Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models

Horvath Dragos,* Marcou Gilles, and Varnek Alexandre

Laboratoire d'InfoChime, UMR 7177 Université de Strasbourg − CNRS Institut de Chimie, 4, rue Blaise Pascal, 67000 Strasbourg, France

Received February 17, 2009

The present work proposes a unified conceptual framework to describe and quantify the important issue of the Applicability Domains (AD) of Quantitative Structure−Activity Relationships (QSARs). AD models are conceived as meta-models $\mu\mu$ designed to associate an untrustworthiness score to any molecule M subject to property prediction by a QSAR model μ. Untrustworthiness scores or "AD metrics" $\Psi^{\mu}(M)$ are an expression of the relationship between M (represented by its descriptors in chemical space) and the space zones populated by the training molecules at the basis of model μ. Scores integrating some of the classical AD criteria (similarity-based, box-based) were considered in addition to newly invented terms such as the consensus prediction variance, the dissimilarity to outlier-free training sets, and the correlation breakdown count (the former two being most successful). A loose correlation is expected to exist between this untrustworthiness and the error $|P^{\mu}(M)\text{-}P^{expt}(M)|$ affecting the property $P^{\mu}(M)$ predicted by μ. While high untrustworthiness does not preclude correct predictions, inaccurate predictions at low untrustworthiness must be imperatively avoided. This kind of relationship is characteristic for the Neighborhood Behavior (NB) problem: dissimilar molecule pairs may or may not display similar properties, but similar molecule pairs with different properties are explicitly "forbidden". Therefore, statistical tools developed to tackle this latter aspect were applied and lead to a unified AD metric benchmarking scheme. A first use of untrustworthiness scores resides in prioritization of predictions, without the need to specify a hard AD border. Moreover, if a significant set of external compounds is available, the formalism allows optimal AD borderlines to be fitted. Eventually, consensus AD definitions were built by means of a nonparametric mixing scheme of two AD metrics of comparable quality and shown to outperform their respective parents.

1. INTRODUCTION

Any rules inferred from a set of observations—including the "laws of nature" unveiled by scientific research—are eventually shown to apply to a limited subset of reality. Major crisis in science always start with the discovery of phenomena outside the applicability domain of to-date accepted theories. A critical step toward the new, more general paradigm, embedding the old as a special case, is the discovery of the specific conditions required for the old theory to work: its Applicability Domain[1,2] (AD). Note that the discovery of its AD is always the last contribution brought to a theory on the verge of losing its dominant status and getting downgraded to a special case of a more general approach. AD definition is however not easy, for the attributes/descriptors allowing to delimit the circumstances where the old theory still holds are not straightforward and cannot be predicted by the old theory: Newtonian mechanics cannot predict its own failure at high velocities (the attribute delimiting it from relativistic physics) or at low size scale (border to quantum mechanics).

Machine learning, aimed at extracting empirical mathematical rules that optimally explain a set of observations as a function of their attributes, faces exactly the same AD problems. It is good practice to tentatively consider the definition of an AD of a model as part of the model building procedure *per se*. Especially in the field of Quantitative Structure Activity Relationships (QSAR), models fitted to minimize discrepancies between predicted and observed property values within a training set of molecules are known[3,4] to depend on the peculiar choice of training molecules. If the observed correlations between molecular properties and specific molecular descriptors were to express an objective "law of nature", they should hold for each of the possible "druglike" chemical structures. However, they were established on hand of a training set which, no matter how large, may never represent a significant sample of the chemical structure space. Given a peculiar training set, what are the other molecules for which the thereon trained models can be successfully applied to predict their properties? The question is not trivial, for the answer is by no means contained in the concerned models. Defining the AD of a model actually amounts to the calibration of a meta-model based on its own specific attributes and equations, and returning, upon input of a molecular structure and a QSAR model equation, a "predictability score" of the molecule by the model—a measure of trust to be associated with the QSAR model output for that compound. The QSAR model prediction should then be amended by this trustworthiness score, i.e. the QSAR model and its applicability domain predictor are combined into a consensus model returning the

* Corresponding author e-mail: horvath@chimie.u-strasbg.fr.

10.1021/ci9000579 CCC: $40.75
Published on Web 06/16/2009

QSAR prediction value if trustworthiness if high or “not predictable” otherwise.

1.1. Literature Review. AD definition research is nowadays a hot topic in the QSAR field. Typically, state-of-the art AD models are used to make a binary decision whether or not a QSAR prediction should be trusted or discarded. They can be roughly classified into range-based, distance-based, and density-based approaches.

Range-based AD models consider a prediction as trustworthy if the concerned molecule is located within the chemical space region occupied by training set compounds. This region can be, in its simplest form, defined as the “box” (hyperparallelepiped) of minimal sizes encompassing the entire training set—thus the name of “bounding box methods” attributed to these AD definitions.[5−7] Convex hull approaches[6,8] are refined bounding box methods which delimit an irregularly shaped chemical space zone of minimal volume encompassing all training molecules. They exclude the empty box “corners” from the considered AD (and therefore require a potentially consequent computational cost), but the convexity constraint fails to address the existence of unpopulated cavities within the cloud of training set compounds, within which the behavior of the predictive model is uncertain.

Distance-based AD definitions[6,9−12] assume a prediction to be trustworthy if the concerned molecule is located in the neighborhood of training set compounds. They feature two main degrees of freedom: the choice of the intermolecular distance/metric[13] and the selection of the molecule-to-training set distance criterion (considered equal to the maximal distance over all training molecules, the average distance with respect to the k nearest neighbors, distance to the geometric center of the training set, *etc.*). After selecting the metric and the molecule-to-training set distance criterion, authors need to supply a threshold value above which predictions are to be discarded as untrustworthy. It makes sense to somehow relate this threshold to the typical variance of intermolecular distance values within the training set.[14,15]

Eventually, density methods aim at describing the set of individual training molecules scattered in chemical space as a fuzzy probability density cloud,[16−20] where AD could be then defined as the—topologically contiguous or disjoined—chemical space zones of density exceeding some user-defined threshold. It replaces the “hull” of range-based methods by equidensity surfaces, and lifts the above-mentioned convexity constraint. However, the herein delimited zones largely coincide with the neighborhoods of training compounds, as delimited by intermolecular distance criteria. This complex and heavy mathematical formalism (rich in empirical parameters required to control the width of the Gaussians used to render the spatial density distribution) is not likely to display any overwhelming benefits compared to the two former AD classes.

Last but not least, it should be mentioned that consensus scoring strategies provide a natural estimator of prediction trustworthiness—the variance of prediction according to the various individual models selected for consensus scoring. High variance implies that individual models strongly disagree in terms of that prediction, some of them being obviously wrong—but which? The consensus average value will accordingly be less trustworthy and rather sensitive to the choice of models in the consensus pool.[14,21]

Awareness of the AD problem is increasing in the QSAR community—nowadays, some Web servers offering access to property prediction models also verify whether the submitted molecules fall within the AD of at least some of the various equations used to make consensus property predictions and ignore the untrustworthy ones.[22]

1.2. General Remarks and Definitions Used in AD Monitoring. An AD definition model $\mu\mu$ can be viewed as a classification tool trying to find attributes that discriminate between compounds that were well predicted and respectively mispredicted by a model μ. These compounds should not have participated to the fitting of μ but belong to the external validation set VS—which, *de facto*, becomes the training set for $\mu\mu$.

Tuning of $\mu\mu$ thus starts with the definition of the two classes (well-predicted and mispredicted compounds) it is assumed to tell apart.

Next, a measure of the performance of $\mu\mu$ must be defined as the objective function of the problem.

Further on, any classification model building technique—potentially involving descriptor selection—may be used to train $\mu\mu$. However, for obvious reasons, $\mu\mu$ should be kept as simple as possible, with a minimal number of empirical choices and fittable parameters. Otherwise, $\mu\mu$ itself may become plagued by overfitting artifacts and require yet another external data set for validation, eventually demanding the introduction of a meta-meta-model $\mu\mu\mu$ to describe its applicability domain. Here, $\mu\mu$ will be kept to the simplest possible expression of a classification model, involving two elements:

- an AD metric, or “mistrust score” $\Psi(M)$, function of descriptors of molecule M and, implicitly, of those of training set molecules defining the AD. It returns a measure of the risk of M lying outside the AD of model μ, *i.e.* a measure of the mistrust to be associated with the prediction of the property of M by model μ.
- an unpredictability threshold Ψ denoting the maximally tolerable $\Psi(M)$ value at which the misprediction risk is still acceptable, all the while ensuring that a reasonable fraction of molecules are predictable according to this criterion. However, the existence of a general, compound set independent, optimal cutoff Ψ associated with a metric Ψ (analogous to the famous 0.8 threshold for Tanimoto scores, used and abused as “rule-of-the-thumb” in similarity searching) may not be taken for granted, and this specific question will be addressed. Therefore, the present work mainly focuses on AD metrics as trustworthiness prioritization tools—are compounds at lower Ψ values more likely to be better predicted? Fuzzy applicability domains, with no hard border, are nevertheless useful in medicinal chemistry practice: run property prediction on a set of virtual compounds, select the subset of predicted to match expectation, prioritize by trustworthiness Ψ, then synthesize, and test the top N of this list according to available experimental throughput and the needs of the project.

Monitoring the optimality criterion with respect to varying unpredictability thresholds associated with various unpredictability metrics, for various models predicting different properties, on hand of chemically diverse training sets, will generate benchmarking information relative to the usefulness of the diverse—classical or original—AD definition schemes. There may be no single failsafe AD definition, but if some

Table 1. QSAR Data Sets Considered in This Study[a]

					SQSconsens	best	worst
					R^2_T	R^2_T (*#var*)	R^2_T (*#var*)
ID	description	no. of SQS models	TS size	VS size	$RMS_{T:V}$	$RMS_{T:V}$	$RMS_{T:V}$
logP	water/octanol partition coeff.	2680	3225	9677	0.873 0.66:**0.70**	0.878 (115) 0.65:**0.69**	0.691 (20) 1.03:**1.06**
logS	aqueous solubility	2529	1309	328	0.876 0.77:**0.79**	0.886 (93) 0.74:**0.74**	0.794 (20) 0.99:**1.16**
OrAbs %	oral absorption	2859	187	51	0.344 12.0:**25.9**	0.760 (26) 15.1:**23.0**	0.723 (20) 16.3:**32.2**
DHFR pIC_{50}	dihydrofolate reductase inhib.	2437	237	124	0.834 0.51:**1.05**	0.761 (19) 0.62:**0.81**	0.752 (24) 0.63:**1.76**
TRYPT pIC_{50}	tryptase inhibition	370	3960	11880	0.731 0.14:**0.14**	0.749 (23) 0.14:**0.14**	0.658 (20) 0.16:**0.17**
X pIC_{50}	proprietary affinity	13063	472	62	0.960 0.30:**0.89**	0.899 (9) 0.47:**0.67**	0.764 (24) 0.72:**2.57**
Y pIC_{50}	proprietary affinity	4073	1030	259	0.821 0.71:**0.77**	0.842 (109) 0.67:**0.73**	0.610 (12) 1.05:**1.24**

[a] The number of individual SQS models used to generate the SQSconsens predictions, training (TS) and validation (VS) set sizes and the performance, in terms of R^2 and *RMS* values, of each of the three models used for AD assessment—the SQS consensus (SQSconsens), the best and respectively worst performers with respect to the validation sets. In the three rightmost columns the upper row reports the correlation coefficient of the model with respect to its TS (for the latter two, the number of variables entering the equation is reported in parentheses). Note that since the herein considered training set accommodates both the actual learning molecules and the internal validation subsets, R^2_T does not correspond to the *learning* correlation coefficient but is the synthetic score over both learning and internal validation compounds alike. This R^2_T matches the RMS_T value reported below, followed (:) by the corresponding RMS prediction error for the external validation compounds RMS_V.

of the schemes are seen to systematically outperform others throughout the study, a top set of preferred unpredictability metrics may emerge.

