JOURNAL OF CHEMICAL INFORMATION AND MODELING

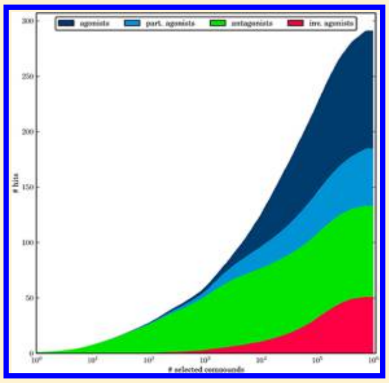
Article

pubs.acs.org/jcim

# Searching for Closely Related Ligands with Different Mechanisms of Action Using Machine Learning and Mapping Algorithms

Jenny Balfer, Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany



**ABSTRACT:** Supervised machine learning approaches, including support vector machines, random forests, Bayesian classifiers, nearest-neighbor similarity searching, and a conceptually distinct mapping algorithm termed DynaMAD, have been investigated for their ability to detect structurally related ligands of a given receptor with different mechanisms of action. For this purpose, a large number of simulated virtual screening trials were carried out with models trained on mechanistic subsets of different classes of receptor ligands. The results revealed that ligands with the desired mechanism of action were frequently contained in database selection sets of limited size. All machine learning approaches successfully detected mechanistic subsets of ligands in a large background database of druglike compounds. However, the early enrichment characteristics considerably differed. Overall, random forests of relatively simple design and support vector machines with Gaussian kernels (Gaussian SVMs) displayed the highest search performance. In addition, DynaMAD was found to yield very small selection sets comprising only ~10 compounds that also contained ligands with the desired mechanism of action. Random forest, Gaussian SVM, and DynaMAD calculations revealed an enrichment of compounds with the desired mechanism over other mechanistic subsets.

## ■ INTRODUCTION

In recent years, machine learning methods have become increasingly popular for ligand-based virtual screening.[1−4] The currently most widely and successfully applied machine learning approaches in virtual screening[3,4] include Bayesian classifiers,[5,6] support vector machines (SVMs),[7,8] and random forests (RFs).[9,10] Bayesian methods are probabilistic approaches that operate on the basis of Bayes' theorem[5] and derive likelihood estimates for test compounds to match feature distributions of reference sets. In virtual screening, naïve Bayesian classifiers are often used to rank database compounds according to their probability to display a specific biological activity.[6] SVMs represent a class of kernel-based learning algorithms[7] that attempt to derive a separating hyperplane for the classification of training examples with different class labels (e.g., active vs inactive) and utilize the resulting linear models to predict class labels of test instances. If a separating hyperplane does not exist in a given feature space, learning sets can be projected into higher-dimensional-space representations where linear separation might be feasible.[7,8] RFs are based on decision-tree structures and combine multiple independently derived trees for a given training set to facilitate consensus predictions.[9] All three methodologies are supervised learning techniques, and they enable ranking of test compounds according to probability scores (Bayesian classification), signed distances of objects from the separating hyperplane (SVM), or consensus classification on the basis of majority voting or regression scores (RF).[10] Compared with other machine learning approaches, a characteristic of SVMs is that they operate in feature spaces of increasing dimensionality. On the other hand, decision-tree-based approaches such as RFs are set apart from many other supervised learning techniques because they do not have black box character. Rather, feature pathways and rules are derived for individual decision trees that are often interpretable in chemical terms.

In ligand-based virtual screening, these machine learning approaches are generally applied to search for novel active compounds on the basis of training sets of known active ligands. In the present work, we have evaluated the performance of machine learning approaches on a more complicated prediction task, namely, the search for specifically active ligands of G protein-coupled receptors (GPCRs) with different mechanisms of action. In this case, a set of ligands is specifically active against a GPCR but consists of subsets of compounds that act by different mechanisms, including agonists, partial agonists, antagonists, and inverse agonists.[11] Agonists bind to the ligand binding site of a receptor and activate it, while partial agonists do not elicit a full functional response. By contrast, antagonists block the ligand binding site of a receptor and prevent signaling. Moreover, inverse agonists stabilize a nonproductive conformation of a receptor and thereby prevent effective ligand binding and ensuing functional effects. Often such specifically active receptor ligands are structurally similar despite their different mechanisms of action.[11,12] In addition, especially antagonistic and inverse-agonistic effects are often difficult to differentiate.[12] Taken together, these rather complex structure/mechanism relationships render the prediction of GPCR ligands with different

**Received:** June 21, 2013
**Published:** August 18, 2013

mechanisms of action an a priori challenging task that has to our knowledge not yet been addressed by machine-learning-based virtual screening.

In addition to the three major supervised machine learning approaches discussed above, we intended to include in the comparison a conceptually different classification method that belongs to a class of mapping algorithms.[13−15] This method, termed dynamic mapping of activity-class-specific descriptor value ranges (DynaMAD),[15] is designed to combine a supervised learning step with a compound mapping operation. First, on the basis of a set of reference compounds, descriptors are scored according to their tendency to adopt activity-class-specific value ranges in comparison with a screening database.[14] Second, preselected descriptors are assigned to different dimension extension layers on the basis of their scores. Third, test compounds are iteratively mapped to prespecified descriptor value ranges. During the mapping process, the dimensionality of the descriptor reference space increases, and compounds that no longer map to all of the class-specific value ranges are omitted.[15] Thus, although they are methodologically completely distinct, SVMs and DynaMAD have in common the characteristic that they operate in feature spaces of increasing dimensionality. In constrast to SVM results, the results obtained by DynaMAD might often be more interpretable, similar to those of RFs, because DynaMAD calculations are also based on explicitly defined sets of descriptors with characteristic value ranges. Taken together, these aspects made it interesting to include DynaMAD in the search for compounds with different mechanisms of action. In contrast to the other machine learning approaches, mapping algorithms have to date been explored only a little in virtual screening.

In the following, we report the results of extensive machine-learning-based search calculations on different sets of receptor ligands containing multiple mechanistic subsets. We also compare the results with those for nearest-neighbor-based fingerprint similarity searching.

## ■ MATERIALS AND METHODS

Further methodological information, compound data set details, and calculation protocols are provided in the following subsections.

**Methodological Information.** *Naïve Bayesian Classifiers.* Naïve Bayesian classifiers utilize Bayes' theorem to predict the *posterior probability* $P(c|x)$ of a feature vector $x$ to belong to class $c$:

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

Here, $P(c)$ is the *prior probability* of class $c$, $P(x|c)$ is the *likelihood* of class $c$ given the *evidence* $x$, and $P(x)$ is the probability of that evidence (i.e., the marginal probability of example $x$).[15,16] The classifier is termed *naïve* because it assumes conditional feature (descriptor) independence, which usually is an approximation in practice.

The prior probability for each class is estimated from the training data as the fraction of examples belonging to that class, and the probability of the evidence is estimated as the fraction of examples having value $x$ for descriptor $d$. Under the assumption of feature independence, the class likelihood is then given as

$$P(x|c) = \prod_{j=1}^{d} P(x_j|c)$$

In our calculations, we used the freely available R package *klaR*.[17,18] Descriptors with a variance of zero for the positive or negative training data were removed.

To account for potential data imbalance and the principally unknown prior probabilities of activity in chemical data, uniform prior probabilities were applied and the test compounds were ranked according to their ratios of posterior probabilities, $R(x)$:

$$P(c) = \frac{1}{2} \quad \text{for } c \in [-1,\ +1]$$

$$R(x) = \frac{P(c = +1|x)}{P(c = -1|x)}$$

*Support Vector Machines.* The fundamental SVM task of constructing an optimally separating hyperplane in feature space for data having different class labels[7] can be rationalized as the optimization of values $w$ and $b$ such that

$$y_i(wx_i - b) \geq 1 \quad \text{for all } x_i \in X$$

where $X$ denotes the set of observations, $y_i = +1$ for all positive training examples, and $y_i = -1$ for all negative training examples. The optimal hyperplane is the one with maximal margins, (i.e., distances to the nearest training examples, representing support vectors). It is obtained by minimizing

$$\phi(w) = \frac{1}{2}w^2$$

subject to the condition $y_i(wx_i - b) \geq 1$. The learned hyperplane is then used to classify each test data according to the side of the hyperplane on which it falls. Furthermore, test instances can be ranked by their signed distances from the hyperplane, that is, they are ordered by the value of $wx_i - b$. For example, for activity prediction, compounds would be prioritized with increasing distance from the hyperplane on the positive ("active") side and deprioritized with increasing distance from the plane on the negative side.

If data points with different class labels (e.g., active vs inactive compounds) are not completely linearly separable, *a soft-margin separating hyperplane* can be derived by introducing *slack variables* $\xi_i \geq 0$ and minimizing

$$\phi(w, \xi) = \frac{1}{2}w^2 + C_{+1}\sum_{y_i=+1}\xi_i + C_{-1}\sum_{y_i=-1}\xi_i$$

subject to the condition

$$y_i(wx_i - b) \geq 1 - \xi_i$$

where $C_{+1}$ and $C_{-1}$ balance the cost between false positives and false negatives.[19]

Furthermore, the dot product $wx_i$ can be replaced by a *kernel function* $K(u,v)$ that implicitly calculates the product of two vectors in higher-dimensional space, hence circumventing the need to apply an explicit mapping function.[20,21] Two examples of popular kernel functions are the Gaussian kernel and the Tanimoto kernel. The Gaussian kernel is defined by

$$K_{\text{Gaussian}}(u, v) = \exp\left(-\frac{\|u - v\|^2}{2\sigma^2}\right)$$

which is often equivalently written as

$$K_{\text{Gaussian}}(u, v) = \exp(-\gamma\|u - v\|^2)$$

where the parameter $\gamma = 1/2\sigma^2$ determines the shape of the kernel, that is, the extent to which neighboring compounds are taken into consideration. The Tanimoto kernel is defined by

$$K_{\text{Tanimoto}}(u, v) = \frac{\langle u, v\rangle}{\langle u, u\rangle + \langle v, v\rangle - \langle u, v\rangle}$$

In our calculations, we used the freely available SVM$^{\text{light}}$ implementation.[22] Compound data were normalized to the range $[-1, +1]$, and $C_{+1}$ and $C_{-1}$ were adjusted to reflect the ratio

$$\frac{C_{+1}}{C_{-1}} = \frac{\text{no. of inactive training compounds}}{\text{no. of active training compounds}}$$

We also compared the linear kernel and the Gaussian kernel with $\gamma = 1$ and standard SVM$^{\text{light}}$ parameter settings.

*Random Forests.* RFs combine several independently derived decision trees that are utilized for consensus scoring to predict class labels of test compounds.[9] Because the generalization error of an RF model typically depends on the strength of its individual trees and the correlation between them, predictive performance can usually be optimized by minimizing the correlation between well-performing trees.[9] We used the R implementation *randomForest* with standard parameters for our calculations.[23] In each case, 500 individual trees were grown on all $n$ training instances and $\sqrt{d}$ variables were randomly sampled from $d$ descriptors at each decision point. The best-performing variables were assigned to each node without tree pruning. Test compounds were ranked on the basis of the fractions of their trees predicting a positive class label.

Since a constant number of trees can predict only a discrete range of probabilities, the rank positions of test compounds with the same fraction of positive tree votes could not be distinguished. In such cases, compounds were randomly ordered. However, our calculations revealed that most of the compounds with the same fraction of positive tree votes were ranked low, with less than 20% of trees predicting positive labels. Hence, random ordering predominantly affected the low-ranked compounds and did not affect the ordering of highly ranked compounds. For selection sets of 100 compounds, there were never more than five compounds having the same activity voting. Hence, the potential effects of partly random ordering could essentially be neglected.

*DynaMAD.* DynaMAD is conceptually different from the other machine learning methods investigated herein. The core of the DynaMAD approach is a mapping procedure in descriptor spaces of stepwise-increasing dimensionality,[15] which represents an unsupervised technique. As a supervised learning component of DynaMAD, descriptors with compound class-specific value ranges are preselected from a large pool of features.[14,15] Given $n$ positively labeled (active) compounds and $m$ unlabeled (database) compounds in a $d$-dimensional descriptor space, each descriptor (dimension) is assigned an individual score $s(d_i)$. This score reflects the fraction of unlabeled data that fall into the value range of positive data in the one-dimensional space $d_i$:

$$s(d_i) = 100(1 - p_i)$$

where

$$p_i = P(\text{classMin}_i \le x_i \le \text{classMax}_i) = \sum_{\text{classMin}_i \le x \le \text{classMax}_i} P(x_i = x)$$

is the probability that descriptor value $x_i$ of an unlabeled instance falls into the range $[\text{classMin}_i, \text{classMax}_i]$, which is bounded by the minimal and maximal values of the positive instances. Therefore, $s(d_i)$ decreases as the number of unlabeled instances falling into the value range of positive instances in $d_i$ increases.

If only few positive training instances are available, the descriptor value ranges $[\text{classMin}_i, \text{classMax}_i]$ can be extended by a value interval $\Delta p_i$ depending on the mapping probability $p_i$. Expanded value ranges are calculated as

$$\left[\text{classMin}_i - \frac{\Delta p_i}{2}, \text{classMax}_i + \frac{\Delta p_i}{2}\right]$$

and the scores are adjusted accordingly:

$$s(d_i)' = s(d_i) - 100\Delta p_i$$

To calculate $\Delta p_i$, we used the default function of DynaMAD:[15]

$$\Delta p_i = \frac{1 - p_i}{100p_i + 10}$$

To classify instances as active or inactive, descriptors are assigned to equally spaced *dimension extension layers* on the basis of their class-specific scores. Thereby, a feature space of increasing dimensionality is obtained for iterative mapping. During this process, test compounds are retained if their descriptor values fall into all prespecified value ranges; otherwise, they are discarded. Hence, the number of mapped test instances decreases from layer to layer. Figure 1 schematically illustrates the DynaMAD approach.

