheraga has also developed PRODOCK for virtual-screening,<sup>9</sup> in the paper of which, it is noted that the two challenges of structure-based screening are to find an energy function that corresponds to the binding energies involved in the binding site, and then to find the minimum energy function. PRODOCK performs this by using Scheraga’s ECEPP 3 forcefield<sup>10,11</sup> and a Mote Carlo method of growing the ligand inside the receptor site.<sup>12</sup>

**Special Issue:** Harold A. Scheraga Festschrift

**Received:** December 14, 2011

**Revised:** February 13, 2012

**Published:** March 1, 2012

The advantages of structure-based docking are that active compounds can be found without any knowledge of existing active compounds, that it can find entirely structurally different drugs with unique scaffolds, providing new directions to search for leads, and that the availability of the 3D protein structure is now very large. Structure-based docking results usually include the correct docking position in their output lists of potential positions but can be poor at ranking the positions.<sup>13</sup>

Ligand-based screening methods identify common patterns and features among known active ligands. This can be done by creating a 2D fingerprint or 3D pharmacophore: a fingerprint is expressed as a string of bits, with each bit indicating the presence or absence of a structural or physiochemical feature, while a pharmacophore is an abstract 3D model of the chemical features of an ideal molecule. The extent of similarity to the fingerprint or pharmacophore is used as a measure of predicted activity.<sup>14,15</sup>

The advantages of ligand-based screening are that it does not require the 3D structure or any other knowledge of the target. The disadvantages are that scaffold-hopping can be difficult, with the most topologically similar molecules to the generated pharmacophore tending to be picked out first, and that there needs to be at least one known bioactive compound as input.

The three methods compared with INDDEX in this paper are eHiTS LASSO, PharmaGist, and DOCK. eHiTS LASSO (electronic High Throughput Screening Ligand Activity by Surface Similarity Order)<sup>16</sup> is a ligand-based virtual screening method based on a neural network algorithm, which produces an exhaustive set of conformations that could fit into an active site without severe steric clash; these are screened using surface property descriptors.<sup>17</sup> It learns from multiple existing compounds. PharmaGist is a ligand-based method available as a web server at <http://bioinfo3d.cs.tau.ac.il/PharmaGist/>,<sup>18</sup> and creates pharmacophore queries by training on up to 32 existing compounds. DOCK<sup>7</sup> is a structure-based screening program that performs a conformational search to dock flexible molecules into the binding sites of protein structures, and scores each position.

The method of machine-learning used by INDDEX is developed from, and inspired by, Inductive Logic Programming (ILP), which was developed by Muggleton.<sup>19</sup> Inductive logic describes the process of inferring the properties of a general population of objects from the recurring properties observed in a sample of those objects. ILP can be programmed with background knowledge, consisting of logical clauses known as predicates defining the properties of, and relationships between, objects. INDDEX implements the ILP approach of constructing relational hypotheses but limits the search space to pairwise distances between features, which were those found to be strongest discriminators for activity in previous studies using ILP.<sup>20</sup> INDDEX learns easily interpretable qualitative logic rules from active ligands, which give an insight into chemistry, and relate molecular substructure to activity and can be used to guide the next steps of drug design chemistry. Each rule is in either in a pairwise format "An active molecule requires fragment A and fragment B, separated by a distance in angstroms" or as a requirement format "An active molecule requires the presence of Fragment C". Rules also define requirements for inactivity. The type of fragments can be varied at the user's discretion, but throughout this paper, a single fragmentation method is used, that considers fragments as groups of atom types, bond types and hydrogenation levels.

Using Support Vector Machines (SVMs), the rules can then be weighted to produce a quantitative model of structure-activity relationships which is used to screen databases of molecules and predict drug activity. SVMs are a powerful learning method used to create a weighted model where each logical rule is a vector. An SVM model is produced in terms of ILP rules. The strength of SV-ILP is not that it performs comparably with pure SVMs, but that it can perform comparably while providing rules that can be understood by organic chemists and used to understand the mechanism of action of molecules binding to a target, and suggest ways to further improve activity.

Earlier testing of combinations of ILP and SVMs to screen molecular bioactivity data and compare performance with 3D pharmacophore and Bayes classification methods were performed on single data set examples.<sup>20–23</sup> This paper presents the largest and most comprehensive test of the method yet.

## METHOD

**Data.** The methodology in this paper is the same as the one used for the assessment of LASSO and DOCK,<sup>17</sup> and for the assessment of PharmaGist and DOCK.<sup>24</sup> This assessment methodology measures performance through the retrieval of active ligands from data sets of decoy compounds from the Directory of Useful Decoys (DUD) data sets,<sup>25</sup> which has become one of the standard benchmarking tools for screening methods.

