JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

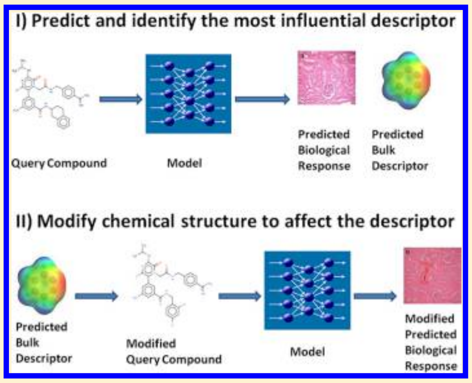
Localized Heuristic Inverse Quantitative Structure Activity Relationship with Bulk Descriptors Using Numerical Gradients

Jonna Stålring,*[†] Pedro R. Almeida,[‡] Lars Carlsson,[†] Ernst Helgee Ahlberg,[†] Catrin Hasselgren,[†] and Scott Boyer[†]

[†]Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Pepparedsleden 1, 431 53 Mölndal, Sweden

[‡]EngInMotion Ltd, Avenida Infante D. Henrique, n. 145, 3510-070 Viseu, Portugal

S *Supporting Information*


I) Predict and identify the most influential descriptor
Query Compound
Model
Predicted Biological Response
Predicted Bulk Descriptor
II) Modify chemical structure to affect the descriptor
Predicted Bulk Descriptor
Modified Query Compound
Model
Modified Predicted Biological Response

ABSTRACT: State-of-the-art quantitative structure−activity relationship (QSAR) models are often based on nonlinear machine learning algorithms, which are difficult to interpret. From a pharmaceutical perspective, QSARs are used to enhance the chemical design process. Ultimately, they should not only provide a prediction but also contribute to a mechanistic understanding and guide modifications to the chemical structure, promoting compounds with desirable biological activity profiles. Global ranking of descriptor importance and inverse QSAR have been used for these purposes. This paper introduces localized heuristic inverse QSAR, which provides an assessment of the relative ability of the descriptors to influence the biological response in an area localized around the predicted compound. The method is based on numerical gradients with parameters optimized using data sets sampled from analytical functions. The heuristic character of the method reduces the computational requirements and makes it applicable not only to fragment based methods but also to QSARs based on bulk descriptors. The application of the method is illustrated on congeneric QSAR data sets, and it is shown that the predicted influential descriptors can be used to guide structural modifications that affect the biological response in the desired direction. The method is implemented into the AZOrange Open Source QSAR package. The current implementation of localized heuristic inverse QSAR is a step toward a generally applicable method for elucidating the structure activity relationship specifically for a congeneric region of chemical space when using QSARs based on bulk properties. Consequently, this method could contribute to accelerating the chemical design process in pharmaceutical projects, as well as provide information that could enhance the mechanistic understanding for individual scaffolds.

■ INTRODUCTION

Nonlinear machine learning (ML) algorithms, applied to quantitative structure−activity relationships (QSAR), are often labeled as black box models offering limited insight into the underlying physical, biochemical, and structural processes giving rise to the biological response.[1] From a practical pharmaceutical perspective, the primary objective of a QSAR is to influence the chemical design such that molecules propagated through the discovery pipeline have optimized biological activity profiles. Such optimization of chemical structures benefits from as much information as possible from the QSAR model. Hence, it is desirable that a QSAR model provides not only a model prediction but also suggestions on how to modify the predicted property in a desired direction.

Assessment of descriptor importance with respect to the response variable is well established within the QSAR community as a method to provide mechanistic insight.[2] ML methods such as partial least-squares,[3] random forest,[4] and artificial neural networks,[5] as well as entropy (Gini index) and near neighbor based (ReliefF) methods, are used to rank the descriptor importance for the full data set. However, such global ranking of descriptors in a structurally diverse data set might not be accurate for individual compounds predicted by a nonlinear ML algorithm, and ultimately the structure activity relationship (SAR) should be understood locally for each scaffold.

Inverse QSAR uses structural fragments of the data set and recombines them into new molecules under constraint equations, assuring chemically relevant structures and potentially more desirable biological properties. The new compounds are predicted by the corresponding QSAR model to assess the modeled biological response. Several approaches have been suggested for inverse QSAR such as genetic algorithms,[6] simulated annealing,[7] enumeration,[8] and mathematical programming.[9] However, these methods are struggling to overcome, for example, computational costs, inaccuracies originating from nonexhaustive searches, unreliable predictions, and local minima.

This work introduces localized heuristic inverse QSAR, representing an extension of the work of Franke et al.[10] and Carlsson et al.,[11] later followed by Marcou et al.,[12] for QSARs

Received: May 10, 2013
Published: July 12, 2013

based on bulk descriptors. Rather than suggesting a range of new compounds relevant to the full data set, as in inverse QSAR, the method provides specific guidance on modifications relevant to individual predicted compounds. The most influential descriptor, in a region localized around a given QSAR prediction, is identified, and the method predicts that modifying the structure to affect this particular descriptor will impact the biological response the most in a desired direction. In contrast to conventional inverse QSAR, users will have to rely upon their own chemical intuition in changing the compound structure to affect the most influential descriptor and thereby affecting the biological response. This methodology offers a variant of inverse QSAR applicable not only to QSARs based on fragments but also to QSARs based on bulk descriptors. In addition to affecting the design, knowledge of bulk descriptors with a substantial impact on the response of a particular compound might offer insights into the mechanistic driving forces of compounds of a particular scaffold. Finally, heuristic inverse QSAR is computationally inexpensive compared to the inverse QSAR problem.

This paper defines the mathematical framework, based on numerical gradients, for the identification of influential descriptors. The method is validated using models based on data sets sampled from analytical functions, which allows for a comparison between the numerical and analytical gradients in the point of prediction. In addition, application of the method to QSAR data illustrates its utility in the pharmaceutical design process and exemplifies successful structural modifications within congeneric chemical series, altering target activity in a desired direction.

■ MATERIALS AND METHODS

Localized heuristic inverse QSAR will predict which descriptor influences the QSAR prediction of a given compound the most. Once the descriptor is identified, the chemical structure can be modified to change the value of the predicted descriptor, which in turn will affect the biological activity, as illustrated in Figure 1. The descriptor considered the most influential of a QSAR prediction is identified by defining a numerical gradient in the point of prediction. For some nonlinear ML algorithms, it is possible to derive an analytical gradient. However, this method intends to be generally applicable to any algorithm returning a decision function value (DFV), and it is therefore not possible to rely upon the existence of an analytical gradient. Furthermore, the method accommodates models based on both discrete and continuous descriptors. The identification of the most influential descriptor uses the relative sizes of the partial derivatives of the DVF with respect to each descriptor. Numerical partial derivatives approximate the analytical partial derivatives of the model response, and the greatest absolute partial derivative corresponds to the descriptor with the greatest influence on the DFV in the point of prediction.

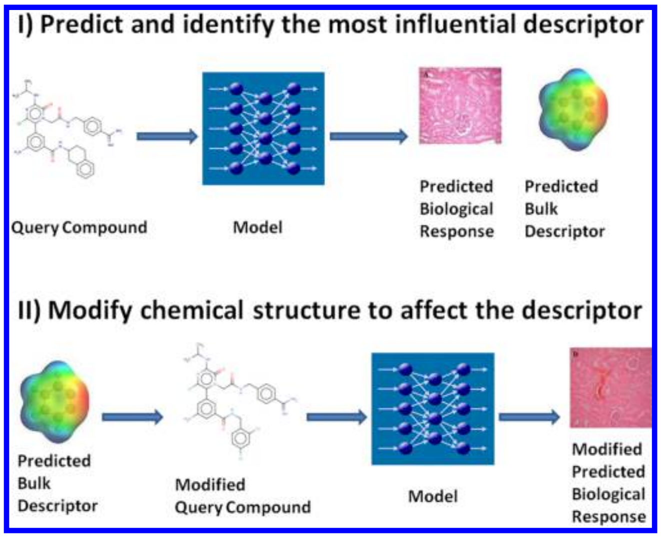

I) Predict and identify the most influential descriptor
Query Compound
Model
Predicted Biological Response
Predicted Bulk Descriptor
II) Modify chemical structure to affect the descriptor
Predicted Bulk Descriptor
Modified Query Compound
Model
Modified Predicted Biological Response

Figure 1. Using localized heuristic inverse QSAR in the chemical design process.

Mathematical Framework and Pseudoalgorithm. For continuous descriptors, the numerical partial derivatives are obtained by calculating the DFVs in two points centered around the original model prediction. More formally, let $(\mathbf{x}_1, y_1)$, ..., $(\mathbf{x}_N, y_N)$ denote the examples of the training set. Each y_n $(n = 1, ..., N)$ is the response of the example with the descriptor vector $\mathbf{x}_n$, where x_{ni} is the value of the ith descriptor of example n. Let $f(\mathbf{x}_n)$ denote the DFV of example n, and let g_{ni}^{C} be the numerical partial derivative of descriptor x_i, evaluated in the point $\mathbf{x}_n$. The derivative is calculated by altering the descriptor value x_{ni} by a descriptor specific step, h_i

$$g_{ni}^{C} = \frac{f(\mathbf{x}_n + h_i\mathbf{e}_i) - f(\mathbf{x}_n - h_i\mathbf{e}_i)}{2h_i} \tag{1}$$

where $\mathbf{e}_i$ is the unit vector with all elements equal to zero except the ith element.

Numerical pseudoderivatives are calculated for discrete descriptors by changing the value of the descriptor x_i into all M occurring values and selecting the greatest change in model prediction, out of the resulting M model predictions, to quantify the pseudoderivative. Hence, the set of absolute differences in predictions, S_M, between predicting the original descriptor vector $\mathbf{x}_n$ with the value x_{il} of the ith descriptor and predicting the same vector with the descriptor value changed into its kth value, x_{ik}

$$S_M = \{|f(\mathbf{x}_n - x_{il}\mathbf{e}_i + x_{ik}\mathbf{e}_i) - f(\mathbf{x}_n)|\}_{k=1}^{M} \tag{2}$$

is used to define the numerical pseudoderivative, g_{ni}^{D}

$$g_{ni}^{D} = \max(S_M) \tag{3}$$

of the discrete descriptor i.

