Check for updates

RESEARCH ARTICLE

REVISED Exploring differential evolution for inverse QSAR analysis [version 2; peer review: 3 approved]

Tomoyuki Miyao[1,2], Kimito Funatsu[1], Jürgen Bajorath[2]

[1]Department of Chemical System Engineering, School of Engineering, The University of Tokyo, Tokyo, 113-8656, Japan
[2]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, D-53113, Germany

v2 **First published:** 31 Jul 2017, **6**(Chem Inf Sci):1285
https://doi.org/10.12688/f1000research.12228.1

Latest published: 06 Sep 2017, **6**(Chem Inf Sci):1285
https://doi.org/10.12688/f1000research.12228.2

Abstract

Inverse quantitative structure-activity relationship (QSAR) modeling encompasses the generation of compound structures from values of descriptors corresponding to high activity predicted with a given QSAR model. Structure generation proceeds from descriptor coordinates optimized for activity prediction. Herein, we concentrate on the first phase of the inverse QSAR process and introduce a new methodology for coordinate optimization, termed differential evolution (DE), that originated from computer science and engineering. Using simulation and compound activity data, we demonstrate that DE in combination with support vector regression (SVR) yields effective and robust predictions of optimized coordinates satisfying model constraints and requirements. For different compound activity classes, optimized coordinates are obtained that exclusively map to regions of high activity in feature space, represent novel positions for structure generation, and are chemically meaningful.

Keywords

Chemical space, active compounds, differential evolution, support vector regression, virtual screening, inverse QSAR

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 06 Sep 2017			
version 1 31 Jul 2017	✓ view	✓ view	✓ view

1. **Hans Matter**, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany
2. **Brian Goldman**, Vertex Pharmaceuticals, Boston, USA
3. **W. Patrick Walters**, Relay Therapeutics, Cambridge, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Chemical Information Science gateway.

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

Author roles: Miyao T: Data Curation, Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Funatsu K**: Conceptualization, Supervision, Writing – Review & Editing; **Bajorath J**: Conceptualization, Formal Analysis, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The project leading to this report has received funding (for TM) from the Japan Society for the Promotion of Science (JSPS) under the JSPS KAKENHI Grant Number 16J05325.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Miyao T, Funatsu K and Bajorath J. **Exploring differential evolution for inverse QSAR analysis [version 2; peer review: 3 approved]** F1000Research 2017, **6**(Chem Inf Sci):1285 https://doi.org/10.12688/f1000research.12228.2

First published: 31 Jul 2017, **6**(Chem Inf Sci):1285 https://doi.org/10.12688/f1000research.12228.1

REVISED **Amendments from Version 1**

In response to the reviewers, we have addressed all issues raised by Hans Matter, incorporated the autoencoder study suggested by Brian Goldman and discussed how families of structures might be evolved. In addition, as suggested by Pat Walters, we make the data sets and descriptors available as an open access deposition (a reference implementation is in the works). Furthermore, we have added an example for deriving the applicability domain and specified the RMSE in Table 2. Finally, we note that structures in Figure 7 were partly compromised due to a reproduction error from a PDF file (the drawings have been corrected) and that standard distance- and potency-based compound rankings were provided as virtual screening reference calculations. In addition, an MMP-based search for potent analogs of compounds proximal to optimized coordinates has also been carried out.

See referee reports

Introduction

Inverse quantitative structure-activity relationship (QSAR) analysis aims to identify values of descriptors used to generate a QSAR model that corresponds to high activity, and build structures of active compounds from these values[1–4]. The inverse QSAR process is challenging since numerical signatures of activity, if they can be determined, must be re-translated into viable chemical structures and active compounds, a task falling into the area of *de novo* compound design[5–7]. A predominant approach to inverse QSAR is the use of multiple linear regression (MLR) models to construct chemical graphs that correspond to an MLR equation[1–4]. Given this equation, a desired y (activity) value constrains relationships between descriptor settings. These constraints make it possible to derive vertex degree or edge sequences, from which chemical graphs might be constructed. For instance, specialized descriptors have been introduced for inverse QSAR on the basis of MLR equations and algorithms for constructing chemical graphs from these descriptors[8–11]. So far only few inverse QSAR studies have employed methods other than MLR. For example, it was attempted to construct chemical graphs from the centroid of activity of a set of compounds in Hilbert space defined by a kernel function[12]. In this case, a pre-image approximation algorithm was used to obtain coordinates in descriptor space and construct chemical graphs from these descriptor coordinates. Alternatively, inverse QSAR was divided into a two-stage process by separating the derivation of preferred descriptor values for a desired activity from the chemical graph construction phase[13–15]. Descriptor information corresponding to a given y value was represented via probability density functions, and regression analysis was performed using Gaussian mixture models in combination with cluster-wise MLR[14]. Subsequently, chemical graphs satisfying a set of descriptor values, or ranges of descriptor values, were generated by assembling ring systems and atom fragments with monotonically changing descriptors[14]. Following this approach, descriptor values must increase when adding an atom, ring system, or other structural fragment to a growing chemical graph. Applying Gaussian mixture models and cluster-wise MLR makes it possible to focus on the applicability domain[14,15] of the underlying models.

The two-stage inverse QSAR process is conceptually based on an important premise adopted from conventional (forward) QSAR, i.e., the higher a predicted activity value is, the more desirable a chemical structure becomes. In two-stage inverse QSAR, this conjecture challenges the descriptor value generation phase because value combinations are ultimately desired that correspond to higher predicted activity than exhibited by any currently available training or test compound. In other words, descriptor settings should be optimized for predicted activity. For this purpose, the use of Gaussian mixture models and cluster-wise MLR left considerable room for improvement, due to its multi-parametric nature and tendency of overfitting if training data were organized into large number of clusters[14]. Recently, autoencoder modeling was proposed as an approach for two-stage inverse QSAR[16]. Continuous latent space, corresponding to a descriptor space, is constructed on the basis of encoding a line notation of a molecule by recurrent neural networks (RNNs). Following this methodology, optimized coordinates in latent space can be directly translated into another line notation by the decoder consisting of RNNs. As such, the approach does not depend on chosen descriptors and has the potential to automatically address two-stage inverse QSAR in a single step. However, the generation of new valid line notations (SMILES strings) for chemical structures corresponding to optimized coordinates was difficult in a case study designing organic light-emitting diodes[16].

In this work, the descriptor optimization challenge of two-stage inverse QSAR has been specifically addressed. We emphasize that the chemical graph construction phase of inverse QSAR is not subject of this work and beyond its scope. Rather, our focal point has been the development of a new methodology for optimizing descriptor settings with respect to higher than observed compound activity, as a prerequisite for candidate structure generation. Therefore, an evolutionary approach is introduced to identify descriptor coordinates that correspond to the highest predicted activity within the applicability domain of a given QSAR model. The methodology and results of proof-of-concept studies are presented in the following.

