JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

OOMMPPAA: A Tool To Aid Directed Synthesis by the Combined Analysis of Activity and Structural Data

Anthony R. Bradley,[‡,†] Ian D. Wall,[∥] Darren V. S. Green,[∥] Charlotte M. Deane,[†] and Brian D. Marsden*[,‡,¥]

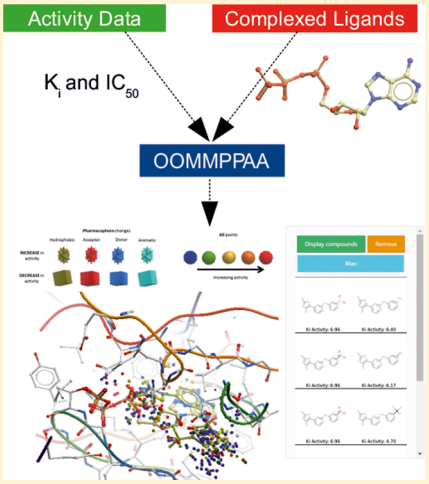
[‡]SGC, Nuffield Department of Medicine, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Headington, Oxford OX3 7DQ, U.K.

[†]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 TG, U.K.

[∥]Computational & Structural Chemistry, GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K.

[¥]Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7FY, U.K.

ABSTRACT: There is an ever increasing resource in terms of both structural information and activity data for many protein targets. In this paper we describe OOMMPPAA, a novel computational tool designed to inform compound design by combining such data. OOMMPPAA uses 3D matched molecular pairs to generate 3D ligand conformations. It then identifies pharmacophoric transformations between pairs of compounds and associates them with their relevant activity changes. OOMMPPAA presents this data in an interactive application providing the user with a visual summary of important interaction regions in the context of the binding site. We present validation of the tool using openly available data for CDK2 and a GlaxoSmithKline data set for a SAM-dependent methyl-transferase. We demonstrate OOMMPPAA's application in optimizing both potency and cell permeability and use OOMMPPAA to highlight nuanced and cross-series SAR. OOMMPPAA is freely available to download at http://oommppaa.sgc.ox.ac.uk/OOMMPPAA/.



INTRODUCTION

In recent years approaches such as high-throughput crystallography and Fragment Based Drug Design have reached maturity, resulting in a rapidly increasing number of available crystal structures and particularly more liganded structures for a given protein. In the pharmaceutical industry it is now common to have access to many tens of liganded crystal structures within a drug discovery program. At the same time, improvements in small-molecule screening throughput and initiatives to consolidate activity data from disparate sources have made thousands of high-quality small-molecule activity data points available for many biologically important protein targets both in the public domain[1] and within the pharmaceutical industry. This data is a key resource in the early stages of drug discovery as it provides information that may aid in the design of small-molecules as part of lead-discovery and lead-optimization.[2,3] However, despite the availability of this wealth of new data, there are few computational tools that are able to systematically exploit it. There is a clear need for novel automated methods that can use these data sets to assist medicinal and computational chemists in the directed synthesis of small-molecules.

Probably the best known method for relating trends in biological data to 3D structure is 3D Quantitative Structure Activity Relationships (3D QSAR).[2,4] 3D QSAR attempts to build statistical models that relate small-molecule bioactivity data with 3D compound properties. There are however well-documented problems with 3D QSAR. The generation of relevant bioactive conformations and dealing with varied binding modes poses a significant problem.[5] Simple models using, for example, linear regression are unable to detect complex and nuanced features of the data.[5] More elaborate models, which include many descriptors or complicated statistical methods, often require large data sets, can be prone to overfitting,[6,7] and can be hard to interpret. Attempts, however, have been made to improve the interpretability of 3D QSAR by developing simplified models, for example by using pharmacophore based abstractions.[8,9]

One way to move beyond the generalizations of typical regression-based QSAR models, to probe the underlying data in more detail, is to carry out the analysis in a pairwise fashion.

Received: April 23, 2014
Published: September 22, 2014

This strategy is adopted in 2D matched molecular pair analysis (2D MMPA).[10,11] A matched molecular pair (MMP) consists of two compounds that are identical apart from one small structural alteration, known as a transformation (the shared part of the molecules is commonly referred to as the context). The impact of a given transformation upon a particular property can be assessed from this. Beneficial transformations can then be applied to a compound series of interest with the aim of improving that property.

Recent work by Geppert and Beck[12] has extended this concept by using pharmacophore-based clustering in "Fuzzy" 2D MMP analysis. This approach enables the clustering of multiple chemically different but pharmacophorically conserved transformations or contexts. This improves the number of observations in each subset. Pharmacophore retyping aims to transfer SAR from chemically different but pharmacophorically identical contexts. However, pharmacophore-based matching overlooks compounds with shared binding modes but slightly different pharmacophores. Further it is possible for compounds with identical pharmacophores to possess different binding modes.

The MMP concept can also be extended to carry out the analysis on biologically relevant 3D conformations. This approach is known as 3D MMPA. A key benefit of 3D MMPA is that SAR can be transferred between structurally and pharmacophorically dissimilar series but with analogous binding modes. Furthermore, it includes information about the local protein environment. This is important since this environment directly affects the nature and magnitude of a given transformation's effect. 3D MMPA also presents a number of key advantages over 3D QSAR. First it offers a simple and easily implemented method for 3D conformation generation.[13] Most importantly it produces models that are directly related to simple pairwise compound transformations. 3D MMPA can thus be used to investigate confounding factors in analysis and show nuanced information in a way that even the most interpretable 3D QSAR models cannot.

Two recent works[13,14] have developed the 3D MMPA field. Both incorporate experimental data on compounds tested in activity assays, but for which no complexed structure is available. They operate on the assumption that similar ligands will have similar binding modes. This is a central tenet of cheminformatics and has recently been shown to be useful in solving complexed ligand crystal structures[15] and docking.[16] The first method, VAMMPIRE,[14] combines data from PDBBind[17] and ChEMBL[1] by aligning MMPs using their shared core. It is presented as a Web application and takes as input (a) a substituent or (b) a transformation or (c) a protein environment. It then displays the effect of this transformation, in the appropriate environment across all available data. VAMMPIRE aggregates data from multiple targets giving the user an overview of commonly tolerated modifications. The method does not, however, look at target-specific trends.

Posy et al.[13] take an alternative approach from VAMMPIRE. They use a target(s)-specific method primarily focusing on the protein kinase p38α. MMPs are used to identify preferred functional groups for a given region of the binding pocket. For example they demonstrate that the addition of cyclopropyl amides in a particular region improves activity in the majority of transformations. They then demonstrate that the addition of this moiety improves the potency of a different compound series. However, considering functional groups (e.g., cyclopropyl amides) limits the applicability of this method to existing chemistry that has been *frequently* explored before. Furthermore, the investigation of average trends across groups and without reference to individual transformations may, as we will show later, miss nuanced effects.

