# Evolving Interpretable Structure−Activity Relationship Models. 2. Using Multiobjective Optimization To Derive Multiple Models

Kristian Birchall,[†] Valerie J. Gillet,*[,†] Gavin Harper,[‡] and Stephen D. Pickett[‡]

Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom, and GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

Received February 13, 2008

A multiobjective evolutionary algorithm (MOEA) is described for evolving multiple structure−activity relationships (SARs). The SARs are encoded in easy-to-interpret reduced graph queries which describe features that are preferentially present in active compounds compared to inactives. The MOEA addresses a limitation associated with many machine learning methods; that is, the inherent tradeoff that exists in recall and precision which is usually handled by combining the two objectives into a single measure with a consequent loss of control. By simultaneously optimizing recall and precision, the MOEA generates a family of SARs that lie on the precision-recall (PR) curve. The user is then able to select a query with an appropriate balance in the two objectives: for example, a low recall-high precision query may be preferred when establishing the SAR, whereas a high recall-low precision query may be more appropriate in a virtual screening context. Each query on the PR curve aims at capturing the structure−activity information into a single representation, and each can be considered as an alternative (equally valid) solution. We then investigate combining individual queries into teams with the aim of capturing multiple SARs that may exist in a data set, for example, as is commonly seen in high-throughput screening data sets. Team formation is carried out iteratively as a postprocessing step following the evolution of the individual queries. The inclusion of uniqueness as a third objective within the MOEA provides an effective way of ensuring the queries are complementary in the active compounds they describe. Substantial improvements in both recall and precision are seen for some data sets. Furthermore, the resulting queries provide more detailed structure−activity information than is present in a single query.

## INTRODUCTION

The application of machine learning methods to the analysis of high-throughput screening (HTS) data has become a widely studied topic in chemoinformatics.[1] The usual aim is to derive a model of activity based on a training set of known actives and inactives with the model subsequently being used to make predictions about previously unseen compounds. Many different techniques have been applied in this context, for example, support vector machines (SVMs),[2] substructural analysis[3] and the related Naïve Bayesian Classifiers,[4] and random forests.[5] As with most applications in chemoinformatics there is no holy grail where one method is superior in all ways to all others, and each technique has its own advantages and limitations. For example, while support vector machines have been shown to be effective in prediction and robust to noise in the data, as is characteristic of HTS data, they are often referred to as "black-boxes" to convey the difficulty of interpreting the models in terms that can be easily understood by a chemist.

In a companion paper,[6] we have described a novel approach to analyzing screening data that aims to generate models of activity that are interpretable. Our method is based on a combination of reduced graphs (RGs) to provide generalized descriptions of the molecules and an evolutionary algorithm (EA) to evolve RG queries (subgraphs). The queries are searched against a training set of actives and inactives, which are also represented by RGs. The aim is to evolve a RG query that is present in as many of the actives as possible, i.e., maximizing recall, while being absent from the inactives, i.e. maximizing precision. This can be thought of conceptually as analogous to approaches for pharmacophore elucidation but is applicable to much larger data sets. However, a limitation of the approach, and other machine learning methods that attempt to classify objects, is that it is not usually possible to simultaneously improve on both recall and precision. For example, a structure−activity relationship (SAR) model can be considered as an encapsulation of descriptor space that includes as many active compounds as possible while excluding inactives. The tradeoff between recall and precision can then be explained by considering a broader model that is able to describe more chemical space and thus retrieve more actives (increasing recall). However, such a model will also tend to increase the retrieval of inactives (resulting in a decrease in precision). Conversely, a model that describes a smaller region of descriptor space may be able to exclude more inactives, albeit at the expense of retrieving fewer actives. Thus, in our previous implementation[6] we transformed this multiple objective

* Corresponding author phone: +44-1142-222652; fax: +44-1142-780300; e-mail: v.gillet@sheffield.ac.uk.
[†] University of Sheffield.
[‡] GlaxoSmithKline.

10.1021/ci800051h CCC: $40.75 
Published on Web 07/19/2008

$$F = \frac{2 \times PR}{P + R} \text{ where } R = \frac{TA}{TA + FI} \text{ and } P = \frac{TA}{TA + FA}$$

| Actual Class | Predicted Class: Active | Predicted Class: Inactive |
|---|---|---|
| Active | *TA* | *FI* |
| Inactive | *FA* | *TI* |

**Figure 1.** Confusion matrix. *TA* is the number of true actives; *TI* is the number of inactives; *FA* is the number of false actives, i.e., inactives that are predicted as active; and *FI* is the number of false inactives, i.e. actives predicted as inactives. *R* is recall and measures the proportion of the actives that are retrieved by a RG query; *P* is precision and measures the proportion of the compounds that are retrieved that are true actives; and *F* is the F-measure which is the harmonic mean of recall and precision.

problem to a single objective optimization using the F-measure[7] which can be calculated from the confusion matrix as shown in Figure 1.

While the F-measure attempts to achieve a balance in recall and precision, it is insensitive as to how the false predictions are distributed between the classes. Thus, the user does not have any control on model specificity. This was illustrated in previous work where the EA was trained to generate models for 5HT3 antagonists (5HT3 Ant) and 5HT1a agonists (5HT1A) activity classes extracted from MDL's Drug Data Reports Database (MDDR).[6] In both cases, models were generated with F-measures of around 52%; however, for the 5HT3 Ants this corresponds to recall of 41.3% and precision of 71.7%, whereas for the 5HT1As the balance is shifted toward recall at 61.6%, with precision of 46.1%. The exact balance achieved is related to the heterogeneity of compounds in the activity class as well as the descriptors used, and the user has no direct control over where the balance will lie.

A further limitation of the EA, which also applies to other machine learning methods, is that it may not be possible to describe all of the actives using a single SAR. This is likely to be the case when the actives are represented by different chemical series or include compounds that represent different binding modes. In the previous work, we attempted to deal with this issue in two ways: first, the representation of the molecules and the queries as RGs enables equivalences to exist between molecules with different chemical graphs, and, second, the use of a subset of SMARTS[8] operators within the RG queries increases their flexibility, for example, by the encoding of alternative nodes. However, while the latter approach can increase the number of true actives identified it may also lead to an increase in false actives, especially when the actives represent different binding modes, i.e. include compounds that make different interactions with the receptor.

Here we have adapted the EA to use multiobjective optimization techniques (as a MOEA) in order to explore the relationship between recall and precision. A family of tradeoff solutions is generated each of which is (near) optimal with respect to recall and precision and each of which represents a different RG query. The result is a Precision-Recall (PR) curve[9] which is related to the ROC curve familiar in virtual screening.[10] The user can then choose an appropriate model according to the intended use. For example, if the aim is to establish the SAR, then queries with high precision are likely to be most useful; however, if the aim is to select compounds for screening, then a model could be chosen with specificity that matches the capacity of the screening assay (cf. the use of a similarity threshold to control the output in a virtual screening experiment).

We then explore two approaches to combining queries in order to derive multiple SARs. One method is based on a two-objective MOEA in which recall and precision form the two objectives, and queries which occupy different parts of the PR tradeoff curve are combined into "teams". The performance of a team is assessed according to the combined performance of the individual queries that comprise the team. In a second approach we include a third objective in the MOEA called uniqueness. This has the aim of ensuring that the queries that are evolved complement one another in terms of the true actives they retrieve. These queries are then combined into teams as for the two-objective MOEA.

The basic methodology of the EA has been described fully in our companion paper[6] and is summarized here for completeness. The Methods section then focuses on the extension of the basic approach to incorporate multiobjective optimization techniques to generate solutions on a PR curve and to generate multiple complementary SARs. The methods are illustrated by application to publicly available data sets and in-house screening data provided by GSK.