Partial numerical derivatives are calculated with respect to all descriptors and compiled in a vector, $\mathbf{g}$

$$\mathbf{g} = \mathbf{g}^{C} + \mathbf{g}^{D} \tag{4}$$

where the elements of $\mathbf{g}^{C}$ (eq 1) are zero for all discrete attributes and, conversely, all elements of continuous attributes are zero in $\mathbf{g}^{D}$ (eq 2). The elements of $\mathbf{g}$ are partitioned into these two groups because the numerical derivatives with respect to discrete and continuous descriptors have to be calculated differently. Hence, they cannot be compared, and the descriptors with the greatest absolute derivatives within each group are identified and reported to the user. In addition to the descriptor, the direction in which the descriptor value should be changed to achieve a user defined impact on the response variable (increase or decrease) is given. The outline of the algorithm below defines in greater detail the calculation of the numerical partial derivatives.

1. For each descriptor (x_i):

(a) If the descriptor is continuous:

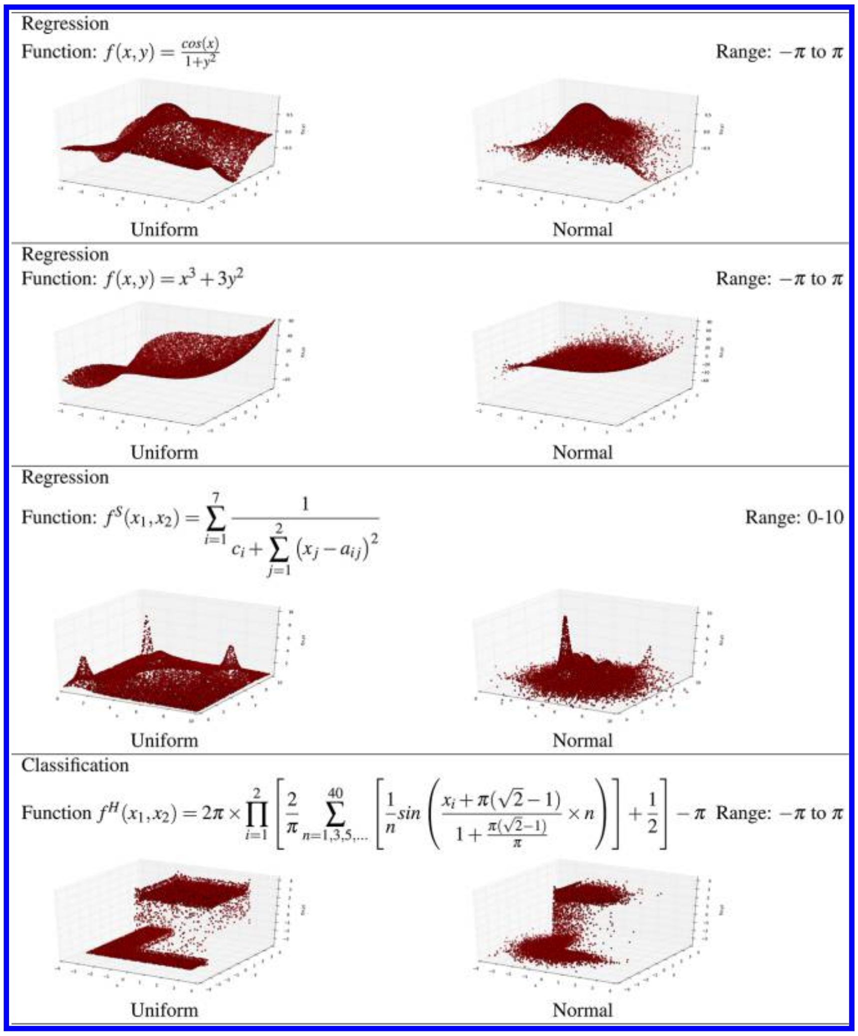


Figure 2. Graphs displaying the 16 000 uniformly and normally sampled data points generated from the given two dimensional functions and constituting the training sets used to tune the method.

   i. Calculate the DFV in the point $\mathbf{x}_n$ with the value of x_{ni} decreased by h_i, $f(\mathbf{x}_n - h_i\mathbf{e}_i)$
   ii. Calculate the DFV in the point $\mathbf{x}_n$ with the value of x_{ni} increased by h_i, $f(\mathbf{x}_n + h_i\mathbf{e}_i)$
   iii. Calculate the partial derivative, g_{ni}^C as defined in eq 1
   iv. Store the partial derivative in the vector, $\mathbf{g}^C$

(b) If the descriptor is discrete:
   i. Calculate the DFV with the original descriptor values, $f(\mathbf{x}_n)$
   ii. For each value of x_i, x_{ik}
      A. Calculate the DFV in point $\mathbf{x}_n$ with the modified value of x_i, $f(\mathbf{x}_n - x_{il}\mathbf{e}_i + x_{ik}\mathbf{e}_i)$
      B. Calculate the difference in model predictions as defined in eq 2
   iii. Choose the partial pseudoderivative for x_i to be equal to the highest absolute value among the M values obtained in step ii, as defined in eq 3.
   iv. Store the partial pseudoderivative in the vector, $\mathbf{g}^D$.

2. If the model is based upon both discrete and continuous descriptors, there will be two vectors ($\mathbf{g}^C$ and $\mathbf{g}^D$) with numerical partial derivatives. The descriptors with the greatest partial derivatives of each vector are presented as the most influential.

Definition of the Step Size. For continuous attributes, a step length must be selected for the partial numerical derivatives (eq 1). This method sets the step size individually for each descriptor by relating the step size to the variation in descriptor values of each descriptor. Hence, the step length, h_i, is defined as the variance of the descriptor x_i, estimated based on all the corresponding values in the training set, times a coefficient, C,

$$h_i = \sigma_{xi} \times C \qquad (5)$$

While the estimated variance is calculated for each data set, the coefficient is optimized in this study to be generally applicable to all data sets. As the pseudoderivatives of the discrete descriptors are calculated based on all possible values of each descriptor, no step size is required.

Tuning for Flat Areas. In flat areas, where the change in function values are small with respect to all descriptors, the relevance of the descriptor corresponding to the greatest partial derivative is questionable. Hence, rather than reporting numerical differences, the method should restrain from identifying a most influential descriptor. Using a gradient threshold reduces the risk of irrelevant numerical differences influencing the chemical design process. Consequently, the threshold is set to a value, ε, below which a partial derivative should be considered zero.

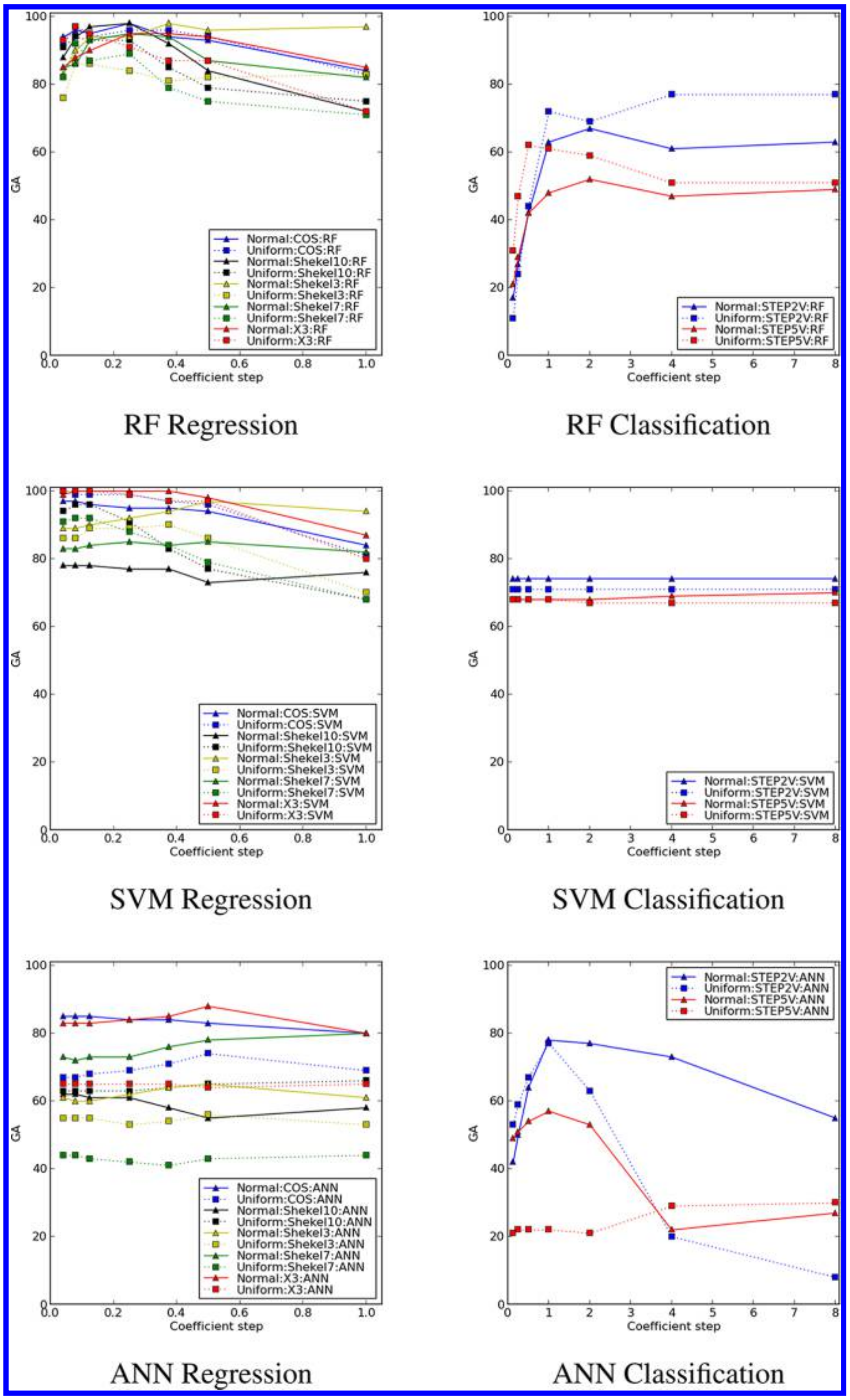
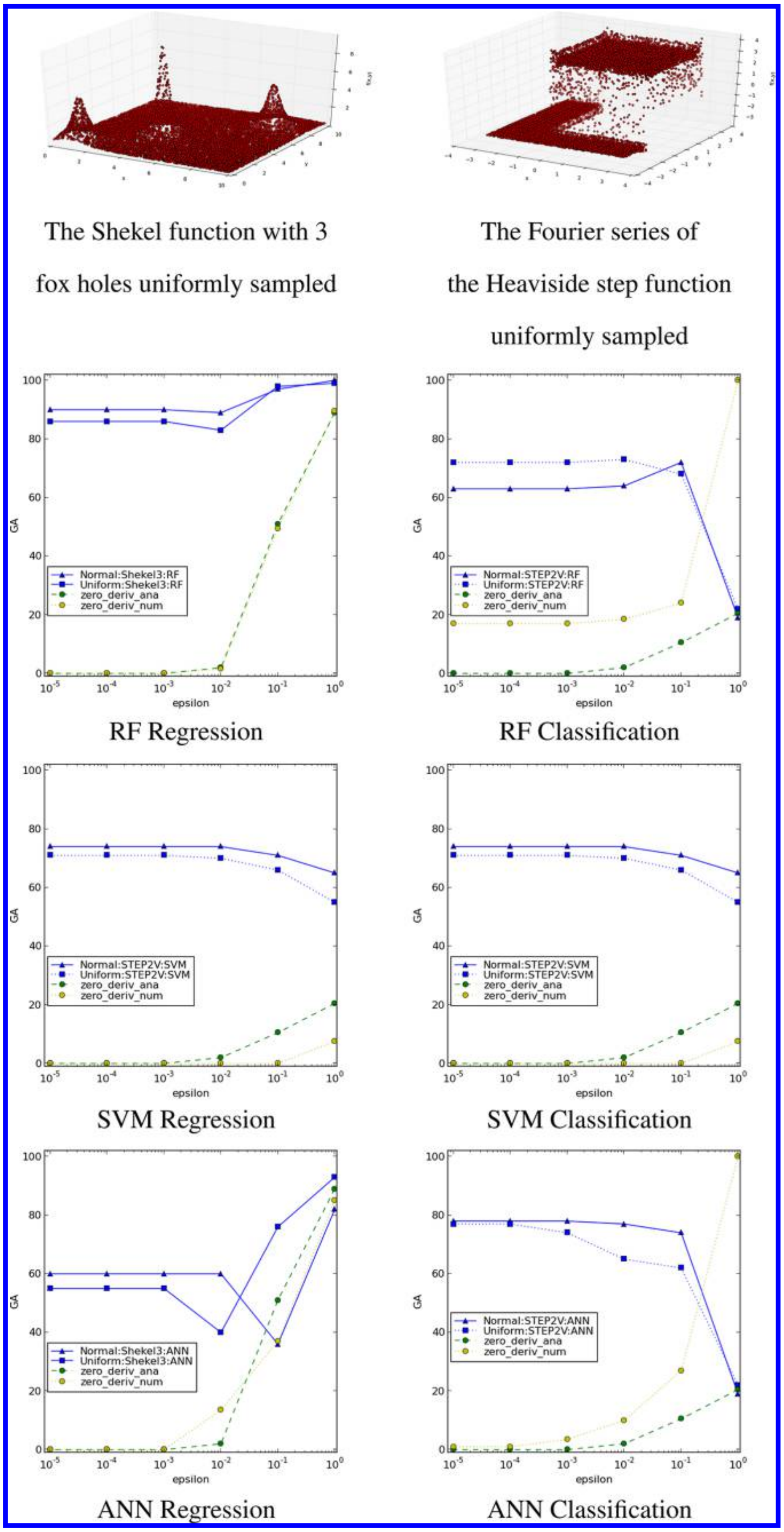


