Published in final edited form as:

J Chem Inf Model. 2016 August 22; 56(8): 1597–1607. doi:10.1021/acs.jcim.6b00248.

Fragment-based Analysis of Ligand Dockings Improves Classification of Actives

Richard K. Belew[†], Stefano Forli[‡], David Goodsell[‡], T. J. O'Donnell[¶], and Arthur Olson[‡]

Richard K. Belew: rbelew@ucsd.edu

[†]Cognitive Science, Univ. California – San Diego, La Jolla CA 92093

[‡]Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla CA 92037

[¶]gNova, San Diego, CA 92104

Abstract

We describe ADChemCast, a method for using results from virtual screening to create a richer representation of a target binding site, which may be used to improve ranking of compounds and characterize the determinants of ligand-receptor specificity. ADChemCast clusters docked conformations of ligands based on shared pairwise receptor-ligand interactions within chemically similar structural fragments, building a set of attributes characteristic of binders and non-binders. Machine learning is then used to build rules from the most informational attributes for use in reranking of compounds. In this report, we use ADChemCast to improve the ranking of compounds in 11 diverse proteins from the Database of Useful Decoys-Enhanced (DUD-E), and demonstrate the utility of the method for characterizing relevant binding attributes in HIV reverse transcriptase.

1 Introduction

Virtual screening methods currently dock large ligand libraries, with hundreds of thousands of compounds, to target macromolecules. However, all but the hits with most favorable docking energies are typically discarded after the experiment. Analysis of virtual screening results across large libraries provides a unique opportunity for fine-grained characterization of a receptor structure in terms of shared features of the ligands that interact with a binding pocket. In general, ligands that bind particularly well have well-defined patterns of interaction with respect to the receptor. ADChemCast is designed to characterize these patterns, by analyzing the “cast” formed as thousands or millions of ligands are “pressed” by computational docking against the same protein binding site. This effectively forms a chemical “negative” image of the protein binding site, in terms of the receptor-ligand interaction features that are consistently shared by tightly-bound ligands and that differ from the bath of weaker-binding ligands.

We can envision building this “cast” in a variety of different ways. We could treat each interaction separately, building a list of relevant ligand-target contacts. This list could then be used to identify contacts that are particularly indicative of active compounds relative to inactive compounds. At the other extreme, we could build a complex cast from each ligand, identifying its constellation of interactions with the target. This could then be used for classification tasks by characterizing similarities between these casts for active compounds. This method suffers, however, from the size of ligand, and the fact that biological ligands are often composed of multiple functional groups that often have different chemical characteristics and bind in neighboring subsites. For ADChemCast we have chosen an intermediate approach, breaking the docked ligands into fragments and building a description of the interactions that multiple, chemically similar fragments share with a specific contact point and binding mode at of the target. These composite target-plus-fragment features are characterized, and used to classify compounds.

The central hypothesis being explored is: do receptor-ligand atomic interactions in the context of ligand fragments provide useful indices to rank and characterize docking results? One way to explore this hypothesis is to use fragment-based attributes as part of a well-defined “active” vs. “decoy” classification task like that posed by DUD-E.¹ After describing our methods in further detail, we describe ADChemCast performance on 11 targets from DUD-E in Section 3.1. In Section 3.2 we focus on the HIV-1 reverse transcriptase target to analyze features of the machine learning classifiers built to provide insight into how ADChemCast representations support their strong performance. In Section 3.4 we exploit this same representation to describe a set of reverse transcriptase inhibitors entirely independent of the ligands included in DUD-E.

1.1 Related work

ADChemCast may be described as an “interaction fingerprinting” method, sharing a common goal of supporting queries over ligand sets for analogue, parallel ligand series with different scaffolds and similar substitution patterns. “Generally, an IFP encodes a presence (1) or an absence (0) of interactions of the ligand with specified amino acids of the binding site, thus forming a binary string (bitstring). Each amino acid of the binding site is described by (some) number of interaction types (hydrophobic, hydrogen donor, hydrogen acceptor, etc.), thus all complexes of the given protein could be described by IFPs of the same length.”²

Sato and Yokoyama³ describe Pharm-IF, a system built from the PLIF residue-based interaction fingerprint tool included in MOE ver. 2007.09, distributed by Chemical Computing Group, Inc. MOE:PLIF incorporates a richer set of interaction features (“Cat”-ion interactions, ionic interaction with ligand anion, and hydrophobic interaction) than used here. Pharm-IF then forms a distance-weighted sum over all ligand atoms, rather than ADChemCast’s focus on dominant interaction features of a docking and smaller individual fragments.

Desaphy et al.⁴ also encode protein-ligand interactions as fingerprints⁴. Their pharmacophoric features of proteins and of ligands are computed independently, rather than for *receptor-ligand interactions with fragments, across many dockings*, as proposed below.

Their method also focuses on *merging* atoms into pseudo-atoms based on bond-type, vs. the RECAP fragmentation via breaking bond-types. The entire interaction is encoded in terms of a “pseudo-atom” located at one of three (center, protein, ligand) positions. They also limit each protein amino acid to be included in only a single interaction. On the two target systems shared between their experiments and those reported here, ADChemCast had much better classification performance: on ADA, Desaphy et al. report AUROC=0.749 while ADCC_LOGR had AUROC=0.912; on PGH2, Desaphy et al. report AUROC=0.626 while ADCC_LOGR had AUROC=0.841. Two other related efforts^{5,6} are mentioned in Section 3.1.

The inSARA project⁷ connects the Matched Molecular Pair⁸ approach of generating fragments by deleting acyclic single bonds, reduced graph methods which consider fingerprints built over pseudo-atoms, and Maximum Common Substructure graphical techniques. The result is an elegant, hierarchic model of how the fragments generated via RECAP or other bond-breaking rule sets might be related.