## METHODOLOGY

**Evolving RG Queries.** The previously described EA evolves what we have called RG queries, which are subgraphs (actually subtrees) of RGs.[11–13] A compound in the training set is predicted active if its RG representation contains the query, otherwise it is predicted inactive. The chromosome representation in the EA is a rooted tree where each node in a tree maps to a node (or a series of alternative nodes) in the RG query. The edges in a tree indicate connections between nodes in the RG query. The tree-based chromosome is parsed into a SMARTS string in which elements from the nonorganic set are used to represent the RG query feature definitions and Daylight toolkit routines are used to conduct the searches on the training set. A subset of SMARTS features is used to enable flexibility in the RG queries (cf generic queries in traditional substructure searching), for example, disconnected nodes, alternative nodes, forbidden nodes, wildcard nodes, the presence of ring fusions, and specification of the degree of a node. These are shown in Table 1, which lists the symbol used in the SMARTS together with a "tag". The tags are attached to nodes in the tree-based chromosome and are used to encode the SMARTS features. The representation of a RG query as both a tree and a SMARTS is shown in Figure 2. On initialization of the EA, a population of tree-based chromosomes is generated at random, and each tree is scored according to its classification rate on the training set. The EA then enters a series of iterations in which individuals are ranked using the F-measure and chosen for reproduction, and offspring are generated and inserted into the population. The EA is configured to maximize the F-measure based on the training set. A test set is used to select the best query evolved over the course of the EA so that the query that forms the output is that which has the best F-measure when applied to the

**Table 1.** Subset of SMARTS Features Used To Represent RG Queries[a]

| symbol | symbol name | property |
|---|---|---|
| | Atom Primitives | |
| * | wildcard | any atom |
| D<n> | degree | <n > explicit connections |
| | Bond Primitives | |
| - | single bond | nonfused connection between RG nodes |
| = | double bond | fused connection between RG nodes |
| ~ | wildcard | can be a fused or nonfused connection |
| | Logical Operators | |
| exclamation | !e1 | not e1 |
| comma | e1,e2 | e1 or e2 |
| | SMILES | |
| . | DOT | disconnected |
| [ and ] | square brackets | used to enclose an atom - required with logical operators and primitives |

[a] Note that the "=" symbol is used to mean ring fusion.

test set. The use of a test set in selecting the query prevents the EA from overtraining.

**Multiobjective EA.** The modifications made to the EA during the development of the MOEA are described below.

**Pareto Ranking.** Recall and precision are handled independently in the MOEA instead of being combined into the single objective F-measure. The recall and precision of each query is calculated based on a training set of active and inactive compounds. The population is then ranked using Pareto ranking, as defined by Fonseca and Fleming,[14] with each individual being assigned a rank according to the number of times it is dominated. A nondominated individual (assigned rank 0) is one for which no other individual is better in all objectives, and one individual dominates another if it is better in at least one objective and is at least equally good in all other objectives. Parent selection is then biased toward individuals with lower ranks.

In order to allow the resolution of the search to be varied, the recall and precision values are binned with the bin number being used in place of the actual values during Pareto ranking: individuals that occupy the same bin are assumed equal. For example, if the number of bins is set to 50, then the resolution is 2%, this corresponding to the maximum difference for two individuals to occupy the same bin.

**Parent Selection.** Roulette wheel parent selection is implemented by dividing the wheel so that each segment represents a distinct rank that exists in the population. Rank 0 is allocated the largest segment of the wheel, and higher ranks are allocated progressively smaller segments; the relative sizes of the segments are dependent on the selection pressure constant. Thus parent selection is biased toward individuals with low rank.

MOEAs are prone to genetic drift, and several strategies have been described in the literature to maintain population diversity.[15] They typically involve comparing individuals in the population. Comparison of the RG queries themselves, either as SMARTS or as trees, would be computationally demanding since they are both nonunique representations, i.e., there are many ways in which an RG query can be mapped to a SMARTS and also to a tree, for example, by starting from a different node or by choosing a different node as root node, respectively. A more rapid way of comparing two queries is to compare the overlap of active compounds that they retrieve from a data set. Thus a retrieval profile is built for each query and is represented as a bit vector of length equal to the number of active compounds in the data set. For a given query, a bit is set to "1" if the query matches the corresponding training set compound, otherwise it is set to "0". Two retrieval profiles can be compared rapidly for identity or for similarity using the Tanimoto coefficient. The retrieval profiles are used as a secondary selection criterion to prevent genetic drift from occurring. The individuals which share a rank are divided into niches based on the similarities of their retrieval profiles: queries with retrieval profiles having Tanimoto similarity greater than a user-defined threshold are classed as being in the same niche. This is achieved using a sphere exclusion algorithm:[16] the first individual in a rank is compared to the remaining individuals at that rank, and any that are within the similarity threshold are placed in the same niche and removed from further consideration; this process is repeated until all individuals have been classified into niches, even if a niche contains only a single individual. Following selection of the rank, the wheel is spun a second time to select a niche at the given rank. The probability of selecting a given niche is equal to $1/N$ where $N$ is the total number of niches at that rank, i.e. it is independent of the number of individuals in each niche. An individual is then chosen at random from the selected niche. Thus, parent selection is primarily dependent on rank, with niching used as a secondary level of selection to control against potential overcrowding.

**Replacement Strategy.** A new offspring has to pass the following tests (applied in the order shown) to be inserted into the population, otherwise it is rejected.

Tree-complexity rules. A series of rules, described in the companion paper, were devised to prevent the formation of overly complex RG queries. Thus an offspring which contains any of the following (>10 nodes, >3 AND tags, >4 wildcards, >4 NOT tags) is rejected.

Dominance rules. The offspring is compared to each individual in the population and is rejected if it does not dominate any of the individuals in the current population.

Retrieval profile: If the retrieval profile of the offspring is unique, then it replaces an individual at the worst rank. If its retrieval profile is identical to another query's, then it replaces that query provided it has better precision, or it has equal precision and less complexity (fewer nodes), otherwise it is rejected.

If an offspring is rejected (for whatever reason), its parents are given a user-defined number of repeat attempts to generate a viable offspring.

**Archiving.** The best individuals generated over the course of the MOEA are stored in an archive which forms a separate population from that used for evolution. This process is equivalent to maintaining a record of the single best individual within the EA; however, in the MOEA there is a family of best individuals to record. Entry into the archive is dependent on performance on the test set (rather than the training set), and on termination of the MOEA the individuals in the archive form the solutions. After each generation, the nondominated individuals in the current population are considered for entry into the archive. First, recall and precision on the test set are calculated, and the individual is compared with those already in the archive using Pareto ranking. If it is nondominated with respect to the archive,

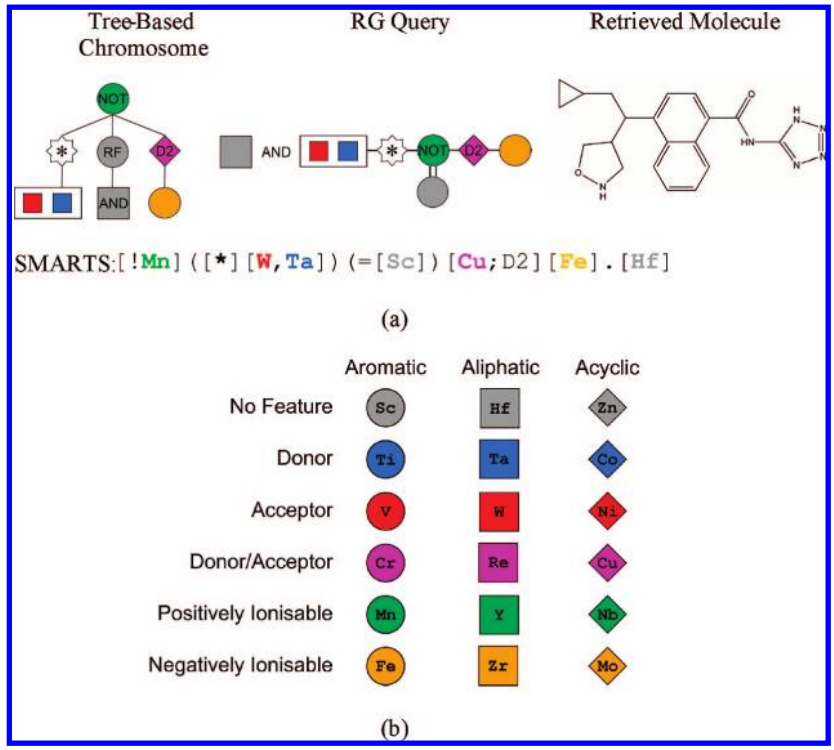


**Figure 2.** (a) An RG query (middle), its representation as a tree-based chromosome (left), and a molecule that contains the query (right). A SMARTS representation of the query is also shown. In the graphical queries shown throughout the manuscript the disconnected operator is illustrated by "AND"; series of alternative nodes are shown enclosed by a rectangle; and nodes that are combined with NOT logic are labeled "NOT". Ring fusions in a EA tree are indicated by "RF". (b) Key to the mapping between node types and element symbols.

then it is entered directly. If it has identical binned recall and precision values to an individual(s) already in the archive, then it is entered only if its retrieval profile is below a threshold similarity value of the retrieval profile of the existing individual(s). This condition is used to prevent overcrowding in the archive due to duplicate or highly similar individuals. If the similarity is greater than the threshold, then the simpler of the two queries is retained, where simplicity is measured by the number of nodes and tags that constitute the query. Finally, any individuals in the archive that are dominated by the new additions are removed.