This paper introduces two elements of originality:

• highlighting the logical relatedness of Neighborhood Behavior[23] and Applicability Domain monitoring problems and adopting the NB Optimality criterion[24−26] to monitor AD model performance.

• introducing novel unpredictability metrics notably based on a count of training set descriptor correlations being violated by the molecule to test.

As a last point, the study will address the issue of consensus AD definitions, trying to assess whether simultaneous use of these preferred criteria might significantly improve "predictability prediction".

2. METHODS

2.1. General Remarks—QSAR Model Building and Associated Data Sets. In the following, by training set (TS) we consider the complete set of N_{TS} molecules M of known activities A_M, together with the initial set of calculated descriptors $d_{M,i}$, with $i=1...N_d$. The external validation set (VS) contains N_{VS} molecules of known activity, used to challenge the models built on hand of the TS. Descriptor selection and fitting of linear and nonlinear equations was performed by the Stochastic QSAR Sampler[3,27] (SQS) and resulted in a pool of models $\{\mu\}$ of top quality in terms of training and internal validation criteria. The initial pool of candidate descriptors included, from case to case, all or some of the following: Fuzzy Pharmacophore Triplets,[28] ISIDA fragment[29−32] counts and ChemAxon[33,34] pharmacophore pairs, BCUT terms, calculated logP/logD, and total polar surface area. Let in the following $\{d\}$ be the set of all the N_d candidate descriptors provided for model building, and $\{d\}_\mu$ the subset of N_μ key descriptors actually used in equation μ.

The herein considered case studies were chosen to be as diverse as possible: they cover both approaches developed in our group for predictive purposes, models that were developed for benchmarking purposes, and models developed in collaborations with industrial partners, with large industry-owned data sets (no details, however, can be mentioned in this regard). The panel of modeled properties (see Table 1) includes logP [Obviously, ChemAxon calculated logP/logD were this time *not* used as candidate descriptors.], logS (physicochemical, extracted from the PHYSPROP[35] database), oral absorption[36] (pharmacokinetic), and ligand binding affinities for dihydrofolate reductase,[37] tryptase[38,39] [data set by courtesy of Prof. G. Schneider, Universität Frankfurt], and other two nondisclosed receptors.

Being issued from different studies, the models were obtained according to slightly different protocols:

• In the *5-fold cross-fitting* approach, the TS was split up into *K=3* or *5* parts, *K-1* of which were iteratively merged to form the actual learning subset, the other serving for internal validation. *2K* parallel SQS runs were thus performed, two for each splitting scheme (one sampling only linear, the second both linear and nonlinear models). Models within training R^2_T and internal validation correlation scores R^2_V close to the respective absolute optima ($R^2_T>R^2_{T,max}-0.1$ and $R^2_V>R^2_{V,max}-0.1$) were co-opted into the final model pool $\{\mu\}$, irrespective of the actual parent splitting scheme.

• Alternatively, *one-step fitting* was used for benchmarking studies using literature data or preliminary studies pending the completion of the lengthy SQS procedure. This used the entire TS for calibration, without any further splitting (recall that internal cross-validation is part of the SQS fitness score). In this case, the model pool regroups all the models with fitness exceeding set-specific thresholds.

Model building is however a secondary issue in this work, which merely requests some (better or worse) models to serve as subject for AD assessment. Therefore, no further details about model building will be given here. In every study, AD assessment was applied to three specific property prediction models:

• SQSconsens: the consensus prediction of VS compounds according to all models in $\{\mu\}$, also allowing to estimate, for each VS molecule, a measure of agreement reached by the individual models expressed by the variance of their prediction. The lower this variance, the larger the confidence in prediction—thus, prediction variance will be used as a putative prediction confidence score in this study.

• best: the model in $\{\mu\}$ returning the most accurate prediction of VS compounds, and

• worst: the model in $\{\mu\}$ committing a maximum of errors with respect to compounds in the VS.

Obviously, in a real QSAR-based virtual screening scheme, the "best" and "worst" models in pool $\{\mu\}$ cannot be known beforehand, but this is irrelevant in this AD-oriented study,

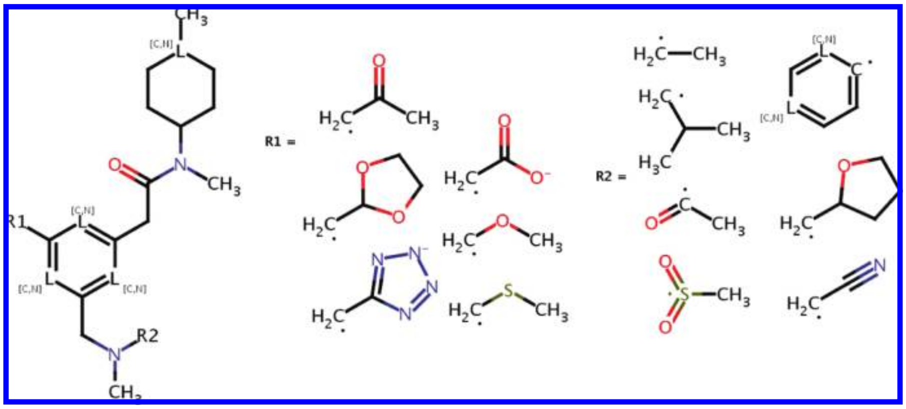

Figure 1. Fictive, diverse combinatorial library (L standing for an atom list of either C or N) of 960 members, in which, for each of its members, the number of cations strictly equals the one of hydrogen bond donors. Any pharmacophore feature-based QSAR derived on hand of such a training set is at risk of mispredicting compounds with “regular” H-bond donors such as −OH.

where the choice of these two extremes plus the consensus is meant to provide a synthetic view on applicability issues.

2.1.1. Correct and Erroneous Predictions. A fuzzy prediction error classifier, $\varepsilon^{\mu}(M)$ is used to quantify the status of property prediction of a VS molecule M by model μ. If the absolute prediction error of property $errP^{\mu}(M)=|P^{\mu}(M)-P^{expt}(M)|$ is inferior to the residual RMS error of the model with respect to training set compounds, $RMST^{\mu}$, then M is well predicted and $\varepsilon^{\mu}(M)=0$. If $errP^{\mu}(M)>1.5\ RMST^{\mu}$, then M counts as mispredicted, with $\varepsilon^{\mu}(M)=1$. In between, the error classifier may smoothly vary between 0 and 1: $\varepsilon^{\mu}(M)=2[errP^{\mu}(M)/RMST^{\mu}-1]$.

2.2. Definition of AD Metrics Ψ. With judiciously chosen $\Psi(M)$, higher values mean increased probability of M falling outside the AD of μ, and thus higher risks to have the property of M mispredicted by μ. This work will distinguish between model-independent (labeled **I**) functions making use of the entire set of candidate descriptors $\{d\}$ and model-dependent (labeled **D**) expressions relying only on the terms $\{d\}_{\mu}$ entering model μ. Model-independent approaches provide a generic estimation of the extrapolation propensity of any model that could be learned on hand of the given training set, in the context of all its available attributes. They generally apply to any trained model μ.

While Ψ will be used for generic references to AD metrics, specific labels for the considered functional forms will be introduced below. Functions including empirical parameters which could be—but were not—fine-tuned to improve their performance were defined with fixed parameter values.

2.2.1. CORRBRK: The Correlation Breakdown Count. Linearly correlated descriptor pairs (d,d') where $d'\approx\alpha d+\beta$ are traditionally discarded from the initial property-attribute matrix in QSAR studies, deemed to be “redundant”, i.e. convey the same information (the one of d and d' with a marginally stronger correlation to the activity value being kept). However, such correlations may often be specific to the considered training set. In as far as two molecular attributes do not have a physical reason to be strongly covariant, the existence of such covariance over a training set is a strong signal of lacking diversity. Correlated descriptors should not be discarded but used to understand the intrinsic boundaries of predictability by an imperfect model derived on hand of an imperfect data set. For example, the combinatorial library enumerated in Figure 1 consists of 960 members and is quite diverse, in the sense that all the pharmacophore features—lipophilic, aromatic, hydrogen bond (HB) donors, acceptors, cations, and anions are represented. Nevertheless, there is a perfect correlation between the number of HB donors and the number of positive charges within this set, as the tertiary ammonium groups are also the only available donors. Supposing that this feature plays a key role in the biological activity of the compounds, a QSAR model will include either the cations count or the HB donor count. Which of the two “redundant” terms will make it into the equation is a matter of chance. When used to predict compounds with neutral HB donors or quaternary cations, the model may perform well if “luck” picked the mechanistically relevant of the two terms or fail otherwise. The initial training set is degenerated and does not allow discrimination between donors and cations—therefore, applying it to compounds in which donors are distinct from cations should, in principle, be forbidden. However, if the equality of cation and donor counts still holds for the compound to be predicted, the initial degeneracy is *not* a problem.