In this work, we present the OOMMPPAA method. OOMMPPAA is a 3D MMPA tool for analyzing target specific features at the pharmacophore level. OOMPPPAA diverges from the 3D MMPPA method of Weber et al. by considering each target's data separately. OOMMPPAA is designed to investigate a particular target's data not a protein environment. In this way target-specific trends in data can be observed. Further, OOMMPPAA differs from the target based 3D MMPA method of Posy et al. in two key ways. First, OOMMPPAA bases its analysis on pharmacophoric changes, not chemical changes, between small-molecules. As demonstrated by Geppert and Beck[12] but for 2D MMPA, the consideration of pharmacophoric features allows for an increase in the number of equivalent substitutions. Second, it considers the differences between molecules not just the destination fragment.

Finally, OOMMPPAA does not aggregate activity changes into general trends but, rather, considers positive and negative activity changes separately. This is a key distinction from many other 3D QSAR and 3D MMP methods and allows for nuanced descriptions of complex data and deeper analysis into confounding factors in possible trends. This makes OOMMPPAA a useful complementary tool for exploration of large and complex structural and activity data. OOMMPPAA is, to our knowledge, the first freely available 3D MMP tool providing an intuitive user interface to interact with the available data and tools to apply the method to a user's own data sets. Here we demonstrate the OOMMPPAA methodology and the tool's application, using examples from publically available cyclin-dependent kinase 2 (CDK2) data and data for a SAM-dependent methyl-transferase from GlaxoSmithKline (GSK).

METHODS AND DATA SETS

Data is input via a user interface or command-line in MacOSX, Windows, or Linux; it is processed by OOMMPPAA and presented as a visual application.

Input of Data. The input required for OOMMPPAA consists of the 3D coordinates of ligands from aligned cocomplexed structures and bioactivity data for all small-molecules of interest for each protein target. Ligand structural coordinates are input as an SD file containing the 3D structures of all cocomplexed ligands. The input ligand structures must be aligned into the same coordinate frame by aligning their complexed proteins. They should also be set to a physiologically relevant protonation state (pH 7.4). Activity data is accepted as a comma-separated variable (CSV) file containing at least two columns; chemical structures as SMILES[18] and activity data as the negative base 10 logarithmic value. If a compound has more than one activity value associated with it, the highest value (most potent) will be chosen for later analysis. Compound SMILES should be entered in the tautomeric form appropriate for the complexed ligands. Currently, the MMP implementation in OOMMPPAA carries out direct string comparisons on SMILES. Since SMILES do not canonicalize tautomeric forms, it is important that consistent tautomeric forms are used; otherwise the tool is likely to miss the affected pairs.

Inactive small-molecules are an important component of the data sets used. Inactive compounds are annotated in the CSV file by adding an "Operator" column. A "<" sign in this column

indicates an inactive compound; the activity value is the reported assay tolerance. An "=" sign indicates a compound for which the activity is known. If the "<" sign is used, the corresponding small molecule can only be less active than other small molecules. For example a small-molecule with activity <5.3 will be assigned 1.1 log unit less active than a small-molecule with activity = 6.4. However, it will not be seen as more active than a small-molecule with activity = 5.0. Combining IC_{50} and K_i data from ChEMBL is also important to maximize the power of this data source. Kalliokoski et al.[19] recently demonstrated that combining these data sources from ChEMBL did not reduce data quality. We do not apply the scaling factor suggested in their work (0.35 log units) but do highlight the activity source in the interactive viewer. This allows the user to exercise their own discretion when using the tool. Since the output is easily scrutinized, disparate data sources can be used, but only the most appropriate comparisons considered in analysis - a clear advantage of the 3D MMP method over conventional regression-based 3D QSAR.

Data Model. OOMMPPAA uses a bespoke Python[20] Django[21] data model to store all data in an SQLite database. The data model does not allow the entry of duplicate data meaning that input files can be appended and re-entered to OOMMPPAA, and only new data will be added. This model enables rapid local access, curation, and connection of large quantities of information.

OOMMPPAA METHOD

The OOMMPPA method is implemented in four steps, as shown in Figure 1. First, the matched molecular pair database is formed using the method of Hussain and Rea.[22] Second all relevant matched molecular pairs are found where one compound of each pair is represented in a crystal structure and the other is not. These pairs are used to predict coordinates of compounds for which no crystal-structure is available. Third, each pair is searched for pharmacophore differences between compounds. Finally the differences found in this last step are displayed and can be queried via an intuitive Web-based and desktop application.

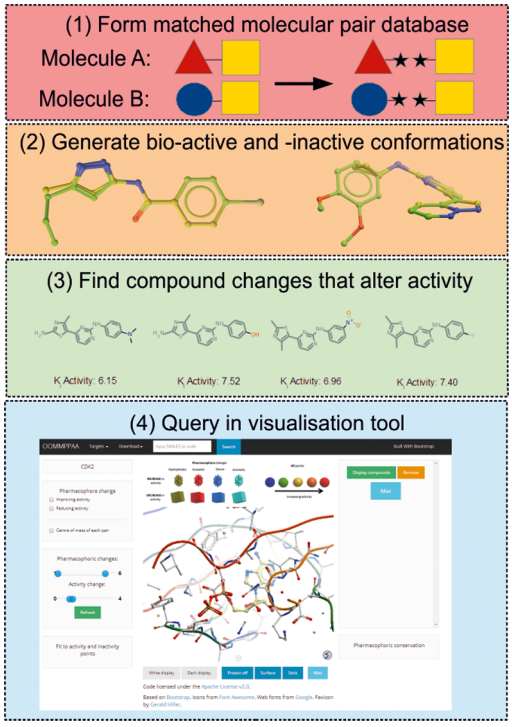


Figure 1. Four stages of the OOMMPPAA method.

Stage 1: Matched Molecular Pair Database. To find matched molecular pairs we use the method developed by Hussain and Rea.[22] All acyclic single bond breaks are exhaustively enumerated, and the pairs formed are stored as canonicalized SMILES. Only single breaks are considered since double and triple cuts would complicate conformation generation. To reduce the total number of pairs, the nonmatching part of the pair does not consist of more than 10 heavy (non-hydrogen) atoms. OpenBabel[23] is employed to assign physiologically relevant protonation states to the compounds (pH 7.4). Matched molecular pairs are found by querying this data model.

Stage 2: Creating Bioactive and -Inactive Coordinates. Figure 2 shows the OOMMPPAA method for generating coordinates for compounds with no crystal structure from each matched molecular pair found in stage 1. The shared core between these two compounds is used as a rigid template to derive 3D coordinates of the compound for which no structural information is available.

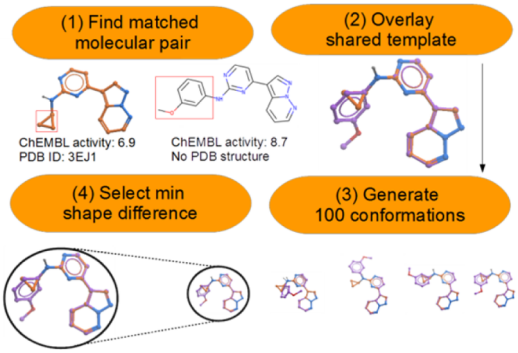


Figure 2. OOMMPPAA's method for generating bioactive and -inactive coordinates. Matched molecular pairs are identified and then superimposed using the shared substructure and the coordinates of the compound with a crystal structure. Conformations of the query molecule are then generated. The conformation with the maximum shape overlap is selected.