Methods

Methodological concept

Inverse QSAR depends on the derivation of descriptor coordinates for a given model and data set. The goal of the methodology presented herein is finding desirable coordinates in a pre-defined descriptor space (**x** space) on the basis of a regression function f(**x**) representing a QSAR model. Confining the search to the applicability domain (AD) of the model translates this task into a constrained optimization problem (COP). The concept of the optimization is illustrated in Figure 1. Newly derived coordinates should be more desirable with respect to pre-defined evaluation criteria than any other data point used to construct the regression model. Accordingly, COP is formulated as follows:

Minimize $f(\mathbf{x})$

s.t. $g_j(\mathbf{x}) \leq 0,\ j = 1, \ldots, q$

$h_j(\mathbf{x}) = 0,\ j = q+1, \ldots, m$

$\mathbf{x} = (x_1, \ldots, x_d),\ l_i \leq x_i \leq u_i,\ i = 1, \ldots, d$

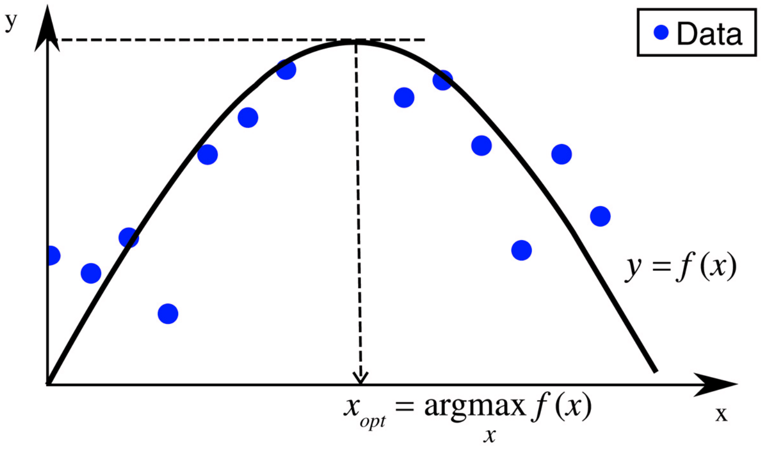

Figure 1. Optimization concept. A regression function f(**x**) fits training data to determine new coordinates in descriptor space. Optimized coordinates based on f(**x**) fall inside the training data range but yield a higher y value than any other data point.

where $\mathbf{x} \in R^d$, f: $R^d \to R$ is the function to be optimized, g_j: $R^d \to R$ is the j-th inequality function, and h_j: $R^d \to R$ is the j-th equality function. The i-th component of **x**: x_i falls into the range $[l_i, u_i]$.

For the purpose of our analysis, the following assignments are made:

x: descriptors;

–f(**x**): QSAR model;

–g_1(**x**): AD model;

g_k(**x**) (k = 2, …, j), h(**x**): constraints for descriptors;

l_i, u_i: lower and upper bounds of i-th descriptor.

Constraints are applied to descriptors to ensure meaningful value ranges. For example, if the 'number of heavy atoms' (x_p) and 'number of hydrogen bond acceptors' (x_q) are selected descriptors, a value of five for the former and six for the latter would be impossible for any given data point (compound). Therefore, in order to prevent such settings, an inequality constraint is required and applied: $x_q - x_p \le 0$.

ε Differential evolution

For addressing COP, the differential evolution (DE) algorithm originally introduced by Stone and Price[17] is investigated herein, which has so far not been considered in inverse QSAR. However, given the conceptual simplicity and computational efficiency of DE, the algorithm has been successfully applied to solve optimization problems in other areas of science and engineering, for example, in scheduling of flow shops[18]. In addition, for deriving a COP solution efficiently, ε differential evolution (εDE) was introduced by Takahama *et al.* as an extension of DE[19], illustrated in Figure 2. A candidate vector **v** for next generation (also called mutant vector) is derived on the basis of three randomly selected vectors:

$$\mathbf{v} = \mathbf{x}_{r1} + F(\mathbf{x}_{r2} - \mathbf{x}_{r3}),$$

where $\mathbf{x}_{r1}$, $\mathbf{x}_{r2}$, $\mathbf{x}_{r3}$ are different vectors from the current generation and F represents a scale parameter for the difference vector. If a variable adopts only discrete values, its value in **v** can be rounded to an integer. If the i-th component of **v**: v_i falls outside the range $[l_i, u_i]$, **v** is updated as follows:[4]

$$v_i = \begin{cases} \min\ \{u_i, l_i + (l_i - v_i)\}, & \text{if } v_i < l_i \\ \max\ \{l_i, u_i - (v_i - u_i)\}, & \text{if } u_i < v_i \end{cases}.$$

An exponential crossover operation with probability-based crossover points is applied to **v** (the probability is called CR). Either $\mathbf{x}_i$ or **v** is selected as $\mathbf{x}_{i+1}$, the individual for the next generation, following ε level comparison of the corresponding vectors.

ε Level comparison

For prioritizing candidates, given constraints and the optimized function are taken into account. The constraint violation Φ(**x**) is defined as follows:

$$\Phi(\boldsymbol{x}) = \sum\nolimits_{j=1}^{q} \max\{0, g_j(\boldsymbol{x})\}^p + \sum\nolimits_{j=q+1}^{m} \left|h_j(\boldsymbol{x})\right|^p,$$

where Φ(**x**) represents the degree of violation, with p set to one. The ε level comparison determines the order between two sets of pair (f(**x**), Φ(**x**)):

$$(f_1, \Phi_1) <_\epsilon (f_2, \Phi_2) \Leftrightarrow \begin{cases} f_1 < f_2, \text{ if } \Phi_1, \Phi_2 \le \epsilon(t) \\ f_1 < f_2, \text{ if } \Phi_1 = \Phi_2 \\ \phi_1 < \phi_2, \ otherwise \end{cases},$$

where t represents a generation in DE. As a decreasing function of t, ε determines the tolerance of constraint violation and ε(t) is determined as follows:[17]

$$\epsilon(t) = \begin{cases} \Phi(\boldsymbol{x}_\theta) & t = 0 \\ \Phi(\boldsymbol{x}_\theta)(1 - \dfrac{t}{T_c})^{cp} & 0 < t < T_c \\ 0 & T_c \le t \end{cases},$$

where $\mathbf{x}_\theta$ is the top θ-th individual, T_c determines the generation in which ε(t) becomes zero, and cp the convergence speed. During the

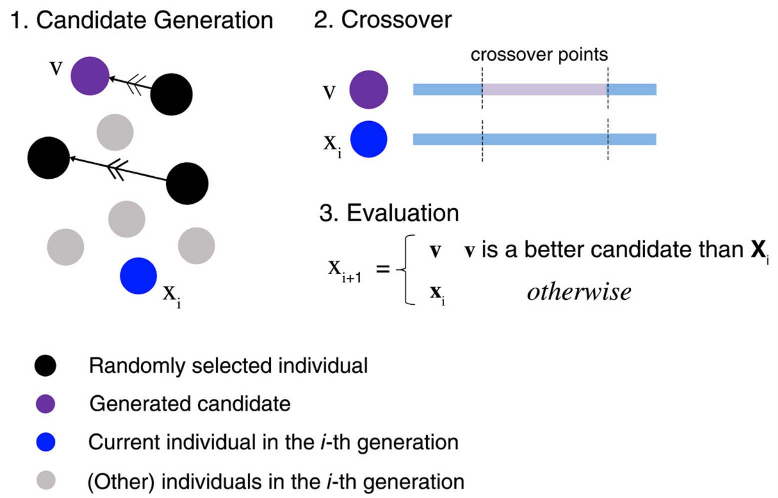
1. Candidate Generation
2. Crossover
v
x_i
crossover points
v
x_i
3. Evaluation
x_{i+1} = { v   v is a better candidate than X_i ; x_i   otherwise
● Randomly selected individual
● Generated candidate
● Current individual in the i-th generation
● (Other) individuals in the i-th generation

Figure 2. Evolutionary algorithm. The steps involved in evolutionary optimization are outlined. First, a candidate $\mathbf{v}$ is obtained from three randomly selected individuals by a differential operation. Second, a crossover operation is applied to an individual $\mathbf{x}_i$ and the candidate. Third, the evaluation step involves ε level comparison of $\mathbf{v}$ and $\mathbf{x}_i$ and results in the next individual.

optimization, $\Phi(\mathbf{x})$ gradually outweighs $f(\mathbf{x})$. In the initial stages of the optimization $\varepsilon(t)$ settings enable the selection of diverse candidates but convergence of the algorithm is determined by $\Phi(\mathbf{x})$ becoming zero. Accordingly, T_c was set to one herein.