The DUD database (<http://dud.docking.org/r2>, release 2, downloaded April 8th, 2011), generated by Huang et al.,<sup>25</sup> contains data sets of between 11 and 444 molecules that are active ligands for each of forty protein targets, selected on the basis of structure and activity data available, giving a total of 3238 ligands (including 164 stereoisomers). Energy minimized 3D structures are provided in DUD for every active molecule. For each active ligand, the DUD database also provides 36 inactive ligands, from a subset of the ZINC database<sup>26</sup> filtered using the Lipinski rules for druglikeness.<sup>27</sup> Each of these inactive decoys is chosen to resemble the ligand in physiochemical properties, but differ from it in topological structure. Some targets have identical decoys.

Active molecules provide positive examples for INDDEX to learn positive rules from, while negative examples produce negative rules to exclude molecules from activity space. Learning negative rules from the decoys in the DUD database would provide an unfair advantage to INDDEX. To provide negative learning examples, 20 molecules were selected, at random, from the whole ZINC database (version 11 on third May 2011), which were classed as inactive compounds, and these same molecules were used with all targets as examples of inactive compounds to learn negative rules from. The ZINC IDs of the molecules used as inactives were 290973, 337363, 1058986, 2973208, 3909444, 4384514, 4982113, 5018499, 5065168, 5356968, 5536756, 5752659, 6938389, 8527733, 8817402, 9102259, 9449998, 9571864, 9950656, and 10139965.

To simulate the screening of a large database of molecules for a small number of active ligands, the decoys from all 40 of the protein targets were merged, with duplicates removed. This gave a data set containing 95 171 of the original 127 679 separate molecule entries.

**Evaluation.** For each of the 40 DUD targets, the retrieval rates of the known active ligands from all the decoy compounds

were measured when INDDEx was trained on 2, 4, 8, 16, and 32 of the known active ligands. Measurements were also made using the standard enrichment factors (EF). Enrichment factors are a measure of performance, expressing the enrichment of the results in a given percentage of the top ranked results produced by a screening process. EFs are calculated as the ratio of the fraction of active compounds found in the sample to the fraction of active compounds in the entire population.

$$\text{enrichment factor}_{\% \text{ sample of population}} = \frac{\left( \frac{\text{actives}_{\text{sample}}}{\text{compounds}_{\text{sample}}} \right)}{\left( \frac{\text{actives}_{\text{population}}}{\text{compounds}_{\text{population}}} \right)}$$

EF<sub>1</sub> is often used as a measure, given the ratio of actives found in the top 1% of results to the actives in the population. However, 1% of a screening database containing perhaps millions of molecules is still an impractical number to test. Because of this, looking at the top 0.1% of results with EF<sub>0.1</sub> is perhaps a more practically relevant comparison measure. Because of the 95 171 decoy compounds in the DUD data set (and between 11 and 444 active compounds depending on the target), finding the EF<sub>0.1</sub> will look at the top ranked 95 compounds.

Enrichment curves were constructed which plot the percentage of known active ligands found as you look down the ranked molecules. In other words, the *y*-axis is  $\text{actives}_{\text{sample}} / \text{actives}_{\text{population}}$  and the *x* axis is  $\text{compounds}_{\text{sample}} / \text{compounds}_{\text{population}}$ . From this it will be seen that an enrichment factor for any sample size can be read off an enrichment curve by dividing the *y*-axis value by the *x*-axis value.

The tests used random sampling of the actives without replacement, and were repeated five times or until there were insufficient actives left to sample. After learning on these randomly chosen molecules, the model built by INDDEx was used to screen a data set containing all the target actives that were not used to learn rules, together with all the inactive molecules over all 40 targets in the DUD (with duplicates removed). Results were ranked by predicted activity, and enrichment curves were produced from the ranked results of all active and decoy molecules. Consensus results were produced by vertically averaging the enrichment curves for all 40 targets. The vertical average is where each point on the curve is the average of all the curve values at that point on the *x* axis.