For each compound 100 *different* conformations are generated and minimized using the Merck Molecular Force Field (MMFF)[24] using the shared core template as a rigid constraint. Hydrogens are added to the molecule before minimization. Conformations within 0.35 Å RMSD from an existing conformation are rejected as not *different*. This is in accordance with the protocol presented by Ebejer et al.[25] using RDKit[26] for varied and accurate conformer generation. The conformation from these local energy minima that maximizes the shape overlap between the two compounds is chosen. Shape overlap is calculated using the RDKit Tanimoto shape distance implementation.

In contrast to previous methods[13,14] we do not consider the contribution of the protein when performing energy minimization of the generated conformer. While inclusion of the protein during this minimization process might produce more accurate compound conformations, it would reduce both the ability to make direct comparisons and the interpretability of the results. For example, Figure 3 shows an example where

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

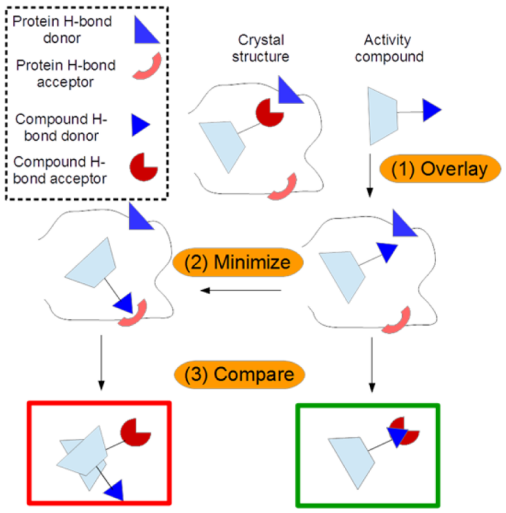


Figure 3. Energy minimization using the protein may lead to missing important protein−ligand interactions. In this instance, a disfavorable change from acceptor to donor would not be directly compared if the compound is minimized within the context of the protein.

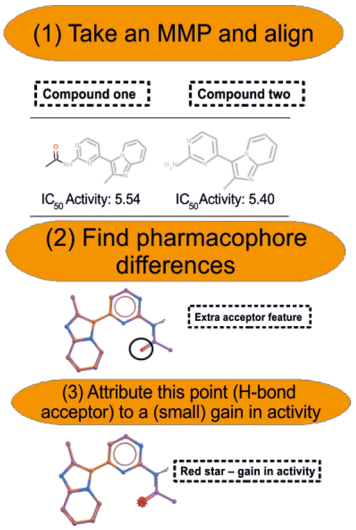


Figure 4. OOMMPPAA's method to find pharmacophoric changes between compounds. All pharmacophore points are found on each compound. Points on one compound but not within 1.5 Å of a like point on the other are found. These are associated with the activity change.

energy minimization would lead to missing an important comparison. In this example replacing a ligand H-bond acceptor for an H-bond donor would result in a clash between the protein and ligand H-bond donors and lead to a loss of activity. Minimizing the ligand to circumvent this might lead to a ring flip forming an interaction with an H-bond acceptor on the protein. This would mean the two groups (H-bond donor and acceptor) would not be directly compared in the analysis outlined in the next section, and so this feature would be missed. When minimization is not carried out, not only is the comparison made, but the unfavorable interaction responsible for the feature can be observed. In this way the OOMMPPAA method presents such extra information from inactive compounds.

It is however important to note that, for active compounds without crystal structure, the pose produced may not be representative of the native binding mode. As such, for these compounds, protein−ligand interactions presented should be treated with caution and considered representative of interaction regions, rather than specific interactions. Furthermore, important protein−ligand interactions may be missed. However, this problem can also persist for compounds minimized within the protein.

Stage 3: Finding Compound Changes That Alter Activity. The above method is able to produce thousands of matched molecular pairs using the data sets tested in this work. In order to highlight important pairs within this data we search both compounds in each pair for compound transformations that are likely to impact protein−ligand binding. In this current work pharmacophoric changes between compounds are considered. SMARTS[27]-based RDKit pharmacophore definitions (H-bond donors, H-bond acceptors, hydrophobic groups, and aromatic groups) are used. Figure 4 shows the process for finding pharmacophoric differences between two compounds. Pharmacophore points for both molecules are generated. All pharmacophore points found on one molecule in the matched molecular pair but not found within a Euclidean distance of 1.5 Å of an equivalent point on the second molecule are identified. These pharmacophore points are stored in the OOMMPPAA database along with the following associated data; the number of pharmacophore point differences between the two molecules and the activity change between the two molecules. This data is then available to be queried by the interactive OOMMPAA visualization tool. Other distances were trialed, and 1.5 Å was qualitatively determined as most appropriate by visual inspection. A greater distance led to points that could be seen to make different interactions being counted as the same. Conversely a smaller distance seemed to lead to like interactions being counted as different.

Stage 4: The Visualization Tool. OOMMPPAA presents the data via an interactive interface that can be explored to direct future compound design and find features of available activity data. A screenshot is shown in Figure 5. The interface is embedded in a Web browser but can be run locally as a standalone program, using a Python Tornado (version 3.0.1) Web server. The ActiveICM (version 1.1−7)[28] Web plugin is used for client-side molecular interactions and visualizations.

3D Visualization of Available Data. Interactive visualization of the matched molecular pair differences and their associated information in 3D is an important component of OOMMPPAA. Two types of points are shown in the viewer. First, the centers of mass of the nonmatching parts of each MMP are displayed as spheres. They are colored on a heat-scale representing the change in activity per heavy-atom difference. As discussed below, these points can summarize the distribution and nature of data available. Second, points representing a pharmacophoric difference between two compounds are shown. Each point is shown as a star if it has improved activity and as a cube if it has reduced activity. They are colored by their pharmacophoric feature.