Figure 3. The gradient accuracy of regression and classification models versus coefficient values, *C*.

■ RESULTS AND DISCUSSION

This section presents the selection of the parameters of the numerical derivatives, performed by maximizing the concordance between the numerical and corresponding analytical derivatives for models based on data sets sampled from analytical functions. The numerical derivatives are further assessed on multidimensional analytical functions of dimensionalities in the range of QSAR models based on bulk descriptors. Finally, localized heuristic inverse QSAR is illustrated on three QSAR data sets, and successful examples of where the identification of influential descriptors could aid in the design process are shown. The analytical validation quantifies the extent to which the numerical gradients accurately describe the local dependence of the response on the various descriptors. Such quantitative assessment is not possible for the QSAR data as the underlying function is unknown. Hence, the QSAR applications complement the quantitative analytical analysis by providing a qualitative assessment in the applied setting.

Analytical Validation. To tune the method and to assess the quality of the descriptors identified as the most influential, models are built on data sets sampled from a set of analytical



The Shekel function with 3 fox holes uniformly sampled

The Fourier series of the Heaviside step function uniformly sampled

RF Regression

RF Classification

SVM Regression

SVM Classification

ANN Regression

ANN Classification

Figure 4. The gradient accuracy of regression and classification models versus derivative threshold values ε.

functions. Using analytical functions allows for validation of the numerical gradients by comparison with the analytical gradients. Hence, let $\pounds_{g^C}$ denote an ordered list of the absolute values of the numerical gradient $\mathbf{g}^C$, while $\pounds_{\nabla f(\mathbf{x})}$ is an ordered

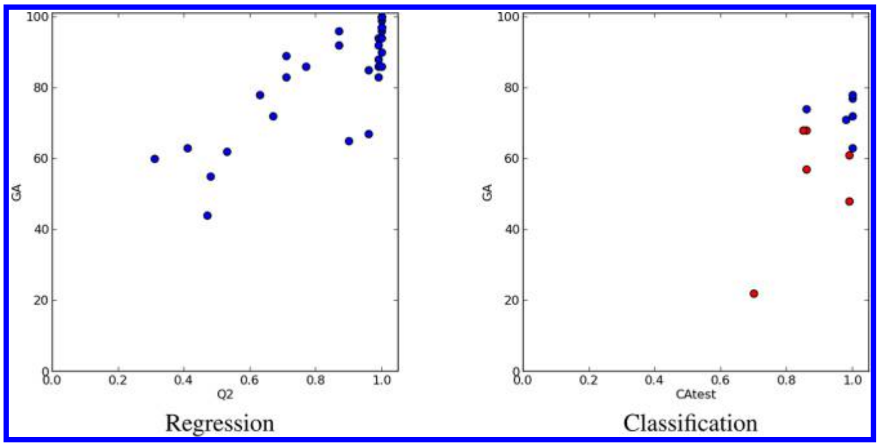


Figure 5. The gradient accuracy versus model accuracy for models based on the two dimensional reference functions.

Table 1. Model and Gradient Accuracies for Multidimensional Functions with the Selected Parameters C = 0.08 (Regression) C = 1.00 (Classification) and $\varepsilon = 10^{-5}$

ML	ML type	equation	Dim	Dist	Q2/CA$_{\text{test}}$	GA (%)	GA enrichment	GA$_2^N$(%)	GA$_3^N$(%)	GA$_2^A$(%)
RF	regression	eq 7, 7 fox holes	10	normal	0.90	45	4.5	70	84	69
			10	uniform	0.84	34	3.4	59	69	61
			20	normal	0.71	24	4.8	42	47	49
			20	uniform	0.68	14	2.8	30	37	29
	classification	eq 8	10	normal	1.00	52	5.2	63	70	64
			10	uniform	0.99	50	5.0	58	65	59
			20	normal	0.98	49	9.8	70	80	75
			20	uniform	0.99	45	9.0	50	56	54
SVM	regression	eq 7, 7 fox holes	10	normal	1.00	96	9.6	100	100	100
			10	uniform	1.00	89	8.9	97	98	98
			20	normal	0.98	90	18	97	99	98
			20	uniform	0.98	72	14.4	93	98	93
	classification	eq 8	10	normal	0.89	71	7.1	91	93	85
			10	uniform	0.78	64	6.4	77	79	82
			20	normal	0.90	74	14.8	91	96	88
			20	uniform	0.67	41	8.2	66	68	58

list of absolute values of the partial derivatives constituting the analytical gradient $\nabla f(\mathbf{x})$. The influential descriptor is considered correctly predicted if the largest element of $£_{\mathrm{g}}^{\mathrm{c}}$ is the largest element of $£_{\nabla f(\mathrm{x})}$ and the signs of these numerical and analytical derivatives are the same. Aside from the training set, an external test set is sampled from the functions and predicted by the model. For each prediction, the most influential descriptor of the function in the test point is identified. The overall quality of the identified descriptors is quantified by a gradient accuracy (GA)

$$\mathrm{GA} = \frac{T_{\mathrm{corr}}}{T} \tag{6}$$

where T is the total number of test instances and T_{corr} is the number of test instances for which the correct descriptor is identified. In addition, gradient accuracies are assessed while considering not only the greatest but also the second and third greatest numerical partial derivatives. Hence, if any of the two or three top ranked numerical partial derivatives in $£_{\mathrm{g}}^{\mathrm{c}}$ correspond to the top ranked analytical partial derivative in $£_{\nabla f(\mathrm{x})}$ and if the signs are the same, the instance is considered predicted with the correct descriptor. These GAs are denoted $\mathrm{GA}_2^{\mathrm{N}}$ and $\mathrm{GA}_3^{\mathrm{N}}$, respectively. Conversely, a more inclusive assessment of the top ranked numerical derivative is obtained by considering it correct if it corresponds to any of the two top ranked analytical derivatives. The corresponding GA is denoted $\mathrm{GA}_2^{\mathrm{A}}$.

Figure 2 defines the set of two-dimensional analytical functions used to tune the step length (h_i of eq 1) and the derivative threshold value of the method.