Regression and applicability domain models

For εDE optimization, any regression function can be employed. In this study, support vector regression (SVR)[20] with ν parameter was applied and the AD was defined by one-class support vector machine (OCSVM)[21] classification with ν parameter. This parameter ranges from zero to one and defines the upper bound of the fraction of margin error and lower bound of the fraction of support vectors. AD consists of regions where the output of OCSVM is greater than or equal to zero. For both SVR and OCSVM, the radial basis function (RBF) kernel: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ was used. A hyper parameter set {C, ν, γ} for ν-SVR was determined by cross validation of training data on the basis of Q^2. For OCSVM model construction, γ was set to maximize the variance of the Gram matrix consisting of the kernel function of the training data[22] and ν was set to 0.01.

Simulation data

Data points on a (x_1, x_2) plane were randomly generated for x_1: [-2 3], x_2: [-1, 4] to yield 50 training and 20 test instances. The corresponding y values were calculated using Mishra's bird function (https://mpra.ub.uni-muenchen.de/2718/) adding a Gaussian error with a mean of zero and variance of one, defined as:

$$f(x_1, x_2) = \sin(x_1)\exp\{(1-\cos(x_2))^2\} + \cos(x_2)\exp\{(1-\sin(x_1))^2\} + (x_1 - x_2)^2.$$

Three independent trials were carried out with different random number generators. Training and test data sets were plotted on the output domain of the bird function with color-coded *true y* values (Figure 3).

Compound data sets

From ChEMBL[23] (version 22), compound data sets were selected using the following criteria: Maximal assay confidence score of '9', interaction relationship type 'D' (direct), activity standard unit 'nM', activity standard type 'K_i', and activity standard relation '='. When multiple K_i values were available for a compound, their geometric mean was calculated to yield its final potency value, provided all measurements fell into the same order of magnitude (otherwise, the compound was discarded). In-house implementation of substructure filters for pan assay interference compounds (PAINS)[24] and other reactive molecules were applied to eliminate compounds with potential chemical liabilities. Filtering was not critical for modeling, but active compounds with sound chemical structures were desired. From qualifying data sets, nine activity classes were randomly selected, as summarized on Table 1.

For each compound, 44 descriptors were initially calculated using RDKit (http://www.rdkit.org). These descriptors included constitutional descriptors (e.g., MW, number of rings, number of rotatable bonds), topological descriptors (e.g., Chi and Kier indices[25]) and partial charge descriptors based on chemical graph's topology (i.e., maximum of Gasteiger/Marsali partial charges[26]). The set of 44 descriptors and the nine compound data sets used herein are made available in an open access deposition on the ZENODO platform[27]. From correlated pairs of descriptors exceeding a correlation coefficient of 0.9, only one was chosen. For each activity class, the final number of descriptors (variables) is reported in Table 1. Compounds from each class were randomly divided into equally sized training and test data sets.

Virtual screening

To test the ability of virtual screening (VS) to identify new active compounds from optimized coordinates, ChEMBL (version 22) was used as a screening source. From 1,414,176 unique compounds passing the substructure filters, training molecules used for modeling of each activity class were removed. All remaining ChEMBL compounds provided a large screening source for VS. For screening compounds, descriptors were calculated as described above and the compounds falling inside the AD of each class-specific model were preselected. Active compounds from each activity class not used for training

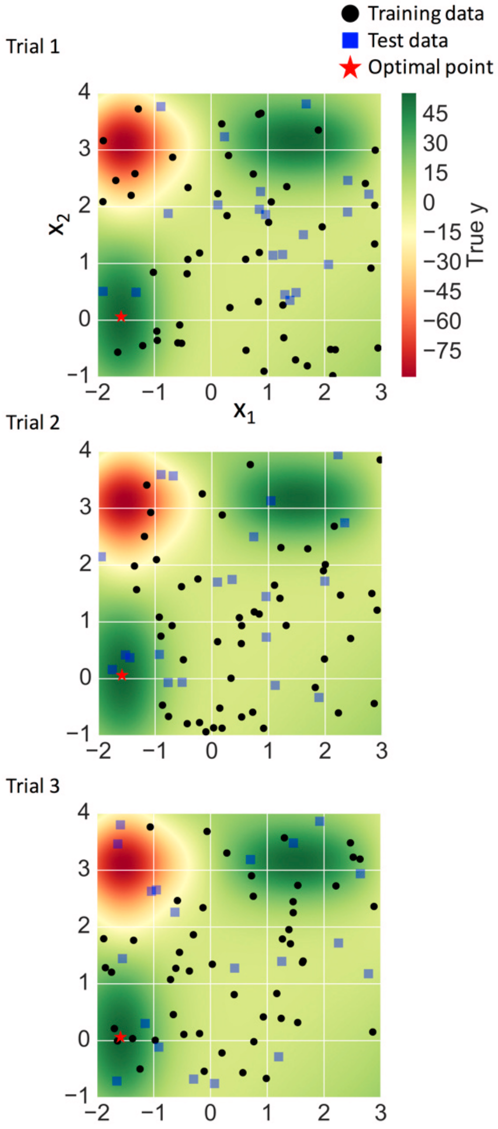

Figure 3. Simulation data sets. In three independent trials, simulation data sets were generated. For each trial, training (black dots) and test (blue squares) data are shown with true y values produced by the bird function f(x_1, x_2).

Table 1. Compound data sets. Nine compound activity classes were taken from ChEMBL (version 22). For each activity class, the target ID (TID), number of compounds (CPDs) for which descriptor values were obtained, and number of descriptors following variable selection are reported.

TID	Activity class	# CPDs	# Descriptors
11	Thrombin inhibitors	1022	26
15	Carbonic anhydrase II inhibitors	2387	26
51	Serotonin 1a (5-HT1a) receptor ligands	1939	26
100	Norepinephrine transporter ligands	1179	26
107	Serotonin 2a (5-HT2a) receptor ligands	1570	26
194	Coagulation factor X inhibitors	1586	25
10193	Carbonic anhydrase I inhibitors	2380	26
12209	Carbonic anhydrase XII inhibitors	1750	26
12968	Orexin receptor 2 ligands	1041	29

represented true-positive test instances, regardless of their potency values. The calculation of descriptors for more than 1.4 million screening compounds was computationally demanding and exceeded the requirements for coordinate optimization.

For ChEMBL screening compounds including test instances, two VS ranking were generated. First, Euclidean distances to optimized coordinates were calculated. In this case, compound potency was not considered for ranking. Second, pK_i values were predicted for all screening compounds using the class-specific SVR models. The latter calculations were carried out to determine if true positives were highly ranked on the basis of activity predictions. The area under the receiver operating characteristic curves (AUC) was calculated as an evaluation criterion.

Analysis protocol

Two proof-of-concept studies were carried out, one using simulation data, the other compounds and their activities. For simulation data, AD and regression models were constructed with the training data from each trial. Training data range scaling within the interval [-1,1] was applied prior to model building. For the SVR models, preferred parameter settings were determined using 10-fold cross validation. Coordinate optimization was carried using individual training data points. Optimized coordinates were evaluated on the basis of true y and maximal training data y values.