Each pharmacophore difference point possesses information based on the matched molecular pair from which it is derived: 1) the number of pharmacophoric differences between the two compounds and 2) the log-change in activity between the compounds. We propose, in accordance with Stumpfe et al.,[29] that these two values, similarity and activity change, are related to how impactful each pharmacophoric difference is on activity. A larger change in activity is a direct indicator of a greater impact on activity. A smaller total number of pharmacophore

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

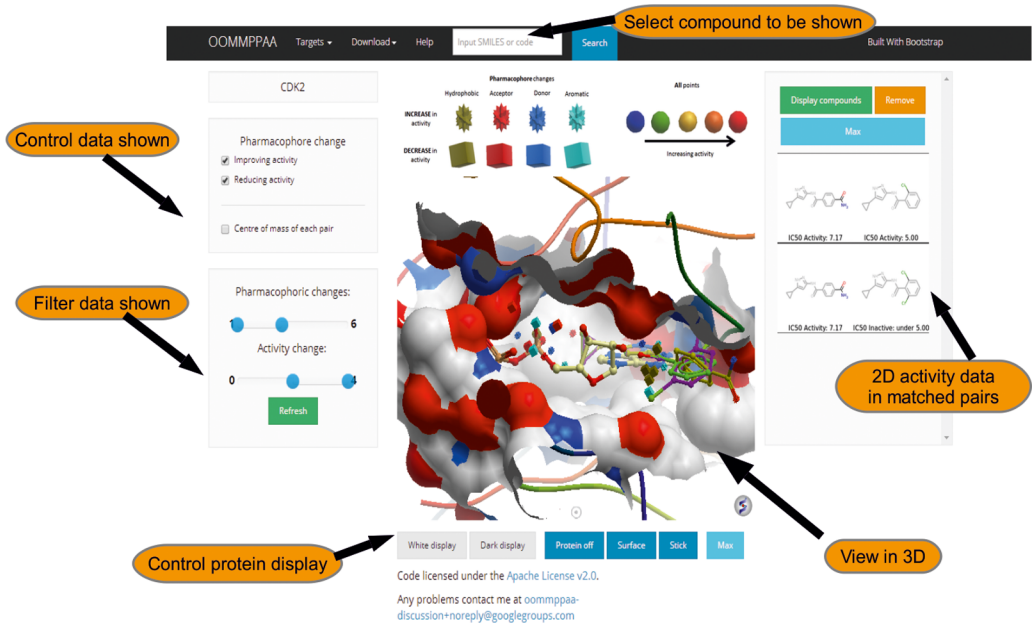


Figure 5. Screenshot of OOMMPPAA's interactive visualization tool. The top left check boxes and sliders control the points shown in the 3D display. The central display shows 3D molecular visualizations. The right-hand bar shows 2D activity data. The bottom left and right are the "Feature maps". Compounds can be queried using the search bar at the top.

point differences suggest that this particular pharmacophoric change is more likely to be responsible for the activity change.

The appropriate values to choose as cut-offs for activity and pharmacophoric differences are dependent on the assay used, the target involved, and the quantity of data considered. Indeed they will most likely differ within regions of the same binding site. The OOMMPPAA interface provides sliders to alter the threshold for each of these values and thus increase the number of points at the expense of noise or *vice versa*.

Viewing Activity Data. Each point in the 3D display is associated with a matched molecular pair and activity changes. The user can interrogate each point for this information by selecting the point and clicking the Display MMPs button. The 3D depictions of the compounds relating to these points will then appear positioned according to the coordinate generation process. The corresponding 2D compound depictions along with appropriate activity information are also shown, in their matched molecular pairs. The matching part of the pair is shown in gray.

Highlighting Key Features. To aid in directed synthesis we provide 2D feature maps for a query compound. These maps highlight 1) compound features that have been shown to reduce activity, 2) regions of the compound from which synthesis might be attempted to improve potency, and 3) compound features that are pharmacophorically conserved across the complexed ligands. The user may enter a compound's SMILES or name, as provided in the input SD file. OOMMPPAA then searches for this ligand. If a SMILES is entered and no exact match is found, the most similar ligand based on Morgan fingerprint[30] similarity (radius 2) is found. The ligand is shown in the 3D display along with two feature maps, as shown in Figure 6. These are colored based on per-atom scores which can be depicted using the RDKit "Similarity Map" compound visualizations recently developed by Riniker and Landrum.[31] The Activity Change map (Figure 6a) provides a qualitative visualization of locations around the query compound that confer changes in activity within the data

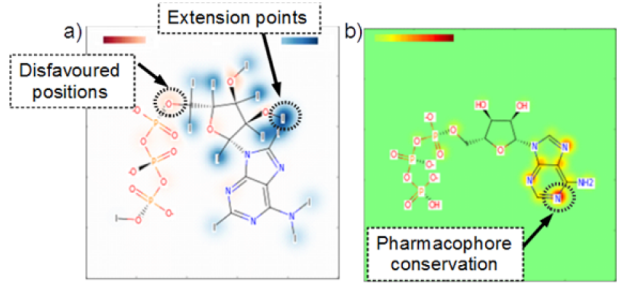


Figure 6. Feature maps for 1QMZ (CDK2 cocomplexed with ATP). a) The Activity Change map. Red indicates the functional groups that might be leading to a drop in activity. Blue indicates the potential to replace a hydrogen with activity improving functionality. b) The Pharmacophore Conservation map shows how pharmacophorically conserved features in a compound are, across all the provided crystallographic data, e.g a highly conserved H-bond acceptor.

provided to OOMMPPAA. All pharmacophore differences conferring an activity difference greater than 0.5 log units and a number of pharmacophore differences less than 4 are used. Extension points to the compound that could be made to increase the compound's potency are highlighted in blue. These are currently defined as hydrogen atoms. A darker blue indicates a hydrogen is near to more pharmacophoric changes that have led to an increase in activity. Atoms in the compound that might be reducing the compound's potency are highlighted in red. A darker red indicates an atom is near to more pharmacophoric changes of the atom's type(s) that have led to any loss in activity.

The Pharmacophore Conservation map (Figure 6b) presents the conservation of pharmacophoric features for the whole molecule against all the cocomplexed ligands for the target. Increasingly pharmacophorically conserved regions of the compound are depicted on a scale from yellow to red. This map indicates pharmacophoric features that are highly

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

conserved across the compounds cocrystallized with this target and thus are putatively important for activity.

Data Sets. OOMMPPAA has been tested and developed using multiple internal GlaxoSmithKline (GSK) and SGC data sets. Here, we demonstrate the application of this tool using two data sets. First, publically available cyclin-dependent kinase 2 (CDK2) data from ChEMBL version 16 consisting of *in vitro* IC_{50} and K_i data was used. We extracted all nonredundant ligands cocomplexed to CDK2 with a resolution less than 3.3 Å from the PDB[32] to use as structural data. Electron density maps for all structures with resolution greater than 3.0 Å were visually inspected to ensure the ligand was correctly modeled. We aligned the ligands into the same coordinate frame via the protein structures using Pymol[33] and visually inspected the alignment. In total this data set consisted of 1,632 unique compounds with activity data and 261 unique cocomplexed ligands. The second data set consisted of internal GSK data for an *S*-adenosyl methionine (SAM)-dependent methyl transferase where all crystallographic data is under 2.5 Å and all activity data was generated using the same IC_{50} bioactivity assay. In total this data set consisted of 2,212 unique compounds with activity data and 92 unique complexed ligands. We suggest this order of data quantity is necessary for undertaking the following depth of analysis. However, we have tested the tool using a smaller data set (15 complexed ligands and ∼1000 activity points from Tm shift). The analysis produced was not as rich; however, useful insights were generated.

RESULTS AND DISCUSSION

OOMMPPAA can be used to demonstrate key features in available activity information and use that information to generate evidence-based hypotheses for compound development. In the following sections we demonstrate its use in both of these areas. The examples used are derived from CDK2 and a SAM-dependent methyl-transferase.