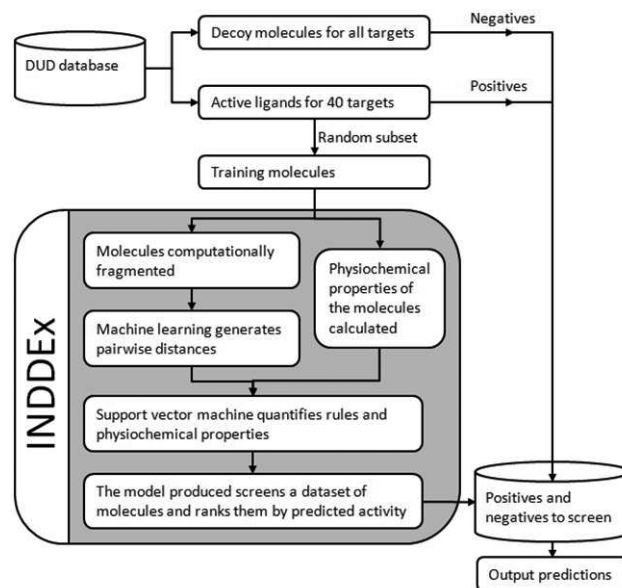
Figures are also given in the results for another measure of performance: the Area Under the Receiver Operating Characteristic (AUROC). The Receiver Operating Characteristic (ROC) curve plots true positive retrieval rate (TPR), also termed sensitivity, on the *y*-axis, against false positive retrieval rate (FPR), equivalent to 1-specificity, on the *x* axis.

$$\begin{aligned} \text{TPR} &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\ &= \frac{\text{actives}_{\text{sample}}}{\text{actives}_{\text{population}}} \end{aligned}$$

$$\begin{aligned} \text{FPR} &= \frac{\text{false positives}}{\text{false positives} + \text{true negatives}} \\ &= \frac{\text{inactives}_{\text{sample}}}{\text{inactives}_{\text{population}}} \end{aligned}$$

It can be seen that as inactives increasingly outweigh actives, the ROC curve will tend toward the enrichment curve. AUROC is the area under the ROC curve, giving a measure of sensitivity performance over the whole scanning process. AUROC measures the retrieval over the entire data set, but because virtual screening is primarily concerned with the earliest retrieved actives, the AUROC measure has been modified as BEDROC<sup>28</sup> (Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic) to use a Boltzmann distribution to weight these results more highly.

**Process.** Figure 1 shows the INDDEx process when training on a random subset of a target from the DUD database. The

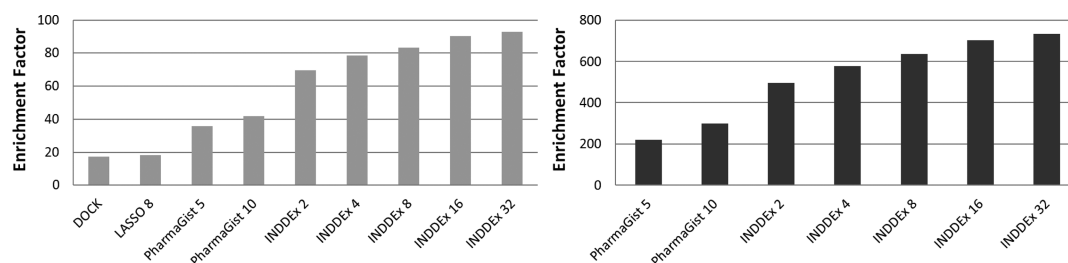


**Figure 1.** Flowchart showing a single run of the INDDEx method and how it is applied to training on a target from the DUD database.

gray-shaded area contains the processes occurring within the INDDEx program, and these processes are described in more detail below:

1. INDDEx takes in 3D minimized structures of molecules. In this case, the 3D structures randomly selected from DUD data sets.
2. Each molecule is computationally decomposed into chemically relevant substructural fragments. "Fragments" are chemically relevant interconnected groups of atoms and bonds within the molecule. INDDEx uses several different fragmentation methods corresponding to atom type, topology, charge, and hydrophobicity which can be calibrated to the bioassay. For the experiments in this paper, all tests were made using a fragmentation method where fragments are made from atom and bond type, topology, and chirality.
3. Rules are induced to describe conditions for activity, by finding presence requirements for, and pairwise distance relationships between, the substructural fragments, the fulfillment of which correlates with ligand activity.
4. A series of physiochemical properties of the molecules are also calculated. They relate to shape, weight, symmetry, log *P*, graph topology and complexity, atom connectivity, polarizability, atom type, rotatable bonds, ionization potential, and aromaticity.





**Figure 2.** Bar graph comparing enrichment factors at the 1% and 0.1% levels for multiple methodologies. Left: EF<sub>1</sub>. Right: EF<sub>0.1</sub>. The number after each method name indicates the number of ligands used as training data.