**Deriving Multiple Queries.** The MOEA can be run with a third objective, in addition to recall and precision, called uniqueness. The uniqueness score is used to encourage the MOEA to evolve queries that retrieve complementary sets of actives for later team formation. The uniqueness score is calculated for each individual in the population using eq 1 which compares its retrieval profile with those of all of the individuals in the archive. For each active $a$ retrieved by query $Q$, the number of queries in the archive that also retrieve $a$ is calculated ($n_a$). The uniqueness of query $Q$ is then the normalized sum over all actives ($A$) retrieved by $Q$.

$$\text{Uniqueness}(Q) = \left(\sum_{a=1}^{a=A} \frac{1}{n_a}\right) / A \qquad (1)$$

A higher uniqueness score is assigned to an individual that has low overlap compared with the other individuals in the archive. The uniqueness score is binned with the number of bins set to 100. As for the recall and precision objectives, the uniqueness score is based on the training set when considering parent selection and on the test set when considering entry into the archive.

**Team Formation.** The combining of queries into teams is implemented as a post-MOEA step. All solutions in the archive are combined pairwise to form two-query teams, and a combined retrieval profile is generated for each team by ORing for their individual retrieval profiles. The combined recall and precision is calculated for each team, the new population of teams is Pareto ranked, and the nondominated teams are retained. Each nondominated individual from the archive is then combined with the two-query teams to form three-query teams, and so on. Team formation can be applied when the MOEA has been run with two objectives (recall and precision); however, the experiments reported later demonstrate that more effective results are obtained when the uniqueness objective is also optimized.

**Parameters.** Preliminary experiments were carried out (results not shown) to establish appropriate parameters for the subsequent runs reported here. These are shown in Table 2.

**Data Sets.** The MOEA has been applied to the Hert data sets extracted from MDL's Drug Data Reports Database (MDDR)[17] and to an in-house screening data set provided by GSK. The MDDR data set consists of the 11 activity classes shown in Table 3.[18] Three sets of compounds of size 3000 were selected from the remainder of MDDR following removal of the 11 activity classes. These form the inactives.

The GSK data set consists of compounds screened for hERG activity. hERG is a protein derived from the human Ether-a-go-go Related Gene and is a potassium ion channel. It is of particular importance in the maintenance of normal heart function.[19] Molecules that bind to and block the channel can lead to arrhythmia of the heart with potentially fatal consequences. The hERG-E assay assesses the disruption of activity from binding to the hERG channel via the conse-

**Table 2.** MOEA Parameter Settings[a]

| | |
|---|---|
| population size | 300 |
| no. of iterations | 1000 |
| parent selection | roulette wheel |
| selection pressure | 1.25 |
| no. of levels in tree at initialization | 3 |
| max no. of child nodes | 3 |
| max attempts | 10 |
| no. of bins for precision and uniqueness | 100 |
| active retrieval profile similarity cutoff | 0.9 |
| genetic operator rates: | |
| chance of modification (else mutation) | 50% |
| chance of mutation (else crossover) | 30% |
| chance of mutation occurring on each node | 50% |
| chance of single vs two-parent crossover | 50% |
| percentage of worst population members replaced each generation | 10% |

[a] See ref 6 for further details.

quential electrophysiological disruption (hERG-E). The activities are recorded as $pIC_{50}$ values with the continuous values being transformed to qualitative classifications by applying an activity cutoff threshold: compounds with $pIC_{50}$ less than or equal to 5.0 were classed as inactive, and those with $pIC_{50}$ above 5.0 were classed as active, as shown in Table 4. The threshold was chosen following experiments with the EA.

## RESULTS

**Generating a PR Curve.** The MOEA was used to generate a PR curve for the 5HT1A data set. The set of actives was divided at random into three subsets, and each was combined with a different subset of inactives to form training, test, and validation sets. Previous results have shown that the EA is robust to different selections of active and inactive compounds, and so the experiments reported here are for three runs using one data set with different random numbers used to seed the MOEA. The MOEA was applied to evolve a family of queries that explore the tradeoff in recall and precision. The progression of the MOEA on the training set is shown in Figure 3 where recall is plotted on the *x*-axis, precision is plotted on the *y*-axis, and each data point represents a different RG query. An ideal solution would be one with perfect recall and precision ($R = 1.0$ and $P = 1.0$) and would be positioned at the top right-hand corner of the plot. On initialization the population is scattered with no solutions that are good in both objectives, Figure 3(a). As the MOEA progresses, the solutions begin to map out a continuous surface that is moving toward the top right. After 1000 generations most of the population now resides on what is known as the Pareto surface, i.e. a large proportion of the population consists of nondominated solutions, Figure 3(c). This indicates that the MOEA is reaching convergence. It is now clear that the ideal solution does not exist, and the population represents a family of tradeoff solutions. Figure 3(d) shows the performance on the test set. The most noticeable feature is that the surface is less smooth; however, the overall performance on the test set is very similar to that seen on the training set and indicates that the queries have good generalizing properties. The solutions output by the MOEA are those in the archive and are nondominated in the test set. They represent a family of equivalent RG queries each of which exhibits a different compromise in recall and precision.

Figure 3 shows that the entire Pareto surface is covered and contains solutions at both extremes: RG queries that retrieve nearly all the actives (recall $\approx$ 1.0), where the precision is very low (i.e., there is a very high proportion of false actives) to solutions at the other extreme, with high precision and very low recall. High precision-low recall queries can be of value when the aim is to extract structure−activity information from a data set since these will describe features that are present in active compounds while being absent from most of the inactive compounds. However, as will be seen below, such queries may not be effective when used for prediction. Conversely, high recall-low precision queries may be of interest when the aim is to derive a predictive model of activity that can be used to select new compounds for testing since a higher rate of false negatives can be tolerated in order to increase the hit-rate. In practice, a user may select a solution where the balance in recall and precision is such that the total number of compounds retrieved by the query (predicted as active) is equal to the capacity of the screen. Such flexibility in the balance of the recall-precision, and the number of compounds retrieved is not possible with our previous approach using a single objective.[6]

Figure 4 shows the performance when each of the solution queries is applied to the validation set. The test set results are shown in red, and the results of applying the queries to the validation set are shown in blue. The performance on the validation set is close to that of the test set for the solutions in the top half of the plot. However, the performance in the validation set deteriorates in the lower half of the plot. These solutions represent high precision and low recall queries; although there are few true positives, the number of false positives is small so that the queries have high sensitivity with respect to the test set; however, the drop in performance on the validation set indicates that they are overtrained and are poor in prediction. The overall performance of the MOEA is more difficult to quantify than the single-objective EA since it is based on a family of solutions rather than a single solution. It is estimated here by calculating the average F-measure (based on the optimized recall and precision values) over all the solutions. Thus, the difference in average F-measure for the nondominated queries applied to the test set and the same queries applied the validation set is 3.8%. However, the performance of queries with test set precision greater than 60% is reduced by 5.0%, compared to a 2.6% reduction for queries with test set precision of less than 60%.