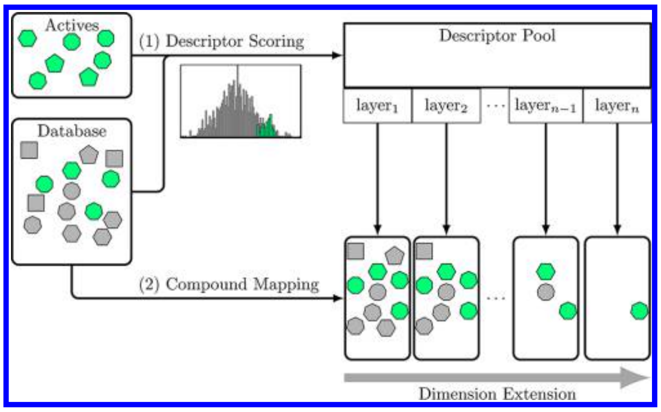


**Figure 1.** Schematic description of the DynaMAD approach. Initially, descriptors are scored on the basis of the reference compounds' value ranges and assigned to different dimension extension layers. Next, compounds are iteratively mapped to subsequent layers (forming chemical reference spaces of increasing dimensionality) for a prespecified number of iterations or until only a few selected compounds remain.

*Control Calculations.* To put the results of the investigated machine learning approaches into perspective, we also carried out a number of control calculations. Similarity searching was performed using the MACCS[24] and ECFP4[25] fingerprints [calculated with the Molecular Operating Environment (MOE)][26] and the $k$-nearest neighbor (kNN) search strategy. In this approach, test compounds are compared to a prespecified number of reference compounds using a given similarity measure and assigned the average similarity to the $k$ most similar references as a final similarity score. In systematic similarity search calculations, we calculated the Tanimoto similarity for the

$k = 1$ and $k = 5$ settings. Furthermore, we carried out SVM control calculations using the Tanimoto kernel and the ECFP4 fingerprint as a molecular representation.

**Compound Data Sets.** Five different sets of GPCR ligands were used for our analysis. They consisted of compounds with different mechanisms of action and that were originally assembled from ChEMBL[27] for the design of molecular-network-based activity landscape representations.[28] The composition and curation of the data sets has been described in detail.[28] These data sets are freely available via the "Downloads" section of http://www.lifescienceinformatics.uni-bonn.de.

The compositions of these data sets, which consist of 148 to 307 ligands, are summarized in Table 1. Three data sets contained four mechanistic subgroups (agonists, partial agonists, antagonists, and inverse agonists), while the two remaining ones consisted of three (all except inverse agonists). Figure 2 shows exemplary compounds from these sets. Within each set, compounds with different mechanisms of action were often structurally similar, which was expected to complicate subgroup predictions.

**Table 1. Data Sets**[a]

| target receptor | mechanism | no. of ligands |
|---|---|---|
| adenosine A1 receptor (AA1) | agonist | 107 |
| | partial agonist | 54 |
| | antagonist | 94 |
| | inverse agonist | 52 |
| muscarinic acetylcholine receptor M1 (AM1) | agonist | 26 |
| | partial agonist | 49 |
| | antagonist | 73 |
| dopamine D2 receptor (DD2) | agonist | 40 |
| | partial agonist | 44 |
| | antagonist | 76 |
| | inverse agonist | 13 |
| histamine H3 receptor (H3R) | agonist | 44 |
| | partial agonist | 46 |
| | antagonist | 92 |
| | inverse agonist | 31 |
| serotonin 1A receptor (S1A) | agonist | 46 |
| | partial agonist | 78 |
| | antagonist | 63 |

[a] The compositions of the five different sets of GPCR ligands and their mechanism-based subsets are reported. Receptor abbreviations in parentheses are used throughout the text. For AM1 and S1A, no inverse agonists were available.

**Calculation Protocols.** For all compounds, 185 numerical descriptors that are available in MOE version 2011.10[26] and can be calculated from molecular graphs were used. This set contains many topological and 2D pharmacophore pattern descriptors. This choice of numerical descriptors was motivated by the requirement of DynaMAD to use such descriptors in order to identify and map class-specific descriptor value ranges and by similar requirements of RFs to establish meaningful decision-tree structures. For DynaMAD, fingerprint descriptors cannot be used, and our descriptor set selection ensured feature consistency across the different methods that were compared. We also deliberately omitted from our analysis 3D descriptors calculated from hypothetical compound conformations, given the uncertainties associated with their use.

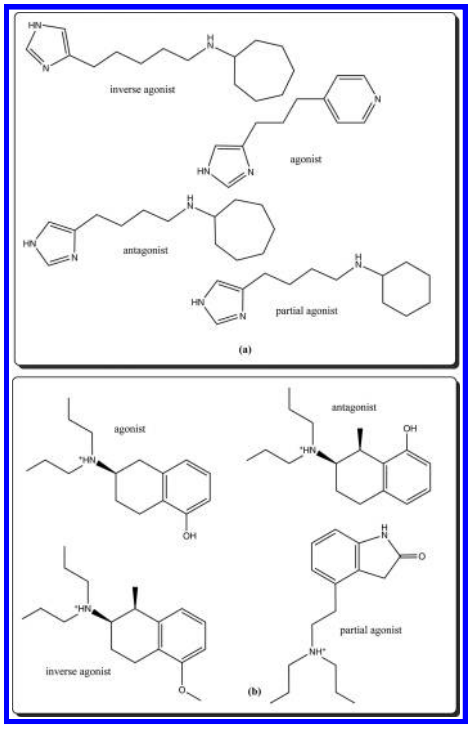


Figure 2. Compounds with different mechanisms of action. Shown are exemplary compounds from the (a) H3R and (b) DD2 data sets. Many compounds with different mechanisms are structurally rather similar. Structural modifications leading to mechanistic changes within these data sets are often only subtle, which renders mechanism-based virtual screening a challenging task.

From each mechanistic subset of the GPCR data sets in Table 1, 100 different reference sets of 12 or 40 compounds with the same mechanism of action were randomly selected. All of the remaining compounds from the data set (including all mechanistic subsets) were added to a background database of 1 000 000 druglike compounds randomly selected from ChEMBL release 14. It should be noted that ChEMBL does not contain confirmed inactive compounds for individual targets (only weakly active ones). This random selection was simply done to obtain a background database of a well-defined large size (fewer or more compounds could also have been used). The druglike character of the background database compounds was important in order to ensure that separation of mechanism-of-action sets from screening database compounds was not artificially favored, for example, through the use of other small organic background molecules such as reagents, as further rationalized below.

For the derivation of each individual classifier, 1000 compounds were randomly chosen from the background database as inactive training examples. Active and inactive training compounds were not included in the test set. Accordingly, there was no bias of performance measures with respect to methods that can "memorize" training data. Subclasses for which sufficient numbers of training and test compounds were not available were omitted from the calculations.

For Bayesian classifiers, SVMs, kNN calculations, and RFs, compound rankings were generated and predictions averaged over 100 independent trials. For DynaMAD calculations, which do not produce compound rankings, mapping statistics were calculated.

**Performance Criteria.** For the evaluation of search performance, the following measures were applied:

$$\text{sensitivity (recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP denotes true positives, FN false negatives, TN true negatives, and FP false positives. Unless specified otherwise, only active ligands with the desired mechanism of action were considered as true positives, while active ligands with other mechanisms were considered false positives in search trials focusing on a specific mechanism.

## ■ RESULTS AND DISCUSSION

**Study Concept.** We have been interested in exploring the ability of supervised machine learning approaches and the DynaMAD algorithm to search for specifically active receptor ligands with different mechanisms of action. Each compound data set consisted of known ligands of a particular GPCR belonging to different mechanistic subsets. Each of these subsets was randomly split into training and test data 100 times, and individual classification models were derived for each reference set in order to obtain statistically sound search results. For DynaMAD, each reference set was used to determine qualifying descriptors with set-specific value ranges that were used for compound mapping. We then carried out systematic virtual screening trials by searching for receptor ligands with a specific mechanism of action that were added as potential hits together with all of the other ligands of the same receptor with different mechanisms to a large database of 1 million ChEMBL compounds. The search/classification task was twofold: compounds with the desired mechanism of action had to be distinguished from structurally similar ligands of the same receptor with different mechanisms (as illustrated in Figure 2) and from other database compounds. In this context, it is important to note that ChEMBL compounds were deliberately chosen as the background database, rather than other small organic molecules, as mentioned above. Druglike compounds are generally more difficult to distinguish from other actives (e.g., the mechanistic subsets of GPCR ligands studied here) than are organic compounds that do not belong to hit or lead optimization series and thus have lower chemical complexity.[3,29] In the following, the results of our simulated virtual screening trials are presented and discussed in detail.

**DynaMAD Search Characteristics.** We first describe the results of a typical DynaMAD calculation to illustrate characteristic features of the mapping procedure. Table 2 reports the average mapping statistics of 100 DynaMAD calculations searching for AM1 partial agonists. Compound mapping to only a few descriptor dimension extension layers resulted in a dramatic reduction in the number of database compounds, for example, from more than 1 million compounds (including the remainder of the AM1 data set) at layer 1 to fewer than 9000 compounds at layer 7. At this stage, 18 of 37 partial agonists (desired hits) remained in the selection set together with 18 of the 99 other AM1 ligands. Over the next few layers, a further steady reduction in the number of compounds was observed. At dimension extension layer 13, only 85 database compounds remained, including eight desired hits and four other AM1 ligands. Only 12 database compounds successfully mapped the final dimension extension layer (layer 20), and this very small selection set contained four desired hits and one other AM1 ligand. Similar mapping statistics were observed in most of the DynaMAD trials. The data revealed the generally high discriminatory power of the underlying descriptor spaces with increasing dimensionality. They also indicated that there was a clear enrichment of compounds with the desired mechanism of action over related ligands and that other database compounds were successfully discriminated on a large scale. For example, at dimension extension layer 12 in Table 2, a database selection set of 155 compounds was obtained, which represents a reasonably sized set for many practical virtual screening applications. This set contained nine of the 37 desired partial agonists (~24%) and five of the 99 other AM1 ligands (~5%). However, this example also illustrates that the final recall rate of desired hits was relatively low. Hence, there was a price to pay for the overall high stringency of the calculations, namely, limited sensitivity to desired hits in high-dimensional descriptor spaces. The overall specificity and recall performance of DynaMAD calculations is further discussed below.

**Table 2. DynaMAD Mapping Statistics**[a]

| layer | size | part. agonists | others |
|---|---|---|---|
| 1 | 1000136 | 37 | 99 |
| 2 | 420917 | 34 | 92 |
| 3 | 382699 | 32 | 78 |
| 4 | 199320 | 30 | 63 |
| 5 | 127341 | 27 | 51 |
| 6 | 19189 | 22 | 29 |
| 7 | 8467 | 18 | 18 |
| 8 | 7433 | 16 | 15 |
| 9 | 1821 | 14 | 10 |
| 10 | 505 | 13 | 8 |
| 11 | 328 | 11 | 7 |
| 12 | 155 | 9 | 5 |
| 13 | 85 | 8 | 4 |
| 14 | 53 | 7 | 3 |
| 15 | 32 | 6 | 3 |
| 16 | 19 | 6 | 2 |
| 17 | 16 | 6 | 2 |
| 18 | 14 | 5 | 2 |
| 19 | 13 | 5 | 2 |
| 20 | 12 | 4 | 1 |

[a] Compound mapping statistics are reported for 100 independent DynaMAD trials searching for AM1 partial agonists. Given are the average numbers of qualifying database compounds that correctly mapped to the descriptor value ranges of each dimension extension layer and the average numbers of targeted partial agonists (part. agonists) and active compounds with other mechanisms (others) contained in these database selection sets.