The same protocol for coordinate optimization was applied to each compound activity class. Furthermore, for hyper-parameter optimization of SVR, five-fold cross validation was carried out. For εDE, predicted pK_i values falling into the AD of each model were used to ensure that optimized coordinates were proximal to compound coordinates, as assessed by distance calculations. Furthermore, optimized coordinates were projected on principal component analysis (PCA) maps of the **x** space formed by the first and second PC. A matched molecular pair (MMP)-based[28] analog search for compounds proximal to the optimized coordinates was also carried out to investigate if analogs exhibited lower (higher) potency when they were distant from (close to) the optimized coordinates. Based on the distance to the optimized coordinates, the top 30 compounds from both training and test data sets were selected. MMPs qualified for this analysis if participating compounds displayed a potency dif-ference of at least one order of magnitude.

The following εDE parameter settings were applied: Number of iter-ations, 1,000 and 10,000 for simulation and compounds data sets, respectively; F, 0.5; T_c,1; p, 1; CR, 0.9 for all data sets. An initial population was obtained using 50 vectors of training instances for simulation data and 511 to 1,193 vectors for training compounds, depending on the size of the compound data sets.

Finally, the ability of distance- and SVR-based VS to predict new active compounds was analyzed. Although *de novo* structure generation was beyond the scope of our investigation, VS might be considered as an alternative way to identify novel active compounds, which was thus examined in the context of our study.

Implementation

All SVR models and ADs were constructed with Scikit-learn[29] 0.18.1 using Python. εDE was implemented in C++. Descriptors were calculated using RDKit interfaced with Python.

All selected compound entries were standardized by removal of ions and solvent molecules and structure regularization, according to the OEChem toolkit (v1.7.7; OpenEye Scientific Software, Inc. Santa Fe, NM, USA).

Results and Discussion

Differential evolution for inverse QSAR

Optimization of descriptor coordinates for preferred values of a given model is a central aspect of the two-stage inverse QSAR process, for which currently only approximate solutions exist. Therefore, a more accurate methodology for coordinate optimization is highly desirable, as investigated herein. The evaluation of εDE as an optimization method for this critical task was inspired by previous results obtained for other types of optimization problems where this approach displayed better performance than alternative evolutionary methods, such as genetic algorithms or particle swarm optimization[30–32]. Moreover, ε-based lexicological comparison of individual feature vectors makes this algorithm straightforward to apply to problems where several constraints must be balanced, as is the case in inverse QSAR. Studies on simulation and compound data sets were designed to evaluate whether εDE was capable of effectively optimizing coordinates on the basis of a regression function.

Investigating simulation data

For initial proof-of-concept, εDE-based search for optimized coordinates was carried out using simulation data generated as described above.

For the three simulation data sets, SVR models were built yielding optimized parameters {γ, ν, C} of {4, 0.25, 2}, {2, 0.25, 1}, {1, 0.125, 16}, respectively. For all OCSVM models, γ was 1. As reported in Table 2, these SVR models accounted for the output of the bird function in a statistically meaningful way.

Standard deviations in the test data sets for trials 1, 2, and 3 were 14.55, 23.21, and 30.42, respectively. Figure 4 shows the

Table 2. Derivation of the support vector regression model for simulation data sets. For each trial, model performance was assessed on the basis of R^2 and root mean square error (RMSE) values for training and test data sets.

Trial	Training		Test	
	R^2	RMSE	R^2	RMSE
1	1.00	1.10	0.87	5.17
2	0.96	3.22	0.70	12.73
3	0.97	3.17	0.84	11.99

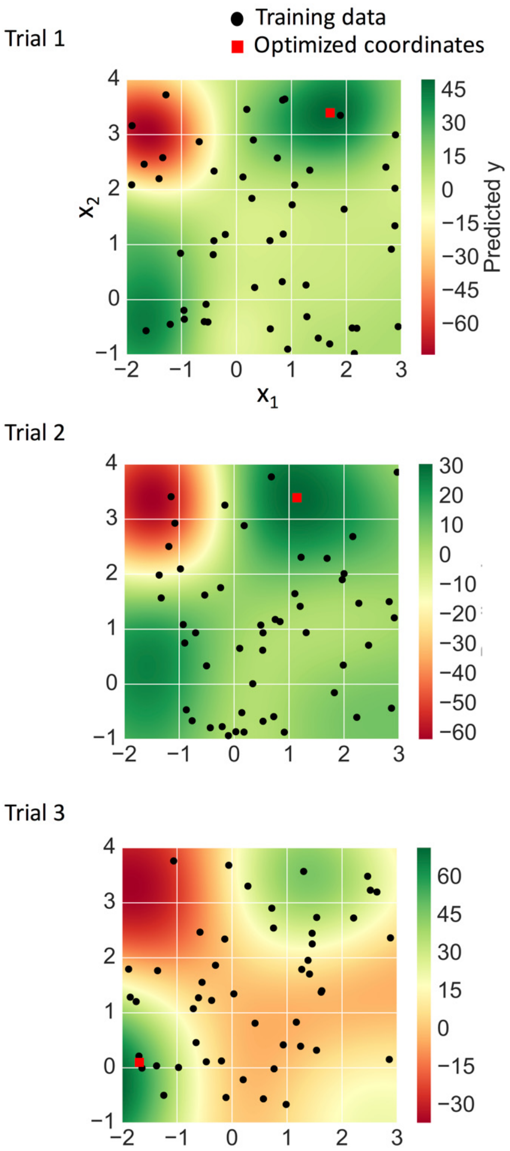

Figure 4. Optimized coordinates. For each of three independent trials, optimized coordinates (red squares) and training data (black dots) were mapped on the SVR prediction surface.

different prediction surfaces of the SVR models for the three trial sets. The surfaces of set one and two were overall similar, whereas the surface of set three differed from the others. In each case, however, individual vectors converged at a single point (Table 3) and optimized coordinates were located in regions of highest predicted y values (Figure 4). In set one, for which the SVR model overall best accounted for the bird function, a training data point was found adjacent to the optimized coordinates, which slightly exceeded the largest predicted y value (Table 3).

For set two, no training data were mapped to local maxima of the bird function, which resulted in a difficult regression scenario. The maximal measured y value in the training data was 29.46 and for optimized coordinates (1.51, 3.16), the predicted y value was 31.14, also slightly exceeding the largest measured y value. However, the predicted value was much smaller than the corresponding true y value of 48.09 (Table 3), due to the inherent regression limitation.

For set three, the maximal true y value within the domain was 56.18 at (-1.59, 0.06). In this case, several training data points were mapped to regions of y values into which optimized coordinates fell (Figure 4), leading to an extrapolative over-prediction of the corresponding y value of 71.94. However, this over-prediction was correctly balanced when the AD of the model was considered instead, leading to a value of 59.40 and a predicted y value for the optimized coordinates of 59.41 (Table 3).

Despite typical regression limitations highlighted by findings for set two and three, the results obtained for simulation data indicated the potential of εDE for predicting optimized coordinates.

Importantly, all solutions converged to single vectors representing novel points in simulation data space with large predicted y values falling inside the AD. Taken together, these results indicated the principal potential of εDE for coordinate optimization on the basis of SVR modeling.

Coordinate optimization for compound data sets

Next εDE optimization was applied to different compound activity classes. In each case, SVR models were derived, optimized coordinates determined, and activity values predicted.

For each compound class, optimized coordinates yielded larger predicted pK_i values than any training or test set compound (Table 4), consistent with the methodological strategy. Optimized coordinates fell inside the AD of each model and were proximal to several active compounds. When setting the ν parameter value to 0.1, hence restricting compound numbers within the AD, optimized coordinates remained unchanged compared to the ν parameter setting of 0.01. Nearest neighbors of optimized coordinates were mostly predicted to be highly active (Table 4), indicating the presence of smooth prediction surfaces in the vicinity of optimized coordinates.