Accordingly, the probability of correctly predicting a novel compound, given a peculiar TS, is expected to decrease with the number of pairs of descriptors in the molecule that fail to respect the correlations observed throughout the TS. Let N_c be the total number of linear interdescriptor correlations that hold with respect to TS molecules m, at correlation coefficient values above 0.7

$$N_0 = size\{d'(m) = \alpha d(m) + \beta|(m\in TS) \text{ and } (d, d'\in \{d\}) \text{ and } (R^2 \geq 0.7) \qquad (1)$$

In the model-dependent approach, only correlations involving descriptors that enter the model are counted

$$N_0^{\mu} = size\{d'(m) = \alpha d(m) + \beta|(m\in TS) \text{ and } (d\in \{d\}_{\mu} \text{ and } d'\in \{d\}) \text{ and } (R^2 \geq 0.7) \qquad (2)$$

For an external molecule M, the linear relations of the sets given in eqs 1 or 2 respectively are considered to be broken, or violated, if the discrepancy between the actual $d(M)$ value and the one extrapolated on hand of $d'(M)$ diverge by more than the RMS deviation on the training set: $|d(M)-\alpha d'(M)-\beta|>RMST_{d(m)=\alpha d'(m)+\beta}$. Let $N_b(M)$ denote the number of interdescriptor relationships broken by molecule M. Then, the model-independent (I) and respectively dependent (D) *CORRBRK* unpredictability scores can be defined as

$$CORRBRK{:}I(M) = \frac{\log[1 + N_b(M)]}{\log[1 + N_c]};\quad CORRBRK{:}D^{\mu}(M) = \frac{\log[1 + N_b^{\mu}(M)]}{\log[1 + N_0^{\mu}]} \qquad (3)$$

The logarithmic form of eq 3 was chosen since in typical data sets used in SQS model building, involving thousands of highly correlated fuzzy pharmacophore triplets and fragment descriptors, both the numbers of initial N_c and broken N_b correlations may span several orders of magnitude.

2.2.2. SQSVAR: The SQS Consensus Prediction Variance. Rather than discarding correlated descriptor pairs prior to model building, the Stochastic QSAR sampler (SQS) is aimed at enumerating a maximum of possible QSAR equations, alternatively using one or the other of apparently redundant descriptors—or even both, if the residuals of their less than perfect correlation "hide" a new, meaningful independent variable. Two apparently equivalent SQS equations μ and μ', based on quasi-identical descriptor choices, with term d in μ being replaced in μ' by the—as far as TS molecules may tell—correlated, equivalent d', may nevertheless return diverging predictions $P^{\mu}(M) \neq P^{\mu'}(M)$ for external compounds with $d(M) \neq \alpha d'(M) + \beta$. While the breakdown of such pairwise attribute correlations is monitored by the *CORRBRK* index, higher order correlation artifacts, similarly leading to divergent predictions, are not. Molecules M for which values predicted by various SQS models are in close agreement are intrinsically less prone to potential descriptor selection artifacts, and thus appear as more trustworthy. The observed variance of $P^{\mu}(M)$ over all representative models μ found by the SQS approach can thus be used as a mistrust score for the SQS average (consensus) prediction values $<P^{\mu}(M)>_{\mu \in \{\mu\}}$. For practical reasons, this variance should be related to the characteristic $RMST^{consens}$ error of the SQS consensus prediction model with respect to training molecules

$$SQSVAR{:}D(M) = \frac{VAR[P^{\mu}(M)]\Big|_{\mu}}{RMST^{CONSENS}} \qquad (4)$$

Note that *SQSVAR* is regarded as a model-dependent AD metric associated with the SQS consensus model and was only used to monitor the trustworthiness of SQS consensus predictions, not the one of individual models μ.

2.2.3. OUTBOUNDS: The out-of-Bounds Descriptor Value Count. This AD metric relies on the classical "box-based" AD definition, assuming that molecules with attributes that are atypical for training set compounds, i.e. lying outside the "box" containing the majority of training compounds in descriptor space, are not predictable by the model. Each descriptor $d \in \{d\}$ can be characterized by its average $<d>$, variance $\sigma(d)$, minimum $min(d)$, and maximum $max(d)$ with respect to molecules $m \in$TS. An upper and lower bound, $u(d)$ and $l(d)$ respectively can be defined as

$$u(d) = \min\{\max(d), \langle d\rangle + 2\sigma(d)\}; l(d) = \max\{\min(d), \langle d\rangle - 2\sigma(d)\} \qquad (5)$$

The *OUTBOUNDS* count for an external molecule M can thus be defined as the number of descriptors $d(M)$ lying outside the $[l(d),u(d)]$ range. The model-independent *OUTBOUNDS:I* is estimated over the entire set of $\{d\}$, whereas the model-dependent *OUTBOUNDS:D*$^{\mu}$ only accounts for $d \in \{d\}_{\mu}$. The working hypothesis is that the likeness to have a model make a correct prediction decreases with the number of descriptors of M having outlying values.

Furthermore, in open-ended molecular feature-count based descriptor sets, such as pharmacophore fingerprints of fragment counts, it is possible to have new features, never seen in TS compounds, appearing in VS molecules. If, for example, the TS consists only of neutral compounds, the presence of carboxylic acids in VS triggers the apparition of new descriptor columns $d^* \notin \{d\}$ in the VS attribute table: the −COOH fragment count, various pharmacophore elements including the anion feature, *etc*. A specific *NEWDESC* counter of structural features that are new to an external molecule M was thus introduced to complement the *OUTBOUNDS* score. *NEWDESC* cannot be said model-dependent or independent, for it is not known whether the new features have an impact on the studied property, as they were not represented in TS. Therefore, *OUTBOUNDS* scores both include *NEWDESC* and are normed with respect to the total number of descriptors in $\{d\}$.

$$OUTBOUNDS{:}1(M) = {}^{1}/_{size\{d\}}[OUTBOX{:}1(M) + NEWDECS(M)]$$
$$OUTBOUDS{:}D^{\mu}(M) = {}^{1}/_{size\{d\}}[OUTBOX{:}D^{\mu}(M) + NEWDESC(M)] \qquad (6)$$

2.2.4. Dissimilarity-Related AD Metrics: AVGDIS, BAVGDIS MINDIS, and Their "Outlier-Free" Versions. Another classical approach to AD definition revisited here is based on monitoring the dissimilarity between M and all or some specific members $m \in$TS, then using this dissimilarity as AD metric—further away from TS members implies higher mistrust for the prediction of M. *AVGDIS* considers the average of Dice dissimilarities Δ between M and all $m \in$TS using z-normed (average/variance-rescaled) descriptor values $z(d)=[d-<d>]/var(d)$ where averages $<d>$ and variances $var(d)$ are calculated over all $m \in$TS

$$AVGDIS{:}1(M) = \langle \Delta(m, M)\rangle_{m \in TS} = \left\langle 1 - \frac{2\sum_{d \in \{d\}} z[d(M)]z[d(m)]}{\sum_{d \in \{d\}} z[d(M)]^2 + \sum_{d \in \{d\}} z[d(m)]^2} \right\rangle_{m \in TS} \qquad (7)$$

Note that, as a consequence of z-rescaling, Δ takes values between 0 (identity) and 2 (anticorrelated z vectors). Alternatively, a biased averaging scheme, where the nearest neighbors m are given higher weight, has been formulated as

$$BAVGDIS{:}1(M) = \frac{\sum_{m \in TS} \Delta(m, M)e^{-3\Delta(m,M)}}{\sum_{m \in TS} e^{-3\Delta(m,M)}} \qquad (8)$$

Eventually, an even stronger bias in favor of near neighbors is considering only the nearest neighbor of M in TS

$$MINDIS{:}I(M) = MIN_{m\in TS}\{\Delta(m,M)\} \qquad (9)$$

The model-dependent versions *$DIS{:}D^{\mu}(M)$ of these scores are calculated only with respect to model-specific descriptors $d\in\{d\}_{\mu}$.

However, not all the TS molecules are necessarily well predicted by the model. In particular, if the point m corresponding to the nearest neighbor of M is an outlier with respect to regression line μ, the (m,M) neighborhood relation would rather imply M also being mispredicted. In order to avoid such pitfalls, an alternative "outlier-free" version of the above-mentioned metrics has been restricted to the TS subset of well-fitted examples. Let TS-ok be the TS where prediction errors are lower than the training RMS: TS-ok=$\{m\in$TS$|P^{\mu}(m)$-$P^{expt}(m)|\leq RMST^{\mu}\}$. Outlier-free metrics denoted *DIS-OK result from the above-mentioned equations by using this restricted TS-ok instead of the entire TS. Since TS-ok is a model-dependent subset, only model-dependent version of the *DIS-OK metrics were considered, i.e. dissimilarity scores were calculated with respect to descriptors $d\in\{d\}_{\mu}$, except for the SQS consensus models. For technical reasons—there is no single equation file listing all the involved descriptors $\{d\}_{SQSconsens}$—the entire set of descriptors is used in this case.

2.2.5. Consensus AD Metrics. Apparently, the most straightforward way to apply the principles of consensus modeling to the AD problem is to consider several AD definitions—in terms of different AD metrics Ψ_k—and to pick only molecules simultaneously fulfilling AD insider criteria $\Psi_k(M)\leq\Psi_k$ for all k definitions. However, as already mentioned, absolute, compound-set and context-independent optimal unpredictability thresholds Ψ_k may not exist for each of the involved AD metric.

By default, the consensus approaches considered here were also designed as "fuzzy" AD definitions based on consensus AD metrics, challenged to provide not absolute accept/reject decisions, but a better relative prioritization in terms of prediction trustworthiness than any of their parent AD metrics. However, "mixing" different AD metrics, potentially covering different value ranges, calls for the introduction of empirical weighing factors—fittable parameters of the AD model, to be avoided. Therefore, prior to mixing the concerned metric values were first normalized and mapped onto a common range [0,1] by conversion to rank-based values: $\Psi(M)$ is replaced by $\Phi(M)$=*fraction of molecules* m *with* $\Psi(m)\leq\Psi(M)$. Consensus AD metrics $\Psi^{12}(M)$= $\Phi^{1}(M) + \Phi^{2}(M)$ were defined as a plain sum of the rank-based scores associated with the parent metrics $\Psi^{1}(M)$ and $\Psi^{2}(M)$. Two such consensus schemes were tested—*SQSVAR+MINDIS-OK:D* the mix of *SQSVAR:D* and *MINDIS-OK:D*, applied in conjunction with SQSconsens predictions, while *MINDIS-OK+CORRBRK:D* was tested in the context of best and worst models.