**Reference Sets for Machine Learning.** Only limited numbers of compounds were available in several mechanistic subsets, as reported in Table 1. Consequently, only relatively small numbers of positive training examples could be utilized. This was a point of concern for machine learning, which typically benefits from the availability of large sets of positive training examples. To enable search calculations for the majority of data sets, we consistently set the number of reference compounds with the desired mechanism for the derivation of classifiers to 12. However, whenever possible, we also built classifiers using 40 positive training examples. As negative training examples, 1000

Table 3. Recall Rates for Different-Sized Training Sets[a]

| | | $SVM^{linear}$ | | $SVM^{Gaussian}$ | | RF | | naïve Bayes | |
|---|---|---|---|---|---|---|---|---|---|
| receptor | mechanism | 12 | 40 | 12 | 40 | 12 | 40 | 12 | 40 |
| AA1 | agonist | 0.00 | 0.00 | 0.15 | 0.18 | 0.43 | 0.64 | 0.03 | 0.00 |
| | part. agonist | 0.00 | 0.00 | 0.31 | 0.43 | 0.55 | 0.86 | 0.00 | 0.00 |
| | antagonist | 0.01 | 0.04 | 0.14 | 0.15 | 0.33 | 0.56 | 0.00 | 0.00 |
| | inv. agonist | 0.00 | 0.00 | 0.30 | 0.25 | 0.56 | 0.83 | 0.00 | 0.00 |
| AM1 | part. agonist | 0.22 | 0.22 | 0.51 | 0.78 | 0.73 | 0.89 | 0.00 | 0.00 |
| | antagonist | 0.14 | 0.27 | 0.66 | 0.76 | 0.69 | 0.85 | 0.53 | 0.00 |
| DD2 | part. agonist | 0.03 | 0.00 | 0.75 | 1.00 | 0.75 | 1.00 | 0.72 | 0.75 |
| | antagonist | 0.35 | 0.11 | 0.70 | 0.86 | 0.69 | 0.89 | 0.65 | 0.86 |
| H3R | agonist | 0.12 | 0.25 | 0.47 | 0.75 | 0.58 | 0.75 | 0.22 | 0.00 |
| | part. agonist | 0.00 | 0.00 | 0.41 | 0.50 | 0.53 | 0.67 | 0.38 | 0.17 |
| | antagonist | 0.00 | 0.00 | 0.45 | 0.52 | 0.42 | 0.60 | 0.00 | 0.00 |
| S1A | agonist | 0.26 | 0.33 | 0.79 | 0.83 | 0.71 | 0.83 | 0.74 | 0.83 |
| | part. agonist | 0.37 | 0.55 | 0.68 | 0.82 | 0.70 | 0.87 | 0.53 | 0.25 |
| | antagonist | 0.34 | 0.23 | 0.74 | 0.91 | 0.78 | 0.95 | 0.68 | 0.86 |

[a] Median recall rates are reported for the top-ranked 100 database compounds and training sets with 12 and 40 reference compounds for SVMs, RFs, and Bayesian classifiers. The database compounds did not include training examples.

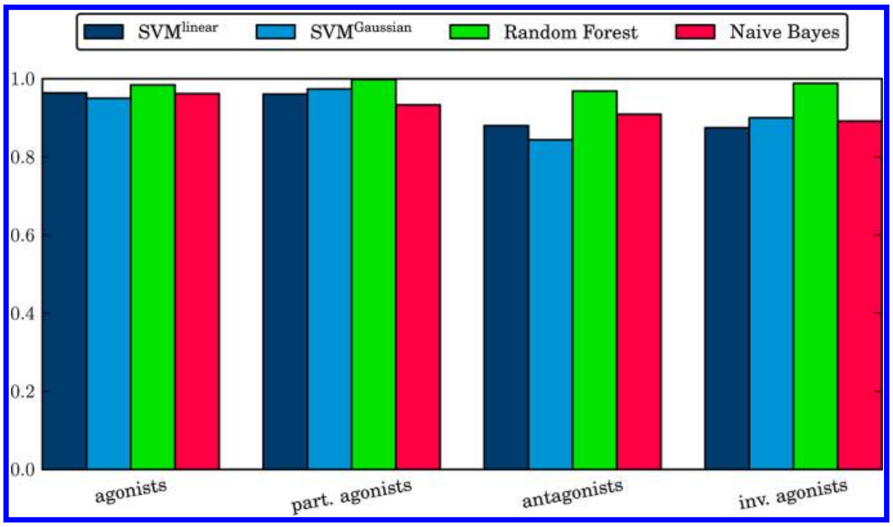


**Figure 3.** Recall performance. For SVMs, RFs, and naïve Bayesian classifiers, AUROC values for the different mechanistic subsets of the AA1 data set are reported. AUROC values cannot be calculated for DynaMAD mapping, which does not produce database rankings. All three classifiers achieved high values. The results are representative of those for the other data sets.

randomly selected ChEMBL compounds were used in each case, and this data imbalance was taken into account by appropriate parameter settings of the methods (cost factor for SVMs; uniform prior probabilities and ranking based on the ratio of posterior probabilities for naïve Bayesian classifiers).

Table 3 provides a comparison of the search results for models derived from the different-sized training sets. Median recall rates for the top-ranked 100 compounds for different mechanistic subsets are reported. In part unexpected observations were made. Although the database selection sets of 100 compounds were rather small, the recall rates greatly varied from 0% (no hits among the first 100 database compounds) to 100% (complete recall), depending on the method and mechanistic subset. For 12 positive training examples, high recall rates above 50% were often observed for RFs, Bayesian calculations, and Gaussian SVMs on three data sets, but only RFs were able to produce high recall rates on all sets. Linear SVMs failed to produce high recall rates. Moreover, while the performance of RFs and Gaussian SVMs consistently increased with a larger number of positive training examples, this was not the case for Bayesian calculations and linear SVMs. For naïve Bayesian classifiers, an increase in median recall was observed for four sets and a decrease for five others, including some substantial decreases (e.g., from 53% to 0% for AM1 antagonists). Most of the variations in the median recall rates for differently sized reference sets were observed in cases where individual trials produced a large range of values. In this context, it should be mentioned that the recall spread of individual trials was usually smaller for RFs and Gaussian SVMs than the other approaches, indicating that Bayesian methods and linear SVMs were more sensitive to changes in training set composition.

Taken together, the results showed that smaller numbers of training examples were not a globally limiting factor in our analysis. In addition, we observed a strong compound set dependence of the method performance. For example, for 12 reference compounds, RF calculations produced recall rated of 32−55% for AA1 ligands, which further increased to 56−86% for 40 reference compounds, whereas Gaussian SVMs produced recall rates of 14−31% for 12 reference compounds and 15−43% for 40 reference compounds. Both linear SVMs and naïve

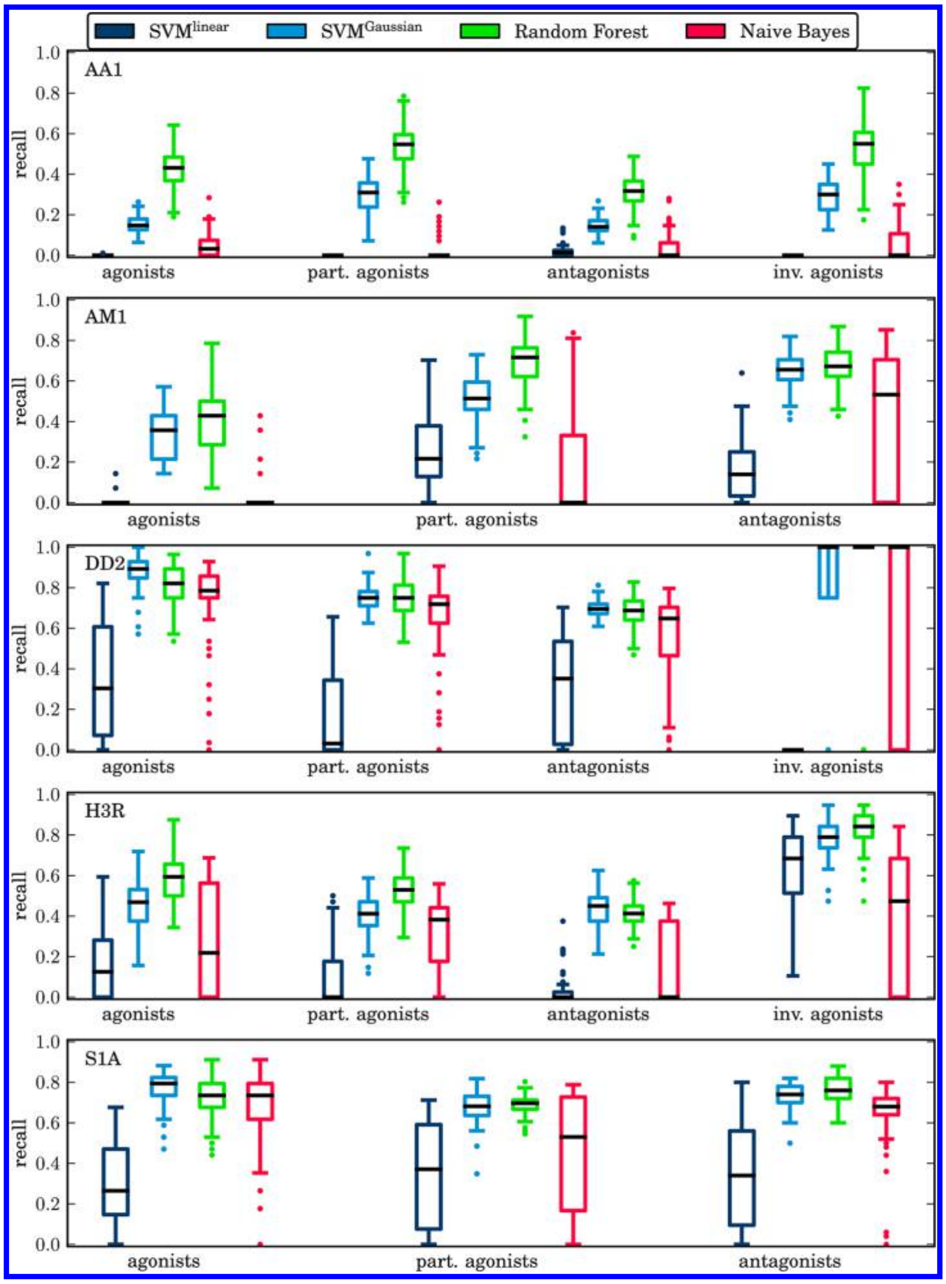


**Figure 4.** Recall rates for machine learning methods on small database selection sets. The distributions of mechanism-specific compound recall rates over individual search trials are reported for small database selection sets of the top-ranked 100 compounds and compared for SVMs, RFs, and naïve Bayesian classifiers. The box plots give the minimal, maximal, quartile, and median values over all 100 individual search calculations.

Bayesian classifiers essentially failed to produce recall for this data set.

**Recall Performance.** For the global assessment of recall performance for ligands with desired mechanisms beyond small database selection sets, we first calculated area under the ROC curve (AUROC) values[30,31] for all trials. Results for the AA1 data set are reported in Figure 3. The results are representative of those for the other data sets. High AUROC values between 0.84 and close to 1 were consistently observed, regardless of the method and data set. Thus, on the basis of global ranking assessments, the calculations were unexpectedly successful. However, as already indicated in Table 3, substantial differences were observed in the early enrichment characteristics of hits, and AUROC values do not account for these differences in rankings of very large imbalanced databases (i.e., containing only small numbers of actives and very large numbers of database compounds). Furthermore, there were clear differences in the dependence of the search calculations on the reference set composition, as already indicated above and revealed in Figure 4. The box plot representations in this figure capture compound recall over all trials for the 100 top-ranked database compounds. Overall, RFs and Gaussian SVMs produced higher recall than linear SVMs and naïve Bayesian classifiers for database selection sets of 100 compounds, as indicated in Table 3 and Figure 4, and RFs performed better than Gaussian SVMs on the AA1 and H3R data set (with the exception of H3R antagonists). Naïve Bayesian classifiers mostly performed better than linear SVMs (with the exception of AM1 partial agonists and H3R inverse agonists). The overall highest recall rates for all methods were observed for the DD2 and S1A data sets.

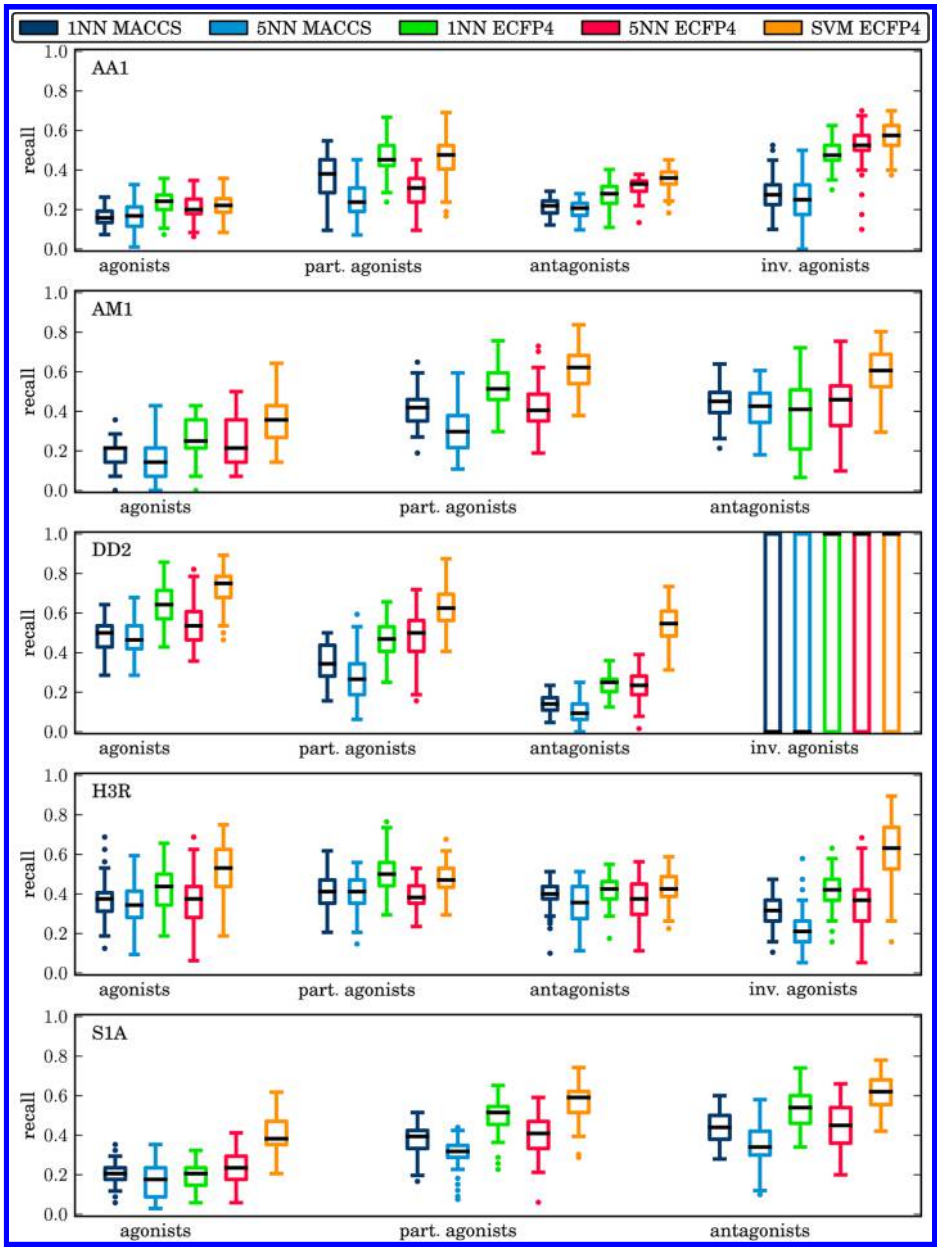


**Figure 5.** Recall rates for fingerprint search calculations. The distributions of mechanism-specific compound recall rates over individual search trials are reported for small database selection sets of the top-ranked 100 compounds and compared with kNN calculations with $k = 1$ and $k = 5$ using the MACCS and ECFP4 fingerprints and SVMs with the Tanimoto kernel using ECFP4 fingerprints. The box plots report the minimal, maximal, quartile, and median values over all 100 individual search calculations.