For comparison, equivalent results are shown for three runs of the single objective EA using different random seeds: each run generates a single solution that represents a (local) maximum of the F-measure. The unfilled circles represent the test set results with the validation set results shown as solid circles. Thus, as well as mapping the extremes of the

**Table 3.** Hert Data Set

| activity class | actives | training set | test set | validation set |
|---|---|---|---|---|
| 5HT3 antagonists (5HT3 Ant) | 752 | 251 | 251 | 250 |
| 5HT1A agonists (5HT1A) | 827 | 276 | 276 | 275 |
| 5HT reuptake inhibitors (5HT-RT) | 359 | 120 | 120 | 119 |
| dopamine D2 antagonists (D2) | 395 | 132 | 132 | 131 |
| renin inhibitors (renin) | 1130 | 377 | 377 | 376 |
| angiotensin II AT1 antagonists (AT1 Ant) | 943 | 314 | 314 | 315 |
| thrombin inhibitors (thrombin) | 797 | 266 | 266 | 265 |
| substance P antagonists (Sub P) | 1246 | 415 | 415 | 416 |
| HIV 1 protease inhibitors (HIV1) | 750 | 250 | 250 | 250 |
| COX inhibitors (COX) | 626 | 212 | 212 | 212 |
| protein kinase C inhibitors (PKC) | 453 | 151 | 151 | 151 |

**Table 4.** hERG-E Data Set

| target | inactive $pIC_{50} \leq 5.0$ | active $pIC_{50} > 5.0$ | total |
|---|---|---|---|
| hERG-E | 1845 | 541 | 2386 |

recall-precision curve, the MOEA solutions cover the space of the EA solutions.

**Identifying Multiple Structural Series Using the MOEA.** In the results shown thus far, each solution aims to capture the SAR in a single query, i.e. each query attempts to retrieve as many of the actives as possible while not being present in the inactives. In this section, we investigate combining individual queries into team-based solutions which are then evaluated on the combined performance of the team-members. Thus, a compound in the training set is predicted as active if it contains any of the queries included in the team, otherwise it is predicted as inactive. It is anticipated that a team will be able to improve on recall by allowing the actives to be described by more than one SAR without seeing deterioration in precision. We investigate two approaches to combining the individual solutions into teams. The first is simply to iteratively combine queries generated by the two-objective MOEA described thus far. In the second

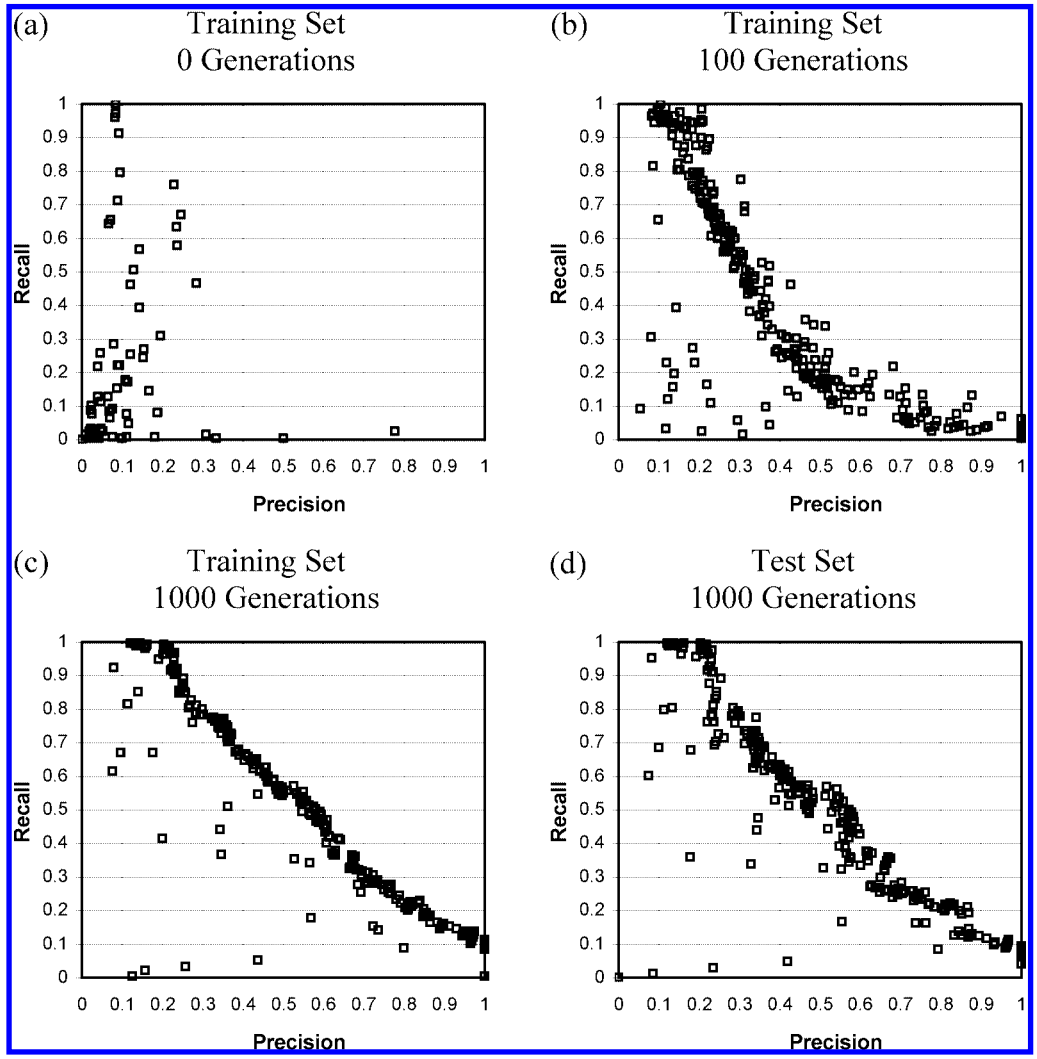


**Figure 3.** PR curves for the 5HT1A data set. The performance of the MOEA on the training set is shown (a) on initialization; (b) after 100 generations; and (c) 1000 generations. Performance on the test set is shown in (d).

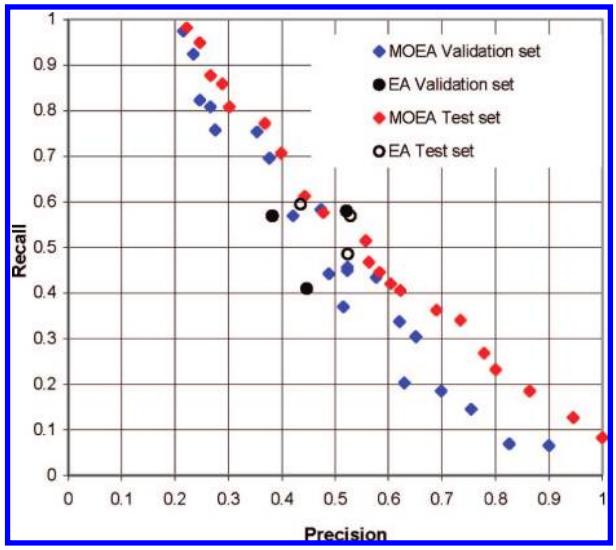


**Figure 4.** PR curve for the 5HT1A validation set. Performance on the test set is shown in red with performance on the validation set shown in blue. Solutions generated from three runs of the single-objective EA are shown in black; the unfilled circles are the test set, and the filled circles are the validation set.

approach, the MOEA is run with a third objective, uniqueness, that is optimized simultaneously with recall and precision. The resulting queries are then combined iteratively as for the two-objective MOEA. The three-objective MOEA aims to evolve queries that complement one another, whereas there is no cooperation between queries during evolution in the two-objective MOEA. In both cases, team formation is applied to the nondominated solutions and proceeds as follows. All pairwise combinations of individual queries are combined to form two-query teams, and their combined recall and precision values are calculated. Pareto ranking is then applied to the teams, and the nondominated teams are retained as solutions. Three-query teams are generated by combining all two-query teams with the individual queries, calculating the combined recall and precision values, Pareto ranking, and retaining the nondominated three-query teams. This process can then be repeated to generate higher order teams.

Results are shown in Figure 5 for the two- and three-objective MOEA applied to the 5HT1A data set. In each case, teams with up to five queries were generated. The performance of the two-objective MOEA is shown at the top of the figure with the test set on the left and the validation set on the right. The corresponding plots for the three-objective MOEA are shown below. In all cases the team solutions are labeled as one-query (the nondominated solutions evolved by the MOEA); two-queries (teams consisting of two queries); and so on. The relative performances of the team-based solutions on the test set are compared quantitatively in Table 5 which gives the maximum F-measure and the average F-measure over all solutions with the same number of queries.