2.3. AD Optimality Scores: The Analogy between Neighborhood Behavior and Model Applicability. In order to be useful, AD metrics must be proven to correlate with the misprediction risk of VS compounds. However, there is no simple linear or nonlinear expression for the expected correlation. Therefore, the validity of AD definitions was typically assessed by comparing the average prediction error (see section 2.1.1) for molecules within: $<\varepsilon^{\mu}(M)>_{\Psi(M)\leq\Psi}$ and outside the AD: $<\varepsilon^{\mu}(M)>_{\Psi(M)>\Psi}$. This approach has two weak points:

- It is not true that the average prediction error for molecules outside the AD must necessarily be high. Indeed, in the case of a binding affinity model, the VS may contain many "exotic" species, predicted to be inactive because this is the default behavior of the model. Being inactive is also the default behavior of molecules—in a very few cases only, predictions will be wrong and the average misprediction rate may thus be very low.
- If a low Ψ threshold is required to minimize $<\varepsilon^{\mu}(M)>_{\Psi(M)\leq\Psi}$, the usefulness of a model with such a narrow applicability domain may become questionable. The above-mentioned selectivity-type criterion must be balanced against the fraction of molecules for which μ is applicable. It is however not easy to choose the VS percentage to be left outside the AD, representing a fair price to pay for letting $<\varepsilon^{\mu}(M)>_{\Psi(M)\leq\Psi}$ decrease by a specified amount.

2.3.1. The NB-AD Analogy. This work will exploit the observed analogy between the Neighborhood Behavior problem and the AD definition problem in order to show that statistical tools developed for the former approach provide a simple and elegant solution for the latter. In NB monitoring, molecule pairs are characterized by their calculated dissimilarity scores, which are expected to correlate with the observed property differences, according to the well-known working hypothesis "Similar molecules have similar properties". This means that among compound pairs of low calculated dissimilarity there should be virtually no pairs with significant property differences ("property cliffs"). If each molecule pair is shown as a point in a 2D-scatterplot with calculated dissimilarity on X and property dissimilarity on Y, the plot must show a "forbidden zone" with few, ideally no points at low X and high Y values. The similarity principle does however not request pairs of compounds with similar properties to be structurally similar—the scatter plot area at high X, low Y is typically heavily populated.

The analogy to the AD monitoring problem is straightforward: molecule pairs in NB correspond to single molecules $M\in$VS, the calculated dissimilarity on X is equivalent to the AD metric $\Psi(M)$, whereas the property difference measure on Y is equivalent to the prediction error classifier $\varepsilon^{\mu}(M)$. Like in NB monitoring, molecules with low mistrust scores $\Psi(M)$ but high prediction errors (low X, high Y) are forbidden, whereas correctly predicted compounds with high mistrust scores may not necessarily be a liability.

Monitoring the quality of an AD approach in terms of its "forbidden zone" area is feasible, by analogy to equivalent approaches in NB studies. However, this work will exploit the synthetic and much more robust NB optimality criterion.[25,26] Optimality in NB was defined on hand of classifying compound pairs into one of the following four categories:

- True Similars—pairs with both low calculated dissimilarity scores and low property dissimilarity values. In AD monitoring, these are **"True Insiders" (TI)**—molecules M with $\Psi(M)\leq\Psi$ and $\varepsilon^{\mu}(M)<0.5$.
- False Similars—forbidden pairs of low dissimilarity but with diverging properties. In AD monitoring, these are **"False**

Insiders" (FI)—*M* with *Ψ(M)≤Ψ* but an intolerable *ε^μ(M)>0.5*—to be imperatively avoided.

• True Dissimilars—pairs with high dissimilarity and high property differences, respectively **"True Outsiders" (TO)**—molecules *M* with *Ψ(M)>Ψ* and an allowable *ε^μ(M)>0.5*.

• Potentially False Dissimilars—(apparently) dissimilar compound pairs, nevertheless displaying similar properties. This includes both pairs that are genuinely dissimilar but still have similar properties and also "false" dissimilars which do display an underlying structural relatedness not being properly accounted for by the dissimilarity metric. In AD monitoring, the **"Potentially False Outsiders" (PFO)** also regroup genuine outsiders (GO) which are nevertheless well predicted and false outsiders (FO) for which Ψ*(M)* overestimates the mistrust level.

The optimal AD definition implies the choice of an unpredictability threshold Ψ simultaneously minimizing the observed occurrences of FI and FO cases. Since genuine and false outsiders in PFO cannot be told apart, the entire PFO set size enters the optimality criterion. Suppose that at Ψ there are *N(Ψ)* "insiders" (*M*∈VS with *Ψ(M)≤Ψ*), which amounts to an insider rate *I(Ψ)=N(Ψ)/N*$_{\text{VS}}$. Using the fuzzy prediction error classifier, the fuzzy counts *γ* of total mispredicted compounds, of false insiders and potentially false outsiders can then be respectively written as

$$\gamma = \sum_{M\in VS} \varepsilon^{\mu}(M) \quad \gamma_{F1} = \sum_{M\in VS}^{\psi(M)\leq\psi} \varepsilon^{\mu}(M)$$

$$\gamma_{PFO} = \sum_{M\in VS}^{\psi(M)>\psi} [1 - \varepsilon^{\mu}(M)] \qquad (10)$$

The AD optimality criterion Ω*(Ψ)* can thus be defined as the ratio between the actual weighted sum of the fuzzy counts from (10)—with higher-weight FI count; *k=3* throughout the current work—and its basis level scored by an equal-size witness subset of randomly picked "insiders"

$$\Omega(\psi) = \frac{k\gamma_{F1}(\psi) + \gamma_{PFO}(\psi)}{k\gamma_{F1}^{(rand)} + \gamma_{PFO}^{(rand)}} = \frac{k\gamma_{F1}(\psi) + \gamma_{PFO}(\gamma)}{kI(\psi)\gamma + [1 - I(\psi)](N_{VS} - \gamma)} \qquad (11)$$

Above, it is assumed that, if a fraction *I* of the VS were picked randomly, the therein retrieved count of mispredicted compounds would be proportional to the total misprediction count *γ*, whereas a fraction *1-I* out of the *N*$_{\text{VS}}$-*γ* correctly predicted molecules would represent the basis level for the PFO count.

It is convenient to represent the optimality criterion as a function of the insider rate *I*, which is, unlike Ω−Ψ plots, independent of the value ranges of individual AD scores. With a series of random numbers as Ψ metric, Ω are expected to fluctuate around the value of 1.0, whereas effective Ψ metrics allow Ω to describe a U-shaped curve and reach a minimum at an optimal threshold Ψ*. The pertinence of the AD model *μμ* defined by the rule "*Ψ(M)≤Ψ**" is given by the deepness of this minimum.

2.3.2. Removing Noise from Optimality Calculations: The Ascertained Optimality Excess Criterion. Whereas the denominator in fraction (11) contains the expectation values of *γ* scores corresponding to a random draw, numerator values obtained with a given series of random Ψ numbers may occasionally diverge from nominal denominator values. Fluctuations gain in importance with decreasing VS size (note that in NB studies the number of instances—molecule pairs—scales as the square of compound numbers, effectively ruling out this source of noise). The other source of noise is the imbalance of correct vs wrong predictions: if there are only 5 correctly predicted examples in a VS of 100, the probability to rank all of them among the top 30 by pure chance is no longer negligible, creating a false impression of effective prioritization of correct predictions. In order to testify real prioritization, Ω-*I* curves must dip not only below the 1.0 threshold but also descend deeper than random curves corresponding to "lucky" ranking of the compounds. For each studied set and every tested AD metric, in addition to the actual Ω-*I* plot corresponding to the assessed Ψ score, *n^{rand}=1000* randomized Ω^{rand}-*I* curves were also generated, using actual *ε(M)* values in conjunction with random AD scores. The observed fluctuations of Ω were monitored in function of *I* and expressed by the variances *var[Ω^{rand}(I)]*. The actual Ω-*I* plot is significant only if it manages to dip, at least for some *I* values, below the confidence limit of *1-var[Ω^{rand}(I)]*. Therefore, the ultimate AD quality criterion monitored in this work—the Ascertained Optimality Excess Ξ is defined as

$$\Xi(I) = \max\{1 - var[\Omega^{rand}(I)] - \Omega(I), 0\} \qquad (12)$$

A positive Ξ*(I)* value means that the AD definition singling out the fraction *I* of VS molecules as "predictable" has some intrinsic merit—of course, the AD threshold Ψ* corresponding to the *I(Ψ*)* maximizing Ξ should be preferentially used. Higher Ξ values stand for better AD definitions. The optimal value Ξ^**(I)* within the range *0.4≤I < 1* was considered as the quality measure of an AD metric in a given case study, and plots were shown with respect to this range only. Ξ*(I)* peaks at lower *I* values are intrinsically unlikely, and can be safely ignored if they occur nevertheless, being synonymous for poor AD definitions.

3. RESULTS AND DISCUSSIONS

The variety of case studies, systematically including the best and the worst performing models, should yield a general idea about the robustness of AD models and their usefulness in prioritizing QSAR predictions according to their trustworthiness score. Obviously, neither "perfect" (100% accurate VS predictions) nor "null" (100% false VS predictions) models *μ* may benefit from the introduction of an AD trustworthiness score—in general, higher Ξ*(I)* values are expectable if *μ* is "average", i.e. has comparable success and failure rates. Optimality scores of different models *μ* are not directly comparable—not even if these models stem from a same training set. Also, it is highly unlikely to discover a single "magic" AD metric systematically outperforming all the others in all situations, nor a "standard" confidence threshold Ψ of universal applicability. Therefore, an empirical "Olympic chart"-based performance scoring scheme will be used to benchmark the relative merits of the different AD metrics. For each of the three models (SQScon-

Table 2. Top Three Ranking AD Metrics in Each of the Benchmarking Competitions, Corresponding to Each Model of Each Case Study, in Terms of Ξ*(I)*

data set	SQSconsens	best	worst
OrAbs	1 - *SQSVAR:D*	1 - *AVGDIS:D*	1 - *BAVGDIS-OK:D*
	2 - *CORRBRK:I*	2 - *AVGDIS-OK:D*	2 - *AVGDIS:I*
	3 - *AVGDIS-OK:D*	3 - *BAVGDIS:D*	3 - *CORRBRK:D*
X	1 - *OUTBOUNDS:I*	1 - *BAVGDIS-OK:D*	1 - *BAVGDIS-OK:D*
	2 - *BAVGDIS:I*	2 - *BAVGDIS:D*	2 - *BAVGDIS:D*
	3 - *BAVGDIS-OK:D*	3 - *MINDIS:I*	3 - *MINDIS:I*
DHFR	1 - *CORRBRK:I*	1 - *MINDIS-OK:D*	1 - *BAVGDIS-OK:D*
	2 - *BAVGDIS-OK:D*	2 - *MINDIS:D*	2 - *BAVGDIS:D*
	3 - *SQSVAR:D*	3 - *BAVGDIS-OK:D*	3 - *MINDIS-OK:D*
Y	1 - *SQSVAR:D*	1 - *MINDIS-OK:D*	1 - *MINDIS-OK:D*
	2 - *MINDIS-OK:D*	2 - *BAVGDIS-OK:D*	2 - *CORRBRK:I*
	3 - *BAVGDIS-OK:D*	3 - *MINDIS:D*	3 - *BAVGDIS:I*
logS	1 - *MINDIS-OK:D*	1 - *MINDIS-OK:D*	1 - *MINDIS-OK:D*
	2 - *SQSVAR:D*	2 - *CORRBRK:D*	2 - *CORRBRK:I*
	3 - *OUTBOUNDS:I*	3 - *OUTBOUNDS:I*	3 - *BAVGDIS-OK:D*
TRYPT	1 - *SQSVAR:D*	1 - *AVGDIS:D*	1 - *AVGDIS:D*
	2 - *BAVGDIS-OK:D*	2 - *CORRBRK:D*	2 - *AVGDIS-OK:D*
	3 - *MINDIS-OK:D*	3 - *AVGDIS-OK:D*	3 - *CORRBRK:D*
logP	1 - *MINDIS-OK:D*	1 - *MINDIS-OK:D*	1 - *CORRBRK:I*
	2 - *SQSVAR:D*	2 - *BAVGDIS-OK:D*	2 - *MINDIS-OK:D*
	3 - *BAVGDIS-OK:D*	3 - *BAVGDIS:D*	3 - *CORRBRK:D*