2 Methods

ADChemCast characterizes ligand-target interaction in several steps, as summarized here and in Figure 1, and described in more detail below. The “cast” is built from two pieces of information: First, computational docking is used to determine bound poses and predicted binding free energies for a Ligand library and the docked conformations are analyzed to identify its RECEPTOR-LIGAND INTERACTION FEATURES (RLIF) with the target (hydrogen bonds, vdw contacts, etc). Second, each of the ligands is decomposed into a set of fragments. This interaction and fragment information may be envisioned as a sparse array of (a) ligand fragments (shown on the vertical axis in the Figure) by (b) RLIF (on the horizontal axis). The set of points falling on vertical lines in this array correspond to multiple fragments that all share the same interaction with the receptor—we will refer to this set as *frag_{RLIF}*. Similarity among fragments is computed (using two-dimensional fingerprints as implemented in OpenBabel; further details in Section 2.3) Each set of points *frag_{RLIF}* is then clustered to identify chemically-similar fragments, resulting in a cluster of fragments with the same docking profile with respect to a particular RLIF, termed a FRAGMENT-QUALIFIED RLIF (RFQ). The very large number of vdW interactions require a separate composite feature combining them into sets we call vdwPatches.

Classification experiments then use docking energy, RLIF, RFQ, and vdwPatches as attributes. For training, RFQ and other attributes are calculated for a set of actives and decoys, either from explicit annotation as in DUD-E or by classifying compounds based on a docked energy threshold, and a classification method identifies particular attributes that distinguish actives and decoys. For classification of unknowns, test compounds are decomposed into fragments, docked, and their attributes are compared to those in the training set.

An overview of the ADChemCast workflow is shown in Figure 2. Python source code for all analysis steps is available at <https://github.com/rbelew/mgl/>

2.1 Fragmentation

The goal of the fragmentation step is to deconstruct a ligand into a series of smaller, chemically-relevant components that retain features of the larger ligand as it interacts with subpockets within a target binding site. For the current version of ADChemCast, we use the RECAP procedure,⁹ which breaks ligands (represented as SMILES strings) based on an ordered list of bond breaking rules specified as SMARTS patterns. Figure 3 shows the types of bonds that can be broken according to the rules used in experiments here, as implemented by one of the authors.^{10,11} The result of applying RECAP is a set of Fragments, also represented as SMILES strings, associated with each ligand. RECAP fragmentation can be applied to any ligand library, characterizing its fragment distribution and independent of virtual screening with respect to any particular receptor structure. As an example, the HIVRT ligand set produced an average of 2.7 (s.d.=0.89) fragments per ligand. Also note that fragmentation need only be done once for each ligand in a library and re-used in analysis of any protein, and also that the bonds broken to create the fragments can be retained as candidates for subsequent chemical synthesis.

2.2 Receptor-ligand interaction features

Receptor-ligand interaction features (RLIF) capture the dominant ligand-target interaction components of the predicted interaction energy such as van der Waals contacts (vdW), hydrogen bonds, metal coordination, and aromatic ring stacking. We characterize RLIF as an ordered sequence of features encoded in a text string:

1. A receptor *chain* of the full PDB structure
2. A *residue* within the chain
3. A *residue atom* within the residue
4. A *type of interaction* between this atom, define with respect to
5. The *ligand atom* involved in the interaction

Figure 4 shows an example of one such interaction. A_025ASP_OD1:HBD_O refers to a hydrogen-bond-donor interaction between the OD1 oxygen atom of the aspartate 25 residue in chain A of a HIV protease structure (PDB:1hxb) and a ligand oxygen atom in a fragment of saquinavir. The colon is used to reinforce the fact that the hydrogen bond donor/acceptor relation is defined with respect to the ligand; i.e., in this example it is the ligand providing the hydrogen bond.

Receptor atoms are identified by the standard PDB atom label from the original protein structure (e.g., “OD1”). For some residues, there is a potential ambiguity in numbering their atom indices: OD1/2 in ASP, OE1/2 in GLU, and CD1/2 CE1/2 in PHE and TYR. These are only of consequence when trying to compare across structures; an example is mentioned in Section 3.3.

Ligand atoms include only the atoms’ type; i.e., the ligand atom is identified by its atomic element (“O”), even though it may be indexed (“O2”) in the PDB file (indicating it is the second oxygen atom listed), and perhaps described as a particular type (“OH”) of oxygen (hydroxyl). The current experiments use just ligand atom names to avoid problems with

inconsistent names when compared across multiple fragments, but extending our methods to incorporate full ligand atom types is an important direction for further work.

The ordering of the features composing a RLIF suggests a natural hierarchy of alternative attributes to describe the interaction. For example, a crude feature set would simply mention a particular residue involved in the interaction; a more refined feature set would include the particular atom in the residue; specifying the type of interaction generates even more features etc. In classification tasks there will therefore be a trade-off between the number of available attributes, their frequencies, and informativeness of each feature. For example, if we ask how many tightly binding ligands involve an interaction with a particular residue, this gross attribute will potentially mask useful discriminations made by attributes specific to particular atoms, interaction types, etc. Despite the obvious correlational structure this generates among attributes, in experiments reported here, two levels of this hierarchy are treated independently: the full RLIF and a short RLIF that includes only information on the target amino acid.

In the current study, RLIF are assigned from docked conformations using AutoDock Raccoon.¹²

2.3 Combining RLIF with fragment clusters

Once ligands are fragmented and RLIF are assigned, ADChemCast then collects all ligands' fragments associated with a particular RLIF, and clusters these fragments to assign one or more RFQ. Chemical similarity is computed over just those fragments sharing the same RLIF binding site, and between fragments, not the full ligands from which they are extracted. A key assumption in this work is that chemical similarity among multiple fragments sharing a binding mode with the same RLIF provides a robust feature to classify members of a ligand library, so RFQ features that do not occur frequently are filtered. In experiments reported here, only singleton fragments are excluded; i.e., only fragments binding to a RLIF shared by more than one ligand are considered. Figure 4 gives one example of what a RFQ feature might look like, combining the RLIF A_025ASP_OD1:HBD_O with the bound fragment NC(C(O)C)Cc1ccccc1 to form the RFQ feature RLIF A_025ASP_OD1:HBD_O+NC(C(O)C)Cc1ccccc1.