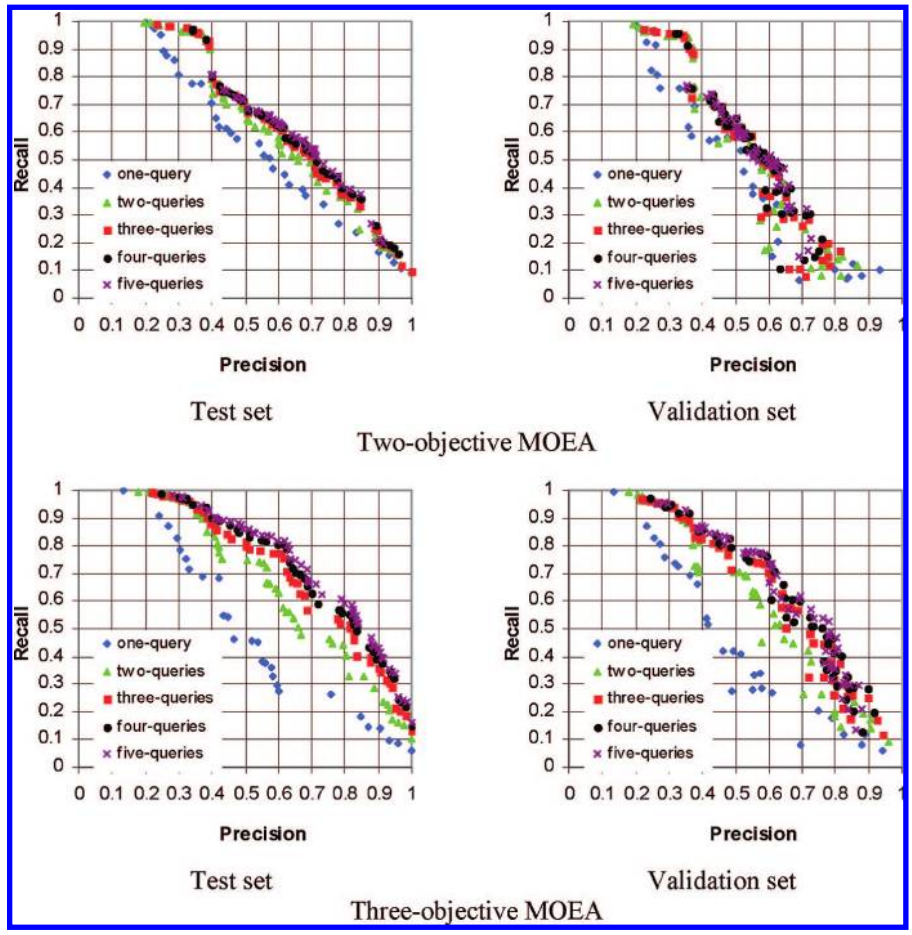


**Figure 5.** Team-formation on the 5HT1A data set. The top plots show performance of the two-objective MOEA on the test set (left) and the validation set (right). The bottom plots show the corresponding performance of the three-objective MOEA. In all cases, one-query solutions are in blue; two-query solutions are in green; three-query solutions are in red; four-query solutions are in black; and five-query solutions are in purple.

**Table 5.** F-Measure Performance of the Two- and Three-Objective MOEA on the Test and Validation Sets

| | two-objective MOEA | | | | three-objective MOEA | | | |
|---|---|---|---|---|---|---|---|---|
| | test set | | validation set | | test set | | validation set | |
| | mean | max | mean | max | mean | max | mean | max |
| one-query | 41.2% | 55.2% | 37.2% | 53.0% | 38.5% | 51.9% | 35.9% | 48.7% |
| two-queries | 47.0% | 58.9% | 42.3% | 55.6% | 49.3% | 63.4% | 46.0% | 60.2% |
| three-queries | 51.2% | 61.0% | 45.8% | 56.5% | 53.4% | 67.8% | 49.7% | 65.1% |
| four-queries | 56.0% | 61.5% | 50.3% | 56.6% | 57.9% | 69.7% | 53.3% | 66.7% |
| five-queries | 58.1% | 61.8% | 52.1% | 56.6% | 60.3% | 70.3% | 55.5% | 66.8% |

Considering the two-objective MOEA, the performance improves as the number of queries that are combined increases; however, the rate of improvement decreases with number of queries. The one-query solutions evolved using the three-objective MOEA tend to exhibit worse performance than the solutions obtained for the two-objective MOEA, particularly for medium precision (0.4 to 0.6). Note, however, that the three-objective solutions actually map out a three-dimensional surface (recall, precision, and uniqueness) of which only two dimensions are shown in Figure 5. Driving the solutions toward ones that complement one another, through the uniqueness objective, clearly has a detrimental effect on recall and precision. The reduced performance in the central region may represent the fact that these queries are more likely to have similar retrieval profiles to either extreme. However, the increase in performance on combining the queries into teams is more significant than was seen for the two-objective MOEA so that the three-objective method outperforms the two-objective method when team formation is applied. This indicates that the uniqueness score is effective in ensuring that the queries are complementary. The maximum validation set F-measure achieved for the two-query solutions using the three-objective MOEA is 60.2% compared to 55.6% for the two-objective MOEA, and its performance exceeds that of any solution generated by the two-objective MOEA.

The performance of the team-based queries diminishes when applied to the validation set, for both the two- and three-objective MOEA runs. For the three-objective MOEA, the difference in average F-measure between the test and validation sets for the individual queries is 2.6% compared to 3.4% for the two-query solutions, with this difference rising to 4.9% for the five-query solutions. However, it should be noted that the F-measure is not being optimized in this case, and so such comparisons are indicative only.

The team formation approach described thus far considers nondominated queries only at each step. While this has advantages as far as speed is concerned, it is possible that teams that are dominated when taken individually may lead to nondominated teams when they are combined in the next iteration. Thus, the more computationally intensive approach of enumerating all teams was investigated. Results are shown in Figure 6, where it can be seen that very few additional solutions are generated.

While mapping the entire Pareto surface allows the performance of the MOEA to be assessed, solutions at the extremes of the Pareto frontier are unlikely to be of interest, and so the use of constraints has also been investigated when forming teams. Figure 7 shows the results of team formation for the three-objective MOEA runs, using a set of constraints on team formation: queries with an enrichment factor of less than 4 (equivalent to low precision), recall of less than 10%, or an active retrieval profile similarity of greater than 0.9 (assessed in an order dependent manner) are discarded. As was seen previously, queries with low recall and high precision are likely to be overtrained and therefore poor in prediction, hence, it may be preferable to exclude these from team formation. The effect of applying the constraints is to reduce the number of teams that are formed and to produce a more even spread of solutions over the region of likely interest.

**Analysis of the 5HT1A Agonist Team Solutions.** Thus far, we have considered the solutions in terms of their quantitative performances only. Here we consider the queries themselves and the structure−activity information encoded therein. The one-query solution evolved by the three-objective MOEA with highest F-measure is compared with the two-query and three-query solutions with highest F-measures generated following team-formation. Although queries were selected based on their test set performance, Table 6 details the performance of each of these solutions on the validation set along with their component queries

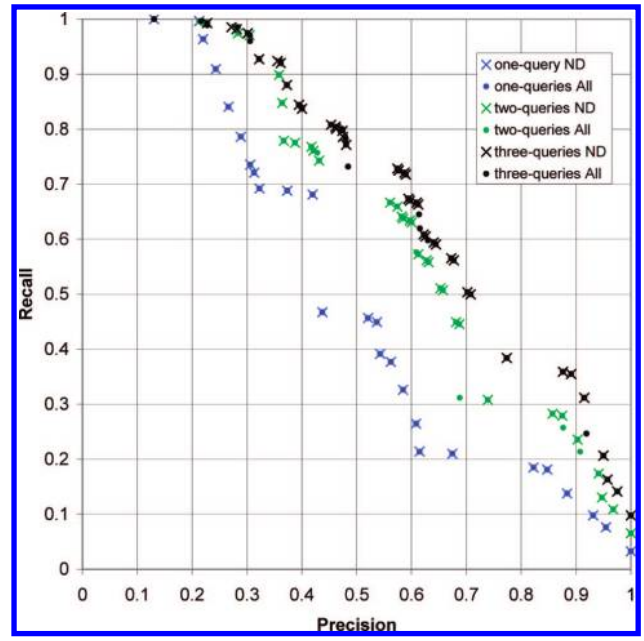


**Figure 6.** Team formation using the three-objective MOEA. The results are shown for team formation using nondominated queries only in each iteration (solid and labeled as ND) and combining all possible queries in each iteration (unfilled and labeled as ALL). In both cases, the starting point is the same−the solutions in the archive (which are all nondominated by default). When combining all queries a few additional solutions are generated, shown by the circles with no counterpart; however, these additional solutions do not justify the additional computational resources required.