- The rules and physiochemical properties are tested for their correlation with activity among the molecules in the learning data, and the ones that have a significant correlation with activity or inactivity are used as a kernel in the next step: a correlation cutoff of 0.1 is normally used and was used to build the models in this paper. The number of rules entered into the kernel depends upon the cutoff and the data set, but typically, when the number of rules (defining both actives and inactives) drops below one hundred, the model produced begins to lose discrimination ability.
- A 2D matrix relating the rules and physiochemical properties to a binary classification of active and inactive molecules is produced, and used as the kernel of a support vector machine. The software used was SVM-Light version 6.02.<sup>29,30</sup>
- In the support vector machine calculation, each rule and each physiochemical property is considered to be a dimension, and each molecule (both active and inactive) in the training data is a vector with a value in each dimension. A support vector machine constructs an (n-1)-dimensional hyperplane through the n-dimensional space, to separate the active from inactive molecules.<sup>31</sup>

The rules that ILP derives can be readily translated into easily understood chemical statements. These intelligible rules can provide chemical information to chemists, and thus informing them about what is important in the compound, and allowing them to use this knowledge when undertaking further synthesis in order to increase the activity of the molecule. Examples of the rules are shown in the Results and Discussion section.

**Similarity Measures.** The active ligands in each DUD target set had their similarities measured using MCSS (Maximum Common Substructure). This was performed using the Small Molecule Subgraph Detector Library,<sup>32</sup> which finds the maximum contiguous common subgraph/substructure shared by two molecules using the Tanimoto coefficient,<sup>33</sup> and grouped using single-linkage agglomerative clustering. For measuring MCSS, the variables correspond to  $N_A$  = number of atoms and bonds in molecule A;  $N_B$  = number of atoms and bonds in molecule B;  $N_{AB}$  = number of atoms and bonds in the maximum common substructure between molecules A and B.

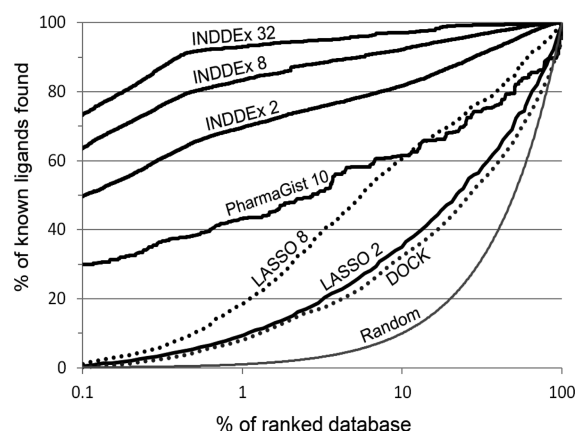
$$\tau = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

## RESULTS AND DISCUSSION

**Performance of INDDEx.** Enrichment curves were constructed for INDDEx's performance in retrieving actives from the pooled set of decoys for all 40 targets in the DUD data set, when trained on 2, 4, 8, 16, and 32 actives. The

enrichment curves were vertically averaged for all random tests on each training number, and then vertically averaged for all 40 targets (Figure 3).

Figure 2 shows the mean enrichment factors, that can be derived from the enrichment curves in Figure 3 (EF<sub>0.1</sub> results



**Figure 3.** Enrichment curves for different methods, showing the recovery of actives in each of the DUD data sets from all decoys in the DUD, vertically averaged across all 40 data sets. Each curve is labeled with the method name and the number of ligands used as training data. Results for LASSO and DOCK from Reid et al.<sup>17</sup> and results for PharmaGist from Dror et al.<sup>24</sup>

for DOCK and LASSO were not given). Figure 3 shows enrichment curves for INDDEx, averaged for all targets, compared with the results for PharmaGist and DOCK<sup>18</sup> and the results for LASSO.<sup>17</sup> The PharmaGist results are derived from data for six representative DUD targets provided in that paper.

Table 1 shows, for each DUD data set, the number of active ligands in the set, the similarity of ligands within the set, and performance of INDDEx for each set. The measure of similarity was given by finding the geometric mean of the Tanimoto similarity for an all-against-all ligand comparison (1 is all molecules identical, 0 is all molecules completely dissimilar). For comparison the geometric mean similarity of 600 compounds randomly selected from the ZINC database of drug-like molecules, and the ZINC database of all existing molecules are 0.180 and 0.163 respectively, indicating that several of the DUD data sets have a level of topological variation that is close to that found when looking at drug molecules in general. The performance is given by the EF<sub>1</sub> and EF<sub>0.1</sub> values for INDDEx when training on 2, 8, and 32 ligands (averaged across all random runs), the AUROC, which is a measure of the overall true positive retrieval rate, or sensitivity, over the entire data set, and the BEDROC, which is a version of