Openbabel and Pybel^{13,14} are used for all calculations. Fragments are encoded as SMILES strings, their fingerprints calculated using `MOLECULE.CALCFP()` and its default FP2 fingerprint type, and the Tanimoto distance metric is applied. Hierarchic clustering is then performed, using the Ward method as recommended by Willett and others¹⁵ (`fastcluster`'s implementation of Ward's method is used for the hierarchic clustering: `FASTCLUSTER.LINKAGE(DISTVEC,METHOD='WARD')`).

Note that while performing Ward clustering across the $\sim 10^4 - 10^6$ fragments associated with *all* ligands in a full library would require significant computational resources, clustering just the fragment subsets associated with each *frag_{RLIF}* generally involves only $\sim 10^1 - 10^3$ of fragments, and so Ward clustering is tractable on typical computers. The hierarchic clustering is then flattened into *NClust_{RLIF}* clusters (`SCIPY.CLUSTER.HIERARCHY.FCLUSTER(CLUSTERING, DISTTHRESH, 'DISTANCE')` `simThresh = 0.7`,

distThresh=0.5, held constant across all experiments.) Each cluster can be identified by its most central fragment Ctr_{RLIF}^i identified as the fragment with maximal similarity across all other fragments.

2.4 van der Waals patches

van der Waals (vdW) forces between receptor and ligand atoms are relatively short range and play an important role in defining shape complementarity as part of calculating docking energy. They pose a challenge, however, since a typical docked pose will provide an order of magnitude more vdW interactions than all other interaction types (hydrogen bond donor, acceptor, etc.) combined. To address this problem, we use a vdWPatch rather than pairwise vdW interaction, defined to be the *set* of those receptor atoms interacting with a particular ligand atom.

While exact membership between two receptor atom subsets might vary a bit, we can relax exact matching. For instance, we could assign a match to subsets overlapping except for a single receptor atom; these could be considered “nearly” the same vdWPatches. Jaccard similarity¹⁶ generalizes this matching between subsets and provides a robust clustering to a smaller set of shared vdWPatch clusters. As an example from the HIVRT target, A106V:[CB,CG1,CG2];A227F:[CD2,CE2] refers to a vdWPatch in which ligands have vdW interactions with the CB,CG1,CG2 atoms of residue A106V as well as vdW interactions with CD2 and CE2 atoms of residue A227F. Using the Jaccard similarity measure, it represents the *centroid of a cluster* of very similar vdW interactions across all HIVRT ligands, including:

- A106V:[CB,CG1,CG2];A227F:[CE2]
- A106V:[CB,CG1,CG2];A227F:[CD2,CE2]
- A106V:[CB,CG1,CG2];A227F:[CD2,CE2];A234L:[CB]
- A106V:[CB,CG1,CG2];A227F:[CD2,CE2,CZ]

These protein features are highlighted in Figure 5, together with all the atoms in a random ligand (ZINC04869408) that (each) share this vdWPatch of interactions with the receptor.

2.5 Classification Tests

We evaluated the effectiveness of ADChemCast in a classification task of 11 proteins from DUD-E. DUD-E provides a set of known actives and a larger set of decoys, chosen to have similar size and chemical characteristics as the active set, and has become a central tool for benchmarking docking and design methods. Ligands and receptors were processed using the default methods in AutoDockTools,¹⁷ and docked using AutoDockVina.¹⁸ A box size was chosen to encompass the binding site as defined by the crystallographic ligand included with the DUD-E data. For the analysis, the docked pose of best predicted interaction energy was used.

ADChemCast classifiers are built over four types of attributes: 1) ligand docking energy, a numeric attribute, 2) individual RLIF, 3) RFQ features, and 4) vdWPatch. In experiments reported here, several classifiers were applied using the WEKA platform.¹⁹ In all cases,

results reported are based on 10-fold cross-validation testing. Results here will focus on the LOGR logistic regression classifier²⁰ (WEKA.CLASSIFIERS.FUNCTIONS.LOGISTIC, default parameters: -R 1.0E-8 -M -1), as an example of an encoding that can be interpreted easily, with broadly robust performance. The result of logistic regression classification is a set of “coefficients” linearly weighting each attribute’s contribution to the cumulative logit (log odds):

$$\ln \frac{Pr(X)}{1 - Pr(X)} = \beta_0 + \sum_a \beta_a a$$

where X represents the vector of attributes a .

3 Results

3.1 Classification of active compounds in DUD-E

The Database of Useful DecoysEnhanced (DUD-E) data set provides 102 separate targets to “... help benchmark molecular docking programs by providing challenging decoys.”¹

Based on sets of known ACTIVE compounds, DUD-E generates DECOY ligands with similar physicochemical properties (e.g. molecular weight, LogP) but dissimilar topology that are likely to be non-binders. The classification method being tested is then given the challenge of labeling of all ligands in the collection as either ACTIVE or DECOY. For validation of the classification ability of ADChemCast, we used a subset of DUD-E. Two HIV systems of interest to the laboratory were included, along with nine additional systems chosen at random.

Docking of the DUD-E sets, similar to results from large ligand libraries typically used in virtual screening (data not shown), shows that the non-specific compounds form a “docking bath”, which together provide a characterization of ambient chemical interactions which are probed by the docking method as it samples the most accessible protein regions for favorable binding. For typical ligand libraries, these large sets of heterogeneous ligands generate a normal distribution of relatively low docking energies. Even for the unusual set of compounds included in DUD-E, Figure 6(a) shows the near-normal distribution of energies for DECOY ligand sets for the HIVRT DUD-E experiment, and also considerable overlap between the distributions for the two sets. Simply using a threshold on docking energy as a single criterion provides one straight-forward, classical signal detection approach to ACTIVE classification.