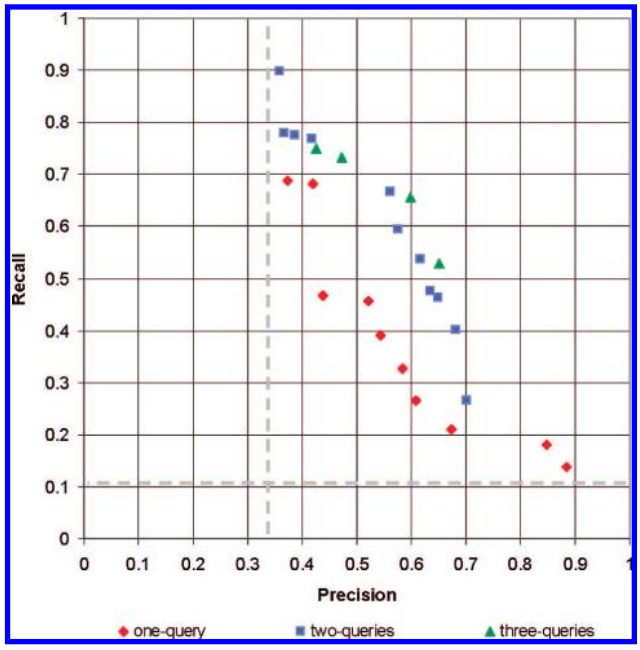


**Figure 7.** The 5HT1A test set results for constrained team formation using the three-objective MOEA. The lower recall limit of 10% and the lower precision limit corresponding to an enrichment factor of 4 are indicated by dashed lines on the plot. A maximum Tanimoto coefficient similarity of 0.9 between the solutions' active retrieval profiles was also imposed.

represented as SMARTS strings, and the queries are shown graphically in Figure 8.

The two-query solution has an improved F-measure relative to the one-query solution. While there is a small overall increase in recall, the improvement is mainly due to an increase in the precision of the constituent queries (A and B) that comprise the team. This is also evident from the queries themselves since the one-query solution defines both aliphatic (Y) and acyclic (Nb) positively ionizable features and hence is able to retrieve similar groups of actives. However, this is at the expense of increased false actives. For example, the aliphatic positively ionizable node (Y) in the one-query solution is specified in an OR list along with two other alternative node types, rather than more specifically on its own and with a requirement for a connectivity of two as in query A.

There are similarities in the SAR information in each query in the two-query team, for example, both include an aromatic featureless node (Sc) and require a positively ionizable node. However, the queries are clearly different, and this is reflected in the fact that there is no overlap in the actives retrieved by the two queries. It seems that the most important difference between the queries lies in their specification of different types of positively ionizable groups, thus allowing the retrieval of two independent subsets of actives within the 5HT1A agonists.

The structural diversity of the compounds that match each of the queries is evident in Figure 8 and is due both to the use of RGs in which different substructures are represented by the same RG node and to the SMARTS operators which are encoded within the queries. For example, the hits consist of different scaffolds ranging from fused and nonfused two-ring structures to three member fused rings to larger linear and branched structures.

While there is further improvement in performance for the three-query solution compared to the two-query solution, the difference in F-measure is smaller than that between the single-query and the two-query solutions. Again, the improvement in fitness is mainly due to an increase in precision, with a small increase in recall. However, while previously each component query of the solution retrieved all its actives uniquely, here approximately 11% of the actives retrieved by the three-query solution are retrieved by more than one of the component queries.

The increased precision of the team-based approach relative to a single query has the obvious benefit of increased performance in classification. Furthermore, there is the additional benefit that the SAR information encoded in the queries is more specific i.e. it is less diluted by features that are present in inactives. However, there is a compromise between the level of specificity of the features encoded in the queries and their generalizability, since highly specific queries (high precision) tend to describe smaller proportions of active chemical space and may have little relevance to the description of the typical features of the activity class as a whole. While such specific SAR information may be regarded as providing a useful starting point for further exploration of a particular region of chemical space, such a query may be of limited use in prediction. For example, when using a query in a virtual screening context, the increasing disparity between the test and validation set results for queries of higher precision may make the predictions less reliable.

**Team-Formation in the MDDR.** The three-objective MOEA was run three times on each of the 11 MDDR activity classes, and constrained team formation was carried out on the resulting sets of nondominated queries. For each class, the best team-based solution was selected on the basis of highest test set F-measure and was applied to the associated validation set. The performances on the test and validation sets are compared on precision and recall, respectively, and the number of queries that comprise the best team solution for each activity class is indicated below the charts. The performances of the best queries evolved using the single-objective EA are also shown. These queries are referred to as single-queries to distinguish them from one-query solutions evolved by the MOEA.

In four of the activity classes (5HT1A, renin, thrombin, Sub P), a team-based solution dominates (i.e. is better in

**Table 6.** Best Team-Based Queries Generated for the 5HT1A Activity Class Using the Three-Objective MOEA

| solution | F | P | R | TA | FA | SMARTS |
|---|---|---|---|---|---|---|
| one-query | 48.7% | 38.6% | 65.9% | 182 | 290 | [Ti,Y,Nb][Ti,Hf,Zn].[Sc,V] |
| two-query (A+B) | 60.2% | 53.0% | 69.6% | 192 | 170 | - |
| A | 45.6% | 51.4% | 40.9% | 113 | 107 | [Sc,V,Mn,Co,Nb]~[Y;D2][*] |
| B | 37.5% | 54.5% | 28.6% | 79 | 66 | [Sc,Cr]=[Hf].[Nb] |
| three-query (A+C+D) | 65.1% | 58.3% | 73.6% | 203 | 145 | - |
| C | 31.9% | 74.7% | 20.3% | 56 | 19 | [W,Nb][Zn][Ti,Ta,Hf,Y;D2]~[*;D1] |
| D | 32.7% | 69.4% | 21.4% | 59 | 26 | [Y,Nb][Hf]=[Sc] |