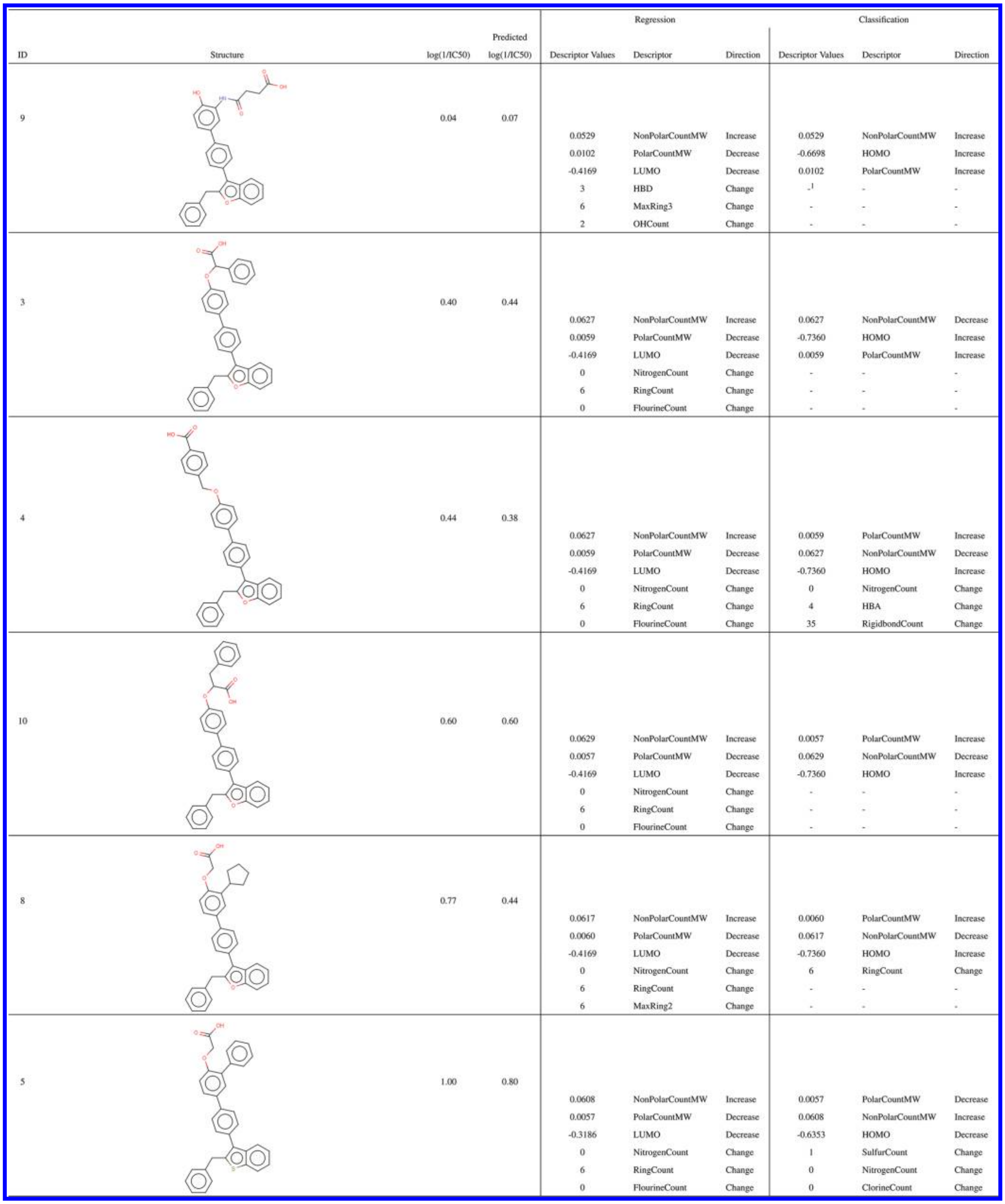
In addition, the figure displays graphs of the uniformly and normally sampled data sets used to build the corresponding regression or classification models. Once the parameters of the method have been selected, the multidimensional functions defined in eqs 7 and 8

$$f^{S}(x_1, ..., x_n) = \sum_{i=1}^{m} \frac{1}{c_i + \sum_{j=1}^{n}(x_j - a_{ij})^2} \tag{7}$$

$$f^{H}(x_1, ..., x_n) = 2\pi \times \prod_{i=1}^{n}\left[\frac{2}{\pi}\sum_{n=1,3,5,...}^{40}\left[\frac{1}{n}\sin\left(\frac{x_i + \alpha}{1 + \alpha/\pi} \times n\right)\right] + \frac{1}{2}\right] - \pi$$

$$\alpha = \pi(2^{n-1/n} - 1) \tag{8}$$

Table 2. The Predicted Most Influential Descriptors within One HPTP Cluster of Eight Compounds, Together with Activity and the Values of the Descriptors



				Regression			Classification		
			Predicted						
ID	Structure	log(1/IC50)	log(1/IC50)	Descriptor Values	Descriptor	Direction	Descriptor Values	Descriptor	Direction
9		0.04	0.07	0.0529	NonPolarCountMW	Increase	0.0529	NonPolarCountMW	Increase
				0.0102	PolarCountMW	Decrease	-0.6698	HOMO	Increase
				-0.4169	LUMO	Decrease	0.0102	PolarCountMW	Increase
				3	HBD	Change	-[1]	-	-
				6	MaxRing3	Change	-	-	-
				2	OHCount	Change	-	-	-
3		0.40	0.44	0.0627	NonPolarCountMW	Increase	0.0627	NonPolarCountMW	Decrease
				0.0059	PolarCountMW	Decrease	-0.7360	HOMO	Increase
				-0.4169	LUMO	Decrease	0.0059	PolarCountMW	Increase
				0	NitrogenCount	Change	-	-	-
				6	RingCount	Change	-	-	-
				0	FlourineCount	Change	-	-	-
4		0.44	0.38	0.0627	NonPolarCountMW	Increase	0.0059	PolarCountMW	Increase
				0.0059	PolarCountMW	Decrease	0.0627	NonPolarCountMW	Decrease
				-0.4169	LUMO	Decrease	-0.7360	HOMO	Increase
				0	NitrogenCount	Change	0	NitrogenCount	Change
				6	RingCount	Change	4	HBA	Change
				0	FlourineCount	Change	35	RigidbondCount	Change
10		0.60	0.60	0.0629	NonPolarCountMW	Increase	0.0057	PolarCountMW	Increase
				0.0057	PolarCountMW	Decrease	0.0629	NonPolarCountMW	Decrease
				-0.4169	LUMO	Decrease	-0.7360	HOMO	Increase
				0	NitrogenCount	Change	-	-	-
				6	RingCount	Change	-	-	-
				0	FlourineCount	Change	-	-	-
8		0.77	0.44	0.0617	NonPolarCountMW	Increase	0.0060	PolarCountMW	Increase
				0.0060	PolarCountMW	Decrease	0.0617	NonPolarCountMW	Decrease
				-0.4169	LUMO	Decrease	-0.7360	HOMO	Increase
				0	NitrogenCount	Change	6	RingCount	Change
				6	RingCount	Change	-	-	-
				6	MaxRing2	Change	-	-	-
5		1.00	0.80	0.0608	NonPolarCountMW	Increase	0.0057	PolarCountMW	Decrease
				0.0057	PolarCountMW	Decrease	0.0608	NonPolarCountMW	Increase
				-0.3186	LUMO	Decrease	-0.6353	HOMO	Decrease
				0	NitrogenCount	Change	1	SulfurCount	Change
				6	RingCount	Change	0	NitrogenCount	Change
				0	FlourineCount	Change	0	ClorineCount	Change

Table 2. continued

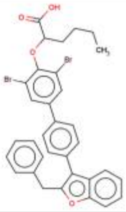
				Regression			Classification		
ID	Structure	log(1/IC50)	Predicted log(1/IC50)	Descriptor Values	Descriptor	Direction	Descriptor Values	Descriptor	Direction
7		1.23	1.07						
				0.0494	NonPolarCountMW	Increase	0.0494	NonPolarCountMW	Increase
				0.0046	PolarCountMW	Decrease	0.0046	PolarCountMW	Decrease
				-0.4158	LUMO	Decrease	29.18	M2M	Decrease
				0	NitrogenCount	Change	6	MaxRing2	Change
				0	FlourineCount	Change	0	FluorineCount	Change
				2	BromineCount	Change	0	ClorineCount	Change

Table 3. The Number of Compounds in the Test Set of Eight Compounds with the Descriptor Identified As the First, Second, or Third Most Influential

influential descriptors	rank 1	rank 2	rank 3	global rank
continuous descriptors				
Increase NonPolarCountMW	8	0	0	33
Decrease PolarCountMW	0	8	0	36
Decrease LUMO	0	0	8	7
discrete descriptors				
NitrogenCount	7	0	0	17
RingCount	0	5	0	14
FluorineCount	0	2	4	20
MaxRing2	0	0	1	31
HBD	1	0	0	21
MaxRing3	0	1	0	39
BromineCount	0	0	2	29
OHCount	0	0	1	50
global descriptors				
polarizability, AREA, MW				

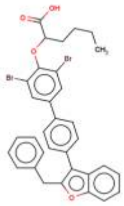
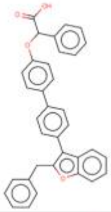
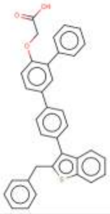
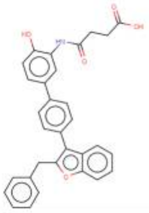
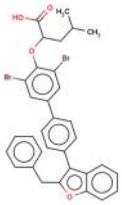
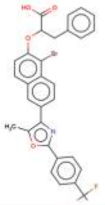
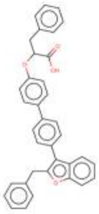
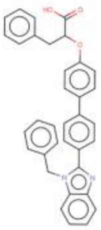
are used to assess the quality of the method. Equation 7 shows the n dimensional Shekel function with m fox holes (local extremes) used to evaluate the method for regression models. Shekel functions are frequently used in optimization problems as test functions because of their multidimensional and multimodal characteristics. They are easily customized to any number of local extremes in any dimension. Hence, the Shekel functions provided a general expression that was manipulated to obtain multiple functions of varying numbers of local extremes, presenting different degrees of challenges to the modeling algorithms. The $\mathbf{C}$ vector and $\mathbf{A}$ matrix defining the c_i and a_ij constants, respectively, are given in the Supporting Information in Tables S2 and S3. The corresponding Shekel function with two variables is displayed in the third row of Figure 2. Equation 8 is a Fourier series expansion used to approximate the discontinuous Heaviside step function. The properties of the Heaviside function are desirable for the classification data sets because of its clear transition in a well-defined area, thereby providing data that should be modeled by a categorical rather than a continuous algorithm. However, analytical partial derivatives are required to use a function for method validation and the Heaviside function had to be transformed into a continuous function, which was accomplished by a Fourier series expansion. Hence, the Heaviside step function was approximated within the variable range from $-\pi$ to π by using 40 Fourier terms. The α coefficient is used to regulate the size of the plateau used to classify the data points to ensure a balanced data set in any dimension.