Table 3. Empirical "Olympic Chart" Monitoring the Performance of the Considered AD Metrics Reports[a]

metric	#Uses	gold	silver	bronze	success
SQSVAR:D	7	3	2	1	1.13
MINDIS-OK:D	21	8	2	2	0.78
BAVGDIS-OK:D	21	4	4	5	0.75
BAVGDIS:D	14	0	3	2	0.40
CORRBRK:D	14	0	2	3	0.39
AVGDIS:D	14	3	0	0	0.32
CORRBRK:I	21	2	3	0	0.31
AVGDIS-OK:D	21	0	2	2	0.21
OUTBOUNDS:I	21	1	0	2	0.17
MINDIS:D	14	0	1	1	0.16
BAVGDIS:I	21	0	1	1	0.10
MINDIS:I	21	0	0	2	0.10
AVGDIS:I	21	0	1	0	0.06
OUTBOUNDS:D	14	0	0	0	0.00

[a] For each metric, the number of times it has been applied (#Uses) compared to the number of "medals" won by the metric. The synthetic success score represents the weighted average of medal counts over the number of participations in benchmarking contests: *Success=(1.5xGold+1.2xSilver+Bronze)/#Uses*.

sens, best, worst) of each study case, Ξ*(I)* curves are plotted and compared. The three metrics scoring top Ξ*(I)* values are awarded "gold", "silver", and "bronze" medals, respectively, as shown in Table 2. Finally, the total number of medals obtained by each metric is reported in the context of its total number of participations, in Table 3. Note that *SQSVAR* metrics only apply to SQSconsens models and are thus utilized only once/case study, i.e. a total of 7 times, where model-dependent approaches—except for **DIS-OK* methods, see section 2.2.4—do not apply, being thus used 14 times, with "best" and "worst" equations. Model-independent and **DIS-OK* metrics compete for each model, for each case study—hence 21 times.

The synthetic success score used to sort metrics in Table 3 (see caption) is highly empirical, as is the entire medal-based ranking which does not keep track of the actual differences in performance of the competing metrics. It was preferred to some empirically weighed average of actual Ξ*(I)* values, which are case-study specific for above outlined reasons (case studies with intrinsically low Ξ*(I)* values would have been underrepresented by such an approach). Nevertheless, some quite obvious and robust trends, largely insensitive to the details of the actual success scoring scheme, seem to emerge from Table 3:

- The SQS variance is by far the most robust indicator of prediction trustworthiness, ranking among the best three performing schemes in all but one (X) of the seven study cases—see, for example, tryptase affinity prediction in Figure 2. X is the case study witnessing the most dramatic discrepancies (up to 3-fold increases) between training and prediction RMS error values, as the external prediction set consisted of a majority of molecules bearing no structural similarity at all to the training compounds. The most of external compounds strongly differ, i.e. populate "exotic" fragments and pharmacophore triplets, absent from training molecules. *SQSVAR* is however insensitive to these new elements, which do not enter any of the models used in consensus scoring and do not add up noise to predictions. Unsurprisingly, the *OUTBOUNDS* score is the top performer here (Figure 3). Also, the smoothness of the curves in Figure 2 is in stark contrast to the ruggedness of the ones in Figure 3—like always in QSAR, data set size is paramount to ensure reliable results.

- Dissimilarity scores with respect to the well-predicted training subset are top performers of general applicability, whereas the *SQSVAR* scheme only works with consensus predictions. Expectedly, *MINDIS-OK* and the average dissimilarity score biased in favor of short distances, *BAVGDIS-OK* are top performers, much better than the plain average dissimilarity. Predictability is best expressed by the shortest distance to the nearest chemical space zone populated by compounds for which the model applies well. Plain average is less well suited, for simultaneous closeness to all of the well-described training molecules is not required. Also, similarity to any outlier of the model training process is not only not an indicator of trustworthiness but also actually an indicator of untrustworthiness. Typically, state-of-the art AD definitions based on distances to the nearest neighbor(s) in the training set do not eliminate training outliers from the eligible nearest neighbors, while current results clearly show the importance of this measure. Of course, outlier elimination affects the AD metric performance if any of the external compounds are located in the chemical space neighborhood populated by these. [For information, their rough number is 900 for logP, 600 for TRYPT, and 80 for X, depending on the considered best/worst model.] In the rare cases where *AVGDIS-OK:D* was outperformed by *AVGDIS:D*, the advantage of the latter over the former is not significant—the removal of outliers had a negligible impact over the average of distances to external compounds. *MINDIS* and *BAVGDIS* are naturally more sensitive to—and clearly biased in favor of—outlier elimination.

- Model-dependent metrics are typically better than their model-independent counterparts, as expected. *OUTBOUNDS* and *CORRBRK* are the only notable exceptions to this rule—while the model-independent version of the former is clearly the most successful one, the performance of the latter is largely independent of the monitored descriptor set.

- The correlation breakdown count is as potent as or more potent than standard similarity-based AD criteria, being outperformed only by outlier-free similarity scoring schemes.

3.1. Absolute or Fittable Trustworthiness Limits? The existence of an AD metric Ψ admitting a globally optimal

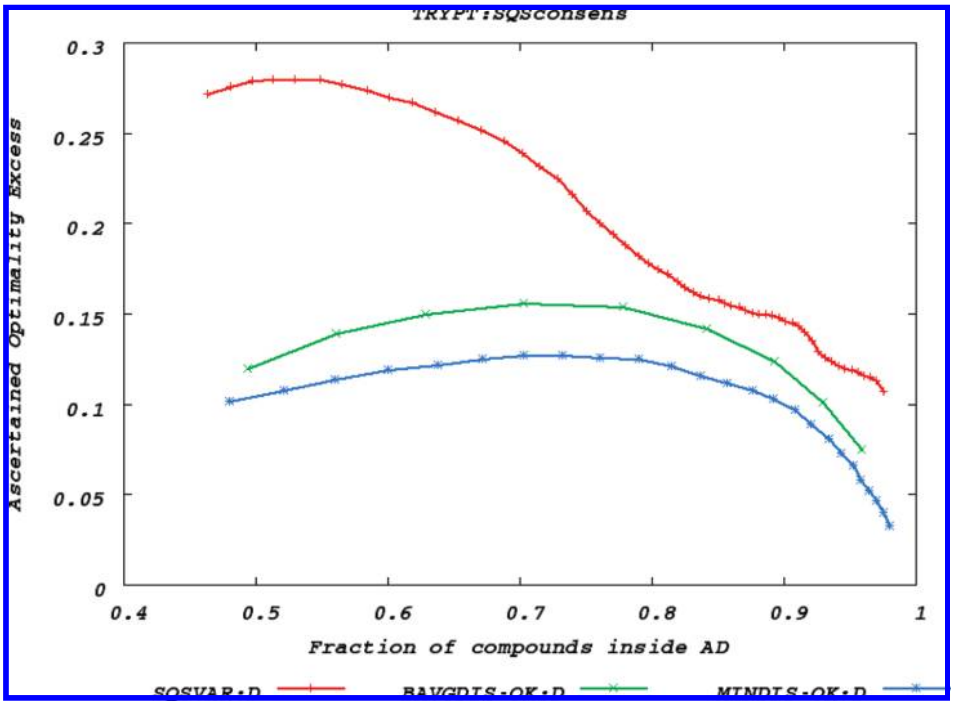

Figure 2. Ξ*(I)* plot illustrating the outstanding performance of the *SQSVAR* metric in prioritizing accurate over erroneous predictions of tryptase affinities by the SQS consensus model.

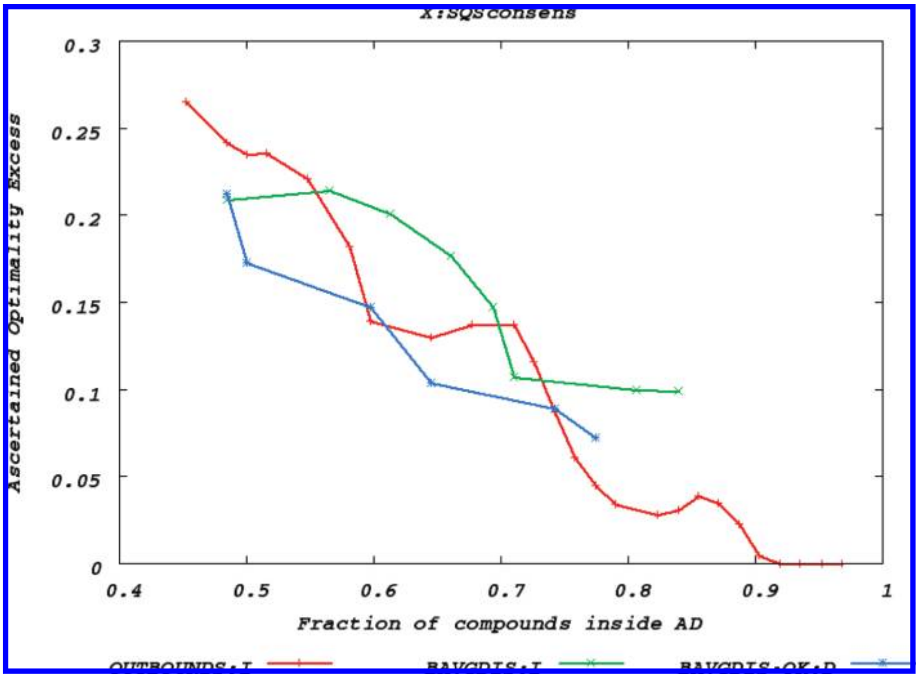

Figure 3. In case of the proprietary case study X, external compounds containing novel, "exotic" functional groups do not stand out by high *SQSVAR* scores but are easily singled out as untrustworthy by the *OUTBOUNDS* count and similarity-based AD metrics.

untrustworthiness threshold Ψ*, to yield an optimal discrimination between right and wrong predictions for no matter which model fitted on an arbitrary data set, is highly unlikely. Even fixed-range metrics such as the Dice-score based **DIS* functions are bound to settle for varying optimal cutoffs, depending on the training set compounds and the descriptors used for dissimilarity scoring. Direct plots of the ascertained excess optimality score with respect to the untrustworthiness threshold of the *MINDIS-OK:D* metric clearly show (Figure 4, top) that no single abscissa value matches high Ξ levels for each of the considered "best" models. Furthermore, the same analysis concerning the SQSconsens approaches (Figure 4, bottom) unsurprisingly reveals a radically different spectrum of individually optimal *MINDIS-OK:D* cutoffs. As in this latter situation *MINDIS-OK:D* is calculated on hand of the entire candidate descriptor set, the entire distribution of intermolecular distances will shift toward higher values. A context-free optimal *MINDIS-OK:D* cutoff value, serving as a rule-of-the-thumb AD delimiter, cannot exist.