Table 1. DUD Screening Results Summary

ID	target data set <sup>a</sup>	active ligands	decoys	mean similarity <sup>b</sup>	EF <sub>1</sub>			EF <sub>0.1</sub>			mean AUROC <sup>d</sup>	mean BEDROC <sup>e</sup>
					2	8	32 <sup>c</sup>	2	8	32 <sup>c</sup>		
1	HIVRT	40	1519	0.219	15.2	35.4	55.6	74	273	556	0.753	0.318
2	VEGFR2	74	2906	0.199	22.4	37.3	67.8	101	192	412	0.815	0.406
3	CDK2	50	2074	0.198	21.5	51.3	80.0	166	360	640	0.832	0.494
4	PDE5	51	1978	0.204	24.2	61.1	82.1	181	408	564	0.847	0.510
5	COX-1	25	911	0.276	27.0	58.8		165	255		0.924	0.586
6	ALR2	26	995	0.255	31.7	48.1		275	426		0.926	0.578
7	PDGFRb	157	5980	0.262	37.1	72.9	95.6	298	512	668	0.875	0.621
8	InhA	85	6532	0.254	58.3	65.9	97.2	529	615	888	0.840	0.637
9	SRC	155	6319	0.286	58.1	76.2	90.7	401	572	703	0.918	0.711
10	COMT	11	468	0.265	60.0	66.7		444	667		0.988	0.844
11	thrombin	65	2456	0.347	67.4	77.8	89.1	620	699	780	0.910	0.755
12	ER agonist	67	2570	0.345	56.3	87.1	94.3	206	508	600	0.982	0.833
13	AChE	105	3892	0.353	65.9	79.8	88.0	537	699	816	0.927	0.777
14	trypsin	44	1664	0.384	75.2	82.6	92.3	684	751	923	0.937	0.808
15	HIVPR	53	2038	0.248	69.3	86.1	96.2	650	755	962	0.959	0.826
16	COX-2	348	13289	0.278	65.9	84.7	91.0	198	218	240	0.970	0.862
17	FGFR1	118	4550	0.327	71.7	91.6	99.2	605	788	969	0.946	0.831
18	ER antagonist	39	1448	0.296	65.4	100.0	100.0	562	976	1000	0.983	0.882
19	ADA	23	927	0.415	67.1	85.1		436	762		0.996	0.938
20	GR	78	2947	0.346	74.5	88.3	98.9	671	780	946	0.976	0.879
21	AmpC	21	786	0.409	78.9	88.5		547	808		0.980	0.886
22	HSP90	24	979	0.365	64.0	94.4		640	887		0.967	0.874
23	TK	22	891	0.554	84.0	85.7		390	464		0.996	0.940
24	P38 MAP	256	9141	0.340	73.9	90.4	98.4	187	215	241	0.945	0.878
25	AR	74	2854	0.311	73.2	92.3	98.8	351	514	615	0.980	0.880
26	MR	15	636	0.452	87.7	85.7		585	857		0.983	0.893
27	PR	27	1041	0.338	78.3	93.0		717	851		0.974	0.936
28	PARP	33	1351	0.498	85.5	87.9	0.0	818	840	0	0.975	0.902
29	EGFR	444	15996	0.412	83.8	91.6	92.8	173	202	213	0.993	0.939
30	Fxa	142	5745	0.266	88.3	91.4	91.6	644	656	711	0.962	0.907
31	PNP	25	1036	0.500	82.2	96.5		778	864		0.992	0.938
32	ACE	49	1797	0.345	82.6	96.6	88.2	434	673	882	0.996	0.942
33	HMGGA	35	1480	0.470	92.1	94.4	66.7	830	917	667	0.985	0.937
34	GPB	52	2140	0.489	92.8	94.1	85.0	304	523	650	0.970	0.909
35	PPAR gamma	81	3127	0.439	95.4	95.6	93.2	896	898	884	0.987	0.961
36	NA	49	1874	0.355	95.7	94.6	100.0	749	893	1000	0.998	0.976
37	DHFR	201	8367	0.353	94.4	99.7	99.9	223	241	275	0.999	0.988
38	GART	21	879	0.465	99.4	98.7		907	936		1.000	0.997
39	SAHH	33	1346	0.436	99.4	99.0	100.0	729	780	1000	0.999	0.986
40	RXR alpha	20	750	0.715	100.0	100.0		989	1000		1.000	0.999
	mean	81.0	3192	0.357	69.2	82.7	86.9	492	631	672	0.950	0.819
	std. deviation	91.3	3432	0.108	23.8	17.1	20.3	253	247	279	0.060	0.171