Prediction surfaces were further characterized graphically by systematically comparing predicted pK_i values of compounds and calculated distances to optimized coordinates. Figure 5 shows the results for two exemplary activity classes, and Figure S1 shows the results for all classes. For set 51 (5-HT1a receptor ligands) in Figure 5, many highly active compounds were located proximal to the optimized coordinates, indicating that these coordinates fell into a well-populated region of activity-relevant chemical space. For set 194 (factor X inhibitors), training and test compounds tended to exhibit higher predicted pK_i with decreasing distance to the optimized coordinates, hence delineating regions of activity progression, which are relevant for compound optimization and exploitation of optimized coordinates.

Data set compounds and optimized coordinates were also projected onto PCA plots of descriptor space (Figure 6, Figure S2). These projections revealed that optimized coordinates were central to

Table 3. Prediction of y values. For each simulation data trial, y values predicted by the SVR model are reported. For training data, the largest measured y value is given in parentheses. "Domain" is defined by x_1 and x_2 with a resolution of 0.005. AD refers to the applicability domain of the OCSVM model. For optimized coordinates, the result of the bird function is given in parentheses as the true y value. (i.e., the y value without error).

Trial	Training (*measured y*)	Domain	AD	Optimized coordinates (*true y*)
1	48.92 (50.02)	50.19	50.19	50.19 (49.54)
2	26.22 (29.46)	31.14	31.14	31.14 (48.09)
3	58.87 (55.18)	71.94	59.40	59.41 (55.15)

Table 4. Optimized coordinates and nearest neighbors. For optimized coordinates, the predicted pK_i value and the output of the OCSVM model for the applicability domain (AD) are reported. For training and test instances, the predicted pK_i value and scaled distance from optimized coordinates are given for the nearest neighbor (NN).

TID	Optimized coordinates		NN in training data		NN in test data	
	Predicted pK_i	AD	Distance	Predicted pK_i	Distance	Predicted pK_i
11	11.49	0.26	0.61	9.50	0.52	10.02
15	12.20	0.06	0.46	10.21	0.45	10.32
51	10.25	0.62	0.39	9.24	0.40	9.09
100	9.50	0.22	0.32	8.63	0.46	7.97
107	11.43	0.32	0.71	9.78	0.71	9.84
194	13.03	0.17	0.82	10.92	0.67	11.90
10193	9.92	0.23	0.33	8.86	0.47	7.56
12209	10.48	0.35	0.72	8.99	0.78	8.69
12968	10.12	0.21	1.06	9.40	1.07	9.39

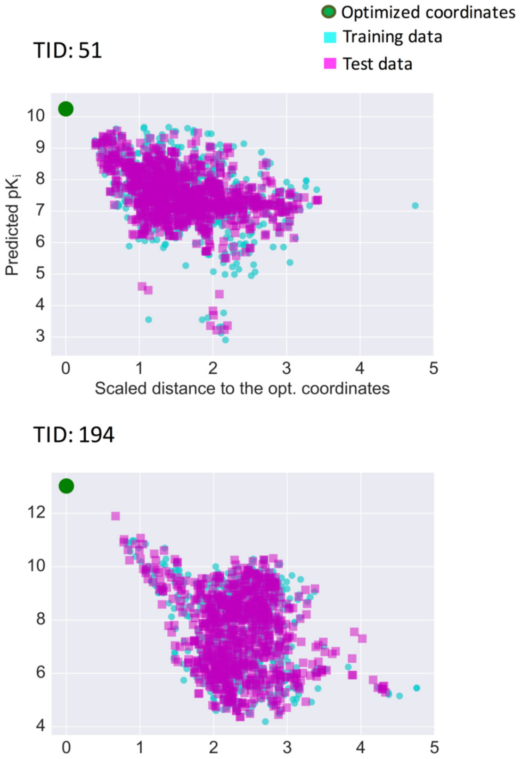

Figure 5. Activity prediction. For two exemplary activity classes, predicted pK_i values are related to the scaled distance of the corresponding compounds to the optimized coordinates. Training data (cyan squares), test data (magenta squares), and optimized coordinates (green circle) are shown.

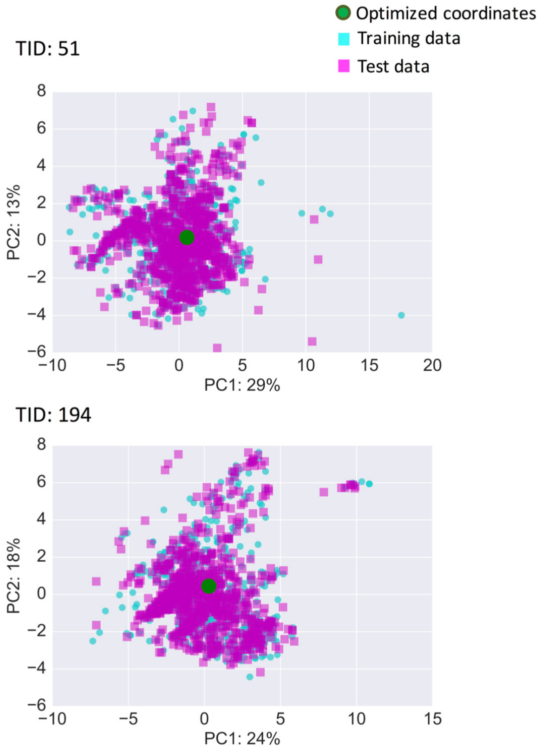

Figure 6. Projection of optimized coordinates. For the two activity classes from Figure 5, training data (cyan squares), test data (magenta squares), and optimized coordinates (green circle) were projected on a principal component (PC) plot derived from training data. For PC1 and PC2, contributions to data set variance are reported in %.

activity class regions in feature space. Furthermore, Figure 7 shows structures of the three nearest neighbors of the optimized coordinates for sets 51 and 194. In both cases, these compounds were structural analogs. Hence, similarity in feature space corresponded to close structural relationships. Consequently, this would also apply to structure generation from optimized coordinates, which would result in additional analog(s), consistent with the principles of QSAR and inverse QSAR.

A search was carried out for analogs of the top 30 compounds based on the distance to the optimized coordinates and the consistency ratio was calculated for each target (Table 5). The number of qualifying MMPs ranged from 14 to 464. The consistency ratios were greater than 0.5 for eight of nine activity classes, implying that structures approaching optimized coordinates may have increasing potency.

The εDE protocol can also be modified to generate a family of solutions. One possibility is splitting a data set into subsets of structurally related compounds, followed by the εDE optimization for finding the optimized coordinates on a per subset basis. Furthermore, since εDE is a methodology for solving a COP, it is also possible to incorporate multiple constraints as solubility and toxicity during the optimization process as long as these constraints are formulated as inequality equations using the same descriptors as for the objective function.