But analysis of these energy distributions revealed that the HIVRT system included in DUD-E had a potential problem: the energy distribution for ACTIVE compounds was seen to be strongly bimodal, with one peak of very favorable energy and a secondary peak with energies worse than typical decoys. Manual curation of the HIVRT ACTIVE set established that this secondary peak corresponded to several nucleoside inhibitors (NRTI), such as AZT; no NRTI should be expected to bind into the non-nucleoside inhibitor binding site of the enzyme. When these were removed from the ACTIVE set for the ADChemCast study, a more typical distribution shown in Figure 6(b) was produced.

This issue with DUD-E's HIVRT target, and another with the AMPC target discussed in Section 3.1, should make us question the validity of all DUD-E based conclusions. In the analysis below, however, we do not address consequences of any such invalidity (e.g., how the false negative rate is effected by matching properties when not excluding 3-D similar compounds). For all its flaws, DUD-E remains a nearly unique, available test set for evaluations such as ours.

The goal of ADChemCast is to improve classification by adding specific information on characteristic interactions between ACTIVE ligands and their targets. Only the 11 DUD-E targets reported were examined; the reported results were not “cherry-picked” from any larger set. All experiments reported here used logistic regression with all parameters ($-R$ 1.0E-8 $-M$ -1) held constant across all targets. While other classification methods were evaluated, logistic regression proved to be the most robust; results on this particular variant are labeled ADCC_LOGR below.

To evaluate ADCC_LOGR, we performed ten-fold cross-validation, in which the 11 active/decoy-labeled DUD-E systems were repeatedly divided into random subsets with training of the classifier performed on 9/10 of the data and then tested on the remaining 1/10. The reported results are the average across these ten experiments. All ADVina and ADCC_LOGR parameters were held constant across all targets.

The overall statistics for each classification are shown in Table 1. As shown in Table 1, by design all DUD-E experiments have only a small number (average 3%) of ACTIVE ligands against a backdrop of primarily DECOYS. Also shown are number of ligands, number of hydrogen bond donors and average number per ligand, number of hydrogen bond acceptors and average number per ligand, number of pi-stacking interactions and average number per ligand, number of RLIF generated, number of vdWatches and average number per ligand, and the number of total attributes used in classification.

One striking feature of these statistics is how different target systems are differentially represented by the various classes of interaction:

- while the number of unique hydrogen bond acceptor and donor acceptor interactions is relatively constant across targets, the number of ligands in which they play a role varies considerably. For example, hydrogen bond acceptor interactions in beta-lactamase (AMPC) occur with approximately 8 times as many ligands as in the dihydroorotate dehydrogenase (PYRD) target. This may be related to the problematic labeling of ACTIVE compounds by the DUD-E authors; see below.
- the number of vdWatches varies considerably across targets, as does the number of ligands making these interactions.

Table 1 also lists the number of ligands with “E-only.” This small fraction of ligands (around 0.5%) had no ADCC_LOGR attributes associated with them beyond the docking energy attribute.

Figure 7 shows results of classification for the 11 DUD-E experiments performed. In this plot (consistent with Figure 5 in the DUD-E reference),¹ the horizontal axis of a standard ROC (receiver operating characteristic) plot has been log-transformed to accentuate performance in the range relevant to biochemical enrichment.

In general, classification results are very good, particularly in the early, very low-decoy fractions which are most useful in biochemical enrichment experiments. The major outlier is AmpC (beta-lactamase), which showed particularly poor performance. Looking more closely at the data provided by DUD-E, we find that 17 of the 48 ACTIVE compounds have reported binding constants of worse than the value of 1 millimolar assigned as the default binding strength of decoys, so perhaps it is not surprising that this system is a particular challenge.

While ADCC_LOGR performance in the low-decoy fractions is most striking, considering classification across the full range allows comparison using the AUROC (area under ROC curve, a.k.a. AUC) classification measure. Table 2 shows AUROC for ADCC_LOGR classifier against several other published results. The first comparison is with respect to those reported with the original DUD-E publication.¹ With the exception of the AMPC target, ADCC_LOGR is superior. Some overlapping results have been reported by Ruiz-Carmona et al.,⁵ using several docking engines including AutoDockVina. At least on these experiments, ADCC_LOGR performance was again superior to both AutoDockVina and the rDock engine they propose; on the HIVRT target and using the GLIDE docking system, however, their results were superior to ADCC_LOGR. Very recently, Lesnik et al have described a range of techniques to calculate ligand similarity using clique algorithms.⁶ In their supporting information, the authors provide AUROC statistics for two alternative two- and three-dimensional methods, across two separate tables S1 and S2 of results. Because it is unclear which method they recommend as generally preferred, Table 2 includes the *best* AUROC value across all their methods, and specifies which variant and table from Lesnik et al.'s supporting information was referenced.

Best results for each target are highlighted in bold. For seven of the 11 targets, ADCC_LOGR performance was superior or equal to all other results.

3.2 Attribute characteristics

While the full range of DUD-E experiments help to demonstrate the robustness of ADChemCast analysis in classification tasks, in this and the next section we focus on the HIVRT experiment to illuminate the fine-grained details of binding characteristics of a receptor structure that the trained classifiers exploit. With this target as with all others, it is first important to realize that the very low fraction of ACTIVES within DUD-E experiments means that any classification method can attain high accuracy by just guessing decoy in all cases. The importance of the single attribute energy in docking experiments has already been mentioned, and most current virtual screening methods employ classifiers using this single feature. A relevant question in these experiments is therefore the relative “apportionment of credit”: beyond just using the variable ENERGY, how much of ADChemCast's performance is due to RLIF-only attributes, to RLIF+fragment attributes, and to vdWatches? Table 3 shows the results of experiments using the same logR classifier on the HIVRT example, varying

only the attributes on which classification was based. As suggested by Figure 6, using only docking energy to discriminate ACTIVES from DECOYS is quite ineffective. Using just RLIF-only attributes, using RLIF+fragment attributes, and using vdwPatches all bring clear improvements, but combining them all gives the best performance. Putting these results for HIVRT in context, Table 1 shows the wide variation in the distribution of these attribute types across the diverse set of DUD-E targets, discussed Section 3.1.