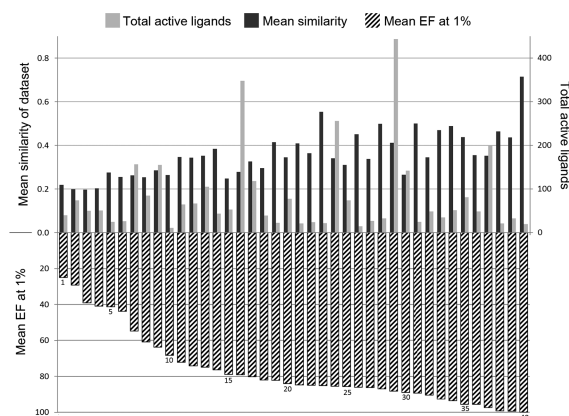
<sup>a</sup>For details of the protein targets, see Huang et al.<sup>25</sup> <sup>b</sup>As the similarity values are ratios, the mean used is the geometric mean. <sup>c</sup>There are no enrichment factor values for training on 32 molecules where there are less than 32 ligands in the target data set. <sup>d</sup>Mean AUROC when training on 2, 4, and 8 active ligands. <sup>e</sup>Mean BEDROC when training on 2, 4, and 8 active ligands. The BEDROC alpha value is set to 20, as recommended by Truchon and Bayly.<sup>28</sup>

AUROC modified to weight the true positive retrieval rate of the earliest retrieved ligands. The table shows that INDDEx performs well on most data sets; with a BEDROC mean and standard deviation 0.819 and 0.171.

Figure 4 considers the relationships of performance on a DUD target compared with the mean similarity and total number of molecules in that target. It shows the expected relationship between mean similarity of the data set and performance as measured by EF<sub>1</sub>, and the lack of a relationship between target set size and performance. INDDEx's EF<sub>1</sub> and the mean similarity of data sets have a Pearson's correlation of

0.70 even when only highly diverse data sets (mean similarity <0.3) are considered.

Table 2 gives figures for the correlation of the various metrics taken, and examines the similarity of success between INDDEx, PharmaGist, and DOCK. The correlation of DOCK's scoring with mean similarity is much weaker than that seen with the ligand-based methods, and there are also significant correlations between the metrics applied to the INDDEx and PharmaGist results. These results show that INDDEx and PharmaGist, both being ligand-based screening methods, have similar areas of success. PharmaGist's areas of success do not correlate

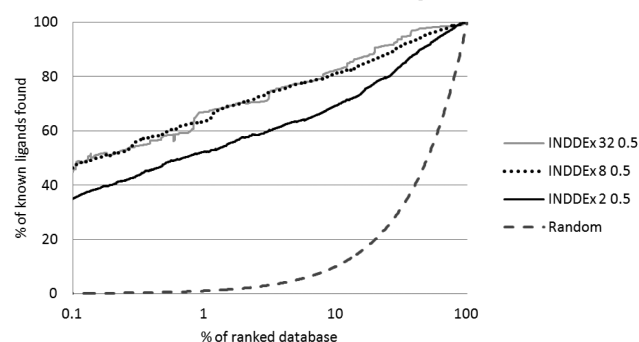


**Figure 4.** Twin bar graphs displaying data about each DUD data set. The upper bar graph shows mean similarity within each set (1 is maximum similarity) and the number of active ligands in each set. The lower bar graph shows INDDEX's  $EF_1$  performance averaged for 2, 4, and 8 actives.

significantly with DOCK, but there is some weak, but significant, correlation between INDDEX and DOCK.

**Similarities and Scaffold-Hopping.** Figure 5 shows a measure of scaffold-hopping capability, by constructing an enrichment curve only considering retrieved molecules that are scaffold-hopping challenges. A scaffold-hopping challenge is defined here as a molecule having Tanimoto similarity coefficient less than or equal to 0.5. It is shown that scaffold hopping is strongly improved when learning from 8 ligands rather than 2, but there was no improvement in performance when training on more than 8 molecules.

**Examples of Rules Found.** Table 3 shows examples of rules that INDDEX found for one of the targets (the PDGFRb, or Platelet derived growth factor receptor kinase), using 16 molecules as training data. Figure 6 shows two of the active ligands from the PDGFRb data set that conform to the rules shown in Table 3.



**Figure 5.** Enrichment factor curves for scaffold hopping challenges. Enrichment is calculated by normal method but ignores all retrieved molecules with greater than 0.5 Tanimoto coefficient of Maximum Common Structural Similarity.

**Speed of INDDEX.** On a 2.3 GHz AMD Opteron processor, INDDEX takes an average of 27 min to build a model from a set of training compounds and generate predictions for all 95 171 decoys, together with the actives in a single data set.