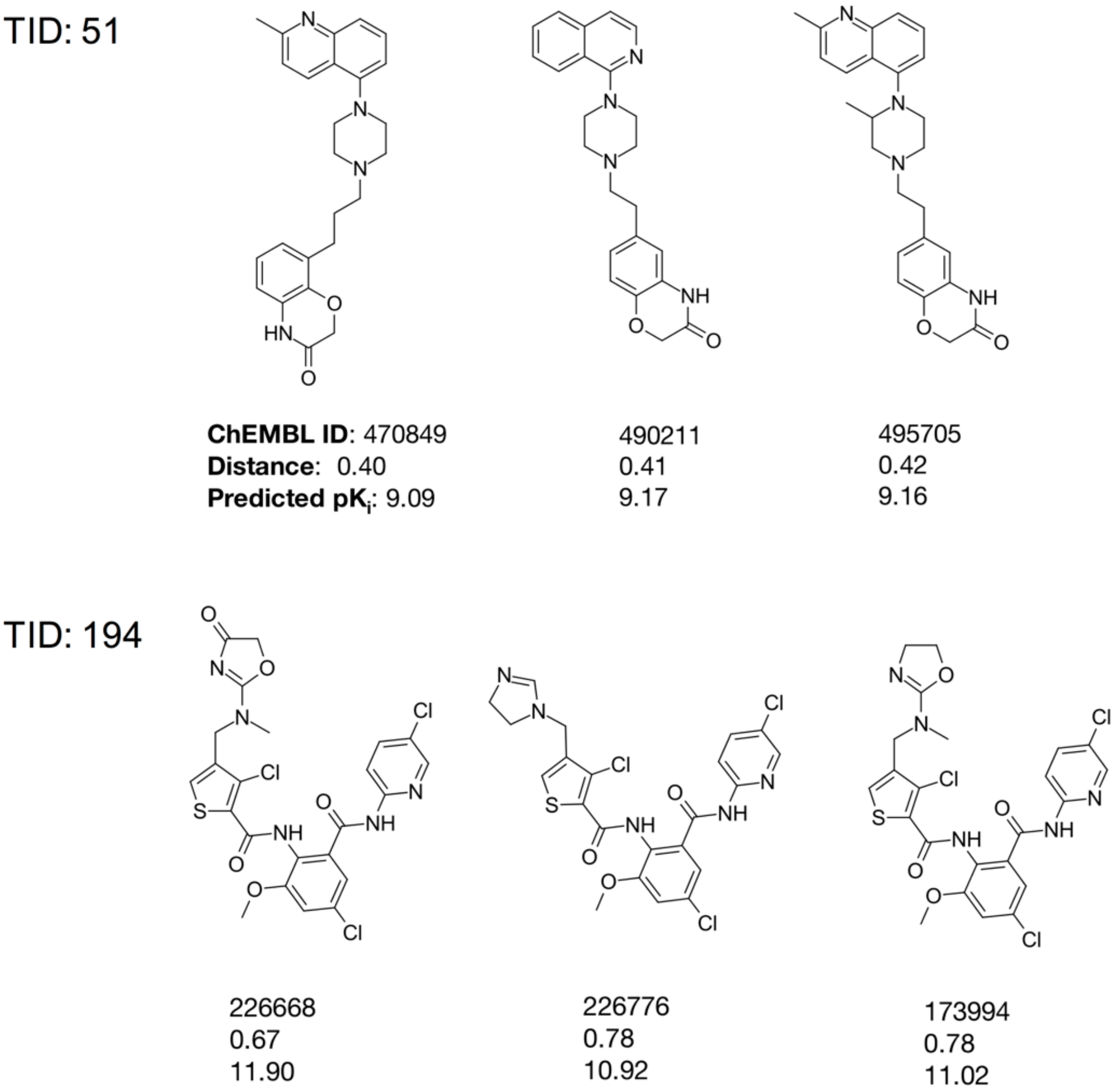

Figure 7. Nearest neighbors of optimized coordinates. For the two activity classes in Figure 5 and Figure 6, structures of the three nearest neighbors of optimized coordinates are shown and their ChEMBL IDs, scaled distances to the optimized coordinates and predicted pK$_i$ values, are reported.

Table 5. MMP-based analog search of the top 30 compounds. The consistency ratio and number of MMPs for the top 30 compounds based on the distance to the optimized coordinates are reported.

TID	Consistency ratio	# MMPs
11	0.62	253
15	0.71	68
51	0.88	108
100	0.64	14
107	0.33	93
194	0.86	464
10193	0.68	130
12209	0.75	60
12968	0.84	86

Table 6. Virtual screening details. Compound (CPD) statistics and VS results for distance-based compound rankings relative to optimized coordinates and potency-based rankings are reported.

TID	# ChEMBL compounds			AUC		True positive ratio (top 30 compounds)	
	Screening CPDs	CPDs in AD	True positive	Distance-based	Potency-based	Distance-based	Potency-based
11	1,413,665	1,004,761	497	0.76	0.51	0.30	0.13
15	1,412,983	1,020,546	1145	0.48	0.77	0.10	0.07
51	1,413,207	822,453	935	0.66	0.69	0.43	0.27
100	1,413,587	835,967	561	0.64	0.75	0.10	0.17
107	1,413,391	865,660	736	0.56	0.61	0.10	0.10
194	1,413,383	535,306	727	0.51	0.64	0.17	0.13
10193	1,412,986	1,172,027	1152	0.62	0.68	0.07	0.07
12209	1,413,301	942,956	838	0.55	0.73	0.00	0.00
12968	1,413,656	217,995	502	0.60	0.50	0.03	0.03

Virtual screening

ChEMBL compounds were screened relative to optimized coordinates from the nine activity classes and Euclidian distances were determined. Furthermore, pK_i values were predicted for screening compounds falling into the AD of each class-specific SVR model. The AD was simply defined as the region where the OCSVM output was greater than or equal to zero. Since the nu value of the OCSVM models was set to 0.01, most of the training and test compounds were classified inside the AD. For example, for data set 51, 97% of the training and 96% of the test compounds fell inside the AD. The VS calculations ultimately led to alternative distance- and potency-based compound rankings. Table 6 reports screening compound statistics and VS results. For each activity class, screening compounds contained large numbers (497–1152) of true positive test instances. Distance-based VS yielded AUC values of at least 0.6 for five of nine activity classes, with a maximum of 0.76. For the four remaining classes, essentially random predictions were observed. Potency-based VS produced AUC values of greater than 0.6 for seven of nine classes, including values above 0.7 for three classes and a maximum of 0.77. Thus, potency-based predictions led to slightly better compound rankings than distance-based VS relative to optimized coordinates, but prediction accuracy was overall moderate at best. Moreover, the true positive ratio among the 30 top-ranked compounds was generally very low for both distance- and potency-based VS. Figure 8 shows exemplary potency prediction landscapes including optimized

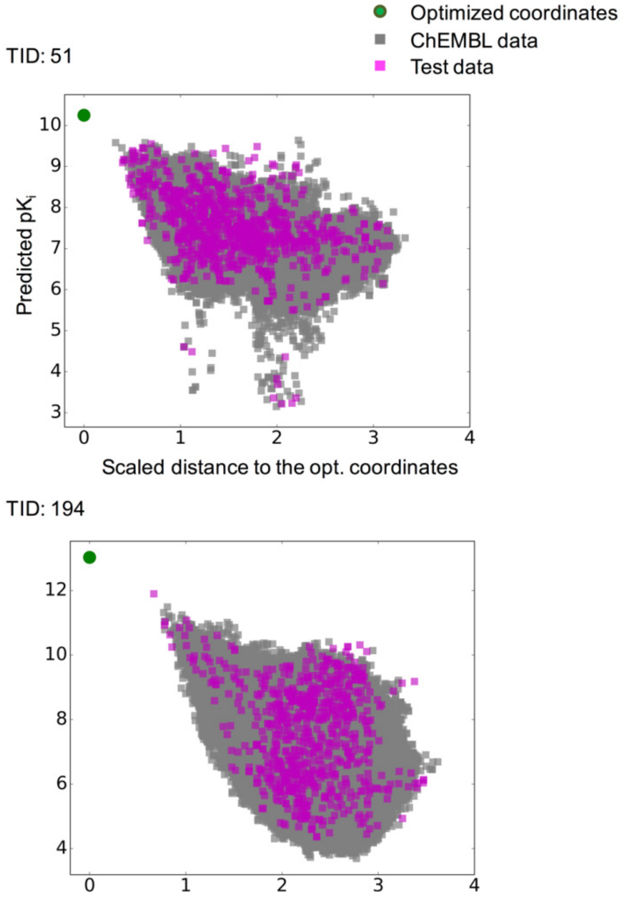

Figure 8. Activity prediction for ChEMBL compounds. For the two activity classes from Figure 5, predicted pK$_i$ values are plotted against the scaled distance of corresponding compounds to the optimized coordinates. ChEMBL compounds (gray squares) and test compounds according to Figure 5 (magenta squares) falling inside the applicability domain are shown. Optimized coordinates are displayed as a green circle.

coordinates as a reference. Figure S3 shows these representations for all activity classes. Highly potent compounds proximal to optimized coordinates were predicted for several activity classes. However, most true positives were not separated from the bulk of ChEMBL screening compounds on the basis of potency predictions. Overall the ability of VS calculations to identify novel active compounds and separate them from false positives was only limited. Thus, although *de novo* structure generation from optimized coordinates is challenging, it would be difficult to replace the structure generation step in two-stage inverse QSAR with standard VS calculations. However, despite limited prediction accuracy, the VS calculations provided support for the chemical relevance of optimized coordinates because for each activity class, at least few true positives were among top-scoring screening compounds and proximal to optimized coordinates.