Definition of Step Size. The step size of the numerical derivatives should be as small as possible to accurately reflect the local changes at the point though, in practice, large enough to affect model predictions. The value of the coefficient, C, of the step size (eq 5) was assessed for several regression and classification machine learning algorithms by considering the GA using various coefficient values. For each of the functions given in Figure 2, 16 000 data points were uniformly and normally sampled to provide training sets. Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) models were built using AZOrange[13] and its automated model hyper-parameter optimization procedure with default values. Similarly, for each function, a test set with 100 data points was sampled and used to quantify the GA. For these two-dimensional functions, a GA above 50% represents an enrichment compared to a random selection of a most influential descriptor.

Regression. Five analytical two-dimensional functions were used to sample 10 training data sets for the regression algorithms. For each model, the GAs were calculated while predicting the test set. Predictions and calculation of the GA were repeated for coefficient values ranging from 0.04 to 1.0, and the results are presented in Figure 3. The accuracy of the selected descriptor with the ANN and SVM algorithms appears to be relatively insensitive to the step length used to calculated the numerical partial derivatives as the GAs are almost constant over the range of C values. However, the RF algorithm has a maximum of the GA around the second point in the graph at $C = 0.08$ for most of the analytical data sets. Hence, for the regression algorithms, the coefficient was uniformly selected to 0.08, regardless of the learner type. Please see Table S1 in the Supporting Information for model accuracies on the external test set (Q2) and the GA of each model with the selected step length $C = 0.08$.

Classification. The Fourier approximation of the Heaviside function, as displayed with two descriptors in the fourth row of Figure 2, was used to evaluate the C value for classifiers. The Heaviside function was used to sample training sets with two and five descriptors, yielding four training sets with uniform or normal distribution. The classification accuracies of the optimized RF, SVM, and ANN models are given in Table S1 in the Supporting Information. Most of the classification accuracies are in the range between 85 and 100%. GAs as a function of C between 0.125 and 8.0 are displayed in the three graphs of Figure 3. The RF algorithm is exceedingly more sensitive to the step size in classification than in regression with GAs increasing up until a C value of around 1. As in the regression case, SVM is relatively robust with respect to the step size, while the ANN algorithm also displays an optimum at 1.0. On the basis of these results, the C value is set to 1.0 for all classification types of learners.

Table 4. Compounds from the Test Set Predicted by the HPTP Regression Model[a]

ID	Structure	Activity log(1/IC50)(μM)	Predicted Descriptor	Descriptor Value	New ID	New Structure	New Activity log(1/IC50) (μM)	New Descriptor Value
7		1.23	Increase NonPolarCountMW Change BromineCount	0.0494 2	3		0.40	0.0627 0
5		1.00	Change NitrogenCount Change RingCount	0 6	9		0.04	1 5
6		1.27	Change NitrogenCount Change FourineCount Change BromineCount	0 0 2	Train1		-0.48	1 3 1
10		0.60	Change NitrogenCount Decrease LUMO	0 -0.4169	Train2		-0.48	2 -0.5431

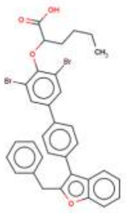
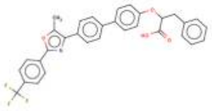
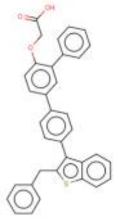
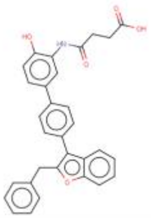
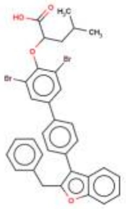
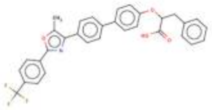
[a] The predicted influential descriptors are used to identify structural modifications resulting in new compounds with lower activity.

Tuning for Flat Areas. An empirical validation of the default value of the derivative threshold ($\varepsilon = 10^{-5}$) was performed using GAs obtained with models trained on data sets sampled from analytical functions with extensive flat areas. Selecting a most influential descriptor based on a set of exceedingly small values would yield GAs corresponding to a random selection, thereby resulting in a low GA. Hence, optimizing the GA of models obtained from training sets spanning flat areas with respect to the threshold can verify the magnitude of the threshold. Instances with all numerical partial derivatives below the threshold ε are considered to be predicted correctly (increasing the GA) if the point is in a flat area of the function where all analytical derivatives are equal to or smaller than ε. The analytical functions, together with the GA as a function of ε, are presented in Figure 4.

Regression. The Shekel function with three fox holes in two dimensions, as displayed in the first graph of Figure 4, was selected to provide a data set for regression modeling with extensive flat regions. The remaining regression graphs in Figure 4 show the GA of these models as a function of the threshold value for the derivatives, ε, ranging from 10^{-5} to 1.0. In addition to the GA, the graphs also display the fraction of test instances for which all numerical (yellow) or analytical (green) partial derivatives were below or equal to ε. Predictions from two different models are assessed, based on normal and uniform data point distribution. For the clarity of the graphs, the fraction of zero numerical derivatives among the test instances is the average for the two models. Because the numerical derivatives should be as close to the analytical as possible, this fraction of zero derivative test instances should also be as similar as possible for the numerical and analytical sets. The RF graph shows that even at the lowest threshold values, the GA is around 90%, limiting the maximal potential effect of numerical inaccuracies to around 10 percentage points. At $\varepsilon = 0.1$, approximately 50% of the test instances are correctly considered to be below the threshold, and the gradient accuracy increases to almost 100%, for the RF algorithm. Conversely, the GA of the SVM algorithm decreases with increasing ε because of too large numerical derivatives, as indicated by the fraction of zero numerical derivatives deviating from the fraction of zero analytical derivatives. For these data sets, the ANN algorithm fails to identify the most influential descriptor with any enrichment offering a notable improvement over a random selection. Furthermore, increasing ε does not improve the performance of the ANN algorithm.

Classification. The Fourier series of the Heaviside function (eq 8) contains extensive flat areas, and it was used in two dimensions to assess the value of ε for classifiers. The GA of the RF classifier is approximately 20 percentage points lower than for the corresponding regression model. The GA is adversely affected by too small numerical derivatives falling under the threshold value for almost 20% of the test instances (at ε =

Table 5. Compounds from the Test Set Predicted by the HPTP Classification Model[a]

ID	Structure	Activity log(1/IC50)(μM)	Predicted Descriptor	Descriptor Value	New ID	New Structure	New Activity log(1/IC50) (μM)	New Descriptor Value
7		POS	Decrease M2M Change FluorineCount	29.18 0	Train3		NEG	16.78 3
5		POS	Decrease HOMO Change SulfurCount Change NitrogenCount	-0.6354 1 0	9		NEG	-0.6698 0 1
6		POS	Decrease M2M Change FourineCount	35.44 0	Train3		NEG	16.78 3

[a]The predicted influential descriptors are used to identify structural modifications resulting in new compounds with lower activity.

10^{-5}), while the analytical derivative is above the threshold. However, this problem is not alleviated by increasing the threshold. Conversely, due to the numerical underestimate of the partial derivatives, greater ε decreases the GA. The SVM classifier does not have the problem of the RF algorithm with too small numerical derivatives, However, with increasing ε, the fraction of zero derivatives does not increase aligned with the analytical derivatives. Hence, the decrease in GA with ε might result from too large numerical derivatives.

In general, ε should be kept small to reduce the risk of omitting predictive information. The relatively small gain in GA for the RF and SVM algorithms in increasing ε implies that further studies are required to be confident that an increased GA, upon increasing ε, generalizes to other functions with different ranges of values in the response dimension. However, the present results are sufficient to support setting ε to 10^{-5}, while it is unlikely that increasing ε would offer substantial improvements in GA.

Dependence on Model Accuracy. Table S1 in the Supporting Information outlines the model and gradient accuracies with the selected C and ε parameters for all functions used to select these parameters. These results indicate a dependence between the GA and the model accuracy, as illustrated by the scatter plots in Figure 5. For the regression models, there is a clear dependency on model accuracy, such that the greater the model accuracy, the greater the GA, with a Pearson correlation coefficient of 0.8. To account for different enrichment factors of varying function dimensionalities, the data points are partitioned into the ones originating from predictions of two (blue dots) and five (red dots) dimensional models, for the classifiers. Yet, no relationship between model and gradient accuracy is evident.

Gradient Accuracy for Multidimensional Functions. Once the step coefficient C and derivative threshold ε have been established, the relevance of the identified descriptors is assessed on multidimensional models, approaching the dimensionality of established QSAR modeling techniques based on bulk descriptors. As in the previous sections, the data points are sampled from analytical functions, and the assessment is quantified by the GAs. An enrichment factor (ε_{GA}) is calculated to obtain a GA measurement unbiased with respect to the number of descriptors (D). Hence,

$$\varepsilon_{GA} = \frac{GA}{p*100} \tag{9}$$

where p is the probability of randomly selecting a descriptor ($p = (1/D)$). The results are presented in Table 1.

Because of the poorer performance of the ANN algorithm, solely the RF and SVM algorithms were tested. For the regression algorithms, the Shekel functions (eq 7) with 10 and 20 descriptors and seven fox holes were used to create 32 000 normally and uniformly distributed data points. The GAs of the SVM regression models are around 90%, while considerably lower for the RF models. Though the GA is as low as 34% for the uniform RF model in 10 dimensions, the descriptor identified by the algorithm still offers at least a 3 fold enrichment over a random selection. Considering the two top ranked numerical derivatives increases the fraction of instances for which the top ranked analytical descriptor was identified by approximately a factor of 2 for the RF regression model (GA_2^N), while extending to the triple GA (GA_3^N) offers a small additional improvement. When the top ranked numerical descriptor corresponds to any of the two top ranked analytical derivatives (GA_2^A), there is also an approximate 2-fold increase in the GA for the RF models. Considering the 3 top ranked descriptors of a regression SVM model, retrieves in almost all cases the most influential analytical descriptor.