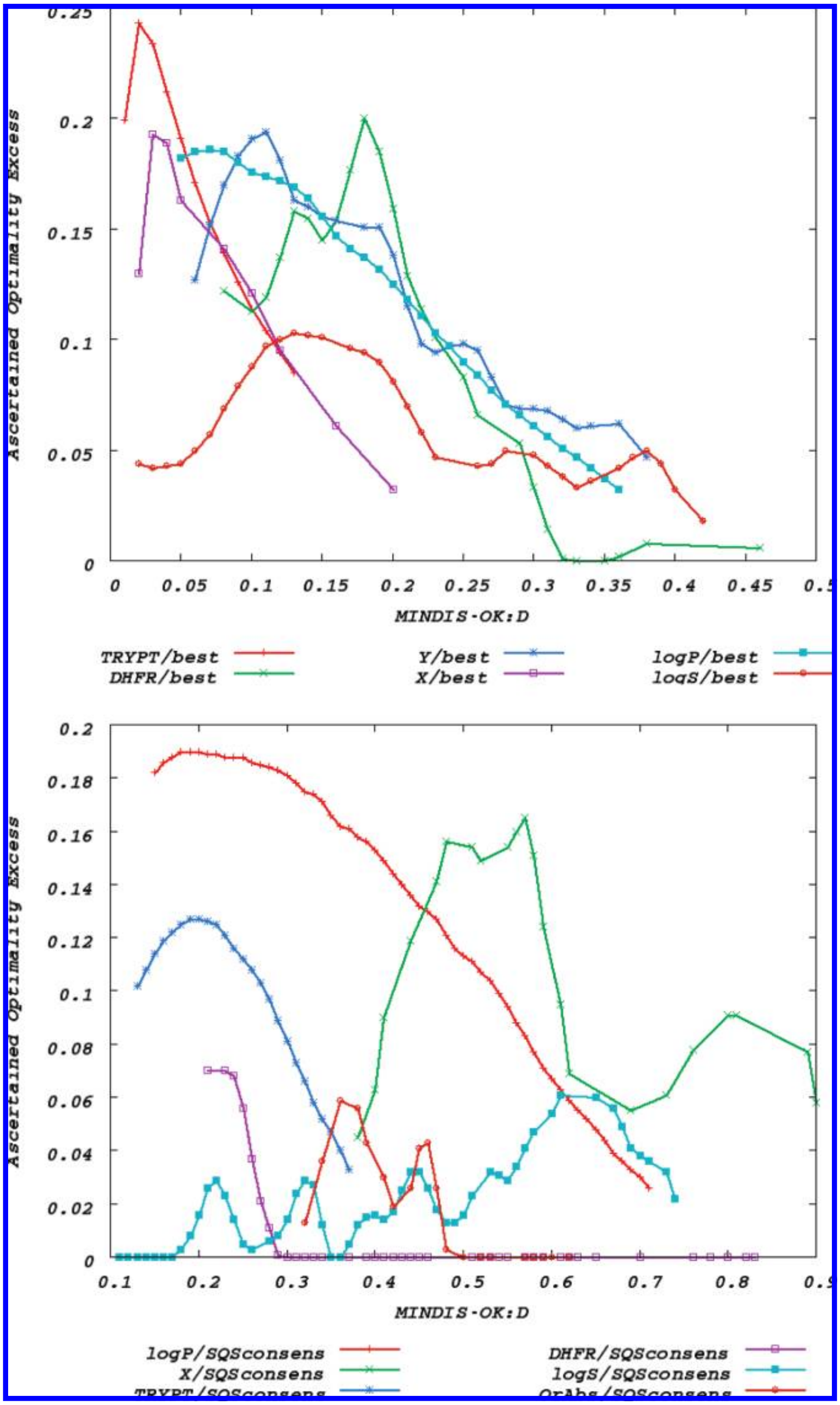

Figure 4. Plots of Ξ with respect to the untrustworthiness threshold of *MINDIS-OK:D* for (above) the best and respectively (below) the SQSconsens models for various case studies show that there is no globally optimal *MINDIS-OK:D* value ensuring simultaneously optimal discrimination between well and wrong predicted external compounds in all the cases.

The situation is hardly different with open-ended scores such as *SQSVAR* which would allow for a generic optimal cutoff value suited as a universal AD delimiter. *SQSVAR* represents the variance of individual predictions by models in the consensus pool, reported to the training set RMS error of the consensus model. It might be argued that *SQSVAR* > *1* is a clear indicator of prediction untrustworthiness: individual models disagree to an extent exceeding the RMS discrepancy between prediction and experiment for training compounds. The plots in Figure 5 basically support this view, with the notable exceptions of DHFR and X (not plotted, but displaying a similar profile). The latter two situations show that unexpectedly high *SQSVAR* values may well occur and show significant optimality (low amplitudes notwithstanding—comparing relative heights of optimality peaks of different data sets makes no sense). However, ignoring these would not make a difference: *SQSVAR* < *1* is hardly a useful AD delimiter. The steep tryptase curve requires the selection of an optimal cutoff value of ∼0.5, down to a precision to 10^{-2}, whereas a precision of 10^{-1} would suffice for logP. Smaller data sets show rugged landscapes that are difficult to interpret.

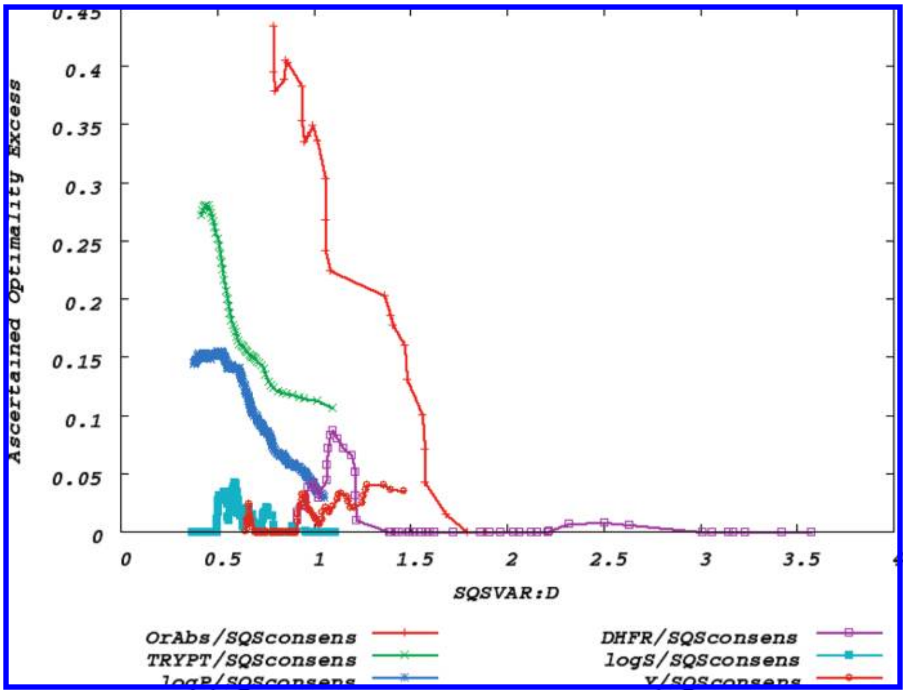

Figure 5. Ξ as a function of the *SQSVAR:D* untrustworthiness threshold, shown for the SQSconsens models of the six case studies having *SQSVAR:D* among the medal-winning approaches.

AD metrics can be therefore successfully applied to prioritize predictions according to their trustworthiness but cannot *a priori* tell whether a prediction is right or wrong (recall that, furthermore, a prediction *must* not be within the AD in order to be correct—hence "*Potentially* False Outsiders"). Optimal untrustworthiness cutoffs are however model and set-dependent and should be considered as the fittable parameter of the AD meta-model $\mu\mu$. Fitting, however, requires not only a training set (as said before, the external validation set of the model μ becomes the training set of the AD meta-model $\mu\mu$) but also an additional meta-model validation set, for which predictions are to be done according to μ, untrustworthiness scores calculated according to the AD metric and predictions rejected if less trustworthy than the fitted optimal cutoff. If these rejections turn out to be judicious, this optimal cutoff might be used as the default rejection criterion of the AD model. Not disposing of additional data, the fitting of the optimal cutoff was addressed by randomly splitting external sets into meta-model training ($TS_{\mu\mu}$, 1/4 of total size) and validation ($VS_{\mu\mu}$, 3/4) subsets. Ascertained optimality curves with respect to AD metrics (Figure 6) were plotted, in parallel, with respect to both $TS_{\mu\mu}$ and $VS_{\mu\mu}$ (labeled, for simplicity, as "1" and "3" respectively, highlighting their relative size). With the two large sets TRYPT and logP, the maxima of the two "best" and, respectively, two "worst" curves coincide, meaning that the optimal untrustworthiness cutoff that would have been learned from subsets "1" are perfectly well suited to ensure optimal separation between correct and false predictions on subsets "3". Not only the optima but also the entire curves overlap fairly well, highlighting the fact that training sets $TS_{\mu\mu}$ are highly representative of the thrice as large $VS_{\mu\mu}$. Note, however, that optimal cutoffs of "best" and "worst" models are not necessarily interchangeable. With DHFR, the fitting runs into the classical training set size problem. The representativeness of $TS_{\mu\mu}$ is no longer granted, and, although the curves no longer resemble each other, with *MINDIS-OK:D* their optima happen to align nevertheless. This is no longer the case with *CORRBRK:D*—in fact, with this metric, training would be impossible on subset "1" (no ascertained optimality effect can be observed). A set of 31 compounds is insufficient to allow fitting of a single parameter—something to consider by authors fitting many-parameter QSARs on even smaller training sets.