Overview of Available CDK2 Bioactivity Data. In Figure 7 an OOMMPPAA visualization shows ATP surrounded by points representing each matched molecular pair for CDK2. Three important features can be seen here. First, there are few points surrounding the adenine core (blue box) indicating that OOMMPPAA finds no matched molecular pairs with transformations in these regions. This might suggest that the compounds available in ChEMBL present highly conserved substructures in this conserved region of kinase binding. However, interrogation of the data shows a number of hinge binding groups are represented, but there are very few MMPs where that group varies and the rest of the molecule is conserved. Second, in the red box there are two lines of points. Each line indicates an SAR series from the same scaffold. These points are predominantly in red indicating that changes in this region have a large effect on activity. Finally, in the green box are a cluster of points largely in blue and purple. These indicate a collection of changes which have generally altered activity only weakly. This cluster sits toward the outside of the protein, and thus groups here may be expected to interact only weakly.

These three examples demonstrate OOMMPPAA's ability to present large quantities of data in a manner which allows the user to identify the nature of and trends in available activity data. Systematically assessing the nature and scale of pairwise compound alterations in different regions of the binding pocket in this way is a novel and useful feature of 3D MMPA.

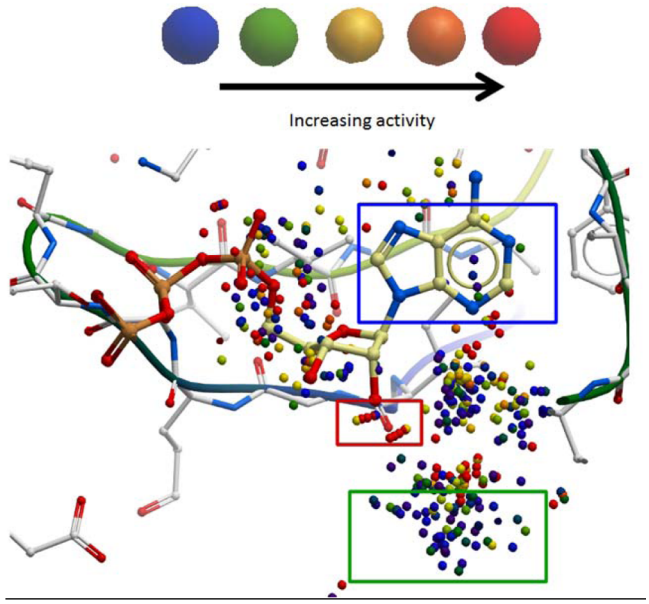


Figure 7. A visualization from OOMMPPAA showing the distribution of all available matched molecular pairs for CDK2. Blue box: the adenine core is surprisingly sparse in data. Red box: lines of spheres indicate compound series. Green box: an area where compounds have not changed activity greatly.

Lead Development Using CDK2. OOMMPPAA's interactive visualizations can be used to suggest changes that may improve the potency of a given compound series. In Figure 8 we demonstrate, retrospectively, how OOMMPPAA could have been used to suggest modifications to the original lead (PDB code 1H1P) from a compound series developed by Davies et al.[34,35] The Activity Change map highlights that changes improving activity could be made from the terminal amine, circled in Figure 8a. Investigating these points in the 3D viewer indicates they are aromatic pharmacophores, and thus adding an aromatic group in this area has previously improved activity. The relevant SAR for these transformations is shown in 2D and is from two different series, as shown in Figure 8b. Davies et al. extended the lead compound by adding an aryl ring in this area leading to an increase in *in vitro* potency from 13 μM to 2.3 μM. This extended ligand was cocrystallized by them and is shown in Figure 8c; the additional aryl group is placed over the cluster of cyan stars, as expected. Further additions surround the aryl group, suggesting future synthetic work that could be performed. The compound comparisons from Davies et al. were not included in this analysis.

In this example the advantage of 3D over 2D MMPA is demonstrated. First, 3D MMPA allows for SAR from one series, binding in the same region of the binding pocket, to be transferred to another pharmacophorically and structural different series. Second, unlike 3D QSAR, OOMMPPAA's model is directly related to pairwise transformations. This allows confounding factors to be scrutinized. For example it can be directly observed whether the aromatic group is simply favored as a linker group for another substituent or whether it is broadly favored irrespective of substitution. Finally it is important to note that OOMMPPAA's focus on pharmacophoric changes, unlike existing 3D MMP methods, means that the different patterns of aryl group substitution in this data set were accurately deconvoluted, spatially.

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

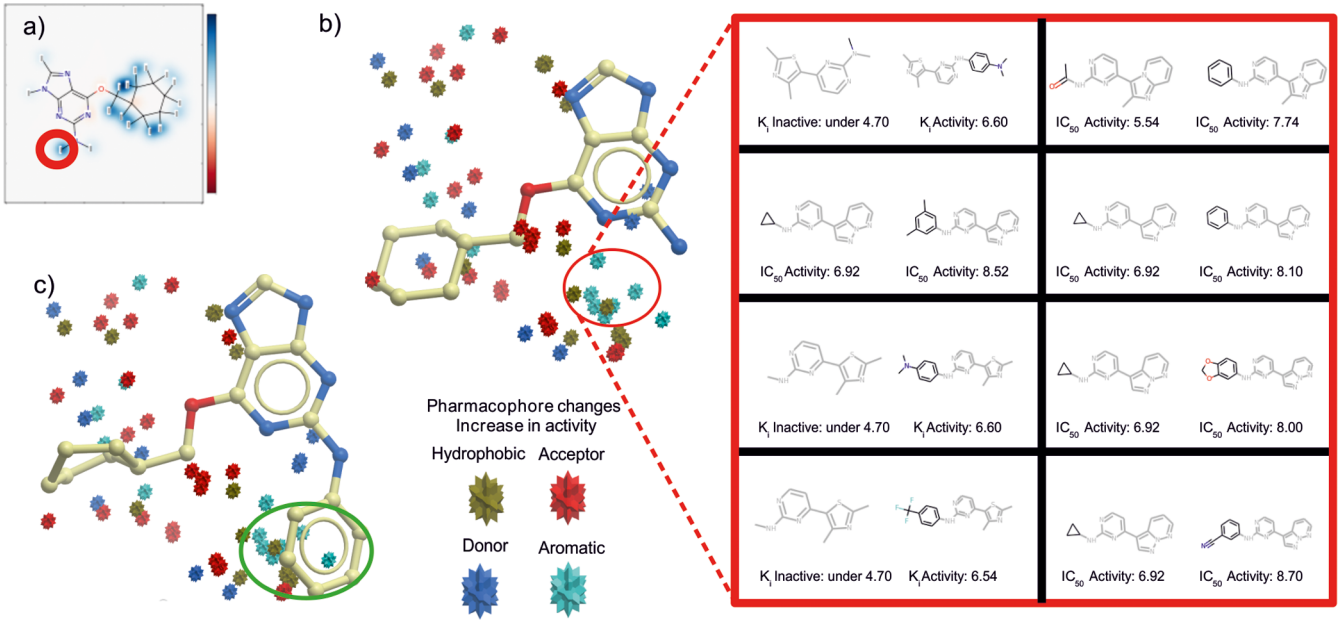


Figure 8. OOMMPPAA used in the optimization of a compound series by Davies et al. a) and b) OOMMPPAA indicated that activity could be improved by addition of an aromatic group in the area circled. b) The associated SAR (red box) indicates a variety of substituted aryl groups improve activity. Experimental data, from Davies et al., shows that addition of an aryl group improves activity (13 μM to 2.3 μM) and c) the aryl group binds in the position predicted.