Table 6. The Predicted Most Influential Descriptors within One PRSS2 Cluster of Eight Compounds, Together with Activity and the Values of the Descriptors

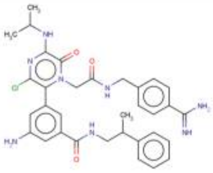
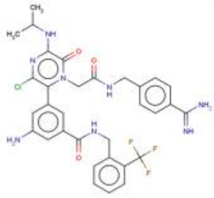
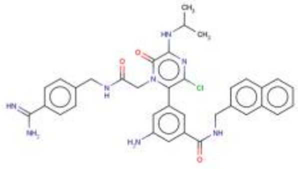
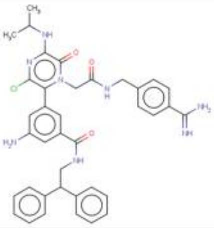
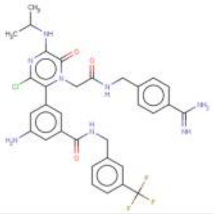
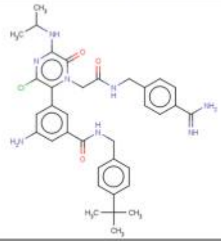
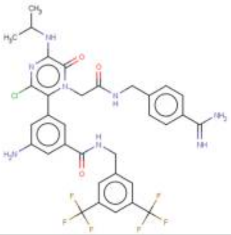
ID	Structure	pIC50	pIC50 Predicted	Descriptor Values	Descriptor	Direction
7		7.29	7.25			
				0.0143	PolarCountMW	Decrease
				-0.7474	HOMO	Decrease
				2.0	MinDistDD	Increase
				8	HBD	Change
				8	NHCount	Change
				0	SulfurCount	Change
6		7.27	7.45			
				0.0135	PolarCountMW	Decrease
				0.0404	NonPolarCountMW	Increase
				-0.7474	HOMO	Decrease
				8	HBD	Change
				8	NHCount	Change
				0	SulfurCount	Change
11		7.11	7.55			
				0.0138	PolarCountMW	Decrease
				0.0415	NonPolarCountMW	Increase
				-0.6180	HOMO	Decrease
				8	HBD	Change
				8	NHCount	Change
				0	FlourineCount	Change
10		6.81	6.97			
				0.0130	PolarCountMW	Decrease
				0.0434	NonPolarCountMW	Increase
				-0.7474	HOMO	Decrease
				8	HBD	Change
				8	NHCount	Change
				0	FluorineCount	Change
9		6.74	7.46			
				0.0135	PolarCountMW	Decrease
				0.0404	NonPolarCountMW	Increase
				-0.7474	HOMO	Decrease
				8	HBD	Change
				8	NHCount	Change
				0	SulfurCount	Change
8		6.64	7.16			
				0.0137	PolarCountMW	Decrease
				0.0411	NonPolarCountMW	Increase
				-0.7474	HOMO	Decrease
				8	HBD	Change
				8	NHCount	Change
				0	SulfurCount	Change
12		6.32	7.48			
				0.0122	PolarCountMW	Decrease
				0.0421	NonPolarCountMW	Increase
				-0.7474	HOMO	Decrease
				0	SulfurCount	Change
				8	NHCount	Change
				8	HBD	Change

Table 6. continued

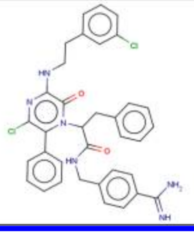
ID	Structure	pIC50	pIC50 Predicted	Descriptor Values	Descriptor	Direction
3		5.13	5.28			
				0.0094	PolarCountMW	Decrease
				0.0485	NonPolarCountMW	Increase
				-0.8174	HOMO	Decrease
				0	FlourineCount*	Change

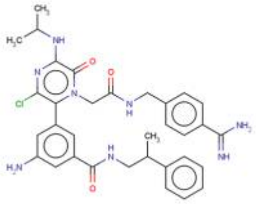
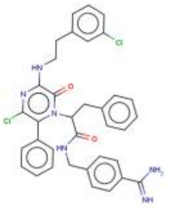
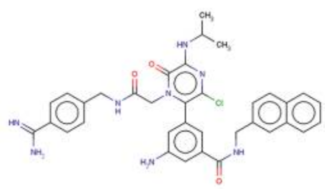
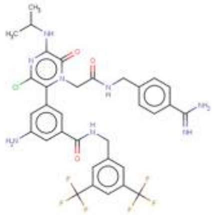
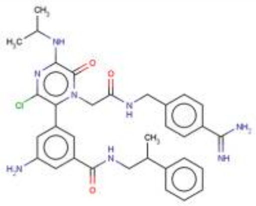
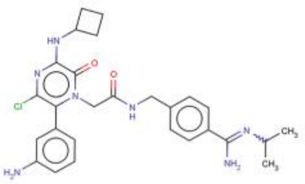
Table 7. The Number of Compounds in the Test Set of Eight Compounds with the Descriptors Identified As the First, Second, or Third Most Influential

influential descriptors	rank 1	rank 2	rank 3	global rank
continuous descriptors				
Decrease PolarCountMW	8	0	0	6
Increase NonPolarCountMW	0	7	0	8
Decrease HOMO	0	1	7	26
Increase MinDistDD	0	0	1	10
discrete descriptors				
HBD	6	0	1	2
NHCount	0	7	0	1
FluorineCount	1	0	2	24
SulfurCount	1	0	4	38
global descriptors				
NHCount, HBD, MWSHDA				

Also for the classification models, based on the Fourier series of the Heaviside function with 10 and 20 descriptors, the SVM algorithm performs substantially better than the RF algorithm, with GA around 70% and 50%, respectively. As for the regression models, considering the two top ranked numerical derivatives increases the chances of identifying the most influential descriptor. It is interesting to note that the GA is similar, while increasing the dimensionality of the classifiers, indicating that the method might be almost equally successful in identifying the most influential descriptor while expanding the number of descriptors. However, a more rigorous study is required to establish the dependency of the GA on the dimensionality.

Modifying Target Activity Using Influential Descriptors. A set of publically available QSAR data sets, containing congeneric chemical series, were used to illustrate the usage of the method for chemical design in a pharmaceutical setting. The data sets were partitioned into a training set and a test set. Test compounds were separated by clustering the full data set with the maximum common substructure clustering procedure by Seeland et al.[14] available in AZOrange and randomly selecting one cluster or set of cluster representatives. This assured structurally similar compounds in a chemical series like test set, offering an empirical and external verification opportunity for the localized heuristic inverse QSAR methodology. Models were trained using AZOrange, selecting the model-hyper parameters using default settings in the parameter

Table 8. Compounds from the Test Set Predicted by the PRSS2 Regression Model[a]

ID	Structure	Activity pIC50	Predicted Descriptor	Descriptor Value	New ID	New Structure	New Activity pIC50	New Descriptor Value
7		7.29	Decrease PolarCountMW Decrease HOMO Change HBD Change NHCount	0.0143 -0.7474 8 8	3		5.13	0.0094 -0.8174 5 5
11		7.11	Decrease HOMO Change FlourineCount	-0.6180 0	12		6.32	-0.7474 6
7		7.29	Increase MinDistDD Change HBD Change NHCount	2 8 8	Train1		4.05	6 6 6

[a]The predicted influential descriptors are used to identify structural modifications resulting in new compounds with lower activity.

Table 9. The Predicted Most Influential Descriptors within One F7 Cluster of Nine Compounds, Together with Activity and the Values of the Descriptors

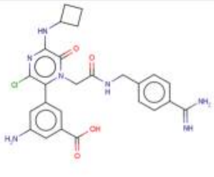
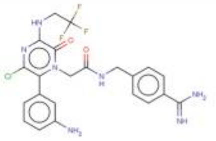
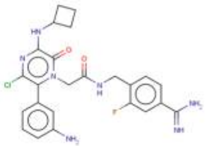
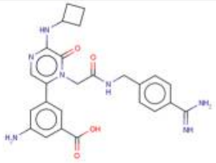
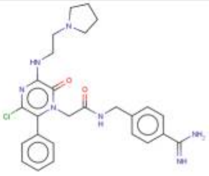
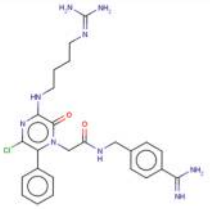
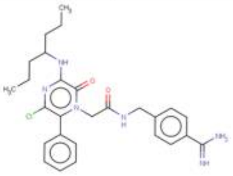
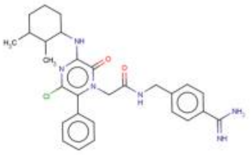
ID	Structure	pIC50	pIC50 Predicted	Descriptor Values	Descriptor	Direction
3		7.70	7.63			
				0.0324	NonPolarCountMW	Decrease
				0.0172	PolarCountMW	Increase
				-0.7476	HOMO	Decrease
				8	HBD	Change
				0	Amine1	Change
				1	ChlorineCount	Change
5		7.52	6.53			
				0.0374	NonPolarCountMW	Decrease
				0.0138	PolarCountMW	Increase
				-0.7433	HOMO	Decrease
				7	HBD	Change
				0	Amine1	Change
				1	ChlorineCount	Change
9		7.40	6.97			
				0.0382	NonPolarCountMW	Decrease
				0.0141	PolarCountMW	Increase
				-0.7433	HOMO	Decrease
				0	Amine3	Change
				7	HBD	Change
				7	NHCount	Change
4		7.28	7.55			
				0.0327	NonPolarCountMW	Decrease
				0.0184	PolarCountMW	Increase
				-0.7477	HOMO	Decrease
				8	HBD	Change
				0	ChlorineCount	Change
				0	Amine1	Change
1		6.57	5.19			
				0.0394	NonPolarCountMW	Decrease
				0.0138	PolarCountMW	Increase
				-0.8174	HOMO	Decrease
				0	FluorineCount	Change
				1	HalogenCount	Change
				2	OxygenCount	Change
2		6.24	5.92			
				0.0324	NonPolarCountMW	Decrease
				0.0172	PolarCountMW	Increase
				-0.8174	HOMO	Decrease
				0	Amine1	Change
				9	HBD	Change
				1	ChlorineCount	Change
7		5.74	5.97			
				0.0432	NonPolarCountMW	Decrease
				0.0118	PolarCountMW	Increase
				-0.8174	HOMO	Decrease
				2	OxygenCount	Change
				0	FluorineCount	Change
				0	Amine3	Change
8		4.84	5.88			
				0.0441	NonPolarCountMW	Decrease
				0.0115	PolarCountMW	Increase
				-0.8174	HOMO	Decrease
				2	OxygenCount	Change
				0	FluorineCount	Change
				0	Amine3	Change