The recalls for the 100 top-ranked compounds of the fingerprint calculations are reported in Figure 5. For the AA1 data set and the H3R partial agonists and antagonists, the kNN similarity search calculations displayed performance comparable to that of Gaussian SVMs and even reached the performance level of RF in some instances. They performed better than linear but worse than Gaussian SVMs for the AM1 and DD2 data sets and H3R agonists. By contrast, for the DD2 antagonists and H3R inverse agonists, linear SVMs outperformed the kNN approach. For the S1A data set, the search performance of kNN calculations was similar to that of linear SVMs. We also note that there was no clear trend in the fingerprint similarity searches favoring either 1NN or 5NN calculations. While 1NN produced higher recall than 5NN in five cases (AA1 partial agonists, AM1 partial agonists, H3R inverse agonists, and S1A partial agonists and antagonists), no significant differences were observed in any other search calculation. Overall, the more sophisticated atom environment fingerprint ECFP4 achieved higher recall rates than MACCS structural keys in most cases. For AA1 inverse agonists, DD2 partial agonists, DD2 antagonists, and H3R inverse agonists, the median recall of ECFP4 was more than 10% higher than that of MACCS for both the 1NN and 5NN search strategies. In view of its superior search performance, ECFP4 was used for the SVM control calculations.

The SVM control calculations using the Tanimoto kernel consistently performed better than or comparably well as the best-performing kNN calculations. However, depending on the data set, the performance of fingerprint-based SVMs was either comparable to or worse than the performance of machine learning methods using numerical descriptors. For the DD2 and

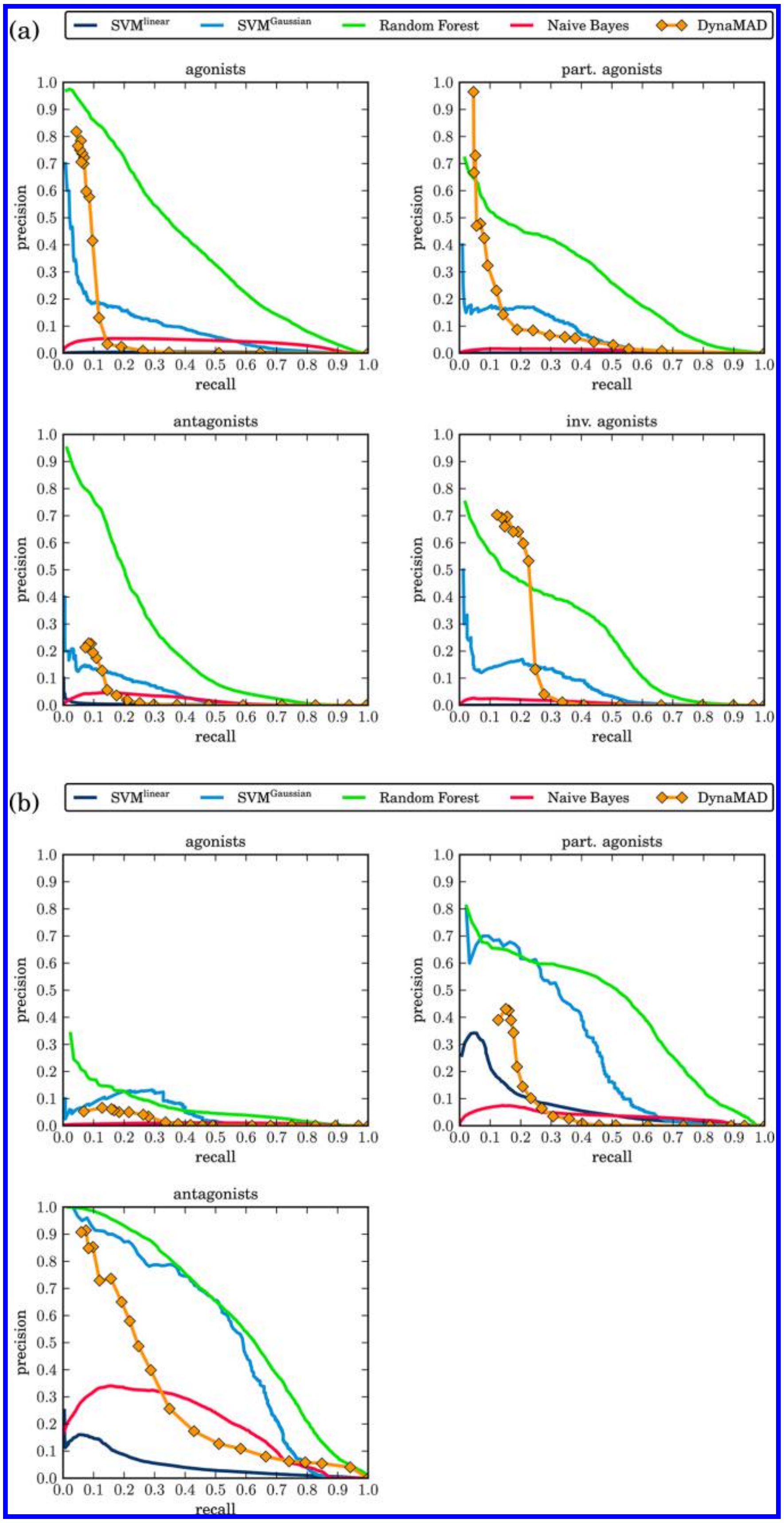


**Figure 6.** continued

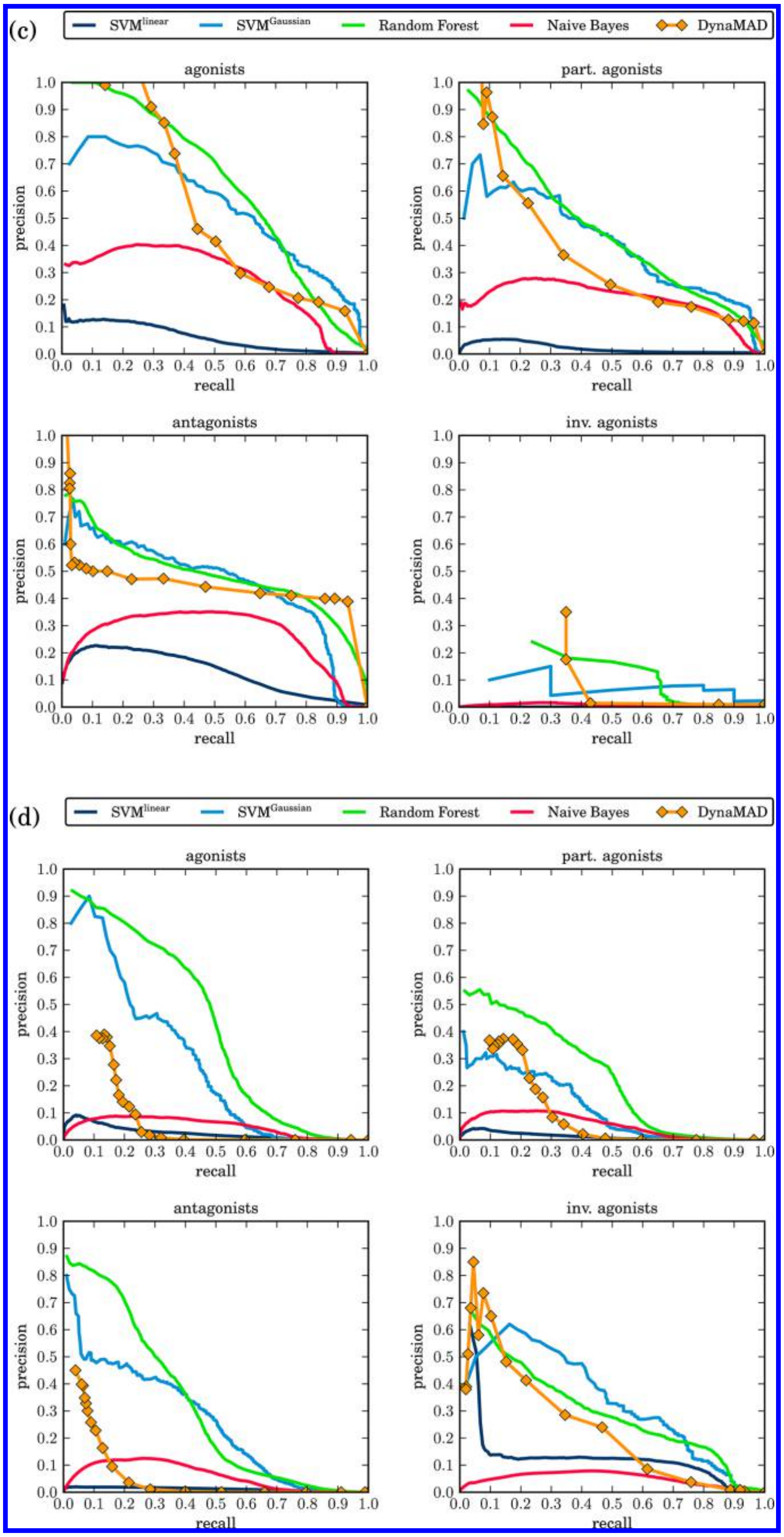


**Figure 6.** continued

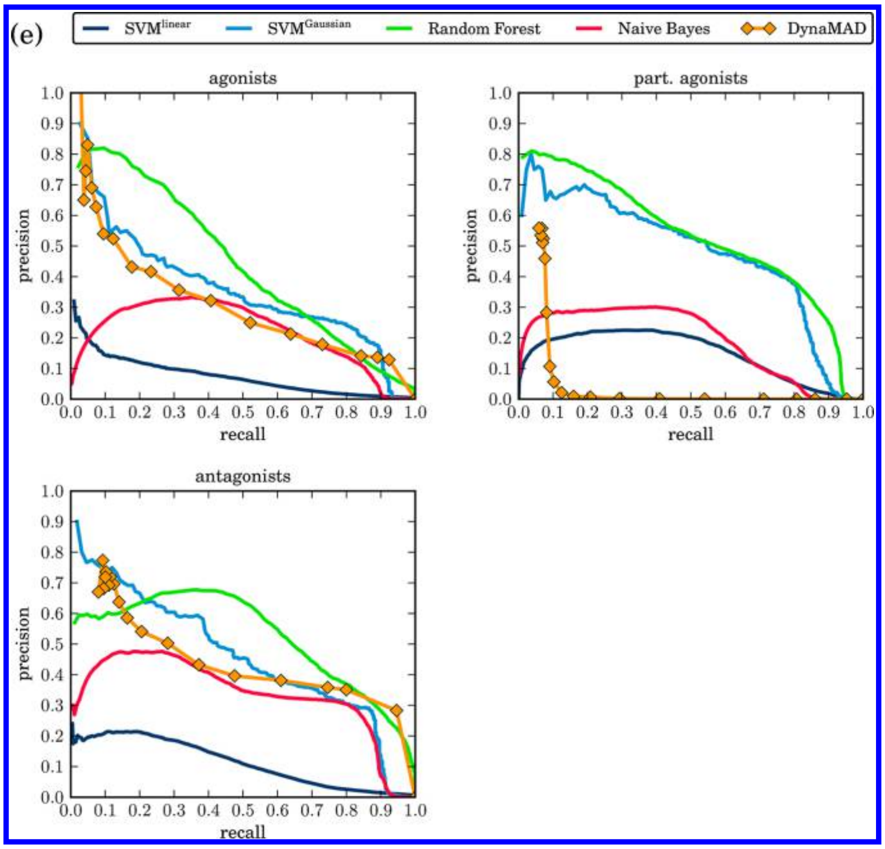


**Figure 6.** Precision−recall curves are shown for all methods and all mechanistic subsets of the (a) AA1, (b) AM1, (c) DD2, (d) H3R, and (e) S1A data sets.