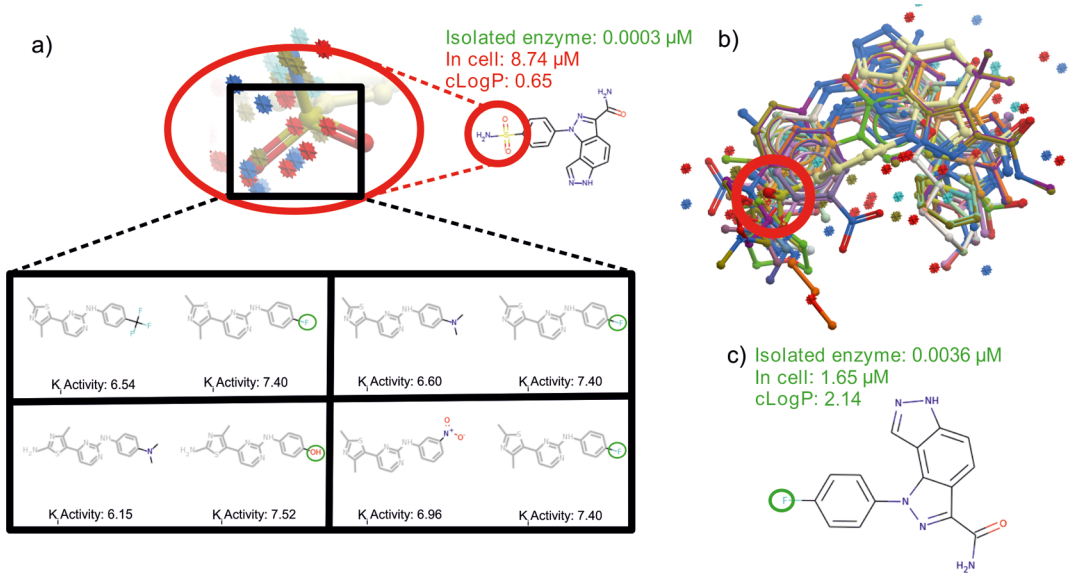


Figure 9. OOMMPPAA can be used to aid in the optimization of a compound with poor cell based activity. a) OOMMPPAA indicates features that could have favorable interactions in the region of the sulfonamide. b) The 3D display demonstrates the scaffolds (rainbow) have similar binding modes to the query compound (cream and thicker). The sulfonamide region is circled in red. c) The fluoride derivative was synthesized by D'Alessio et al. and improved cell based activity, while maintaining a respectable activity against the isolated enzyme. cLogP was calculated using the RDKit implementation of the Crippen[36] method.

Improving Cellular Activity against CDK2. OOMMPPAA can be used to assist in the optimization of properties other than potency. Figure 9 demonstrates the use of OOMMPPAA to suggest ways to improve the cell permeability of a compound series generated by D'Alessio et al.[37] against CDK2. Compound 1 from this work possesses high potency (0.3 nM) against the isolated enzyme but low potency in cells (8.74 μM) probably due to its low computed log partition coefficient (cLogP) (0.09), as shown in Figure 9a. The sulfonamide group contributes strongly to this low cLogP, and so altering this group might improve the cellular activity for this compound series. However, structural data indicates that the sulfonamide forms favorable interactions with the protein.

The pharmacophoric changes that have been explored in this region of the protein pocket that have led to a greater than 0.5 log unit increase in activity with less than four pharmacophoric changes between compounds are shown in Figure 9a. These points indicate features that have improved activity in this region of the binding site. Figure 9b shows the 3D display and demonstrates these scaffolds have similar predicted binding

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

modes to the series being developed. First the substitutions in this region indicate, from two other series, that the sulfonamide is indeed favorable as potency against the isolated enzyme is lost when replaced with other groups (not shown). However, OOMMPPAA also indicates alternative functional groups, fluoride and hydroxyl, that improve activity in this region. Further these groups might also improve cell based potency due to their increased cLogP. D'Alessio et al.[37] demonstrated that replacing the sulfonamide with a fluoride did indeed lead to a 6−7-fold increase in cellular potency while maintaining a reasonable activity (36 nM) against the isolated enzyme. The hydroxyl analogue was not synthesized; however, we postulate it would lead to a similar increase in activity.

In this example OOMMPPAA's methodology presents significantly different insights to 2D MMPA or 3D QSAR. OOMMPPAA's analysis allows the user to explore SAR from different series, binding in the same region of the binding pocket to develop hypotheses for compound alterations. 3D QSAR does not allow exploration of the data in this direct manner. Equally 2D MMPA does not allow facile transfer of SAR between structurally and pharmacophorically different series based on shared binding location. Finally by considering pharmacophoric differences, unlike existing 3D MMP methods, OOMMPPAA is able to highlight pertinent functional changes that have led to improved potency.

Highlighting Nuanced SAR with a SAM-Dependent Methyl-Transferase. OOMMPPAA can be used to identify nuanced SAR. In Figure 10a the Activity Change and Pharmacophore Conservation maps for a SAM-dependent methyl-transferase are shown. They indicate contrasting features, within the context of the data provided. In the left-hand map of Figure 10a the N3 atom within the adenine ring (circled) is shown to be at a position at which an H-bond acceptor results in a reduced activity. Yet in the right-hand map, it is also highly conserved, pharmacophorically, and thus putatively important for protein−ligand binding. This suggests that this H-bond acceptor is beneficial in some series but not in others. Investigating this further using OOMMPPAA showed that the compounds responsible for the loss of activity feature possess a terminal amide group forming H-bonding interactions with the protein.

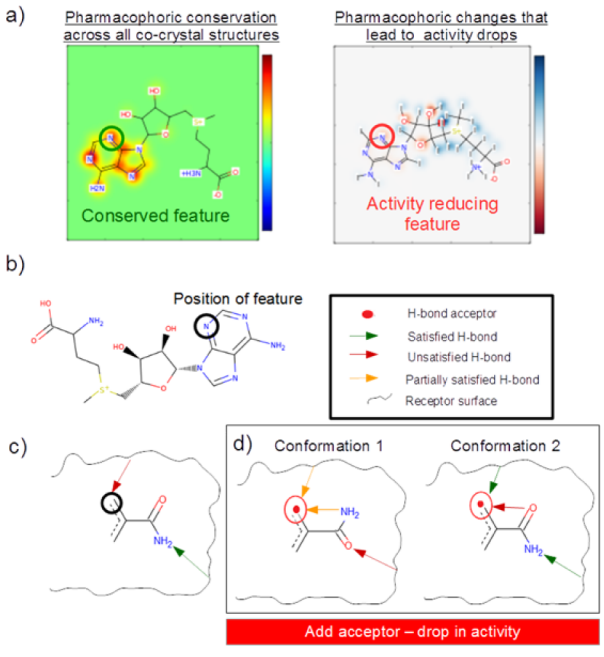


Figure 10. a) Feature maps for SAM bound to the SAM-dependent methyl-transferase, left with the conserved H-bond acceptor feature highlighted. Right, the same feature has also been responsible for falling activity. b) Rationalization of this change. In the presence of a terminal amide. c) Addition of an H-bond acceptor in this position (black circle) leads to an intramolecular H-bond. d) This potentially reduces activity by causing a conflict between conformations.