One broadly useful characteristic of all three varieties of attributes is their *information gain*: the reduction in any distribution's (in this case ACTIVES across all ligands) Shannon entropy H provided by knowledge of the attribute a 's value:

$$\text{InfoGain}_a = H(\text{Active}) - H(\text{Active}|a) = H(X) - \sum_i P(x_i) \log_2 P(x_i)$$

Table 5/Appendix A enumerates the 50 most informative attributes associated with HIVRT. Typically, the energy attribute is ranked first in these lists, followed by mixture of informative RLIF-only, fragment-qualified, and vdwPatch attributes that together tune the classification of compounds that are missed by the energy attribute. For HIVRT, RLIF-only and vdwPatch dominate the list of most informative attributes, but in other systems, fragment-qualified attributes contribute more strongly.

For each attribute, also shown are:

- the amount of information the attribute conveys on this task
- the type of the attribute: RLIF, RFQ (fragment-qualified RLIF), vdwP (vdwPatch)
- the receptor atom, interaction type and ligand atom (RA_I_LA) of the attribute
- RFQ also include the attribute's fragment qualifier, separated by a circumflex (RA_I_LA^fragment)

3.3 Characteristics of DUD-E systems

Given the problematic labeling of ACTIVE compounds for the HIVRT target mentioned in Section 3.1, it is worthwhile noting two other aspects of the complexes provided by DUDE. First, the DUD-E decoy sets are skewed in the sense that they have been developed to make a particularly difficult test for classification, not to span the space of potential ligands and fragments. In future work, we intend to explore the ability of larger ligand libraries to provide a more general description of the interaction attributes.

Second, our analysis is necessarily dependent upon the protein structure chosen by DUDE's authors, which can be problematic for target structures with significant mobility. A good example is provided by contrasting results on HIVPR using the PDB:1XL2 structure included in DUD-E vs. an alternative structure, PDB:3KF0. As mentioned in the paper introducing the PDB:1XL2 structure,²¹ this structure supports "an unexpected binding mode" relative to most other structures and known inhibitors. PDB:3KF0 is much more

typical. Figure 8 compares classification results across these two systems using a simple Naive Bayes classifier, and demonstrates that the PDB:3KF0 receptor structure provides a much more useful reference frame for distinguishing DUD-E's HIVPR ACTIVE vs. DECOY ligands. Table 4 gives an indication of how this is possible, by listing some of the RLIF which were differentially engaged by the same set of 37672 HIVPR ligands across the two structures. For example, residue A_025D (specifically A_025D_OD2_hba_O, etc.) proved *inaccessible* to all of the ligands with respect to PDB:1XL2, but was involved in nearly a third of the dockings against PDB:3KF0; the dimeric B_025D residue was a bit more accessible. Conversely, A_008R_NH2_hba_O was involved in more than twice as many dockings with respect to PDB:1XL2 than PDB:3KF0. Differential analysis of alternative structures for the same protein, and sensitivity of ADChemCast methods to structural variations will be one of the most important areas for further work.

The potential ambiguity generated by the isomorphism in amino acid side chain atoms mentioned in Section 2.2 arises in the comparison of PDB:1XL2 and PDB:3KF0. It happens the two structures have used the same naming conventions for the ASP atoms in the HIVPR active site, and while their labeling of several GLU, PHE and TYR residues are not consistent, these do not generate potentially inconsistent RLIF.

3.4 Extrapolating beyond DUD-E

A collection like DUD-E – biologically validated ligands against chemically similar but inactive ones – is invaluable in the development of screening technologies. ADChemCast analysis can also be used to create an interaction-based description of the binding site, for use in drug optimization or design. Figure 9 shows the binding site of HIVRT, highlighting the ADChemCast attributes shown in Table 5/Appendix A to provide the most information for classification. The top RLIF, RFQ and vdWPatch all correspond to the interaction of ligands with a small pocket with residue K101 at its base. Structures of two approved NNRTI (Etravirine and Rilpivirine, PDB:3MEE and PDB:3MEC), not included in the DUD-E set of HIVRT ligands, reveal that they take advantage of this same pocket, binding with an aromatic group that fills the space and forms a hydrogen bond with the K101 carbonyl. ADChemCast provides a descriptive vocabulary of RLIF and fragments, generated from a ligand collection developed for much different purposes, but recapitulating key features of other inhibitors' bindings in technically precise terms.

4 Discussion

The results presented here are primarily testing ADChemCast as a tool for classifying trial compounds in systems that have significant prior work, which have provided a set of known actives. Combination of these known actives with a bath of decoys is used to build a description of attributes that capture unique aspects of specificity that are not represented in the general scoring function used during the initial docking analysis. As demonstrated in the results presented here, inclusion of these specific attributes can improve classification, providing a more accurate ranking when performing screening or drug optimization tasks. A major goal of future work will be to generalize the method to remove the need for labelled

actives, for instance, by assigning the top-ranked compounds from a traditional virtual screen as the actives, and using this to generate the classification attributes.