S1A data sets, the fingerprint-based SVMs achieved higher recall than the linear SVM using numerical descriptors but lower recall relative to all other methods. For H3R inverse agonists, the recall of the fingerprint-based SVM was lower than that of the linear SVM. For the other H3R ligands and the AA1 and AM1 data sets, fingerprint-based SVMs reached the recall rates of Gaussian SVMs and even RFs, thereby outperforming linear SVMs and naïve Bayesian classifiers.

**Precision−Recall Comparison.** We then related the observed recall performance to the precision of the calculations, which made it possible to include DynaMAD in the comparison. In Figure 6, precision−recall curves are reported for the search results obtained with all methods and data sets. For the AA1 set (Figure 6a), linear SVMs and naïve Bayesian classifiers yielded only very low precision, while RFs and DynaMAD produced precision values of 50% or higher (with the exception of DynaMAD searches for antagonists, with 25% precision). Gaussian SVMs produced precision values of 70% for agonists and intermediate values of 40−50% for others. Only low precision was observed for AM1 agonists using all methods (Figure 6b). For the other mechanistic subsets, RFs produced the highest precision values, closely followed by Gaussian SVMs and then DynaMAD. Similar trends were observed for DD2 (Figure 6c), H3R (Figure 6d), and S1A (Figure 6e). In these cases, naïve Bayesian classifiers also frequently yielded precision values of 30−40%, while linear SVM precision was mostly lower (i.e., at the 20% level). Hence, RFs overall dominated the recall and precision performance, followed by Gaussian SVMs and DynaMAD, which displayed higher precision than Bayesian classifiers and linear SVMs.

**Specificity Analysis.** Thus far, we have mostly focused on recall and precision of the database search calculations on the basis of global rankings and small database selection sets (to monitor early enrichment characteristics). As discussed above, the receptor ligand data sets under study made it possible to determine the ability of different methods to distinguish between related ligands with different mechanisms of action. In Figures 7 and 8, we report the average cumulative recall curves for compounds belonging to all mechanistic subsets for a given class and search calculations focusing on a particular mechanism of action. The graphs reveal the ability of the search calculations to distinguish between ligands with the desired mechanism of action and other mechanisms and hence provide insights into the intraset specificity of the classifiers. Because DynaMAD does not produce compound rankings, the corresponding graphs report discrete selection sets for different dimension extension layers. The four examples in Figures 7 and 8 represent the different search phenotypes we observed. For example, in the search for AA1 agonists (Figure 7a), RF, Gaussian SVMs, and DynaMAD calculations were essentially specific for small database selections of 10 to 100 compounds. However, for larger selection sets, only Gaussian SVMs, DynaMAD, and naïve Bayesian classifiers exhibited a certain degree of selectivity for agonists over ligands with other mechanisms. Fingerprint calculations (Figure 8a) displayed a similar level of selectivity compared with RFs but yielded lower compound recall for small database selection sets. In the example in Figures 7b and 8b, all of the methods displayed a strong tendency to specifically select desired antagonists over AA1 ligands with other mechanisms. Taken together, the results in Figures 7a,b and 8a,b illustrate the presence of rather different search characteristics for mechanistic subsets within the same