In Figure 10c we show that introduction of an H-bond acceptor in the highlighted position (in Figure 10b) for a compound series containing a terminal amide would cause an intramolecular H-bond to form. This would alter the energetically preferred conformation of the terminal amide to one resulting in a mismatch of H-bonding features with the protein (as shown in Figure 10d). We postulate that this is responsible for the net-loss in binding affinity for this compound series.

The information shown by OOMMPPAA would therefore dissuade future lead development from using this particular combination of amide and H-bond acceptor. Since the overall trend in this region is for H-bond acceptor groups to improve activity (there are more activity improving "stars" than "cubes" in this region), this would have been obscured if the effects had been aggregated into a trend, as is carried out in existing 3D MMP and 3D QSAR methodologies.

Highlighting Potential Dead-Ends with SAM-Dependent Methyl-Transferase. OOMMPPAA can also identify areas of compound design that have consistently led to a reduction in activity. Figure 11 shows SAM in its bound

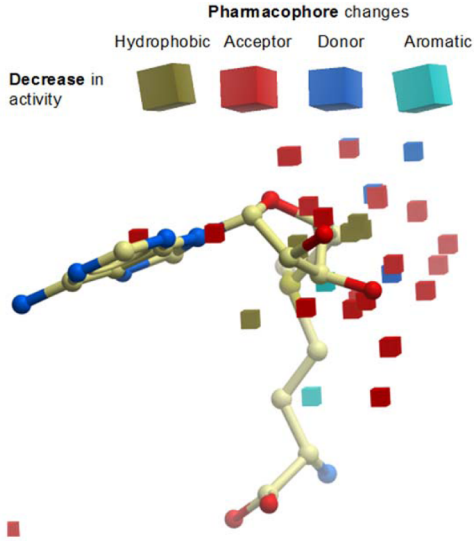


Figure 11. Pharmacophore changes associated with at least 1.5 log unit losses in activity with the native cofactor (SAM) shown for context.

conformation to the SAM-dependent methyl-transferase, surrounded by all points that are associated with a loss of activity, colored by their pharmacophore type. A cluster of acceptors is visible around the ribose ring of the SAM molecule. Investigation of the underlying matched molecular pairs of this cluster showed that adding a carbonyl group to compounds in this region consistently saw greater than 0.5 log unit activity losses, and hence future analogous alterations may have a similar consequence.

dx.doi.org/10.1021/ci500245d | J. Chem. Inf. Model. 2014, 54, 2636−2646

The carbonyl groups were added with the aim of forming H-bonding interactions with the protein. By looking at one compound series, or multiple compound series in 2D, this feature may not have been apparent. However, OOMMPPAA produces flexible 3D visualizations of composite activity data within the context of the protein environment, generating the above hypothesis and informing prospective compound design. 3D QSAR methodologies may have indicated this feature of acceptors reducing activity. However, OOMMPPAA provides this analysis transparently, based on simple pairwise comparisons allowing simple analysis of potentially confounding factors and nuanced SAR, as discussed in the previous example.

COMPARISON TO EXISTING METHODS

In each of the above examples OOMMPPAA highlights interesting features in existing SAR that could be used to inform future compound design.

In these examples the advantage of 3D over 2D MMPA is shown. 3D MMPA allows SAR information from one series to be transferred to another pharmacophorically and structurally different series, binding in the same region of the binding pocket. Furthermore, OOMMPPAA differentiates itself from existing QSAR methods. The models produced are directly related to simple pairwise SAR. As a consequence hypotheses inferred from OOMMPPAA can be directly scrutinized. This transparency enables rationalization of otherwise potentially confounding factors. A key example of this is the influence of the amide group combined with an H-bond acceptor in the examples above. Finally OOMMPPAA develops upon existing target specific 3D MMP methods by clustering changes into pharmacophoric groupings and by not aggregating changes into trends. The use of pharmacophores is crucial in the hypotheses generated in each of the above examples, thus demonstrating the value of this approach. The pharmacophore approach also provides a broader range of suggested modifications for future compounds than a more specific functional group analysis. Further, by not aggregating into trends, nuanced features were observed.

FUTURE DEVELOPMENTS

By design, OOMMPPAA currently takes a simple view of protein−ligand binding, considering only four pharmacophore features. This approach enables OOMMPPAA to present simple, user-friendly visualizations. However, the method can be readily extended to consider any group that can be expressed by a SMARTS pattern. It would be possible to assess OOMMPPAA's ability to use other groups (e.g., methyl groups, halogens and basic/acidic groups) to aid in lead optimization.

Further to this, OOMMPPAA is currently a ligand-centric tool. It implicitly incorporates protein information by aligning complexed ligands. Future development will consider potential protein−ligand interactions. This will allow OOMMPPAA to assess whether an overlaid compound, which has not been cocrystallized, might be sterically hindered by the protein itself. If such a clash consistently leads to an activity decrease, this would indicate an inflexible region of the protein. Conversely a region where this is not the case might indicate the protein is conformationally flexible in this region.

We propose OOMMPPAA could be extended to be used for experimental design. A number of current methodologies propose novel ligands, focusing on increasing compound potency.[38,39] However, OOMMPPAA is perhaps more powerful if used to improve knowledge regarding protein−ligand binding. First OOMMPPAA could be used to propose experiments that would enhance the information density of the model. For example it might highlight whole SAR series that currently have no X-ray structure available and thus suggest which compounds should be prioritized for cocrystallization. Second it might highlight available, untested, compounds to probe underexplored regions of the protein binding site. In the example of CDK2 little information was available around the adenine core. OOMMPPAA might find compounds from a large (for example, corporate) database that are matched pairs with complexed ligands and would produce structural variations in this region. Since the MMP method used is optimized for large data sets, this would be computationally inexpensive.

Finally OOMMPPAA could be further extended to test proposed hypotheses. As demonstrated above, OOMMPPAA is able to visualize regions of conflicting data, e.g. where addition of an acceptor both increases and decreases activity in different compound series. It might then be able to find all untested compounds that would probe this area and suggest these as future experimental candidates.