Conclusions

The optimization of coordinates in feature space for high activity values predicted with a regression model is a central task in two-stage inverse QSAR. For this multi-constraint optimization problem, no generally applicable approach is currently available. The evaluation of differential evolution for coordinate optimization, as reported herein, was motivated by the successful application of this algorithm in areas of science other than chemistry. The study has provided proof-of-concept evidence that εDE is suitable for generating optimized coordinates in given feature spaces. For different compound classes, consistent predictions were obtained for εDE in combination with SVR, displaying robust convergence behavior and yielding optimized coordinates that not only met statistical and data set requirements, but were also chemically relevant, as indicated by compound mapping and distance- or potency-based VS calculations. However, due to limited prediction accuracy, distance-based VS relative to optimized coordinates would not be suitable to replace the *de novo* structure generation step in inverse QSAR, at least not on the basis of our reference calculations. Regardless, encouraging results were obtained for coordinate optimization. Taken together, the findings reported herein indicate that εDE optimization has the potential to further advance inverse QSAR analysis.

Data availability

ZENODO: Compound data sets for inverse QSAR and a descriptor list[27].

Competing interests

No competing interests were disclosed.

Grant information

The project leading to this report has received funding (for TM) from the *Japan Society for the Promotion of Science* (JSPS) under the JSPS KAKENHI Grant Number 16J05325.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

TM is a Fellow of the *Japan Society for the Promotion of Science*.

Supplementary material

Supplementary Figures S1-S3: Prediction surfaces and PCA projections for different activity classes and prediction surfaces for the ChEMBL screening data set, respectively.

Click here to access the data.

References

1. Kier LB, Hall LH, Frazer JW: **Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts.** *J Chem Inf Comput Sci.* 1993; **33**(1): 143–147.
Publisher Full Text
2. Hall LH, Kier LB, Frazer JW: **Design of Molecules from Quantitative Structure-Activity Relationship Models. 2. Derivation and Proof of Information Transfer Relating Equations.** *J Chem Inf Comput Sci.* 1993; **33**(1): 148–152.
Publisher Full Text
3. Skvortsova MI, Baskin II, Slovokhotova OL, *et al.*: **Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indexes Characterizing Molecular Shape (Kier Indices).** *J Chem Inf Comput Sci.* 1993; **33**(4): 630–634.
Publisher Full Text
4. Skvortsova MI, Fedyaev KS, Palyulin VA, *et al.*: **Inverse Structure-Property Relationship Problem for the Case of a Correlation Equation Containing the Hosoya Index.** *Dokl Chem.* 2001; **379**(1–3): 191–195.
Publisher Full Text
5. Schneider G, Baringhaus KH: ***De Novo* Design: From Models to Molecules.** In: *De novo Molecular Design.* Wiley-VCH Verlag GmbH & Co. KGaA Weinheim Germany; 2013; 1–55.
Publisher Full Text
6. Speck-Planche A, Cordeiro MNDS: **Fragment-based *in silico* modeling of multi-target inhibitors against breast cancer-related proteins.** *Mol Divers.* 2017; **21**(3): 511–523.
PubMed Abstract | Publisher Full Text
7. Speck-Planche A, Dias Soeiro Cordeiro MN: **Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based *in Silico* Approach for the Design of Virtual Anti-Hepatitis C Leads.** *ACS Comb Sci.* 2017; **19**(8): 501–512.
PubMed Abstract | Publisher Full Text
8. Faulon JL, Visco DP Jr, Pophale RS: **The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies.** *J Chem Inf Comput Sci.* 2003; **43**(3): 707–720.
PubMed Abstract | Publisher Full Text
9. Faulon JL, Churchwell CJ, Visco DP Jr: **The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences.** *J Chem Inf Comput Sci.* 2003; **43**(3): 721–734.
PubMed Abstract | Publisher Full Text
10. Churchwell CJ, Rintoul MD, Martin S, *et al.*: **The signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides.** *J Mol Graph Model.* 2004; **22**(4): 263–273.
PubMed Abstract | Publisher Full Text
11. Weis DC, Faulon JL, LeBorne RC, *et al.*: **The Signature Molecular Descriptor. 5. The Design of Hydrofluoroether Foam Blowing Agents Using Inverse-QSAR.** *Ind Eng Chem Res.* 2005; **44**(23): 8883–8891.
Publisher Full Text
12. Wong WW, Burkowski FJ: **A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem.** *J Cheminform.* 2009; **1**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Miyao T, Kaneko H, Funatsu K: **Ring-System-Based Exhaustive Structure Generation for Inverse-QSPR/QSAR.** *Mol Inform.* 2014; **33**(11–12): 764–778.
PubMed Abstract | Publisher Full Text
14. Miyao T, Kaneko H, Funatsu K: **Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x).** *J Chem Inf Model.* 2016; **56**(2): 286–299.
PubMed Abstract | Publisher Full Text
15. Miyao T, Kaneko H, Funatsu K: **Ring system-based chemical graph generation for *de novo* molecular design.** *J Comput Aided Mol Des.* 2016; **30**(5): 425–446.
PubMed Abstract | Publisher Full Text
16. Gómez-Bombarelli R, Duvenaud D, Hernández-Lobato JM, *et al.*: **Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.** arXiv:1610.02415.
Reference Source
17. Storn R, Price K: **Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces.** *J Global Optim.* 1997; **11**(4): 341–359.
Publisher Full Text
18. Onwubolu G, Davendra D: **Scheduling Flow Shops Using Differential Evolution Algorithm.** *Eur J Oper Res.* 2006; **171**(2): 674–692.
Publisher Full Text
19. Takahama T, Sakai S: **Constrained Optimization by the ε Constrained Differential Evolution with Gradient-Based Mutation and Feasible Elites.** In: *2006 IEEE International Conference on Evolutionary Computation.* IEEE; 2006; 1–8.
Publisher Full Text
20. Smola AJ, Schölkopf B: **A Tutorial on Support Vector Regression.** *Stat Comput.* 2004; **14**(3): 199–222.
Publisher Full Text

21. Schölkopf B, Platt JC, Shawe-Taylor J, *et al.*: **Estimating the support of a high-dimensional distribution.** *Neural Comput.* 2001; **13**(7): 1443–1471.
PubMed Abstract | **Publisher Full Text**

22. Tang Y, Guo W, Gao J: **Efficient Model Selection for Support Vector Machine with Gaussian Kernel Function.** In: *2009 IEEE Symposium on Computational Intelligence and Data Mining*; IEEE, 2009; 40–45.
Publisher Full Text