Table 9. continued

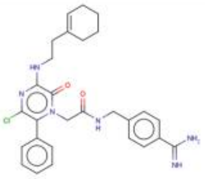
ID	Structure	pIC50	pIC50 Predicted	Descriptor Values	Descriptor	Direction
6		4.56	5.74			
				0.0443	NonPolarCountMW	Decrease
				0.0116	PolarCountMW	Increase
				-0.8174	HOMO	Decrease
				2	OxygenCount	Change
				0	FluorineCount	Change
				0	Amine3	Change

Table 10. The Number of Compounds in the Test Set of Nine Compounds with the Descriptors Identified As the First, Second, or Third Most Influential

influential descriptors	rank 1	rank 2	rank 3	global rank
continuous descriptors				
Decrease PolarCountMW	9	0	0	25
Increase NonPolarCountMW	0	9	0	8
Decrease HOMO	0	0	9	5
discrete descriptors				
HBD	3	2	0	1
Amine1	1	2	1	45
ChlorineCount	0	1	3	37
Amine3	2	0	2	40
NHCount	0	0	1	2
FluorineCount	1	3	0	39
HalogenCount	0	1	0	41
OxygenCount	3	0	1	29
global descriptors				
HBD, NHCount, MWSHDA				

selection algorithm, and instructions on how to use localized heuristic inverse QSAR as implemented in AZOrange are provided in Supporting Information Table S4. It should be noted that the method would be difficult to use in ensemble modeling where predictions for several underlying models are combined. A set of 54 two- and three-dimensional bulk property based descriptors including descriptors for molecular size, lipophilicity, hydrogen bonding, electrostatics, and topology were calculated with the AstraZeneca in-house descriptor engine. These descriptors were selected because of the relative ease with which knowledge of their importance can be translated into ideas on structural modifications. The details of the descriptors are described elsewhere[15] and summarized in the Excel file in the Supporting Information. The set of test compounds are predicted together with the most influential descriptors. In addition to reporting the descriptors, the algorithm also provides information on the direction in which continuous descriptor values need to be changed to achieve a user defined desired impact on the response. For example, if the energy of the highest occupied molecular orbital is selected for a QSAR prediction of hERG activity and the user defined that the desired change is a decrease in activity, the algorithm will also predict whether the orbital energy should be increased or decreased in order to decrease hERG activity.

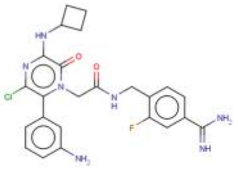
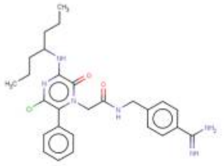
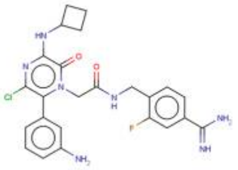
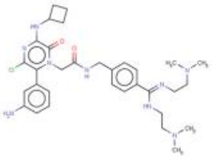
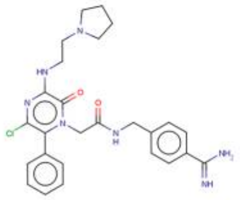
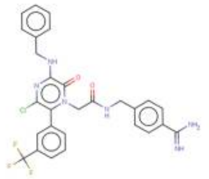
As the numerical partial derivatives of descriptors of different types cannot be compared (eqs 1 and 3), top ranked discrete as well as continuous descriptors are presented. Discrete variables are treated as enumeration variables in Orange.[16] Hence, the algorithm does not recognize any order between discrete values, and because of this unordered nature of the enumeration variables, discrete attributes can only be attributed by “Change,” rather than “Increase” or “Decrease” as for continuous descriptors. Compounds structurally similar to the predicted compound, with the descriptor value modified in the predicted direction and with responses differing in the desired direction, are considered examples of where the methodology could have been able to aid in the design process.

Modifying Activity of Human Protein Tyrosin Phosphatase Inhibitors. Protein tyrosin phosphatase (PTP) and, in particular, PTP 1B could be a useful target for the treatment of type II diabetes.[17] Taha et al.[18] developed a QSAR model with 137 human PTP inhibitors with a leave-one-out accuracy of 0.68. This data set was used to illustrate how the identification of influential bulk descriptors can be used to alter target activity. However, in this study, HPTP is considered a secondary target, and molecular changes resulting in decreased activity is sought. The data set was clustered with the threshold and minimum cluster size parameters set to 0.7 and 10, respectively, resulting in seven clusters. Eight compounds were randomly selected as an external test set from one randomly selected cluster. The training set was used to build an SVM model with radial basis functions ($C = 1.0$ and $\gamma = 0.05$) with a 10-fold CV accuracy of 0.68 (Q2). The test set was predicted with a coefficient of determination (R^2) of 0.93.

Table 2 lists actual and predicted activities and the three top ranked influential descriptors of discrete and continuous type.

Table 3 displays the incidence by which these top ranked descriptors are predicted as the first (rank 1), second (rank 2), and third (rank 3) most influential within the cluster. As a comparison, the three top ranked descriptors retrieved with global ranking over the training set with the RF variable importance method are presented at the bottom of the table. Finally, the last column of Table 3 presents the global rank of the locally most influential descriptors. Though related, the descriptors top ranked by the global method are different from those predicted to be locally important. Among the continuous descriptors, the PolarCountMW, NonPolarCountMW, and LUMO descriptors are consistently predicted as the first, second, and third most influential, respectively, for all test compounds. PolarCountMW is a count of the number of polar atoms (O, N, S, and P, with S and P in high oxidation states) normalized by the molecular weight, while NonPolarCountMW is the number of carbons and hydrogens with the PolarCount subtracted and the resulting sum normalized by molecular weight. There is a greater diversity among the top ranked discrete descriptors with eight different descriptors occurring among the three top ranked descriptors of the eight test compounds. Please see the Excel file in the Supporting Information for a comprehensive list of descriptors and their definitions. Table 4 shows examples of how the identified influential descriptors can be used to change the molecular structure, resulting in lower HPTP activity.

Table 11. Compounds from the Test Set Predicted by the F7 Regression Model[a]

ID	Structure	Activity pIC50	Predicted Descriptor	Descriptor Value	New ID	New Structure	New Activity pIC50	New Descriptor Value
9		7.40	Change HBD	7	7		5.74	5
			Change NHCount	7				5
9		7.40	Amine3	0	Train2		4.20	2
			Change HBD	7				5
			Change NHCount	7				5
1		6.57	Change FluorineCount	0	Train1		5.15	3
			Change HalogenCount	1				4

[a]The predicted influential descriptors are used to identify structural modifications resulting in new compounds with lower activity.

Modification of structure 7 into 3 increases the NonPolarCountMW, predicted to be the most influential descriptor, and reduces the activity from 1.23 to 0.40 (log(1/IC50)). This structural modification also represents a decrease in the converse property PolarCountMW, predicted to be the second most influential descriptor. In addition, changing structure 7 into 3 and thereby removing two bromines also reduces the BromineCount, which is the third most influential discrete descriptor of compound 7. The NitrogenCount is predicted to be the most influential discrete descriptor of structure 5, and changing the molecular structure into the only nitrogen containing structure of the test set reduces the activity to 0.04. Furthermore, this structural modification alters the compound with respect to the second most influential discrete descriptor while reducing the ring count from six to five. The test set does not contain any compounds with fluorine, and the LUMO energy is almost the same for most compounds. Hence, compounds were sought in the training set which could illustrate such structural modifications. Changing structure 6 into structure Train1 adds one nitrogen and three fluorine atoms, while removing a bromine, which corresponds to alteration of all three top ranked discrete descriptors, and the activity is reduced from 1.27 to −0.48. Similarly, changing structure 10 into structure Train2, in accordance with the predicted descriptor NitrogenCount and decreasing LUMO, decreases target activity.

The assessment with analytical functions indicated that the performance of the method might in general be poorer for classifiers. To illustrate the difference between the performance of the method for regression and classification models in an applied setting, the response was categorized as POS if log(1/(IC50)) was above 0.5 and NEG otherwise, which resulted in a balanced data set (POS 51%, NEG 49%). For comparison with the regression model, the same external test set was predicted. Furthermore, the same set of descriptors was used, and the compounds were classified by the categorical SVM algorithm in AZOrange ($C = 8193$, $\gamma = 3.05$), resulting in a classification accuracy of 72% in 10-fold CV. All compounds in the test set except compound 8 were predicted correctly. Table S4 compiles the descriptors predicted by the classification model for the test set.

As observed with the regression model, PolarCountMW and NonPolarCountMW are among the three top ranked continuous descriptors for almost all test compounds. However, the HOMO rather than the LUMO energy is also top ranked. The top ranked discrete descriptors are less consistent in the comparison between regression and classification models. In particular, for several of the test compounds, no discrete descriptors are identified. For these compounds, it was confirmed that a change in any discrete descriptor was predicted to increase rather than decrease the activity. The accuracy of this prediction could be challenged and illustrates the expected lower reliability of influential descriptors identified by classifiers.Table 5 summarizes the remaining successful structural modifications of test compounds 7, 5, and 6.

However, the descriptors predicted for compound 10 were too unspecific to guide any structural modifications, even if the fourth and fifth (total number of atoms and the number of π

atoms) top ranked continuous descriptors were also considered. In accordance with the predicted third top ranked discrete descriptor of compound 7, changing its structure into compound Train3 of the training set reduces the moment of inertia with respect to the second principal axis, because of the elongated, unbranched structure of compound Train3, while decreasing target activity. Furthermore, three fluorine atoms are added, which aligns with the predicted change in FluorineCount. The modifications of compound 5 into compound 9 of the test set are still supported by the decrease in HOMO energy and the change in the number of sulfur and nitrogen atoms. Support for changing compound 6 into Train1 is not provided by the predictions from the classifier. However, similarly to compound 7, compound 6 can be modified into Train3 while decreasing the second moment of inertia and changing the fluorine count, though this decreases the third moment of inertia, contradictory to the second most influential descriptor M3M.