CONCLUSION

OOMMPPAA is a novel and freely available computational tool to aid in directed synthesis by analysis of large structural and activity data sets, comprising tens of liganded structures and hundreds of activity data points from K_i and IC_{50} data. OOMMPPAA uses 3D MMP analysis to infer possible binding geometries of compounds for which crystal structures are not available. Structural changes that alter the compound binding mode are not predicted in this method. This means poses produced should be used with caution. However, the method has the advantage of providing more relevant comparisons of transformations, particularly when considering those that are activity reducing.

OOMMPPAA then builds upon existing 3D matched molecular pair methods by including pharmacophore based concepts from 2D MMP and 3D QSAR analysis. Critically, OOMMPPAA considers the pharmacophoric differences between fragments not just the final molecule and does not aggregate activity changes into general trends but, rather, considers positive and negative effects separately. An intuitive and interactive interface plays an integral role within OOMMPPAA, providing an easy-to-use tool for analyzing large quantities of structural and bioactivity data. Diverse examples of OOMMPPAA's use have been given, supporting our belief that OOMMPPAA is an important new tool in any computational and medicinal chemist's arsenal.

OOMMPPAA is freely available released under the Apache Version 2.0.

A demo version of the Web application using the CDK2 data set can be found at http://oommppaa.sgc.ox.ac.uk/OOMMPPAA/.

The source code is available here: https://bitbucket.org/abradley/oommppaa/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: brian.marsden@sgc.ox.ac.uk. Corresponding author address: SGC, NDM, Old Road Campus Research Building, Roosevelt Drive, Headington, Oxford OX3 7DQ, U.K.

Author Contributions

Work was carried out by Anthony Bradley. First draft and editing was carried out by Anthony Bradley. Drafting, editing, and comments were carried out by Charlotte Deane, Darren Green, Ian Wall, and Brian Marsden.

Funding

Anthony Bradley receives funding from the EPSRC, SABS-IDC, and from GlaxoSmithKline [grant number EP/G037280/1]. The SGC is a registered charity (number 1097737) that receives funds from Abbvie, Bayer Pharma AG, Boehringer Ingelheim, the Canada Foundation for Innovation, the Canadian Institutes for Health Research, Genome Canada, GlaxoSmithKline, Janssen, Lilly Canada, the Novartis Research Foundation, the Ontario Ministry of Economic Development and Innovation, Pfizer, Takeda, and the Wellcome Trust [grant number 092809/Z/10/Z].

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Stephen Pickett and Jameed Hussain for useful discussions of the method. We would also like to thank David Damerell for extensive help in generating the distributable and for useful discussions on the project. We would like to thank everybody at the SGC, OPIG, and GlaxoSmithKline for facilitating this work.

■ REFERENCES

(1) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−7.

(2) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design−a Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95−115.

(3) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*, 425−436.

(4) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(5) Scior, T.; Medina-Franco, J. L.; Do, Q.-T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How To Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297−4313.

(6) Hawkins, D. M.; Basak, S. C.; Shi, X. QSAR with Few Compounds and Many Features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663−670.

(7) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 1−12.

(8) Manchester, J.; Czerminski, R. SAMFA: Simplifying Molecular Description for 3D-QSAR. *J. Chem. Inf. Model.* **2008**, *48*, 1167−1173.

(9) Kotani, T.; Higashiura, K. Comparative Molecular Active Site Analysis (CoMASA). 1. An Approach to Rapid Evaluation of 3D QSAR. *J. Med. Chem.* **2004**, *47*, 2732−2742.

(10) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(11) Kenny, P. W.; Sadowski, J. *Structure Modification in Chemical Databases*; 2004.

(12) Geppert, T.; Beck, B. Fuzzy Matched Pairs: A Means To Determine the Pharmacophore Impact on Molecular Interaction. *J. Chem. Inf. Model.* **2014**, *54*, 1093−1102.

(13) Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. 3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* **2013**, *53*, 1576−1588.

(14) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. AMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* **2013**, *56*, 5203−5207.

(15) Klei, H. E.; Moriarty, N. W.; Echols, N.; Terwilliger, T. C.; Baldwin, E. T.; Pokross, M.; Posy, S.; Adams, P. D. Ligand Placement Based on Prior Structures: The Guided Ligand-Replacement Method. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *70*, 134−143.

(16) Kawabata, T.; Nakamura, H. 3D Flexible Alignment Using 2D Maximum Common Substructure: Dependence of Prediction Accuracy on Target-Reference Chemical Similarity. *J. Chem. Inf. Model.* **2014**, *54*, 1850−1863.

(17) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(18) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(19) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PLoS One* **2013**, *8*, e61007.

(20) Foundation, P. S. Python Language Reference, Version 2.7.3.

(21) Django Software Foundation. Django (Version 1.5), 2013.

(22) Hussain, J.; Rea, C. Computationally Efficient Algorithm To Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(23) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(24) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(25) Ebejer, J. P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

(26) Landrum, G. *RDKit: Open-Source Cheminformatics.*

(27) Inc.:, SMARTS; Daylight Chemical Information Systems, Santa Fe, NM, Vol 471.

(28) Raush, E.; Totrov, M.; Marsden, B. D.; Abagyan, R. A New Method for Publishing Three-Dimensional Content. *PLoS One* **2009**, *4*, e7394.

(29) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2013**, 130827130253002.

(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(31) Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminf.* **2013**, *5*, 43.

(32) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(33) Schrödinger, LLC. *The {PyMOL} Molecular Graphics System, Version~1.3r1*; 2010.

(34) Gibson, A. E.; Arris, C. E.; Bentley, J.; Boyle, F. T.; Curtin, N. J.; Davies, T. G.; Endicott, J. A.; Golding, B. T.; Grant, S.; Griffin, R. J.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Tucker, J. A.; Whitfield, H. J. Probing the ATP Ribose-Binding Domain of Cyclin-Dependent Kinases 1 and 2 with O(6)-Substituted Guanine Derivatives. *J. Med. Chem.* **2002**, *45*, 3381−3393.

(35) Davies, T. G.; Bentley, J.; Arris, C. E.; Boyle, F. T.; Curtin, N. J.; Endicott, J. A.; Gibson, A. E.; Golding, B. T.; Griffin, R. J.; Hardcastle, I. R.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Tucker, J. A.; Wang, L.; Whitfield, H. J. Structure-

Based Design of a Potent Purine-Based Cyclin-Dependent Kinase Inhibitor. *Nat. Struct. Biol.* **2002**, *9*, 745−749.
(36) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Model.* **1999**, *39*, 868−873.
(37) D'Alessio, R.; Bargiotti, A.; Metz, S.; Brasca, M. G.; Cameron, A.; Ermoli, A.; Marsiglio, A.; Polucci, P.; Roletto, F.; Tibolla, M.; Vazquez, M. L.; Vulpetti, A.; Pevarello, P. *Benzodipyrazoles: A New Class of Potent CDK2 Inhibitors*; **2005**; Vol. *15*, pp 1315−1319.
(38) Zhou, P.; Tian, F.; Shang, Z. LigEvolutioner, a New Strategy for Modification and Optimization of Lead Compounds in Receptor/ligand Complexes. *Chem. Biol. Drug Des.* **2008**, *72*, 525−532.
(39) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.