## CONCLUSIONS

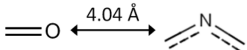
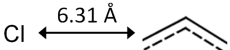
The rules formed by INDDEX (examples shown in Table 3) are in the form of pairwise distances between fragments. The intention behind using this rule format was to look for common spatial relationships, rather than creating rigid 3D pharmacophores, in order to better tolerate conformational diversity. In some targets DUD has limited topological diversity in the active ligands. In others, the diversity is close to that found in the whole drug-like molecule subset of the ZINC database. As expected, limited topological diversity gives better performance for ligand-based methods, but even in the more topologically diverse target sets, which require a greater amount of scaffold hopping, INDDEX can achieve a good rate of retrieval. Looking at scaffold hopping in isolation, INDDEX performs strongly,

**Table 2.** (Top) Pearson Product-Moment Correlation Coefficients between the Various Performance Measures, Similarity, and Total Ligands for All 40 Targets; (Bottom) Corresponding  $P$  Values for the Pearson Correlation Values<sup>a</sup>

Pearson product-moment correlation coefficients							
	mean similarity	INDDEX AUROC	INDDEX BEDROC	INDDEX $EF_1$	INDDEX $EF_{0.1}$	PharmaGist $EF_1$	DOCK $EF_1$
no. of actives	-0.17	0.02	0.07	0.08	-0.50	-0.09	-0.28
mean similarity		0.64	0.70	0.71	0.58	0.50	0.45
INDDEX AUROC			0.95	0.89	0.53	0.43	0.41
INDDEX BEDROC				0.98	0.61	0.50	0.37
INDDEX $EF_1$					0.64	0.53	0.37
INDDEX $EF_{0.1}$						0.61	0.22
PharmaGist $EF_1$							0.09
corresponding $P$ values for the Pearson correlation coefficients							
	mean similarity	INDDEX AUROC	INDDEX BEDROC	INDDEX $EF_1$	INDDEX $EF_{0.1}$	PharmaGist $EF_1$	DOCK $EF_1$
no. of actives	0.297	0.926	0.673	0.633	$1.1 \times 10^{-3}$	0.600	0.076
mean similarity		$4.1 \times 10^{-6}$	$2.8 \times 10^{-7}$	$1.4 \times 10^{-7}$	$4.8 \times 10^{-5}$	$5.7 \times 10^{-4}$	$1.9 \times 10^{-3}$
INDDEX AUROC			$5.8 \times 10^{-21}$	$1.0 \times 10^{-14}$	$2.5 \times 10^{-4}$	$2.5 \times 10^{-3}$	$4.6 \times 10^{-3}$
INDDEX BEDROC				$2.0 \times 10^{-29}$	$1.7 \times 10^{-5}$	$5.1 \times 10^{-4}$	$9.5 \times 10^{-3}$
INDDEX $EF_1$					$5.0 \times 10^{-6}$	$2.1 \times 10^{-4}$	$8.6 \times 10^{-3}$
INDDEX $EF_{0.1}$						$1.6 \times 10^{-5}$	0.085
PharmaGist $EF_1$							0.290

<sup>a</sup> $EF$ , AUROC, and BEDROC for INDDEX are the means of the training on 2, 4, and 8 ligands. Data for PharmaGist and DOCK taken from Dror et al.<sup>24</sup>

Table 3. Rule Examples<sup>a</sup>

Rule (all distances have a tolerance of 1 Ångström)	Fit to training data	Interpretation
	0.574	A nitrogen-containing aromatic ring separated from a carbonyl group by more than the distance across one ring (which is between 2.4 and 2.7 Ångströms).
	0.662	A chlorine atom separated from a phenyl ring by about twice the distance across one phenyl ring.

<sup>a</sup>Fit to training data is the Pearson correlation coefficient of the rule to the active and against the inactive molecules: 1.0 would indicate coverage of all actives only, 0.0 equal coverage of actives and inactives, -1.0 coverage of all inactives only.

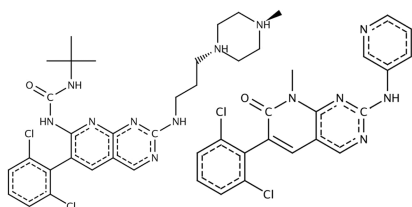


Figure 6. Examples of two active ligands from the PDGFRb target data set that conform to the rules shown in Table 3.

even when there are low numbers of active compounds to learn from.

A powerful next step, which is the subject of continuing research, is to combine docking energy calculations with the results of INDDEX to create a consensus score. A structure-based method such as DOCK has been shown to have different areas of success, and would be a good choice to create a consensus score from. The pioneering work of Harold Scheraga in developing energy forcefields has been embedded in many of these docking methods.

INDDEX has been shown to be a powerful new approach to virtual screening, whose strength lies in learning topological descriptors of multiple active compounds. This method shows potential benefits for pharmaceutical discovery and insight.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: chris\_r\_reynolds@yahoo.com.