3.2. Consensus AD Metrics. The two considered consensus scenarios—mixing *MINDIS-OK:D* with *SQSVAR:D* for SQSconsens models, and respectively with *CORRBRK:D* for "best" equations—were plotted next to their parent AD metrics in Figure 7, in order to evidence potential synergy effects: consensus metrics leading to taller *Ξ(I)* peaks than any of their parents. ["Worst" models were also monitored under the same circumstances as "best" and the study included all target, not only the shown—plots had to be left out, in order to enhance the readability of Figure 7.] Even if the consensus metric did not score the highest absolute optimum, specifically higher *Ξ(I)* values at large fractions of compounds inside the AD may still count as a benefit of metric mixing. The figure clearly shows that the parameter-free metric mixing scheme proposed here specifically shows—sometimes very strong—synergy effects, virtually every time when the mixed metrics are of comparable quality (see framed plots). Mixing a high-quality and a low-quality AD score leads to a mediocre consensus, but this is typical for consensus modeling in general. Also, mixing of strongly intercorrelated metrics would bring no major benefit—combining *MINDIS-OK:D* with other **DIS* scores makes little sense, while *SQSVAR* and *CORRBRK*, based on different principles, nicely complement the dissimilarity-based metric.

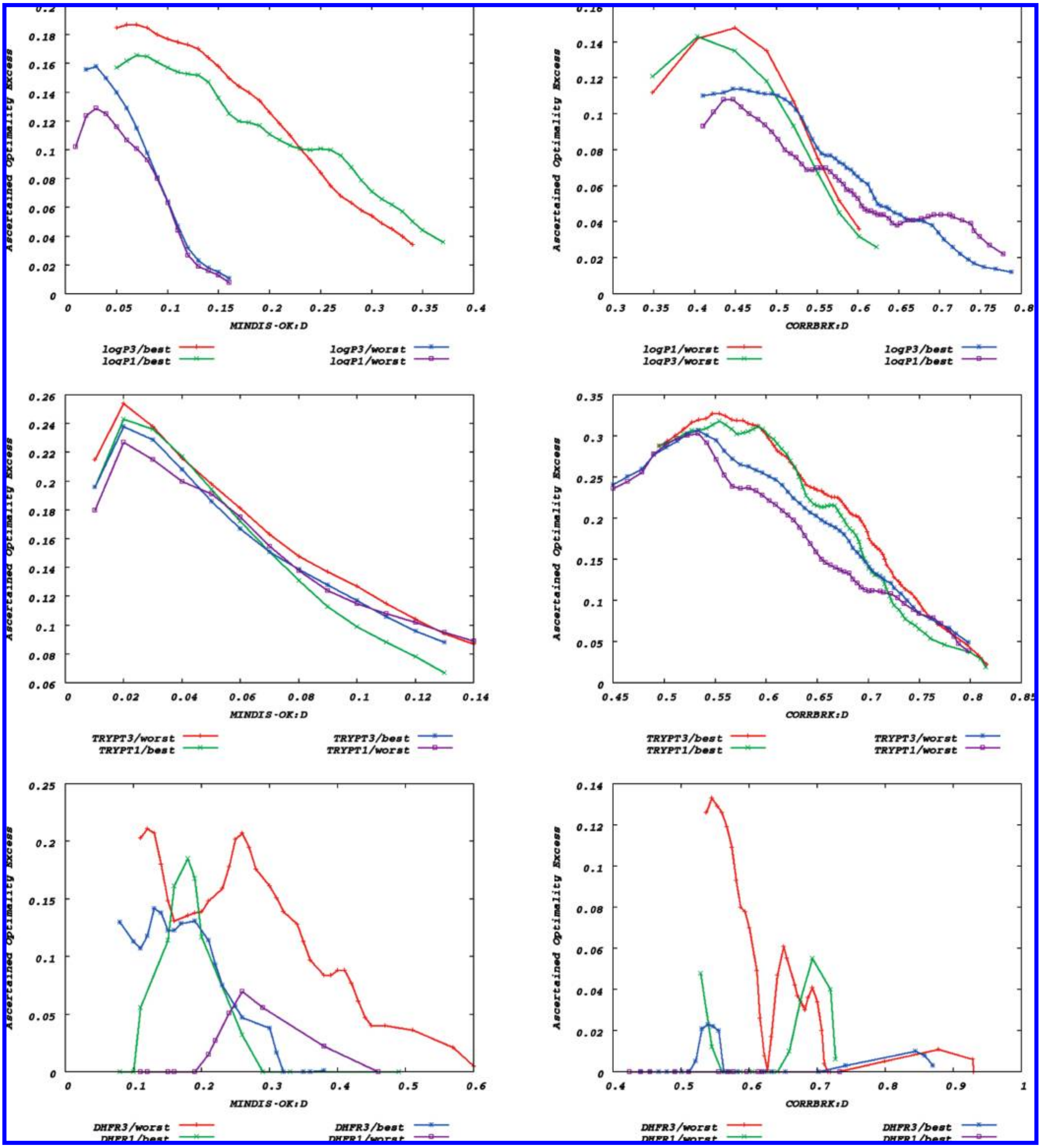

Figure 6. Ξ as a function of two untrustworthiness thresholds (*MINDIS-OK:D* and *CORRBRK:D*) in three study cases (only best and worst models shown), drawn with respect to two disjoined subsets amounting to one and respectively three-quarters of the entire set and accordingly labeled "1" and "3".

4. CONCLUSIONS

The application of NB monitoring tools to the applicability issue of QSAR models to predict external compounds lead to a unified conceptual framework, supporting both the use of untrustworthiness scores for prediction prioritization purposes as well as the classical definition of AD borders. The propensities of meaningful prioritization of the trustworthiness of predictions according to various AD metrics can be straightforwardly read from Ξ*(I)* curves. It was shown that the variance of consensus predictions based on large families of SQS models is the most robust trust indicator explored in this work, followed by outlier-free dissimilarity measures between the external predicted compound and its nearest neighbor(s) among the properly fitted training set compounds. Another original untrustworthiness score discussed here, the correlation breakdown count, appeared to be more robust than the classical dissimilarity measures toward all training set compounds, irrespective of their outlier

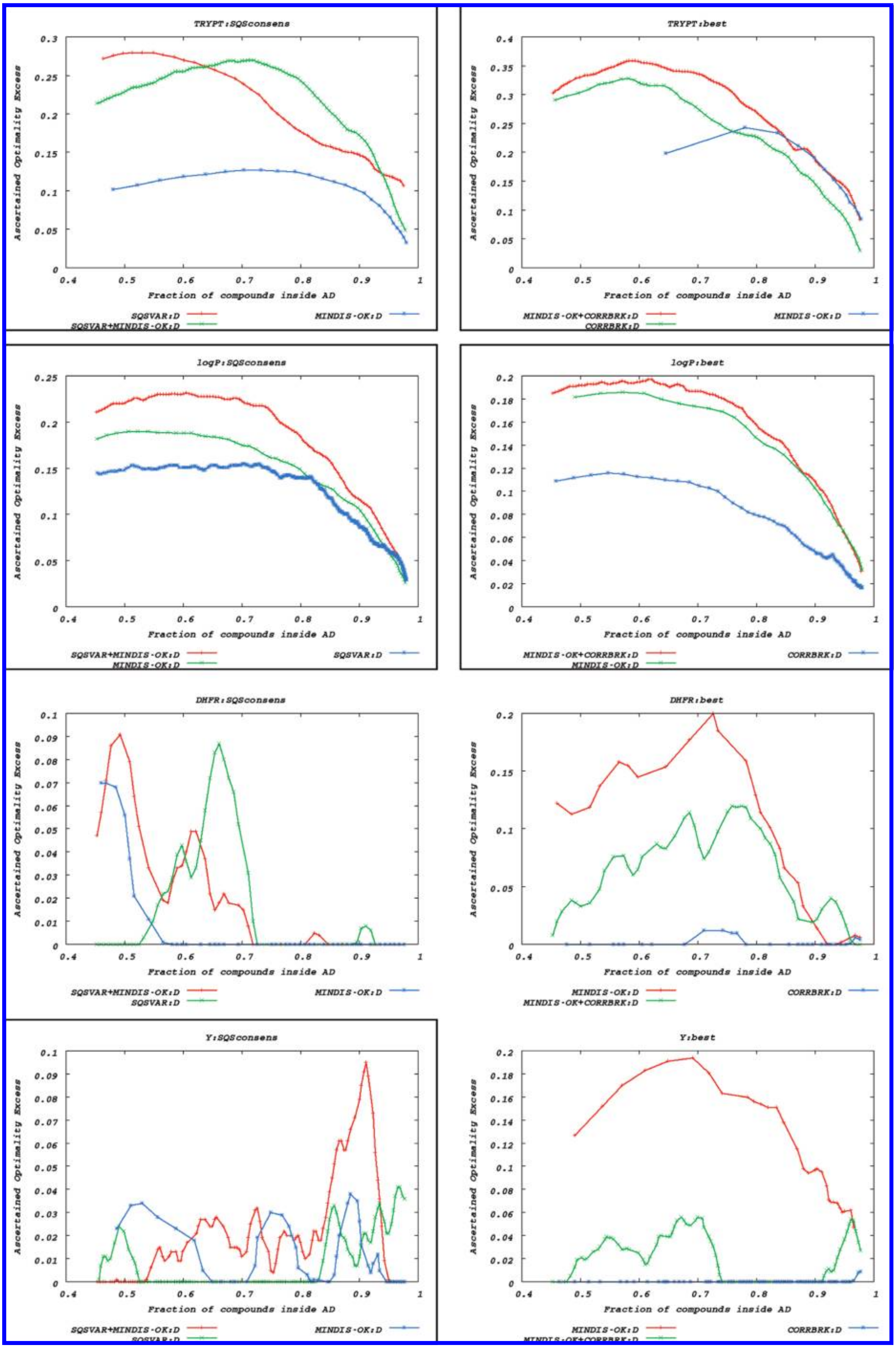

Figure 7. Comparative Ξ*(I)* plots of AD consensus metrics and their parents, for various case studies and models (for SQSconsens models, *MINDIS-OK:D* has been mixed with *SQSVAR*, while for "best" equations *MINDIS-OK+CORRBRK:D* was used). Situations evidencing a strong synergy of the mixed metrics are framed.

status with respect to the prediction model. Being located in the neighborhood of a molecule failing to have its property well reproduced although it participated in model training or internal validation is not good news in terms of predictability. Detection of compounds with descriptor values outside the chemical space "boxes" populated by training molecules was the less sophisticated—and the less effective—AD metric discussed here.