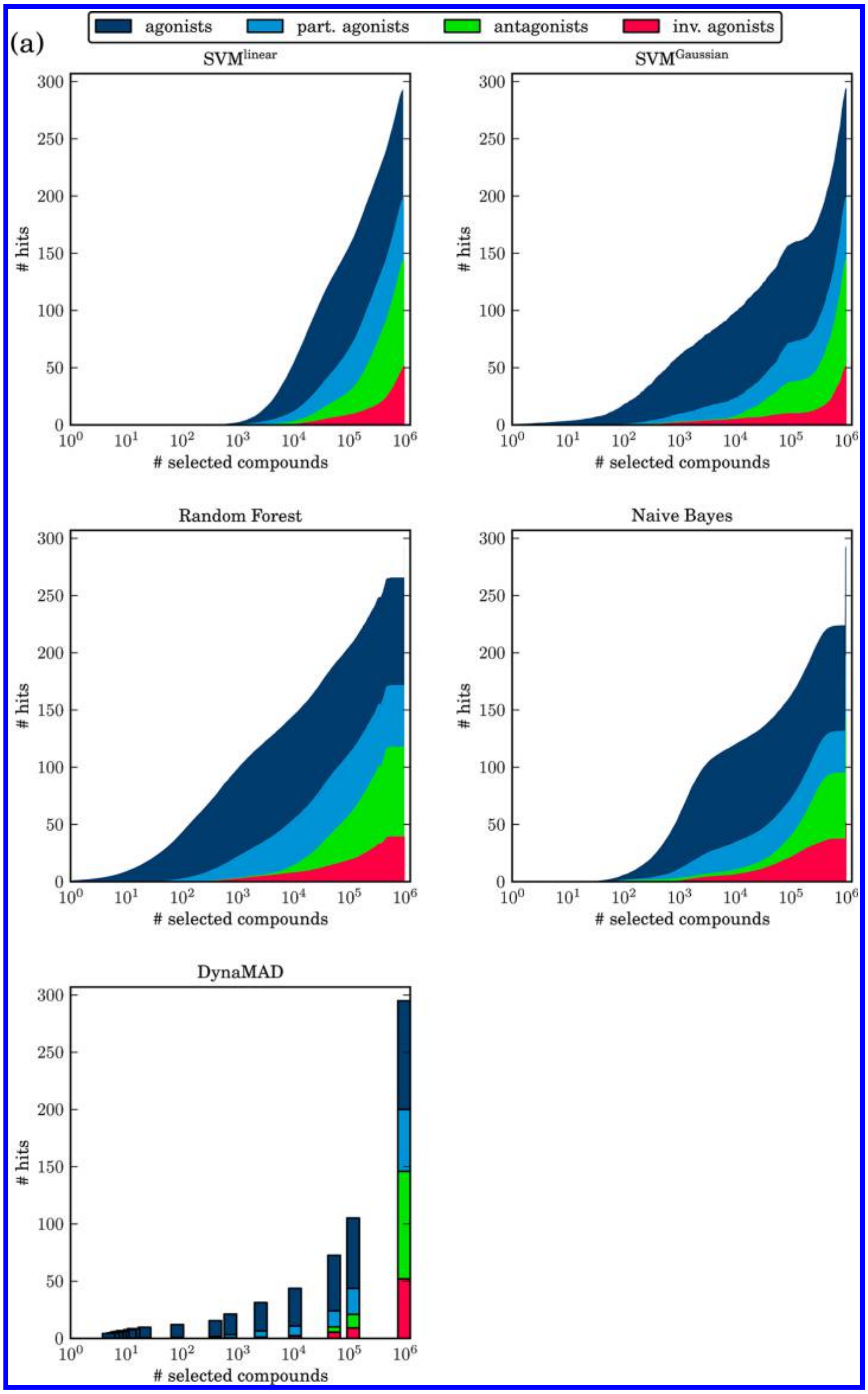


**Figure 7.** continued

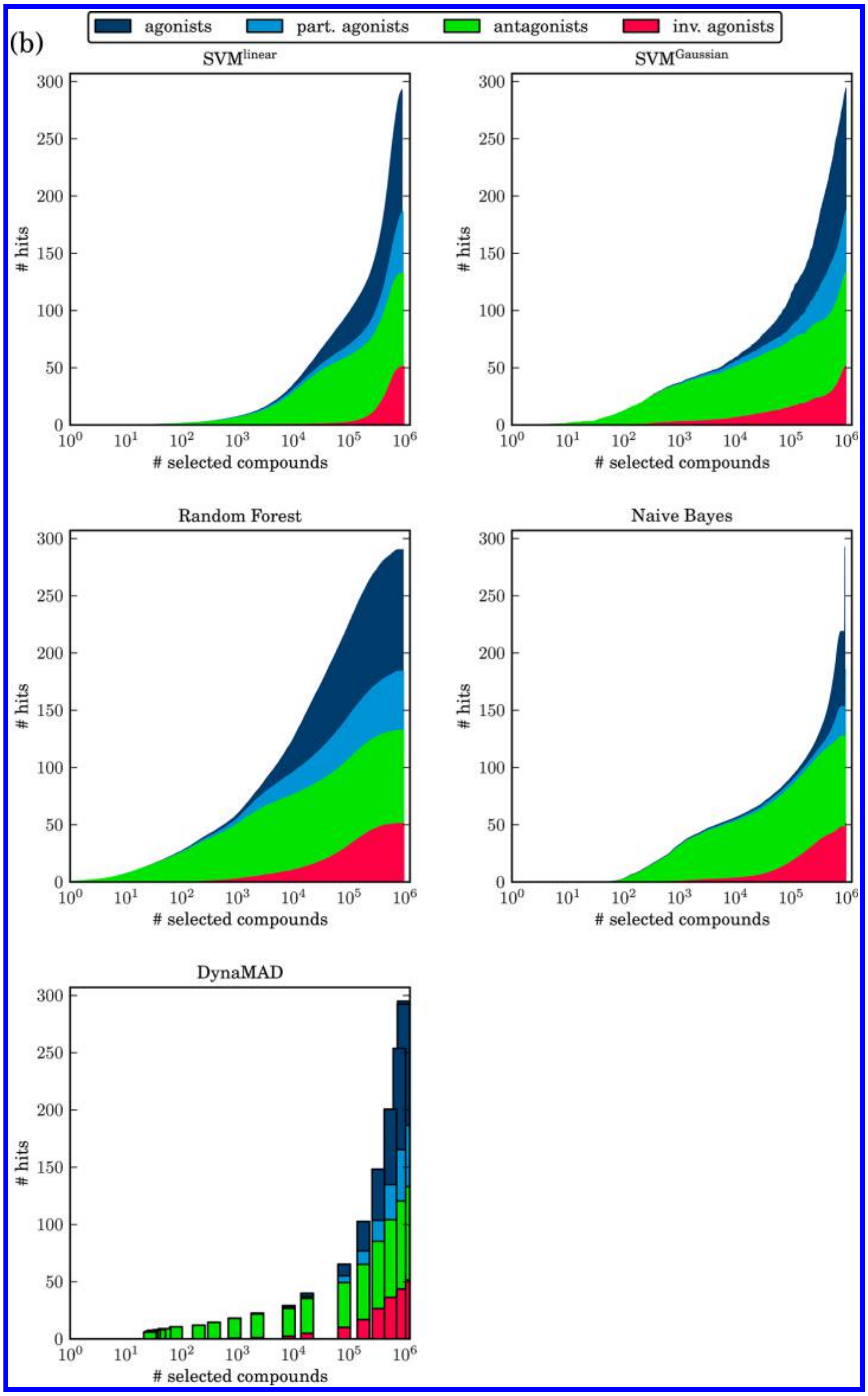


**Figure 7.** continued

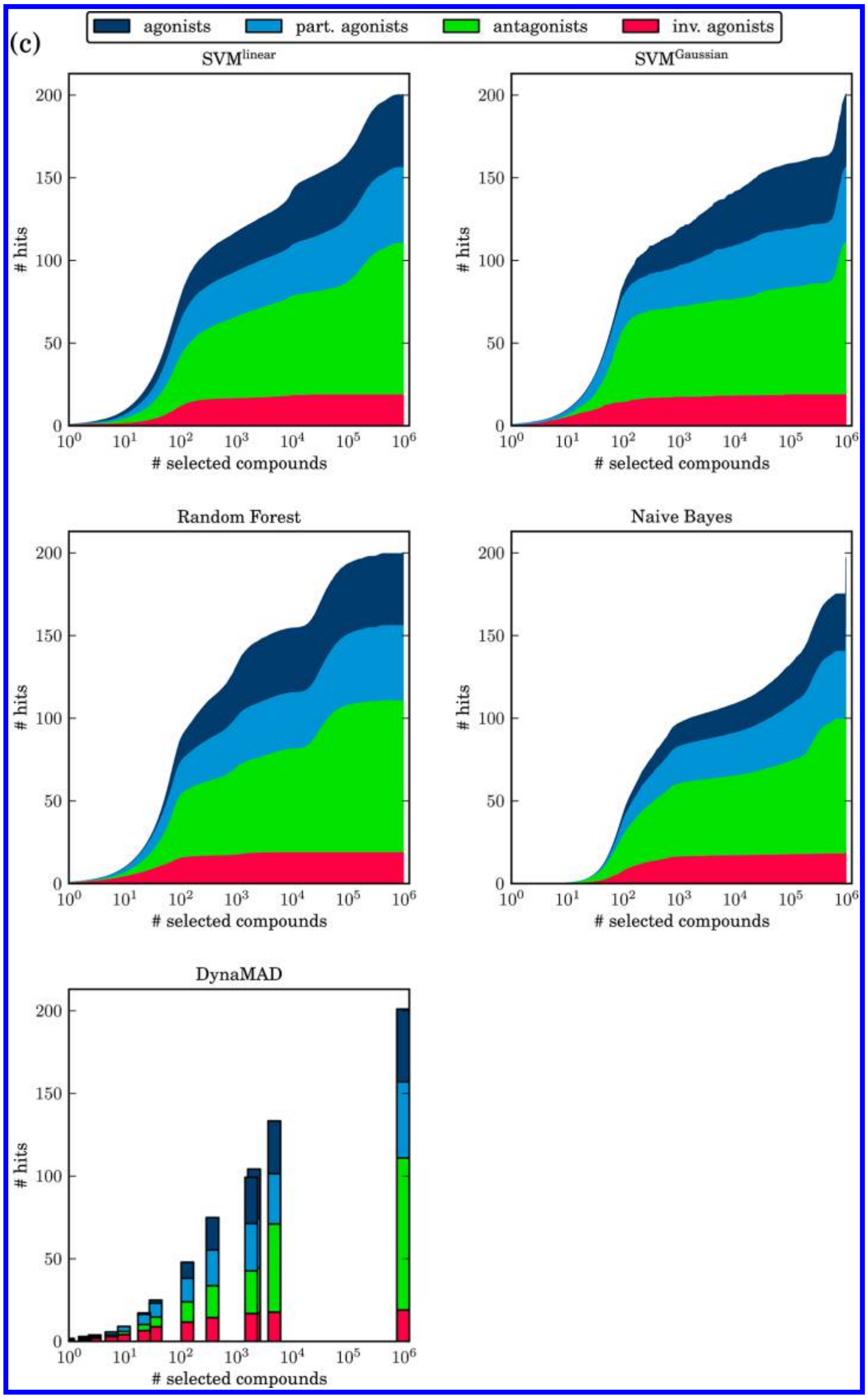


**Figure 7.** continued

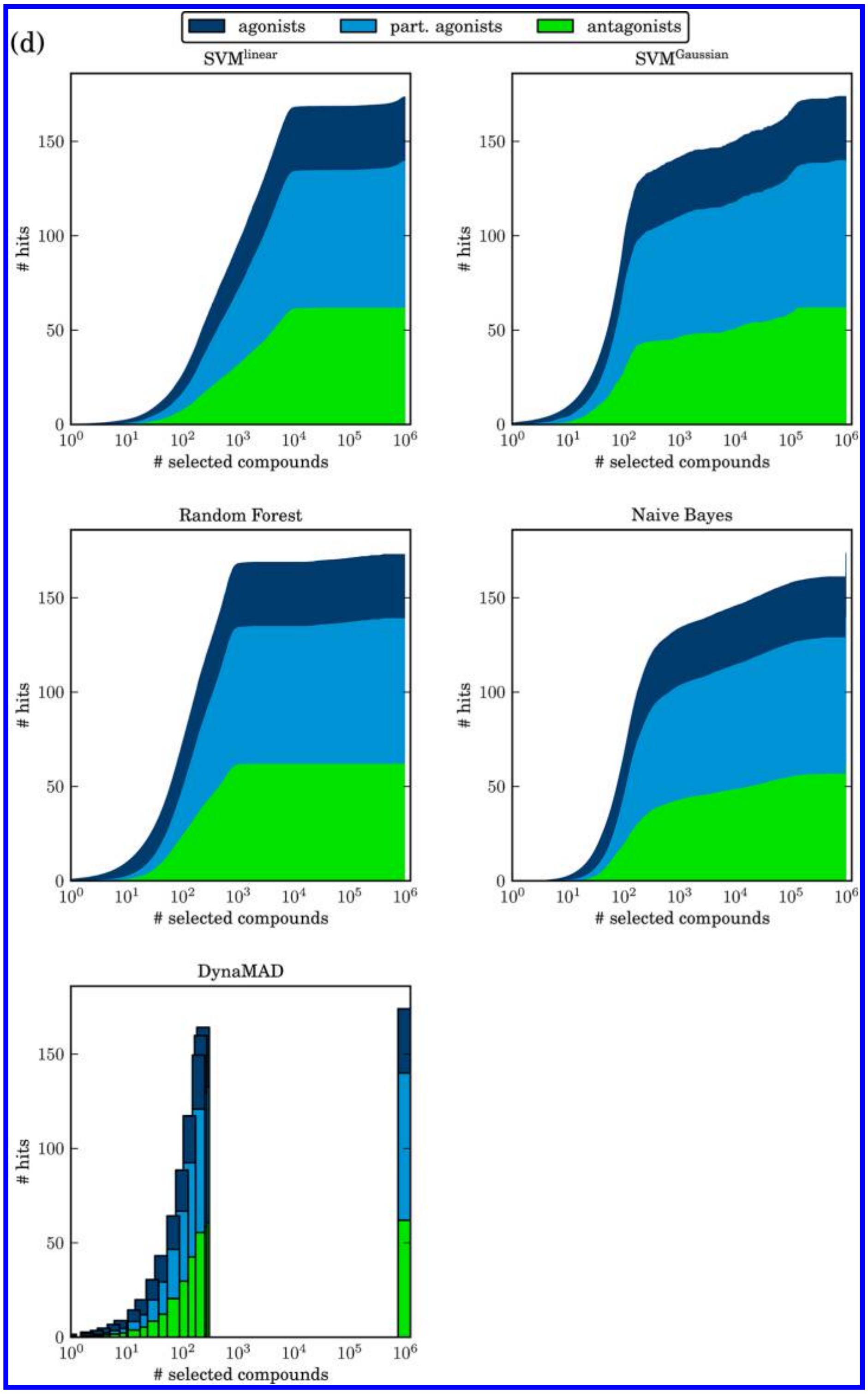


**Figure 7.** Specificity analysis using numerical descriptors. For search calculations on four exemplary mechanistic subsets, including (a) AA1 agonists, (b) AA1 antagonists, (c) H3R inverse agonists, and (d) S1A agonists, the average recalls of compounds belonging to all mechanistic subsets per data set are compared to calculations using numerical descriptors.

data set, which we frequently observed. In contrast to Figures 7b and 8b, none of the methods displayed specificity for H3R inverse agonists (for which only 19 potential hits were available), as reported in Figures 7c and 8c. However, despite this small number of potential hits, H3R inverse agonists were still found by DynaMAD, Gaussian SVMs, RF, and fingerprint search calculations in selection sets comprising only 10 database compounds. As shown in Figures 7d and 8d, none of the methods had notable specificity for S1A agonists. However, the large gap between discrete selection sets for the first and second DynaMAD layers indicated that S1A ligands are structurally closely related and have chemical features that set them apart from many database compounds, consistent with the generally high recall rates observed for S1A ligands.