### Notes

The authors declare the following competing financial interest(s): Michael Sternberg and Stephen Muggleton both serve as board members of Equinox Pharma Ltd. Equinox Pharma Ltd also provided some of the funding for this work. Christopher Reynolds was the recipient of the BBSRC CASE studentship award.

## ACKNOWLEDGMENTS

This work was supported by a BBSRC CASE studentship in conjunction with Equinox Pharma Ltd. INDDEX is a trademark of Equinox Pharma Ltd.

## ABBREVIATIONS

AMD, Advanced Micro Devices; AUROC, Area Under the Receiver Operating Characteristic; BEDROC, Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic; DUD, Directory of Useful Decoys; ECEPP, Empirical Conformational Energy Program for Peptides; EF, enrichment factor; eHiTS, electronic High Throughput Screening; FPR, false positive retrieval rate; ILP, Inductive Logic Programming;

INDDEX, Investigational Novel Drug Discovery by Example; LASSO, Ligand Activity by Surface Similarity Order; MCSS, maximum common substructure; PDGFRb, Platelet derived growth factor receptor kinase; ROC, Receiver Operating Characteristic; SVM, support vector machines; TPR, true positive retrieval rate; ZINC, ZINC is not commercial

## REFERENCES

- (1) Schneider, G. *Nat. Rev. Drug Discovery* **2010**, *9*, 273–276.
- (2) Muggleton, S. H.; Lodhi, H.; Sternberg, M. J. E.; Amini, A. *Support Vector Inductive Logic Programming*; U.S. Patent no. 12/066,689.
- (3) Lyne, P. D. *Drug Discovery Today* **2002**, *7*, 1047–1055.
- (4) Cavasotto, C. N.; W. Orry, A. J. *Curr. Top. Med. Chem.* **2007**, *7*, 1006–1014.
- (5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
- (6) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. *J. Med. Chem.* **1988**, *31*, 722–729.
- (7) Ewing T. J. A.; Makino S.; Skillman A. G.; Kuntz I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases <http://www.ingentaconnect.com/content/klu/jcam/2001/00000015/00000005/00333130> (accessed May 26, 2011).
- (8) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. *Proteins: Struct., Funct., Bioinform.* **2004**, *57*, 225–242.
- (9) Trosset, J.; Scheraga, H. A. *J. Comput. Chem.* **1999**, *20*, 412–427.
- (10) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361–2381.
- (11) Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 6472–6484.
- (12) Maurer, M. C.; Trosset, J.; Lester, C. C.; DiBella, E. E.; Scheraga, H. A. *Proteins: Struct., Funct., Bioinform.* **1999**, *34*, 29–48.
- (13) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (14) Willett, P. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- (15) Eckert, H.; Bajorath, J. *Drug Discovery Today* **2007**, *12*, 225–233.
- (16) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. *J. Mol. Graph. Modell.* **2007**, *26*, 198–212.
- (17) Reid, D.; Sadjad, B. S.; Zsoldos, Z.; Simon, A. *J. Comput. Aided Mol. Des.* **2008**, *22*, 479–487.
- (18) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. *Nucleic Acids Res.* **2008**, *36*, W223–228.
- (19) Muggleton, S. *NGCO* **1991**, *8*, 295–318.
- (20) Amini, A.; Muggleton, S. H.; Lodhi, H.; Sternberg, M. J. E. *J. Chem. Inf. Model.* **2007**, *47*, 998–1006.
- (21) Amini, A.; Shrimpton, P. J.; Muggleton, S. H.; Sternberg, M. J. E. *Proteins: Struct., Funct., Bioinform.* **2007**, *69*, 823–831.
- (22) Cannon, E. O.; Amini, A.; Bender, A.; Sternberg, M. J. E.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. O. *J. Comput Aided Mol. Des.* **2007**, *21*, 269–280.

- (23) Tsunoyama, K.; Amini, A.; Sternberg, M. J. E.; Muggleton, S. H. *J. Chem. Inf. Model.* **2008**, *48*, 949–957.
- (24) Dror, O.; Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. *J. Chem. Inf. Model.* **2009**, *49*, 2333–2343.
- (25) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (26) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (27) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (28) Truchon, J.-F.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (29) Joachims, T. *SVM-Light*; University of Dortmund, Informatik, AI-Unit Collaborative Research Center, 2008.
- (30) Joachims, T. *Learning to classify text using support vector machines*; Springer: New York, 2002.
- (31) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273–297.
- (32) Asad Rahman, S.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. *J. Cheminform.* **2009**, *1*, 12.
- (33) Tanimoto, T. T. *An elementary mathematical theory of classification and prediction*; IBM, 1957.