Somehow expectedly, allowing untrustworthiness scores to focus on molecular descriptors entering the model rather than exploiting all the candidate descriptors provided as input to the model builder.

Untrustworthiness-based ranking allows the user to prioritarily pick, from a virtual library, compounds with probably valid predictions according to a given model but does not impose a hard limit beyond which no prediction needs to be trusted. None of the studied AD metrics seem to admit some "joker" cutoff value applicable in all situations, but model-specific AD borders can be easily fitted, i.e. chosen equal to the untrustworthiness score which maximizes Ξ. If the external validation set of the QSAR model, serving as training set of the AD model, is large and diverse enough, this optimal AD border parameter can be used as rule-of-the-thumb for acceptation/rejection of further predictions.

Eventually, it was shown that nonparametric, rank-based mixing of nonredundant AD metrics of comparable quality is highly likely to exhibit synergy effects, leading to consensus untrustworthiness scores being more robust than any of their parents.

While previous work in the domain tended to focus on specific AD approaches, the current paper introduced a general framework for the unified benchmarking thereof. Its strongest feature is the objective definition of the locally optimal untrustworthiness cutoff, for any considered AD metric, with respect to given training/validation sets. The objective selection of this cutoff implicitly provides an answer to the recurrent problem and major source of confusion in many of the previous AD-related studies, where an AD definition typically featured two independent degrees of freedom, which had to be discussed and evaluated separately: (1) the choice of the AD metric and (2) the choice of the untrustworthiness cutoff. Given an objective way to fix (2), the discussion may now focus on the key problem (1): a straightforward benchmarking exercise suffices to propose the best suited AD scheme for each situation.

As there are no generic rules to predict what type of descriptors or what learning method is best suited to build a QSAR model, there are no reasons to expect that the winning AD definition could be foretold. Albeit only the modeling of continuous properties was reported here, the methodology can be applied to categorical models as well (work in progress). Furthermore, if, supposedly, all the molecules of the validation set were well predicted or (more likely) mispredicted by the model, then any thereon based AD definition attempt is bound to fail.

It should not be forgotten that AD definitions are, themselves, nothing else but empirical models trained on hand of external validation sets which therein play the role of a learning set. Reducing the number of degrees of freedom to the essential one—the choice of the AD metric—is of paramount importance. The space of possible AD metrics is, however, huge. It may therefore not be excluded that one of these happens to explain "by pure chance" the differences between well-predicted and poorly predicted validation set members. The goal of the current work was not to define absolute Applicability Domains for the studied QSAR cases but to encourage chemoinformaticians to systematically apply this AD paradigm in their studies. For each of the studied cases, some even better AD delimitations might have been "fitted"—and then likely invalidated upon confrontation with novel molecules. The only (weak) guarantee the current methodology may offer against overfitting is its consistently minimalistic policy of fitting and/or parameter guessing. Introducing meta−meta-models to predict the AD of an AD model makes little sense. Only large scale applications by the entire community, including virtual screening followed by experimental validation, may show whether this formalism will be found useful—in the sense that therewith defined AD models will be proven, statistically speaking, more robust and more difficult to invalidate than others. As far as we can tell, based on diverse and as large as possible data sets, selected for no other reason than the intrinsic interest in developing predictive models for those properties, we did not yet manage to disprove the above working hypothesis.

Supporting Information Available: For five of the seven studied targets, the provided.zip file includes, on one hand, predicted property values for external validation compounds according to "best", "worst", and "SQSconsensus" models, and, on the other, associated AD metric scores. Unzipping will thus extract multiple directories, each containing Unix text files. For example, extraction will create "logP/best.pred" containing, for each logP model validation compound, the current compound number (first column), its predicted property value according to the considered model "best" (column 2), and its experimental property (**last** column. Note that a prediction variance column may be inserted between predicted values in column 2 and experimental values, which thus form column 4 if this variance score is present, which is always the case in SQSconsens.pred files. Since best.pred and worst.pred are each outputs of a single QSAR model, there is no prediction variance—if a third column is present, depending on the used prediction software version, then it is filled with zeroes). Also,.pred files contain, at the end, #-prefixed comment lines, featuring some statistical parameters: use grep -v # to remove these.

AD metric scores are given in.admet files, labeled METRIC.MODEL.admet, where METRIC is one of the herein considered scores, MODEL stands for the specific model ("best", "worst", SQSconsens, or "nomodel" for the model-independent versions of the metrics). They are one-column files containing the raw trustworthiness scores, in the order matching the.pred files. The following Unix command line example pastes the considered trustworthiness scores and prediction values, sorts the merged table according to increasing untrustworthiness and uses an awk command to calculate the prediction error for each compound, by subtracting predicted (now column $3) from experimental (column $NF):

```
paste TRYP/AVGDIS.worst.admet TRYP/worst.pred|grep −v #| sort −g|awk '{err=$3-$NF;if (err < 0) err=-err;print $1,$3,$NF,err}'.
```

It creates a four-column output consisting of untrustworthiness, predicted property, experimental property, and absolute prediction error, sorted by the former. Ideally, this sorting should deny large prediction errors to appear among the top lines of the list. You may pick the top *N* most trustworthy predictions of the list:

```
paste TRYP/AVGDIS.worst.admet TRYP/worst.pred|grep −v #| sort −g|awk '{err=$3-$NF;if (err < 0) err=-err;print $1,$3,$NF,err}'|head −N
```

and calculate the average prediction error (average of column nr. 4). In as far "potentially false outsiders" are not the dominant contribution in the set, this average should—statistical fluctuations notwithstanding—increase with set size *N*. Of course, the user may reproduce optimality calculations on hand of provided data. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D. Varnek, A., Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(2) Stanforth, R. W.; Kolossov, E.; Mirkin, B. A measure of domain of applicability for QSAR modelling based on intelligent K-means clustering. *QSAR Comb. Sci.* **2007**, *26*, 837–844.

(3) Bonachera, F.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2008**, *48* (2), 409–425.

(4) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; WILEY-VCH Verlag GmbH: Weinheim, 2004; pp117−137.

(5) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 155–173.

(6) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA, Altern. Lab. Anim.* **2005**, *33* (5), 445–459.

(7) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *ATLA, Altern. Lab. Anim.* **2004**, *44* (6), 1912–1928.

(8) Fernandez Pierna, J. A.; Wahl, F.; de Noord, O. E.; Massart, D. L. Methods for Outlier Detection in Prediction. *Chem. Int. Lab. Syst.* **2002**, *63*, 27–39.

(9) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADMET Predictions. *Drug Discovery Today* **2006**, *11* (15/16), 700–707.

(10) Bruneau, P.; McElroy, N. R. LogD7.4 Modeling Using Bayesian Regularized Neural Networks. Assessment and Correction of the Errors of Prediction. *J. Chem. Inf. Model.* **2006**, *46* (3), 1379–1387.

(11) Papa, E.; Villa, F.; Gramatica, P. Statistically Validated QSARs, Based on Theoretical Descriptors for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales Promelas (fathead minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256–1276.

(12) Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and Validation of k-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates. *J. Med. Chem.* **2003**, *46* (14), 3013–3020.

(13) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983–996.

(14) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Ob̈erg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733–1746.

(15) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22* (1), 69–77.

(16) Eijkel, G. C. v. d.; Jan, C. A. v. d. L.; Backer, E., A Modulated Parzen-Windows Approach for Probability Density Estimation. In *Proceedings of the Second International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data*; Springer-Verlag: London, UK, 1997; pp 479−489.

(17) Fukumizu, K.; Watanabe, S. In Probabililty Density Estimation by Regularization Method, *Proceedings of the International Joint Conference on Neural Networks*; Nagoya, IEEE: Nagoya, 1993; pp 1727−1730.

(18) Schioler, H.; Hartmann, U. Mapping Neural Network Derived from the Parzen Window Estimator. *Neural Networks* **1992**, *5* (6), 903–909.

(19) Duda, R.; Hart, P. *Pattern Classification and Scene Analysis;* John Wiley & Sons: New York, 1973.

(20) Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.

(21) Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful "in silico" design of new efficient uranyl binders. *Solvent Extr. Ion Exch.* **2007**, *25* (4), 433–462.

(22) Marcou, G. ISIDA Predictor. http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi (accessed May 2009).

(23) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39* (16), 3049–3059.

(24) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces - A Benchmark for Neighborhood Behavior Assessment of Different In Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691–698.

(25) Horvath, D.; Jeandenans, C. Neighborhood Behavior of In Silico Structural Spaces with respect to In Vitro Activity Spaces - A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680–690.

(26) Papadatos, G.; Cooper, A. W. J.; Kadirkamanathan, V.; Macdonald, S. J. F.; McLay, I. M.; Pickett, S. D.; Pritchard, J. M.; Willett, P.; Gillet, V. J. Analysis of Neighborhood Behavior in Lead Optimization and Array Design. *J. Chem. Inf. Model.* **2008**, DOI:10.1021/ci800302g.

(27) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation - How much effort may the mining for successful QSAR models take. *J. Chem. Inf. Model.* **2007**, *47*, 927–939.

(28) Bonachera, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy Tricentric Pharmacophore Fingerprints. 1 - Topological Fuzzy Pharmacophore Triplets and adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **2006**, *46*, 2457–2477.

(29) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191–198.

(30) Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M., Jr.; Solov'ev, V. P.; Varnek, A. QSAR modeling of blood:air and tissue:air partition coefficients using theoretical descriptors. *Bioorg. Med. Chem.* **2005**, *13*, 6450–6463.

(31) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(32) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.

(33) ChemAxon Screen User Guide. http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Feb 2009).

(34) ChemAxon pKa Calculator Plugin. http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (accessed Feb 2009).

(35) SRC PHYSPROP database. http://www.srcinc.com/what-we-do/product.aspx?id=133&terms=Physprop (accessed Feb 2009).

(36) QSARWorld Percentage of Human Oral Absoption. http://www.qsarworld.com/qsar-datasets.php?mm=5 (accessed Feb 2009).

(37) Sutherland, J. J.; OBrien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(38) Schuller, A.; Fechner, U.; Renner, S.; Franke, L.; Weber, L.; Schneider, G. A pseudo-ligand approach to virtual screening. *Comb. Chem. High Throughput Screening* **2006**, *9* (5), 359–364.

(39) Schuller, A.; Schneider, G. Identification of Hits and Lead Structure Candidates with Limited Resources by Adaptive Optimization. *J. Chem. Inf. Model.* **2008**, *48* (7), 1473–1491.

CI9000579