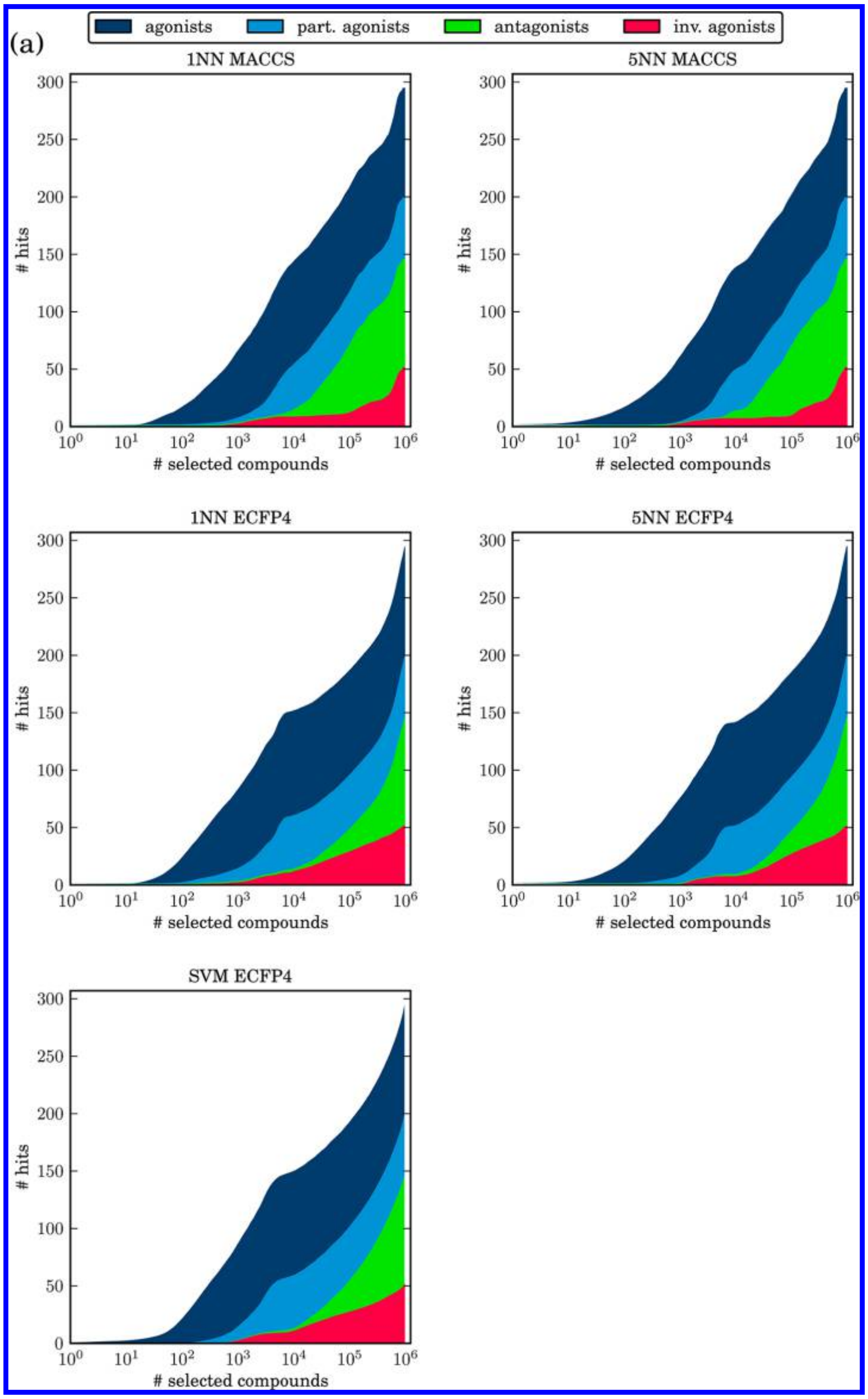


**Figure 8.** continued

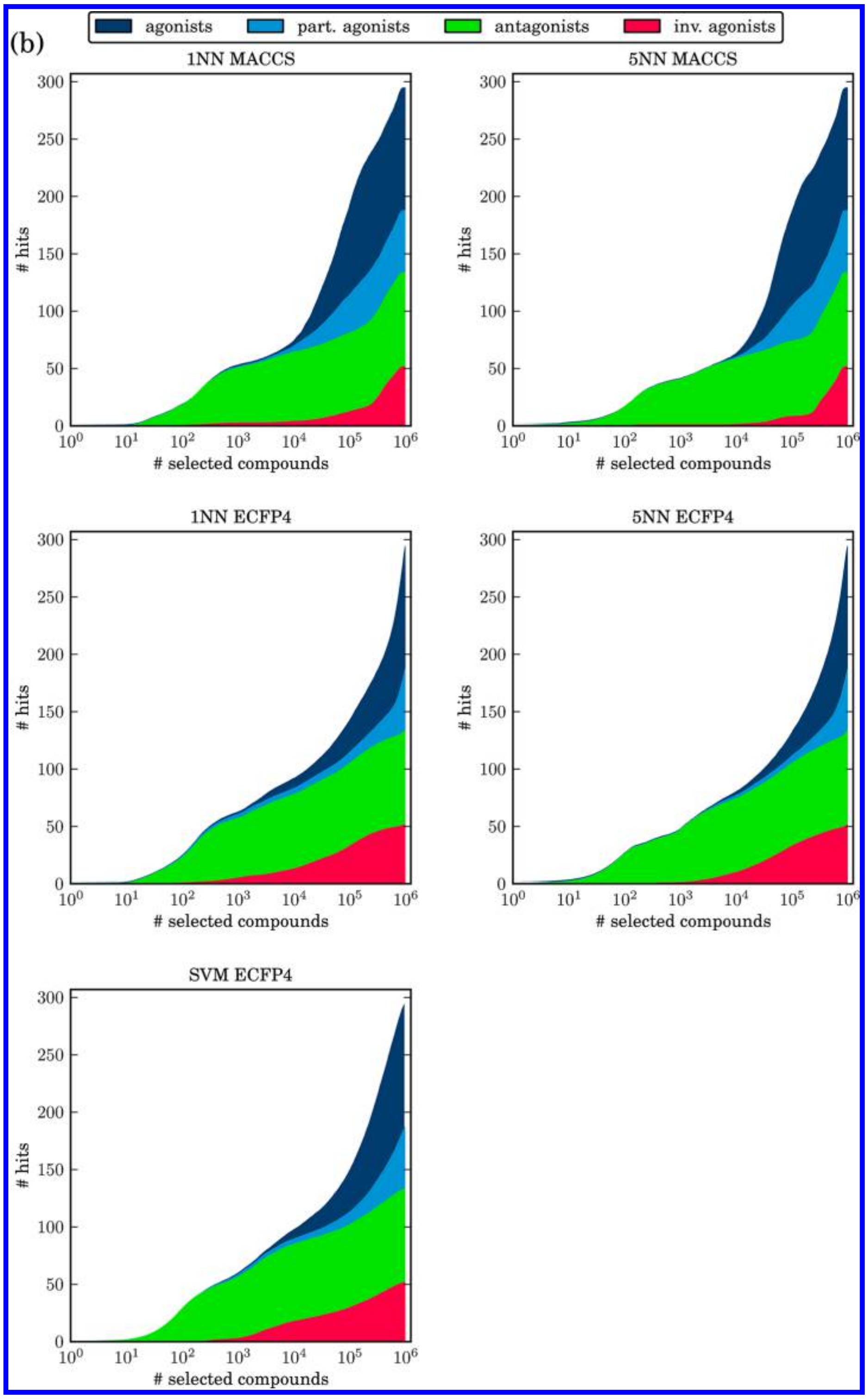


**Figure 8.** continued

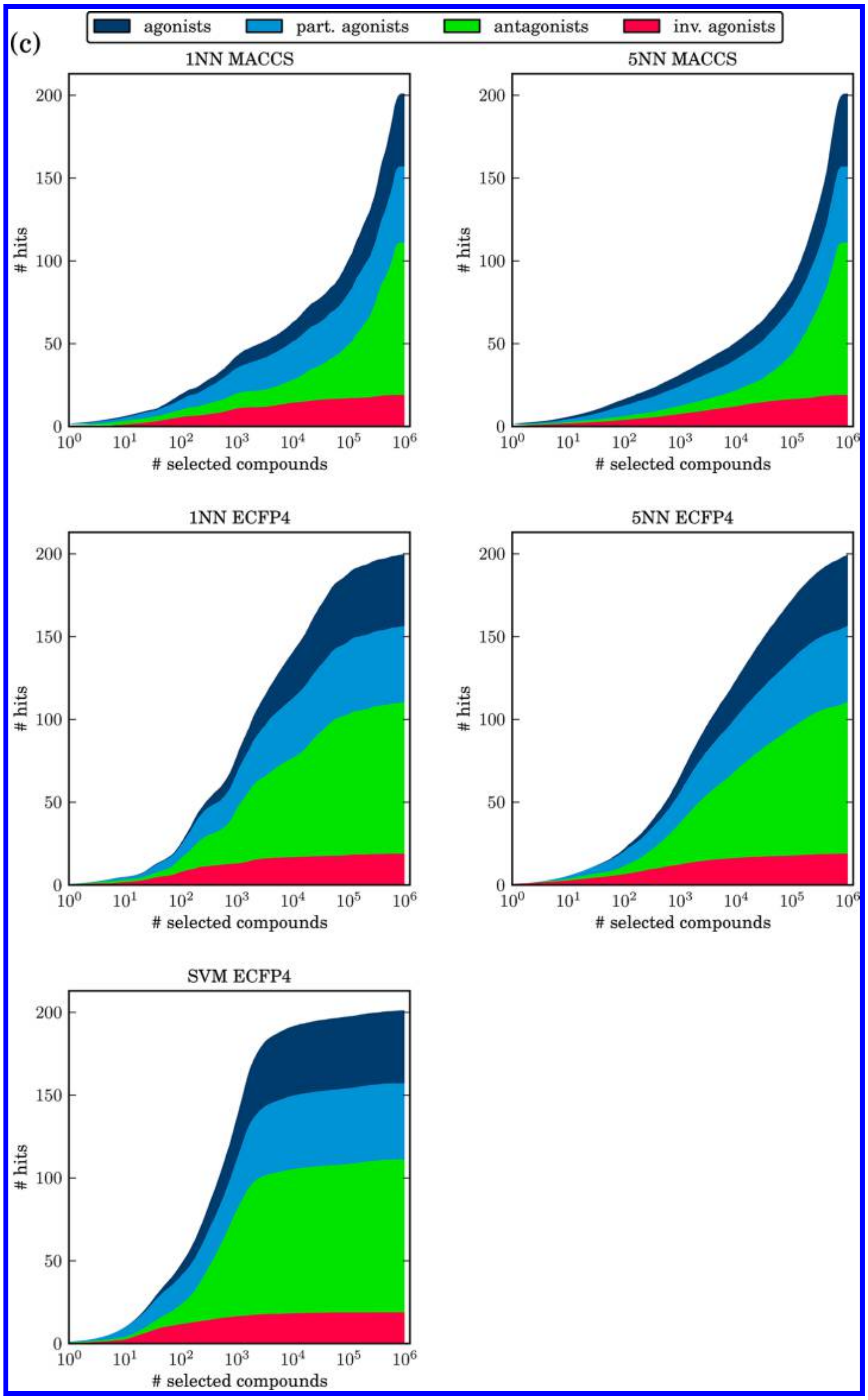


**Figure 8.** continued

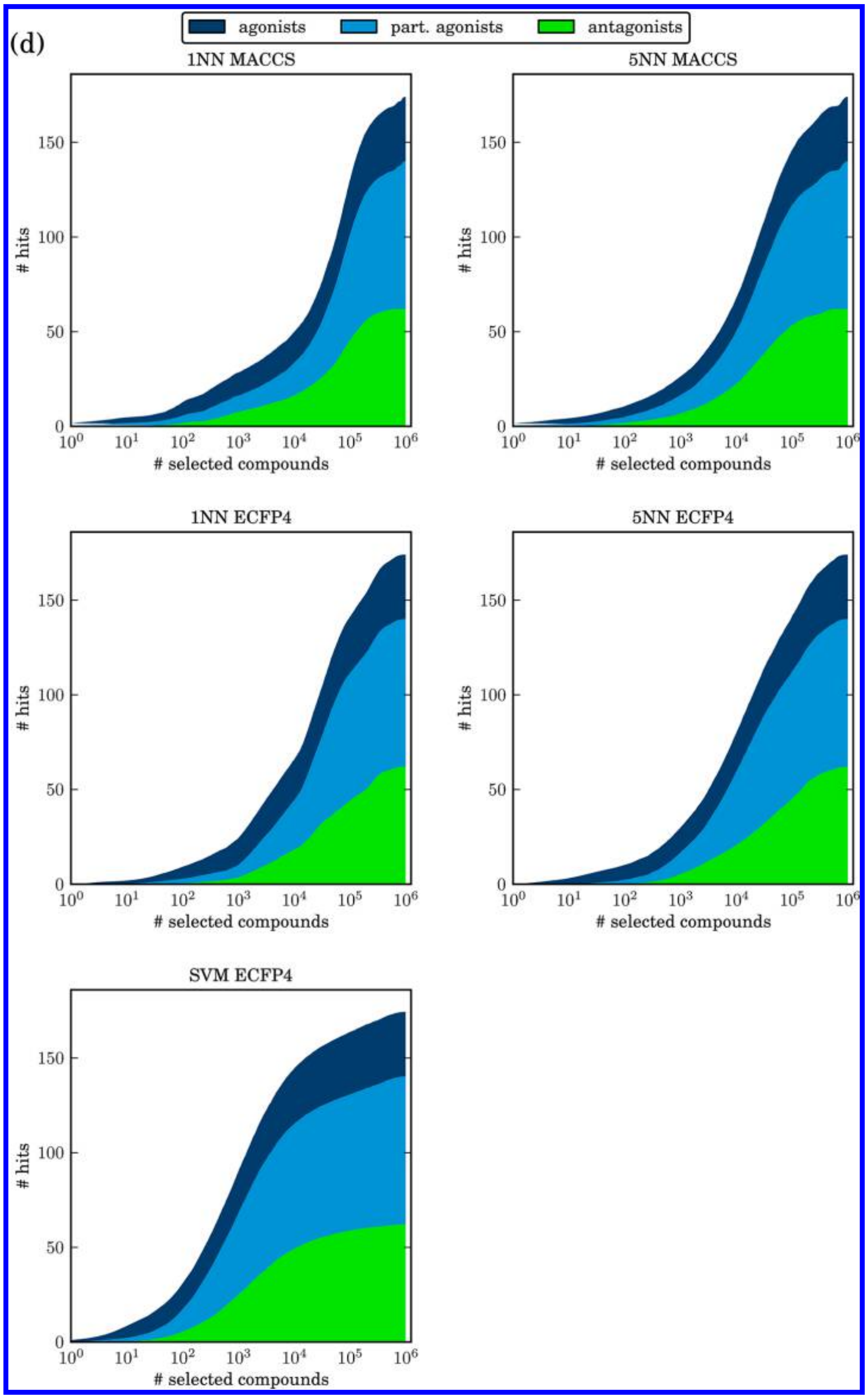


**Figure 8.** Specificity analysis using fingerprints. For search calculations on four exemplary mechanistic subsets, including (a) AA1 agonists, (b) AA1 antagonists, (c) H3R inverse agonists, and (d) S1A agonists, the average recalls of compounds belonging to all mechanistic subsets per data set are compared for kNN fingerprint search calculations.

Table 4 reports the recalls of ligands with desired and other mechanisms for DynaMAD calculations and database selection set sizes closest to 100 compounds together with standard deviations for the calculations. Because the standard deviations between layer sizes were often high, the DynaMAD results could not be directly compared to the corresponding results obtained for top-ranked 100 compounds from SVM, RF, and Bayesian calculations, which are reported in Table 5. For the AA1 data set, DynaMAD displayed a strong tendency to enrich ligands with the desired mechanism over all others. For the other sets, such clear trends were not observed because the cumulative numbers of ligands with other mechanisms were often larger than the number of ligands with the desired mechanism, especially for the structurally very homogeneous S1A and DD2 data sets.

Table 4. Identification of Related Ligands with DynaMAD[a]

| | | layer size | | desired mechanism | | other mechanisms | |
|---|---|---|---|---|---|---|---|
| receptor | mechanism | mean | std | mean | std | mean | std |
| AA1 | agonist | 84 | 33 | 26 | 9 | 4 | 4 |
| | part. agonist | 90 | 30 | 12 | 5 | 12 | 8 |
| | antagonist | 79 | 41 | 17 | 8 | 1 | 2 |
| | inv. agonist | 69 | 38 | 13 | 4 | 3 | 2 |
| AM1 | agonist | 93 | 20 | 4 | 2 | 10 | 5 |
| | part. agonist | 100 | 29 | 18 | 9 | 24 | 27 |
| | antagonist | 91 | 25 | 32 | 13 | 16 | 18 |
| DD2 | agonist | 105 | 34 | 22 | 3 | 75 | 27 |
| | part. agonist | 84 | 29 | 21 | 5 | 54 | 21 |
| | antagonist | 103 | 21 | 43 | 8 | 55 | 12 |
| | inv. agonist | 97 | 0 | 1 | 0 | 96 | 0 |
| H3R | agonist | 84 | 30 | 9 | 4 | 12 | 11 |
| | part. agonist | 84 | 30 | 11 | 4 | 15 | 7 |
| | antagonist | 89 | 32 | 14 | 6 | 4 | 5 |
| | inv. agonist | 89 | 38 | 13 | 3 | 44 | 25 |
| S1A | agonist | 94 | 25 | 22 | 4 | 65 | 23 |
| | part. agonist | 90 | 33 | 26 | 18 | 30 | 26 |
| | antagonist | 93 | 19 | 35 | 7 | 57 | 14 |

[a]Reported are average (mean) sizes of database selection sets and standard deviations (std) for dimension extension layers of individual DynaMAD calculations that yielded selections sets closest to 100 compounds and the corresponding numbers of correctly identified hits with the desired mechanism as well as active compounds with other mechanisms.

However, in all cases, compounds with the desired mechanism were identified in small database selection sets. Similar observations were made for RF calculations (Table 5), which also displayed a strong tendency to specifically detect AA1 ligands with the desired mechanism over others. Only AA1 partial agonists were occasionally misassigned to AA1 ligands having other mechanisms. The AA1 partial agonists are structurally rather diverse, which might explain these findings. Other mechanistic compound subsets were structurally more homogeneous and thus easier to classify. These different types of structural relationships among compounds with high and lower prediction accuracy are illustrated in Figure 9.

In addition, comparable RF results were obtained for AM1 and H3R antagonists. Table 5 also shows that linear SVMs and naïve Bayesian classifiers did not detect more ligands with the desired mechanism than others. However, taking into account that there generally were much larger numbers of compounds with other mechanisms available than with the desired one, the majority of the search calculations displayed an enrichment of ligands with the targeted mechanism of action over others (albeit at varying magnitudes), consistent with the findings discussed above.