When designing a fragment-based approach, we can imagine two different methods: we could dock and analyze all the fragments separately, or we could dock the ligands and then break them into fragments. ADChemCast takes the latter approach, docking ligands, then analyzing them in terms of their component fragments. Generally, this is considered an advantage over methods considering the binding affinity of short fragments in isolation^{22,23}. For example, Barelier et al deconstructed known inhibitors into fragments and conclude:

“...even if the fragments bind in the same overall region as the initial inhibitors, they do not necessarily retain the exact position they had in the lead they were extracted from. ...The fragment molecules can (1) exhibit no detectable affinity for the protein, (2) bind to the protein in a similar position as in the lead they come from, or (3) bind to the protein in binding sites different from the original lead.”²³

This supports the hypothesis that it is advantageous to analyze fragments' positions and interactions as determined by their location within larger molecules. One future goal for ADChemCast is to support fragment-based construction of new compounds, by recombining fragments from different ligands that optimize local interactions, and using the existing linkages of these fragments within the ligand to guide recombination.

4.1 Future work

Current ADChemCast methods described certainly have limitations. First, the methods build and currently depend on use of AutoDock and AutoDockTools software.²⁴ It has been cited more than 2500 times, and was downloaded more than 14,000 times in 2014 alone. We provide full Python source code for ADChemCast analysis steps (<https://github.com/rbelew/mgl/>), to allow this existing community of users to build upon the available open source AutoDock tools. However, the ADChemCast methods described do not depend on any specific details of AutoDock, and applying them to other open source docking systems, and/or to the docked results of proprietary docking systems, may be important extensions. Second, these experiments reveal some ligands that contained particularly large fragments associated with macrocycles, and a procedure to break these prior to fragmentation is being explored. Third, incorporating richer pharmacophoric representations (e.g., hydrophobic/-philic interactions, and other features mentioned in Section 1.1) will be an important extension. Fourth, while all RLIF, RLIF+fragment and vdwPatch attributes have been treated independently in the current analysis, correlation among them is evident and much more effective and insightful representations of intra-residue atom interactions are possible. Other experiments (not reported here) suggest that more elaborate modeling of correlated attributes, other classifiers (particularly support vector machine and deep neural networks), mixtures of classification methods, and other machine learning techniques can improve reported classification performance considerably. Fifth, experiments across much larger ligand libraries, such as those provided by the FightAIDS@Home project,^{25,26} will allow us to include a much broader set of fragments than those in DUD-E, and provide more precise guidance towards chemical synthesis. Sixth, the dichotomous classification of ACTIVES vs DECOYS as posed by the DUD_E experiment is an academic exercise in comparison to the

more realistic *regression* tasks that ADChemCast can perform with respect to varying activity measures from biological assays. We are also experimenting with using simply the distinction between very high-affinity docking solutions in contrast to the rest of a virtual ligand library to apply ADChemCast methods when no biological data is available, as well as using existing ligands (e.g., approved inhibitors or lead compounds) as query probes against the indices built by ADChemCast. Finally, we recognize all *in silico* docking and cheminformatic methods can currently capture only some aspects of the criteria required for high affinity ligand binding, let alone biologically relevant viral inhibition and drug approval.

In summary, ADChemCast introduces a novel form of protein-ligand interaction description connecting recurrent features of a protein's docking with shared ligand fragments. This vocabulary of features has been shown to be successful at distinguishing active compounds from DUD-E decoys across multiple targets, and in the case of HIV-1 reverse transcriptase, useful as a way of characterizing exploitable features of known inhibitors. The extraction of this additional information, beyond the typical docking energy score, bodes well for large-scale *in silico* searches through the massive ligand libraries now available.

Acknowledgments

The authors dedicate this paper to our memory of C. David Stout, a long-term colleague who has patiently explained what the structures have meant for many years. We acknowledge Daniel N. Santiago's help with DUD-E AutoDock runs, and Max Chang's comments on an earlier manuscript. The authors were supported by NIH #P50GM103368.

References

1. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 2012; 55:6582–6594. [PubMed: 22716043]
2. Chupakhin V, Marcou G, Gaspar H, Varnek A. Simple Ligand-Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* 2014; 10:33–37. [PubMed: 25210596]
3. Sato T, Honma T, Yokoyama S. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for *in Silico* Screening. *J. Chem. Inf. Model.* 2010; 50:170–185. [PubMed: 20038188]
4. Desaphy J, Raimbaud E, Ducrot P, Rognan D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* 2013; 53:623–637. [PubMed: 23432543]
5. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* 2014; 10:e1003571. [PubMed: 24722481]
6. Lešnik S, Štular T, Brus B, Knez D, Gobec S, Janežič D, Konc J. LiSiCA: A Software for Ligand-Based Virtual Screening and Its Application for the Discovery of Butyrylcholinesterase Inhibitors. *J. Chem. Inf. Model.* 2015; 55:1521–1528. [PubMed: 26158767]
7. Wollenhaupt S, Baumann K. inSARA: Intuitive and Interactive SAR Interpretation by Reduced Graphs and Hierarchical MCS-Based Network Navigation. *J. Chem. Inf. Model.* 2014; 54:1578–1595. [PubMed: 24850242]
8. Wawer M, Bajorath J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* 2011; 54:2944–2951. [PubMed: 21443196]

9. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Model.* 1998; 38:511–522.
10. O'Donnell, T. Design and use of relational databases in chemistry. CRC Press; 2008.
11. O'Donnell, TJ. [Accessed: 28 June 2016] Python implementation of the RECAP fragmentation algorithm. <http://gist.github.com/95387>
12. Forli, S. [Accessed: 2015-07-14] AutoDock-Raccoon2: virtual screening analysis tool. <http://autodock.scripps.edu/resources/raccoon2>
13. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G. Open Babel: An open chemical toolbox. *J. Cheminf.* 2011; 3:1–14.
14. O'Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the Open-Babel cheminformatics toolkit. *Chem. Cent. J.* 2008;2. [PubMed: 18269743]
15. Willett P. The Calculation of Molecular Structural Similarity: Principles and Practice. *Mol. Inf.* 2014; 33:403–413.
16. [Accessed: 27 Apr 2016] Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index
17. Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* 2016; 11:905–919. [PubMed: 27077332]
18. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 2010; 31:455–461. [PubMed: 19499576]
19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter.* 2009; 11:10–18.
20. Le Cessie SJCVH. Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* 1992; 41:191–201.
21. Specker E, Böttcher J, Lilie H, Heine A, Schoop A, Müller G, Griebenow N, Klebe G. An Old Target Revisited: Two New Privileged Skeletons and an Unexpected Binding Mode For HIV-Protease Inhibitors. *Angew. Chem., Int. Ed.* 2005; 44:3140–3144.
22. Babaoglu K, Shoichet BK. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* 2006; 2:720–723. [PubMed: 17072304]
23. Barelier S, Pons J, Marcillat O, Lancelin J-M, Krimm I. Fragment-based deconstruction of Bcl-xL inhibitors. *J. Med. Chem.* 2010; 53:2577–2588. [PubMed: 20192224]
24. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* 2009; 30:2785–2791. [PubMed: 19399780]
25. [Accessed: 28 June 2016] Fight AIDS @ Home. <http://fightaidsathome.scripps.edu/>
26. Chang MW, Lindstrom W, Olson AJ, Belew RK. Analysis of HIV Wild-Type and Mutant Structures via in Silico Docking against Diverse Ligand Libraries. *J. Chem. Inf. Model.* 2007; 47:1258–1262. [PubMed: 17447753]

A: Informative attributes

Table 5

Informative attributes

Info	Type	RLIF / RFQ / vdwp
7.62E-3	e	e
3.58E-3	RLIF	A_101K_O_hbd_N
3.04E-3	RLIF	A_101K_O_hbd
1.95E-3	RFQ	A_101K_O_hbd_N ^{CCc1cc(N)c(=O)nc1C}
1.81E-3	RLIF	A_101K_NZ_hba_O

Info	Type	RLIF / RFQ / vdwP
1.53E-3	RLIF	A_101K_NZ_hba
8.83E-4	vdwP	A+100L:[CA,CB,CD1,CD2,CG]; A+101K:[C,O]; A+103K:[CG]
6.71E-4	vdwP	A+100L:[CD1]; A+234L:[CB,CD1]
6.15E-4	vdwP	A+188Y:[CB,CD1,CD2,CE2,CG];A+234L:[CD1]
5.95E-4	RLIF	A_101K_N_hba
5.75E-4	vdwP	A+106V:[CG1]; A+227F:[CD2,CE2]
5.66E-4	vdwP	A+103K:[CG]; A+106V:[CB,CG1]; A+189V:[C]; A+190G:[CA,N]
5.30E-4	RLIF	A_103K_NZ_hba
5.14E-4	vdwP	A+100L:[CD1,CD2]; A+181Y:[CD2]
4.42E-4	RLIF	A_101K_N_hba_O
4.28E-4	vdwP	A+188Y:[CB,CG]
4.08E-4	RLIF	A_103K_NZ_hba_O
3.98E-4	vdwP	A+100L:[CD2]; A+181Y:[CB,CG]
3.93E-4	RLIF	A_189V_N_hba_S
3.93E-4	vdwP	A+100L:[CB,CD1,CG]; A+101K:[O]; A+103K:[CG]; A+106V:[CG1]
3.72E-4	RLIF	A_103K_N_hba_S
3.66E-4	vdwP	A+100L:[CB,CD1]; A+101K:[C,O]; A+102K:[CA]; A+318Y:[CD1,CD2,CE1,CE2,CG,CZ,OH]
3.44E-4	vdwP	A+100L:[CB,CD1,CG]; A+101K:[C,O]; A+103K:[CG]
3.29E-4	RLIF	A_101K_NZ_hba_S
3.26E-4	RLIF	A_190G_N_hba
2.98E-4	vdwP	A+103K:[CG]; A+106V:[CG1]; A+190G:[CA]
2.94E-4	RLIF	A_318Y_OH_hba_O
2.91E-4	RLIF	A_101K_N_hba_S
2.88E-4	vdwP	A+100L:[CD2]; A+181Y:[CB,CD2,CG]; A+188Y:[CB,CG]
2.88E-4	vdwP	A+100L:[CD1]; A+234L:[CB,CD1,CG]; A+318Y:[CE1,CZ,OH]
2.81E-4	RLIF	A_190G_N_hba_O
2.80E-4	vdwP	A+100L:[CD2]; A+179V:[CG1]; A+181Y:[CA,CB,CG]
2.80E-4	RLIF	A_318Y_OH_hbd
2.62E-4	vdwP	B_138E_OE2_hbd
2.57E-4	vdwP	A+101K:[O]; A+103K:[CG]
2.55E-4	vdwP	A+179V:[CG1]
2.51E-4	vdwP	A+104K:[O]; A+106V:[CG2]; A+236P:[CA,CB,CG]
2.50E-4	RLIF	A_318Y_OH_hbd_N
2.47E-4	vdwP	A+099G:[O]; A+101K:[CD,CE,HZ2,HZ3,NZ]; B+138E:[OE2]
2.47E-4	vdwP	A+106V:[CG1,CG2]; A+227F:[CD2,CE2]; A+234L:[CB]; A+235H:[C,O]; A+236P:[CA]
2.43E-4	vdwP	A+106V:[CG2]; A+225P:[CB,CG]; A+236P:[CA,CB,CG]
2.38E-4	vdwP	A+106V:[CB,CG1]; A+188Y:[C,CB]; A+189V:[C,CA,N]; A+190G:[CA,N]

Info	Type	RLIF / RFQ / vdwP
2.35E-4	vdwP	A+095P:[CB,CG]; A+181Y:[CD2,CE2,CG,CZ]; A+229W:[CD2,CE2,CE3,CH2,CZ2,CZ3]
2.34E-4	vdwP	A+106V:[CB,CG1]; A+188Y:[CB]
2.31E-4	vdwP	A+100L:[CB,CD1]; A+101K:[O]; A+103K:[CG]
2.18E-4	vdwP	A+095P:[CB,CG]; A+100L:[CD1,CD2,CG]; A+229W:[CH2,CZ3]
2.16E-4	RLIF	A_188Y_O_hbd_N
2.16E-4	vdwP	A+188Y:[CD2,CE1,CE2,CG,CZ]; A+227F:[CB,CD2,CG]; A+229W:[CB,CG]; A+234L:[CD1]
2.09E-4	vdwP	B_138E_OE2_hbd_N
1.94E-4	vdwP	A+100L:[CD2]; A+103K:[CE,CG]; A+179V:[CB,CG1,CG2]