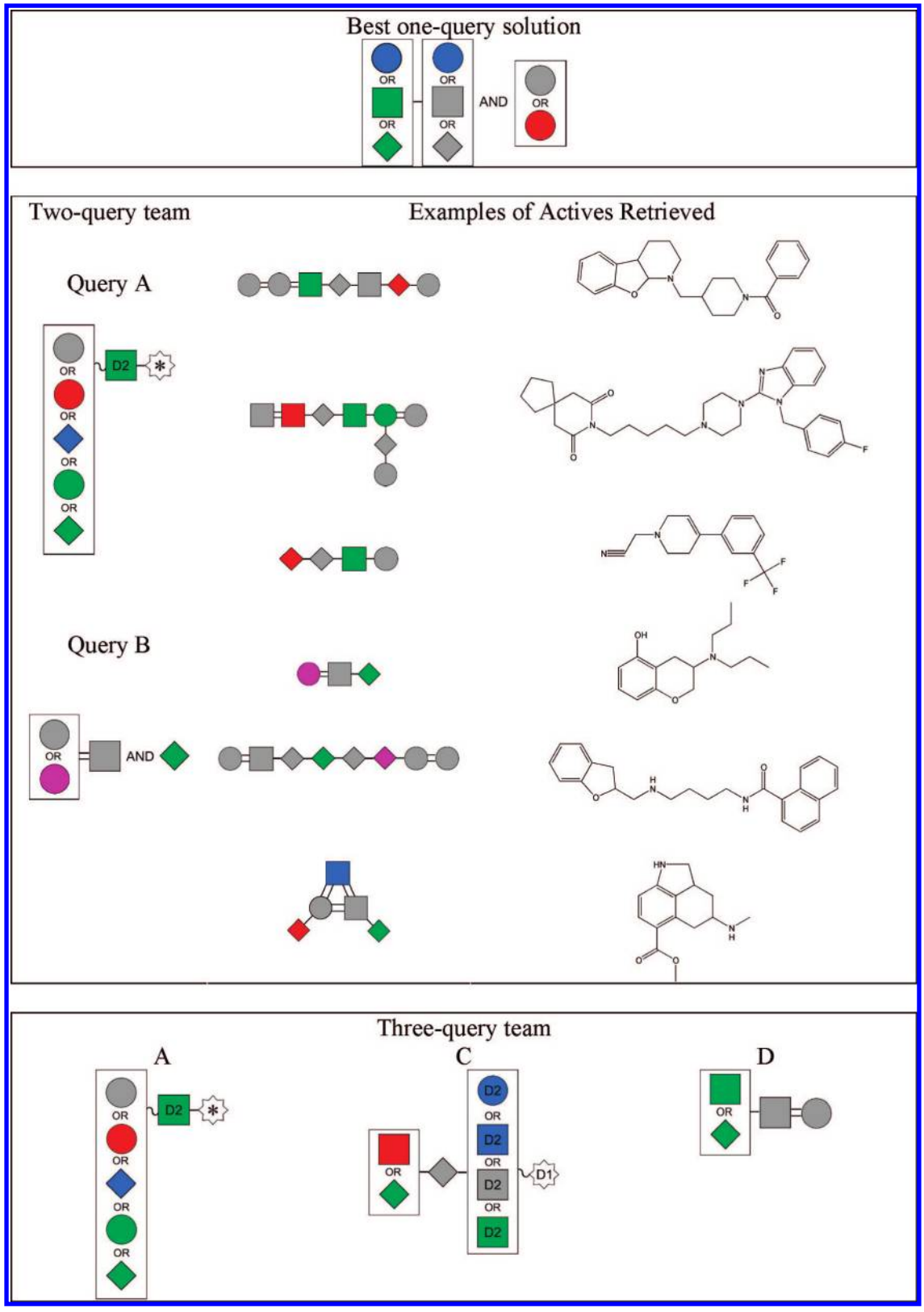


**Figure 8.** The best team-based queries generated for the 5HT1A activity class using the three-objective MOEA.

both precision and recall) the single-query solution in both the test and the validation sets. There is only one activity class (HIV1) where the team-based solution is dominated by the single-query solution. For the remaining activity classes, either the team-based solution dominates the single-query in recall and is dominated in precision or vice versa. The actual difference in the F-measure performance between the single-query and team-based solutions is small: the mean increase in validation set F-measure is 3.3% taken over the activity classes where the single-queries dominate but is larger, 7.8%, for the activity classes where the team-based queries dominate. Taken over all the activity classes, the team-based solutions are better by 2.8% on the test set and 2.1% on the validation set.

The maximum team-size found over all activity classes is four which indicates that team formation (at least as

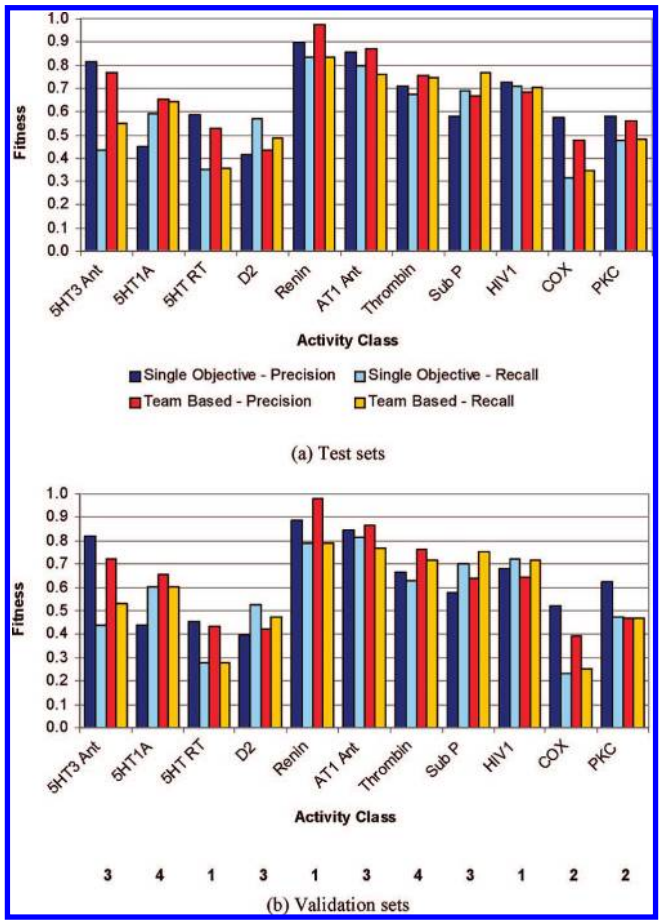


**Figure 9.** Results of team formation on the 11 MDDR data sets. Performance on the test sets is shown in the top half of the figure, and performance on the validation set is shown in the bottom half—with the number of teams that gave the best solution in terms of the F measure for each data set shown below.

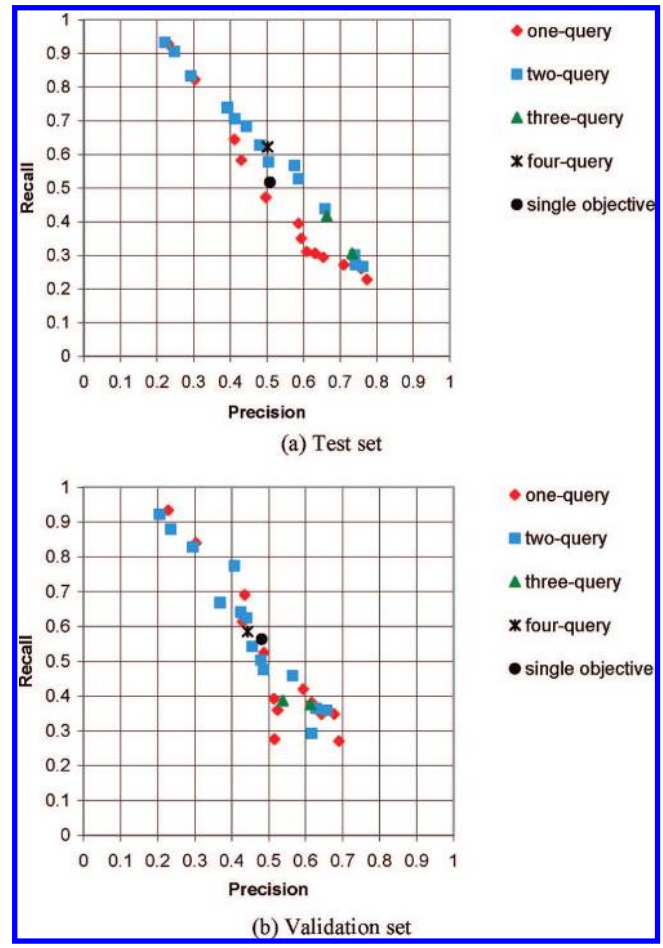


**Figure 10.** The MOEA applied to the hERG data set: (a) performance on the test set and(b) performance on the validation set.

implemented here) has natural limits. Furthermore, it is also interesting that in three of the cases (HIV1, renin, and 5HT-RT) the best team-solution consists of just one query. Differences in the optimal number of queries to describe an activity class may be a reflection of the natural grouping of compounds within the classes. For example, the renins have relatively high structural homogeneity and may be adequately described by one query, whereas classes with greater structural diversity may be better described using multiple queries, with each query describing a different structural class. However, this rationale does not explain why the class with the lowest structural similarity (COX inhibitors) is best described using just two queries, and it may be simply that the structure−activity information is too dilute. A further consideration is the extent to which the RGs are able to capture different structural series that may be present. Finally, the "best" team-solution has been selected using the F-measure which tends to favor solutions with more equal compromises between precision and recall. While this allows performance across the different data sets to be compared, it may not always result in the most desirable solution. For example, as discussed earlier, queries with relatively high precision may be more useful when attempting to derive a SAR.

**GSK Data Set.** The results of team-formation applied to the hERG-E data set using the three-objective MOEA are shown in Figure 10. Performance on the test set is in the top of the figure and the validation set is at the bottom. The best query derived from three runs of the single-objective EA is also plotted for comparison.