## ■ CONCLUDING REMARKS

In this work, we have investigated the previously unconsidered question of whether ligands of a given receptor with different mechanisms of action can be distinguished in simulated virtual screening trials using state-of-the-art supervised machine learning techniques, including SVMs, RFs, and Bayesian classifiers. This has been the main goal of our investigation, given that machine learning approaches in virtual screening are usually applied to facilitate activity predictions for single targets without further differentiation. Also included in the comparison were DynaMAD, a conceptually distinct mapping algorithm operating in descriptor reference spaces of increasing dimensionality, and a simple kNN search strategy using fingerprint representations. Together with RF, DynaMAD is conceptually simpler than SVMs or Bayesian methods, yet it is unusual in its design and not widely applied. The data sets we studied consisted of several subsets of receptor ligands with several different mechanisms, including agonists, partial agonists, antagonists, and inverse agonists. As we have shown (and as can be reconciled on the basis of the freely available data sets), compounds with different mechanisms of action are usually structurally similar or analogous and often are distinguished only by subtle chemical modifications. Thus, discerning these structure–activity relation-

Table 5. Recall of Related Ligands[a]

| | | $SVM^{linear}$ | | $SVM^{Gaussian}$ | | RF | | naïve Bayes | |
|---|---|---|---|---|---|---|---|---|---|
| receptor | mechanism | desired | other | desired | other | desired | other | desired | other |
| AA1 | agonist | 0 | 0 | 15 | 1 | 40 | 3 | 4 | 2 |
| | part. agonist | 0 | 0 | 12 | 5 | 23 | 13 | 1 | 2 |
| | antagonist | 2 | 0 | 12 | 1 | 26 | 2 | 3 | 0 |
| | inv. agonist | 0 | 0 | 12 | 1 | 21 | 7 | 2 | 2 |
| AM1 | agonists | 0 | 2 | 5 | 14 | 6 | 15 | 1 | 1 |
| | part. agonist | 9 | 9 | 19 | 35 | 26 | 39 | 7 | 11 |
| | antagonist | 10 | 12 | 40 | 17 | 41 | 32 | 27 | 17 |
| DD2 | agonists | 10 | 14 | 24 | 73 | 23 | 32 | 20 | 47 |
| | part. agonist | 5 | 25 | 24 | 73 | 24 | 71 | 21 | 59 |
| | antagonist | 20 | 22 | 45 | 54 | 44 | 51 | 35 | 38 |
| | inv. agonist | 0 | 0 | 1 | 99 | 1 | 89 | 1 | 38 |
| H3R | agonists | 5 | 8 | 15 | 21 | 19 | 24 | 8 | 13 |
| | part. agonist | 4 | 9 | 14 | 22 | 18 | 45 | 10 | 28 |
| | antagonist | 2 | 4 | 35 | 13 | 33 | 32 | 12 | 5 |
| | inv. agonist | 12 | 66 | 15 | 74 | 16 | 72 | 8 | 35 |
| S1A | agonists | 10 | 17 | 26 | 70 | 24 | 49 | 22 | 45 |
| | part. agonist | 23 | 28 | 45 | 46 | 45 | 53 | 29 | 28 |
| | antagonist | 17 | 19 | 37 | 63 | 39 | 57 | 32 | 57 |

[a]Reported are the average numbers of correctly identified hits with the desired mechanism and ligands with other mechanisms for the 100 top-ranked database compounds from SVM, RF, and Bayesian calculations.

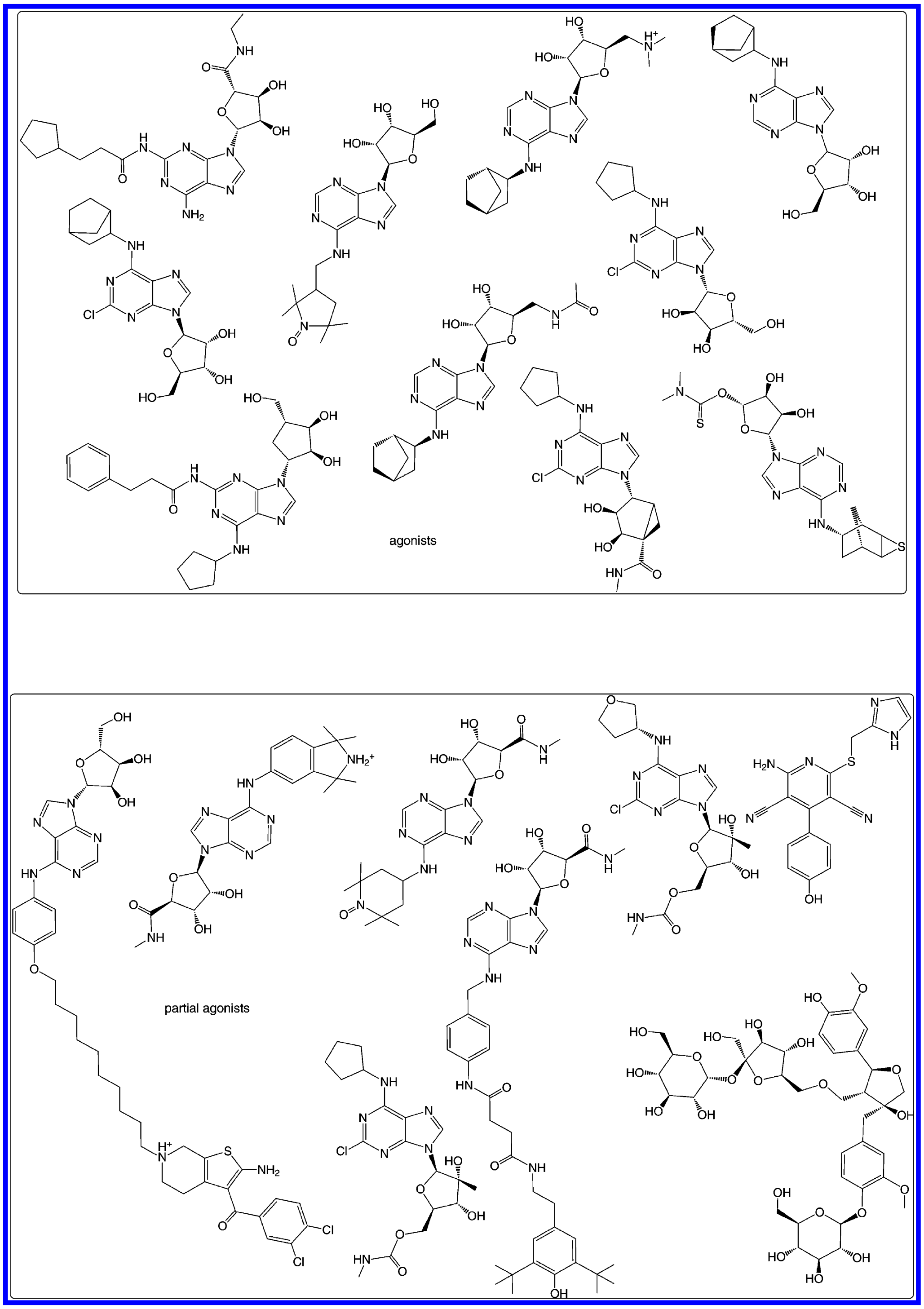


**Figure 9.** Agonist and partial agonist compounds of the AA1 receptor. Shown are (a) agonists and (b) partial agonists from the AA1 data set. Agonists are structurally more similar to each other than partial agonists, which likely explains the higher prediction accuracy observed for AA1 agonists.

ships presented a more complex and challenging prediction task for machine learning than the single-activity predictions usually carried out in the context of virtual screening, in accord with the major goal of our study specified above.

Given the large number of search calculations carried out on these data sets, we mostly adhered to standard parameter settings of publicly available implementations of machine learning approaches to render the calculations transparent and easily reproducible. Thus, at least in some cases, it might be possible to further improve the search performance by exploring other parameters and settings. For interested investigators, this opportunity is provided by making the data sets freely available. Because of the differentiated composition of the data sets, only limited numbers of positive training examples were available, as well as limited numbers of potential hits that were "hidden" in a large background database of 1 million druglike compounds with various biological activities.

Our systematic search calculations yielded in part unexpected results. The machine learning methods consistently generated compound rankings with high AUROC values. Hence, receptor ligands were generally preferentially detected over large numbers of database compounds. However, early enrichment characteristics for which AUROC values did not account substantially differed between methods. Surprisingly, a major finding of our study is that RF models of simple design yielded best search results for relatively small data set selection sizes of 100 compounds. In addition, RFs also had higher precision than Gaussian SVMs or naïve Bayesian classifiers. We consider these findings significant because of the conceptual simplicity and computational efficiency of the RF approach compared with other popular machine learning approaches. Furthermore, DynaMAD calculations, which produced discrete selection sets over different dimension extension layers, were also less sensitive than RF models but had comparable precision. A characteristic feature of the DynaMAD calculations was that they consistently identified ligands with the desired mechanism in very small database selection sets of only ~10 compounds, which was also observed in a number of RF searches. Hence, for practical applications where the identification of new interesting compounds takes center stage (rather than the optimization of compound recall), DynaMAD should also merit further investigation; this is another relevant take-home message from our comparison. We have also shown that Gaussian SVMs performed better overall than Bayesian classification. In addition, it should be emphasized that 2D molecular representations were sufficient to yield generally high prediction accuracy in mechanism of action studies. We used a large set of 185 descriptors that are calculated from molecular graphs. For models generated by RF and DynaMAD, whose descriptor selections are readily interpretable, we observed that implicit molecular surface descriptors with a variety of mapped surface properties[26] were especially predictive across the different compound sets and recurrent in the best-performing models (although other types of topological and molecular property descriptors were also selected). By design, these composite surface property descriptors are orthogonal and rich in information.[26] However, they are not vulnerable to uncertainties in generating hypothetical bioactive conformations because they are calculated from molecular graphs in a consistent manner.

While mechanism-based subsets of ligands for which classification models were trained were effectively retrieved from the database, they were only incompletely separated from related ligands with other mechanisms. Hence, there was no true "mechanism specificity" of the machine learning calculations, as we would expect for these structurally closely related ligands. Nevertheless, many calculations notably enriched ligands with the desired mechanism of action over others in small database selection sets, especially in the case of RFs, Gaussian SVMs, and DynaMAD, thus demonstrating the feasibility of the approach in principle. For one of the data sets and additional mechanistic subsets from other sets, RF, Gaussian SVM, and DynaMAD calculations indeed displayed a high degree of specificity.

Control calculations with kNN and SVM searching using fingerprints revealed selectivity for different molecular mechanisms comparable to complex machine learning approaches using numerical descriptors. However, the early enrichment ability of fingerprint similarity searching was not comparable to that of machine learning approaches. In addition, the performance of fingerprint-based SVM models was strongly data-set-dependent. For the structurally homogeneous DD2 and S1A data sets, the search results using these SVM models were inferior to those using naïve Bayesian, Gaussian SVM, and RF calculations, whereas they displayed comparable performance to Gaussian SVMs and even RFs on the other structurally more heterogeneous compound sets.

Taken together, the findings imply that machine-learning-based virtual screening for ligands with a specific mechanism of action should have potential for practical applications. In practice, identifying a few interesting candidates in database selection sets of limited size would be considered a success, even if one or more related compounds with other mechanisms are also identified. In conclusion, on the basis of our detailed comparisons, especially RF models should be of high interest for mechanism-of-action-oriented compound screening, in addition to Gaussian SVMs. Moreover, very small database selection sets from DynaMAD calculations should be well worth considering, as they are likely to contain attractive candidate compounds.

## ■ AUTHOR INFORMATION

### Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors thank Kathrin Heikamp for helpful discussions and Preeti Iyer for support with ligand sets.

## ■ REFERENCES

(1) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 53−62.

(2) Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Comb. Chem. High-Throughput Screening* **2009**, *12*, 332−343.

(3) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(4) Vogt, M.; Bajorath, J. Chemoinformatics: A view of the field and current trends in method development. *Bioorg. Med. Chem.* **2012**, *20*, 5317−5323.

(5) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000; pp 20−83.

(6) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(7) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(8) Burbidge, R.; Trotter, M.; Holden, S.; Buxton, B. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *J. Comput. Chem.* **2001**, *26*, 5−14.

(9) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(10) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *43*, 1947−1958.

(11) Kenakin, T. Principles: Receptor theory in pharmacology. *Trends Pharmacol. Sci.* **2004**, *25*, 186−192.

(12) Greasley, P. J.; Clapham, J. C. Inverse agonism or neutral antagonism at G protein-coupled receptors: A medicinal chemistry challenge worth pursuing? *Eur. J. Pharmacol.* **2006**, *553*, 1−9.

(13) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21−29.

(14) Eckert, H.; Bajorath, J. Determination and mapping of activity-specific descriptor value ranges (MAD) for the identification of active compounds. *J. Med. Chem.* **2006**, *49*, 2284−2293.

(15) Eckert, H.; Vogt, I.; Bajorath, J. Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: Design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006**, *46*, 1623−1634.

(16) Alpaydin, E. *Introduction to Machine Learning*, 2nd ed.; MIT Press: Cambridge, MA, 2010.

(17) R Foundation for Statistical Computing, Vienna, Austria. The R Project for Statistical Computing. http://www.R-project.org (accessed Aug 26, 2013).

(18) Weihs, C.; Ligges, U.; Luebke, K.; Raabe, N. klaR: Analyzing German business cycles. In *Data Analysis and Decision Support*; Baier, D., Decker, R., Schmidt-Thieme, L., Eds.; Springer: Berlin, 2005.

(19) Morik, K.; Brockhausen, P.; Joachims, T. Combining statistical learning with a knowledge-based approach—A case study in intensive care monitoring. In *Proceedings of the 16th International Conference on Machine Learning*; Morgan Kaufmann: San Francisco, 1999; pp 268−277.

(20) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; ACM: New York, 1992; pp 144−152.

(21) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(22) Joachims, T. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods—Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, MA, 1999; pp 169−184.

(23) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18−22.

(24) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.

(25) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742−754.

(26) *Molecular Operating Environment (MOE)*, version 2011.10; Chemical Computing Group: Montreal, QC, 2011.

(27) 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−D1107.

(28) Iyer, P.; Bajorath, J. Mechanism-based bipartite matching molecular series graphs to identify structural modifications of receptor ligands that lead to mechanism hopping. *Med. Chem. Commun.* **2012**, *3*, 441−448.

(29) Wang, Y.; Bajorath, J. Advanced fingerprint methods for similarity searching: Balancing molecular complexity effects. *Comb. Chem. High Throughput Screening* **2010**, *13*, 220−228.

(30) Witten, I. H.; Frank, E. *Data Mining—Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, 2005; pp 161−176.

(31) Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145−1159.