Modifying Activity of Anionic Trypsin 2 Inhibitors. A set of trypsin 2 (PRSS2) inhibitors was collected from published chemical patents (US7119094) through the GOSTAR[19] database, yielding 339 compounds with annotated pIC50 values for PRSS2 inhibition. It is assumed that PRSS2 is a secondary target and thus that it is desirable to change the chemical structure to reduce activity.

The data set was clustered with the threshold and minimum cluster size parameters set to 0.5 and 10, respectively, resulting in eight clusters. Eight compounds were randomly selected as an external test set from a randomly selected cluster. The training set was used to build an SVM model with radial basis functions ($C = 0.8$ and $\gamma = 0.03125$) with a 10-fold CV accuracy of 0.57 (Q2). The test set was predicted with an R^2 of 0.74.

Table 6 lists actual and predicted activities and the three top ranked influential descriptors of discrete and continuous type.

Table 7 displays the incidence by which a descriptor is predicted as the first, second, and third most influential within the cluster. As a comparison, the three top ranked descriptors retrieved with global ranking over the training set with the RF variable importance method are presented at the bottom of the table. Finally, the last column of Table 7 presents the global rank of the locally most influential descriptors. Though most of the influential descriptors are not prominently ranked by the global method, two of the locally top ranked discrete descriptors are also suggested as the most important by the global variable importance method. Among the continuous descriptors, the PolarCountMW, NonPolarCountMW, HOMO, and MinDistDD descriptors are predicted as the most influential for the test compounds. The top ranked discrete descriptors are the number of hydrogen donors (HBD), the number of NH groups, and the number of fluorine and sulfur atoms.

Table 8 shows examples of how the identified influential descriptors can be used to change the molecular structure, resulting in lower PRSS2 activity.

Changing structure 7 into 3 decreases PolarCount MW while removing an amide group. This structural modification also reduces the number of hydrogen bond donors and NH groups from 8 to 5, and the highest occupied orbital becomes more stable. Altering structure 11 into 12 adds six fluorine atoms in accordance with the FlourineCount being identified as the third most important descriptor. Furthermore, the HOMO energy is reduced. All compounds in the external test set have a minimal distance between the hydrogen donors of two bonds. Hence, compounds had to the sought in the training set to find examples of modifications to the amidine group. In addition to increasing the minimal distance between hydrogen donors from 2 to 6, changing structure 7 into Train1 also reduces HBD and NHCount from 8 to 6.

Modifying Activity of Factor VIIb Inhibitors. A set of Factor VIIb (F7) inhibitors were collected from published chemical patents (US20050043313) through the GOSTAR[19] database, yielding 365 compounds with annotated pIC50 values for F7 inhibition. It is assumed that F7 is a secondary target and thus that it is desirable to change the chemical structure to reduce activity.

The data set was clustered with the threshold and minimum cluster size parameters set to 0.7 and 10, respectively, resulting in 11 clusters. Nine compounds were randomly selected as an external test set from a randomly selected cluster. The training set was used to build an SVM model with radial basis functions ($C = 8$ and $\gamma = 0.03125$) with a 10-fold CV accuracy of 0.71 (Q2). The test set was predicted with an R^2 of 0.48.

Table 9 lists actual and predicted activities and the three top ranked influential descriptors of discrete and continuous type, while Table 10 displays the incidence by which a descriptor is predicted as the first, second, and third most influential within the cluster.

As a comparison, the three top ranked descriptors retrieved with global ranking over the training set with the RF variable importance method are presented at the bottom of the table. Finally, the last column of Table 10 presents the global rank of the locally most influential descriptors. Though most of the influential descriptors are not prominently ranked by the global method, two of the locally top ranked discrete descriptors are also suggested as the most important by the global variable importance method.

As displayed in Table 11, changing compound 9 into compound 7 reduces the pIC50 value from 7.40 to 5.74.

The structural modification is supported by the change in the number of hydrogen bond donors and in the number of NH groups predicted to be among the most influential descriptors. The structural modifications suggested for compound 9 are further supported considering compounds in the training set. Changing compound 9 into compound Train2 includes also the addition of two tertiary amines and reduces target activity even further to 4.20. Modification of compound 1 into compound Train1 is another example of structural modifications aligning with the predicted descriptors and changing target activity in the desired direction. Three fluorine atoms are added, and the pIC50 is reduced from 6.57 to 5.15.

General Considerations. Although it has been demonstrated that the method can retrieve descriptors of individual QSAR predictions that can be used to alter the structure resulting in desired changes in the response, comparison of numerical derivatives of properties with different units and different variance is inherently difficult. Further optimization of the step size might be possible by detecting potential outliers in the training set, as well as including several data sets in the estimation of the variance of a particular descriptor.

Comparing the top ranked descriptors across the three studied QSAR data sets shows that the size normalized number of polar atoms is frequently among them. This might indicate that this descriptor generally has a high impact on biological activity, aligning with the commonly recognized principle that lipophilicity promotes activity. However, it might also be an artifact of how the numerical derivatives are calculated, and further studies are required to deduce the localized influence of the polar count. Despite this uncertainty, a practical remedy would be to consider a greater number of top ranked continuous descriptors to promote the diversity and generate additional, more localized and specific design ideas.

Though the current implementation is reasonably successful in identifying the descriptors with the greatest analytical partial derivatives, the threshold value of the derivatives could probably be further improved by individual tuning for each data set, considering the magnitude of the response variable. Furthermore, for experimental data, it might be possible to quantify the magnitude of an irrelevant change in the biological response and thereby use end point specific thresholds.

■ CONCLUSION

Localized heuristic inverse QSAR has been developed to aid in chemical design by identifying the most influential bulk descriptors for individual QSAR predictions. The method is generally applicable to many machine learning algorithms and to both regression and classification models. The numerical derivatives were validated with two-dimensional analytical reference functions, displaying excellent agreement with the corresponding analytical derivatives, in particular for regression models. Furthermore, the analytical validation showed that the greater the regression model accuracy, the greater the accuracy of the numerical gradients. For multidimensional analytical functions, approaching the dimensionality of QSAR models based on bulk descriptors, SVM regression models are highly successful in identifying the greatest analytical derivatives, while the RF models exhibit lower performance with approximately half the enrichment factor. In general, concordance between the analytical and numerical derivatives is poorer for classifiers. Hence, the performance of localized heuristic inverse QSAR can be expected to be better if the response can be described as continuous. Three QSAR data sets have been used to illustrate how localized heuristic inverse QSAR can guide chemical design. Examples of structural modifications, promoted by the predicted descriptors and reducing the biological activity in the desired direction, are shown. Localized heuristic inverse QSAR is an alternative to global descriptor ranking when structural alterations and mechanistic insight within a congeneric chemical series are sought. The QSAR examples qualitatively show that the locally important descriptors in most cases suggest design alterations different from those obtained by the global ranking method. Hence, the global ranking would not have been able to guide structural modifications of the test set compounds in the directions of the locally identified variables. The heuristic element reduces the computational complexity as compared to inverse QSAR as well as permits design support from QSARs based on bulk descriptors.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1 summarizes the model and gradient accuracies for models based on data sets sampled with the two-dimensional functions. The Excel file complies information about the bulk descriptors used by the QSAR models. Tables S2 and S3 specify the parameters used in the Fourier series in eq 4, while Table S4 displays the predicted response and descriptors of the three QSAR test sets.

This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jonna.stalring@astrazeneca.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge the Open Source community and in particular the Orange developers at the Artificial Intelligence Laboratory at the University of Ljubljana. We also acknowledge Dr. Peter Varkonyi and Dr. Hongming Chen, computational scientists at AstraZeneca, for compiling the patent data sets.

■ REFERENCES

(1) Helma, C.; Kazius, J. *J. Curr. Comput.-Aided Drug Des.* **2006**, *2*, 123.

(2) Shao, L.; Wu, L.; Fan, X.; Cheng, Y. *J. Chem. Inf. Model.* **2010**, *50*, 1941.

(3) Tsygankova, I. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 132.

(4) Svetnik, V.; Liaw, A.; Tong, C.; Wang, T. *Multiple Classifier Syst. Lect. Notes Comput. Sci.* **2004**, *3077*, 334.

(5) Guha, R.; Jurs, P. *J. Chem. Inf. Model* **2005**, *45*, 800.

(6) Douguet, D.; Thoreau, E.; Grassy, D. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449.

(7) Kvasnicka, V.; Pospichal, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 516.

(8) Churchwell, C.; Rintoul, M.; Martin, S.; Visco, D.; Kotu, A.; Larson, R.; Sillerud, L.; Brown, D.; Faulon, J. *J. Mol. Graph. Model.* **2004**, *22*, 263.

(9) Churi, N.; Achenie, L. *Ind. Eng. Chem. Res.* **1996**, *35*, 3788.

(10) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. *J. Med. Chem.* **2005**, *48*, 6997.

(11) Carlsson, L.; Helgee, E. A.; Boyer, S. *J. Chem. Inf. Model.* **2009**, *49*, 2551.

(12) Marcou, G.; Horvath, D.; Solov'ev, V.; Arrault, A.; Vayer, P.; Varnek, A. *Mol. Inf.* **2012**, *31*, 639.

(13) Stålring, J.; Carlsson, L.; Almeida, P.; Boyer, S. *J. Cheminform.* **2011**, *3*, 28.

(14) Seeland, M.; Girschick, T.; Buchwald, F.; Kramer, S. *ECML PKDD'10 Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*; Springer: New York, 2010; p 213.

(15) Paine, S.; Barton, P.; Bird, J.; Denton, R.; Menochet, K.; Smith, A.; Tomkinson, N.; Chohan, K. *J. Mol. Graphics Modell.* **2010**, *29*, 529.

(16) J., D.; Zupan, B. *Orange: From Experimental Machine Learning to Interactive Data Mining*; Faculty of Computer and Information Science, University of Ljubljana: Ljubljana, Slovenia, 2004. www.ailab.si/orange.

(17) Na, M.; Oh, W.; Kim, Y.; Ci, X.; Kim, S.; Kim, B.; Ahn, J. *Bioorg. Med. Lett.* **2006**, *16*, 3061.

(18) Taha, M.; Bustanji, Y.; Al-Bakri, A.; Yousef, A.; Zalloum, W.; Al-Masri, I.; Atallah, N. *J. Mol. Graphics Modell.* **2007**, *25*, 870.

(19) *GOSTAR databases 2012*; GVK Bioscieces Private Ltd.: Hyderabad, India, 2012.