As seen for the 5HT1A data set, the one-query solutions evolved using the three-objective MOEA include queries that are of comparable performance to the best query generated using the single-objective EA while also mapping out more of the PR curve. Furthermore, combining the individual queries gives substantial improvements in performance on the test set over the individual queries. The largest improvement in fitness is observed when comparing the one-query solutions to the two-query team solutions, with much smaller improvements in fitness resulting from the addition of third and fourth queries. Unlike for the 5HT1As, there is little benefit to be had from increasing the number of queries per solution above two, particularly since the increased complexity of larger solutions makes them less interpretable. As expected, the overall performance on the validation set is worse than the test set, with a greater difference in performance generally being observed for solutions of higher precision. The deterioration in performance is greatest for the team solutions to the extent that they become comparable to the individual queries, i.e., there is no longer any advantage in combining queries in terms of prediction.

However, the two-query solutions provide valuable structure−activity information. For example, the two-query solution (A + B) with the highest F-measure is shown in Figure 11. Quantitative results are in Table 7 together with

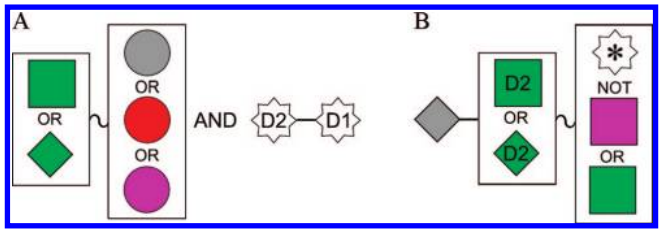


**Figure 11.** The best two-query solution generated for the hERG-E data set using the three-objective MOEA.

**Table 7.** Best Two-Query Solution Generated by the MOEA Is Compared with the One-Query Solution That Retrieves a Similar Number of Compounds (C) and the One-Query Solution That Has Closest Recall (D)

| | test set | | | validation set | | | |
|---|---|---|---|---|---|---|---|
| | F | P | R | F | P | R | SMARTS |
| A | 41.2 | 60.9 | 31.1 | 36.0 | 51.5 | 27.6 | [Y,Nb]~[Sc,V,Cr].[*;D2][*;D1] |
| B | 47.2 | 58.7 | 39.4 | 44.5 | 51.4 | 39.2 | [Zn][Y,Nb;D2]~[!Re;!Y] |
| A+B | 57.1 | 57.6 | 56.7 | 49.1 | 47.9 | 50.3 | |
| C | 48.4 | 49.7 | 47.2 | 50.5 | 48.7 | 52.5 | [Y]~[*].[Zn] |
| D | 51.4 | 47.0 | 56.7 | 52.7 | 46.8 | 60.2 | [Y]~[*][!Cr;!Co].[Sc,Mn] |

the one-query solution (C) that retrieves the closest number of compounds as A+B (true actives and false inactives: TA + FI) and the one-query solution (D) that has closest recall to A+B (true actives: TA).

The two-query solution (A+B) has greater precision (56.7%) on the test set than both C and D. Although a similar gain in performance is not seen in the validation set where A+B is similar in performance to query C, the structure−activity information present in the queries A and B is richer than that in query C, which simply specifies an aliphatic positively ionizable group connected to a wildcard node (i.e., any node type) and a linker node, which may or may not be connected to the other two.

Queries A and B share some common features, for example, the presence of a positively ionizable aliphatic or acyclic node. However, they differ in several important ways. In query B, the positively ionizable node is required to be of degree two: there is no such restriction in query A. Query A requires a minimum of four RG nodes, whereas query B contains three nodes (with alternatives specified). Query A requires that specific aromatic node types be present adjacent to the positively ionizable group, whereas query B is much less specific about this requirement. Query B does however require that a linker node be adjacent to the positively ionizable group, whereas query A does not. Consequently, it is not surprising that approximately 80% of the compounds retrieved by the two-query solution are retrieved uniquely by either one of the queries. Considering those actives ($pIC_{50} > 5$) retrieved uniquely by each query, query B retrieves compounds of a higher mean activity than query A−5.62 and 6.01 for queries A and B, respectively. In addition, while only 2 out of the 20 unique actives retrieved by query A are above the high activity cutoff ($pIC_{50} > 6$), this number rises to 18 out of 41 for query B, further emphasizing the fact that query B tends to retrieve compounds of higher activity than query A. The relatively low retrieval overlap between the queries of the two-query solution and the fact that the queries are retrieving actives of different potency is desirable as it adds to the detail in the SAR information and is perhaps indicative of the potential for retrieving a greater diversity of actives when using the multiquery approach. This detailed analysis has generated easily understandable SAR information consistent with the currently held beliefs relating to the hERG pharmacophore.[20,21] Thus, although in terms of the validation set performance there may be little difference between using a single query and using a multiquery solution, the additional benefits discussed here demonstrate the value of the multiquery approach.

## CONCLUSIONS

We previously reported a method for evolving RG queries with the aim of capturing the structure−activity information contained in a screening data set using a representation that is readily interpretable by a chemist. Here we have extended the approach to evolve multiple RG queries using multiobjective optimization techniques. By simultaneously optimizing recall and precision it is possible to investigate the tradeoff that generally exists in these conflicting objectives. The result is a family of queries that lies on a PR curve. The multiobjective approach offers significant advantages over using a single combined objective, such as the F-measure, as is typically used in machine learning methods, since the user is now able to select a query with an appropriate balance in recall and precision. For example, a low recall-high precision query may be preferred when establishing the SAR, whereas a high recall-low precision query may be more appropriate in a virtual screening context.

Although a family of solution queries is evolved, each query aims at capturing the structure−activity information into a single representation, and each can be considered as an alternative (equally valid) solution. We then investigated the combining of individual queries into teams with the aim of capturing multiple SARs that may exist in a data set. Team formation is carried out iteratively as a postprocessing step following the evolution of the individual queries. The inclusion of uniqueness as a third objective within the MOEA was found to be an effective way of ensuring that the queries in the final population are complementary in terms of the active compounds that they describe. Substantial improvements in both recall and precision were seen for the higher-order teams for some data sets; however, it is clear that the optimum number of queries required to describe an activity class will vary depending on several factors including the size and diversity of the activity class. For example, a relatively homogeneous set of actives may be described best using a single query. Nevertheless, attempting multiquery formation up to a reasonable number of queries such as four or five allows the best solution to be selected according to the users' criteria for simplicity, accuracy, and SAR information.

## ACKNOWLEDGMENT

We acknowledge Eleanor Gardiner for helpful comments on this manuscript; Daylight Chemical Information for software support; and MDL Information Systems Inc. for the provision of the MDDR database. The work was funded by GlaxoSmithKline and BBSRC via an industrial CASE studentship.

## REFERENCES AND NOTES

(1) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11*, 694–699.

(2) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead hopping using SVM and 3D pharmacophore fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 1122–1133.
(3) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead- discovery programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469–474.
(4) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.
(5) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.
(6) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving interpretable structure-activity relationships. 1. Evolving reduced graph queries. *J. Chem. Inf. Model.* **2008**, *48*, 1543–1557.
(7) van Rijsbergen, C. J. *Information retrieval*; Butterworth: London, 1979.
(8) James, C. A.; Weininger, D.; Delaney, J. *Daylight theory manual 4.82*; Daylight Chemical Information Systems, Inc.: Los Altos, 2003.
(9) Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves In *Proceedings of the 23rd International Conference on Machine Learning*; Pittsburgh, PA, 2006; pp 233−240.
(10) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the "Receiver operating characteristic" Curve approach. Application to high-throughput docking on metabotropic glutamate receptor type 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
(11) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145–2156.
(12) Barker, E. J.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Willett, P. Scaffold-hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
(13) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
(14) Fonseca, C. M.; Fleming, P. J. Multiobjective optimization and multiple constraint handling with evolutionary algorithms - part 1: A unified formulation. *IEEE Trans. Systems, Man, Cybernetics* **1998**, *28*, 26–37.
(15) Goldberg, D. E. *Genetic algorithms in search, optimization and machine learning*; Addison-Wesley: Wokingham, 1989.
(16) Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
(17) MDL Information Systems Inc. 2440 Camino Ramon, Suite 300, San Ramon, CA 94583.
(18) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
(19) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463–469.
(20) Song, M.; Clark, M. Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
(21) Aronov, A. M.; Goldman, B. B. A model for identifying hERG K+ channel blockers. *Bioorg. Med. Chem.* **2004**, *12*, 2307–2315.

CI800051H