Figure 1.
Ligand fragments - RLIF interaction model

Figure 2.
ADChemCast workflow

1: Amide

2: Ester

3: Amine

4: Urea

5: Ether

6: Olefin

7: Quaternary
nitrogen 8: Aromatic nitrogen
aliphatic carbon 9: Lactam nitrogen
aliphatic carbon 10: Aromatic carbon -
aromatic carbon

11: Sulphonamide

Figure 3.
RECAP bond-breaking rules

Figure 4.
RLIF-Ligand fragment interaction model

Figure 5.
vdwPatch A106V:[CB,CG1,CG2];A227F:[CD2,CE2] associated with PDB:3LAN

(a) Including NRTI

(b) Without NRTI

Figure 6.

Energy distribution of docking results of HIVRT ligands. Energy is in kcal/mol.

Figure 7.
Classification performance across all experiments

Figure 8.
Classifier performance on alternative HIVPR structures

Figure 9. Rilpivirine bound to the NNRTI site of reverse transcriptase (upper left), and the most informative examples of RLIF, RFQ and vdwPatch attributes found by ADChemCast for this site. One atom in rilpivirine is close to the site of interaction identified in all three attributes.

Table 1

Classification experiment summary: Number of ligands, number of hydrogen bond donors and average number per ligand, number of hydrogen bond acceptors and average number per ligand, number of pi-stacking interactions and average number per ligand, number of RLIF generated, number of vdW Patches and average number per ligand, number of ligands using only the energy attribute, number of ligands that are ACTIVE, number of DECOYS and number of total attributes used in classification.

Mysinger12	Description	Ligands	HBA	HBA-per	HBD	HBD-per	PPI	PPI-per	RLIF	vdwPatch	VdwP-per	E-only	Active	Decoy	Attribute
ACES	Acetylcholinesterase	27020	51	0.95	57	0.74	4	0.62	440	328	21.82	114	656	26356	1108
ADA	Adenosine deaminase	5733	31	0.70	38	1.30	2	0.28	172	98	14.29	24	262	5471	421
AMPC	Beta-lactamase	2961	65	3.83	41	0.72	2	0.27	157	49	7.91	1	62	2899	316
CDK2	Cyclin-dependent kinase 2	29102	45	1.13	40	1.33	1	0.01	392	306	22.64	30	798	28304	1273
EGFR	Epidermal growth factor receptor erbB1	36242	51	0.97	47	0.54	1	0.00	460	361	25.68	50	832	35410	1242
HIVPR	HIV-1 protease	37672	49	0.95	43	0.55	0	0.00	415	323	26.01	103	534	37136	1268
HIVRT	HIV-1 reverse transcriptase	19586	39	0.43	44	0.32	4	0.32	295	208	17.47	75	453	19133	469
HMDH	HMG-CoA reductase	9171	66	2.14	56	0.52	0	0.00	269	147	19.20	33	299	8872	674
PGH2	Cyclooxygenase-2	23924	60	1.07	48	0.16	3	0.00	366	255	18.93	152	524	23393	626
PYRD	Dihydroorotate dehydrogenase	6775	30	0.51	33	0.11	1	0.14	173	109	14.52	116	132	6641	236
ROCK1	Rho-associated protein kinase 1	6577	34	0.57	46	0.94	1	0.00	179	98	15.08	102	201	6374	316

Table 2

Comparison of AUROC across alternative DUD-E classifiers

Expt	ADCC_logr	Mysinger12	RC14_rdock	RC14_Glide	RC14_Vina	Lesnik15 (max)
aces	0.851	0.810	0.600			0.64 (2d-S1)
ada	0.912	0.760	0.660	0.650	0.390	0.91 (2d-S1)
anpc	0.721	0.790	0.590	0.400	0.480	0.87 (3d-S2)
cdk2	0.866	0.790	0.800	0.780	0.670	0.68 (3d-S1)
egfr	0.862	0.840	0.800	0.840	0.630	0.9 (2d-S1)
hivpr	0.794	0.600	0.770	0.800	0.770	0.81 (3d-S2)
hivrt	0.777	0.640	0.710	0.840	0.690	0.7 (3d-S2)
hndh	0.950	0.740				0.95 (2d-S1)
pgh2	0.841	0.620				0.78 (3d-S1)
pyrd	0.906	0.830				0.84 (3d-S1)
rock1	0.855	0.740				0.65 (3d-S1)

Table 3

Relative performance for attribute subsets of HIVRT

AttributeSet	N	AUROC
Energy only	1	0.522
RLIF only	123	0.635
RLIF+fragment only	168	0.604
vdwPatches only	208	0.707
All	500	0.777

Table 4

Contrasting RLIF across alternative receptor structures for HIVPR

RLIF	x1XL2 Freq	x3KF0 Freq
A_008R_NH2_hba_O	4638	2188
A_025D_OD2_hba_N	0	3019
A_025D_OD2_hba_O	0	13244
A_025D_OD2_hba_S	0	1151
A_027G_O_hbd_N	1153	2854
A_048G_O_hbd_N	882	2628
A_050I_N_hba_O	23	3823
B_008R_NH2_hba_O	1778	139
B_008R_NH2_hba_S	166	1445
B_025D_OD1_hbd_N	227	2611
B_025D_OD2_hbd_N	1936	51
B_027G_O_hbd_N	2244	4351
B_030D_N_hba_O	3305	2089
B_048G_O_hbd_N	4324	3044