23. Bento AP, Gaulton A, Hersey A, *et al.*: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1083–D1090.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Baell JB, Holloway GA: **New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.** *J Med Chem.* 2010; **53**(7): 2719–2740.
PubMed Abstract | **Publisher Full Text**

25. Hall LH, Kier LB: **The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling.** John Wiley & Sons, Inc.; 2007; **2**: 367–422.
Publisher Full Text

26. Gasteiger J, Marsili M: **Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges.** *Tetrahedron.* 1980; **36**(22): 3219–3228.
Publisher Full Text

27. Miyao T, Funatsu K, Bajorath J: **Compound data sets for differential evolution calculation and a descriptor list.** *Zenodo.* 2017.
Data Source

28. Griffen E, Leach AG, Robb GR, *et al.*: **Matched molecular pairs as a medicinal chemistry tool.** *J Med Chem.* 2011; **54**(22): 7739–7750.
PubMed Abstract | **Publisher Full Text**

29. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-Learn: Machine Learning in Python.** *J Mach Learn Res.* 2011; **12**: 2825–2830.
Reference Source

30. Dong X, Liu S, Tao T, *et al.*: **A Comparative Study of Differential Evolution and Genetic Algorithms for Optimizing the Design of Water Distribution Systems.** *J Zhejiang Univ Sci A.* 2012; **13**(9): 674–686.
Publisher Full Text

31. Tušar T, Filipič B: **Differential Evolution versus Genetic Algorithms in Multiobjective Optimization.** In: *Evolutionary Multi-Criterion Optimization.* Springer Berlin Heidelberg: Berlin, Heidelberg; 2007; 257–271.
Publisher Full Text

32. Iwan M, Akmeliawati R, Faisal T, *et al.*: **Performance Comparison of Differential Evolution and Particle Swarm Optimization in Constrained Optimization.** *Procedia Eng.* 2012; **41**: 1323–1328.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓

Version 1

Reviewer Report 29 August 2017

https://doi.org/10.5256/f1000research.13238.r24662

✓ **W. Patrick Walters**

Relay Therapeutics, Cambridge, MA, USA

This paper describes a new method for approaching one aspect of the inverse QSAR problem. As the authors point out, the inverse QSAR problem can be divided into two steps.

1. The generation of a set of coordinates, in some multidimensional space, that corresponds to one or more optimal compounds.
2. The generation of molecular graphs that would produce chemical structures with descriptor values corresponding to these coordinates.

This paper focuses on the first problem and does not address the second.

The paper is well written and the topic will be of interest to computational chemists working in both industry and academia. The methodology is described well and an individual with some QSAR expertise should be able to reproduce this work. However, in the interest of making it easier to reproduce the work described here, and making the method more widely available, it would be useful for the authors to make a reference implementation available. On a similar note, the authors mention that the ChEMBL datasets are available from the original source. As a service to those readers who would like to reproduce the work, it would be useful if the authors provided the specific datasets used in this paper as a download. As part of this download, the authors could also include a list of specific descriptors used and annotate which compounds were rejected due to PAINS filters.

The definition of the applicability domain used in this paper could be expanded. It would be useful for the authors to provide a specific worked example demonstrating how the applicability domain is defined for one of the ChEMBL examples described in the paper. This example could be expanded to demonstrate how a set of optimal coordinates is defined.

One other beneficial addition to this paper would be a comparison with established methods. The authors provide the results of virtual screening based on their method but do not provide a comparison with commonly used techniques. One way to do this would be to provide a simple comparison with an SVM model for activity calculated based on the same descriptions. A comparison with activity calculated based on nearest neighbors in a simple principal component space could also be provided.

A few specific comments:

It is unclear what the RMSE value is in Table 2. This is on an arbitrary dataset, how should RMSE be interpreted?

There appear to be a number of errors in the chemical structures in Figure 7.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: Prof Bajorath and I are Gateway Advisors for the channel in which the article is published.

Reviewer Expertise: computational chemistry, cheminformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 25 August 2017

https://doi.org/10.5256/f1000research.13238.r24700

✓ **Brian Goldman**
Modeling & Informatics, Vertex Pharmaceuticals, Boston, MA, USA

This is an extremely well written submission by Bajorath *et. al* detailing an important aspect of the inverse QSAR problem, namely the generation of a feature vector optimizing the output of a QSAR equation. An interesting and novel component of the work is that constrained optimization is utilized ensuring generated solutions lie within the domain of applicability of the QSAR model. This paper is primarily of theoretical interest in that due to inherent limitations with inverse-QSAR the technique is rarely adopted as a means of drug discovery. The primary reason for the lack of adoption revolving around the de-novo design of compounds matching optimized descriptor values.

The introduction to the paper covers the relevant literature but could be made stronger by discussing the recent work of Gómez-Bombarelli *et. al.*[1] who use generative models to approach the inverse-QSAR problem. Their work concerns building and optimizing QSAR equations in the latent space of an autoencoder and subsequently decoding optimized points into molecular structures. Although the work by Gómez-Bombarelli is preliminary, it addresses both QSAR optimization and structure generation and would most likely be interesting to readers of the current paper.

An aspect of the current technique that could be discussed in the paper is the generation of a family of solutions. Currently, the presented technique produces one optimized feature vector. It would strengthen the paper if the authors discussed how a family of diverse feature vectors (structures) could be evolved using the presented methodology.

Overall, the paper in its current form is more than acceptable. The methodology is clearly delineated and sufficiently supported by the included results.

Minor points:

- Figure 3: it would be informative to highlight the optimal point with a particular glyph (perhaps a red star?)

- Figure 7: Chemical structures look suspect and should be checked

References

1. Gómez-Bombarelli R, Duvenaud D, Hernández-Lobato JM, Aguilera-Iparraguirre J, et al.: Automatic chemical design using a data-driven continuous representation of molecules. *arXiv*. 2017. Reference Source

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: machine learning, cheminformatics, algorithms

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 August 2017

https://doi.org/10.5256/f1000research.13238.r24665

✓

Hans Matter
R&D IDD / Structure, Design & Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany

This interesting contribution by Bajorath *et al.* addresses an important part of the inverse QSAR problem towards automated generation of structures with high activity from QSAR models. Inverse-QSAR, while intellectually appealing, does not find significant applications in modelling in the pharmaceutical industry. Main limitations are linked to de-novo structure generation due to issues with synthetic accessibility. Another hurdle is the challenge to identify the optimal descriptor setting for a model. The authors focus on this latter point, namely a novel and accurate approach for coordinate optimization. They demonstrate how to generate optimal descriptor coordinates under certain constraints like model applicability domain and meaningful descriptor values. A convincing validation study using virtual screening in ChEMBL 22 is also presented.

The report title and abstract cover the content well. The chemoinformatics approach is well conducted and clearly described. The results are presented in a clear and interesting way and capture the interest of F1000 readers. The authors might also want to mention, whether software tools and subroutines from their study are available. Therefore this contribution is an essential view on interpretation of QSAR models and should be indexed in its present form.

Some points could be addressed to highlight further aspects of their work.

- How does such an optimization approach handle typical types of variables from real-life models, e.g. two-level variables, variables with small or no SD?

- Often model analysis should not result in a single solution, but multiple related structures.

Could the optimization approach find multiple descriptor regions to offer options for monitoring secondary properties (e.g. solubility)?

- It might be illustrative for one ChEMBL target to systematically generate analogs for potent leads close to the descriptor optimum by applying simple MedChem transformations and check, whether some analogs come closer to the optimum in descriptor space. Chemical changes are minimal here and one could access their impact to the descriptor optimum.

- The moderate VS success using QSAR models might suggest a non-optimal approach to define the applicability domain. Some details on the AD definition and the descriptor space might be useful. Do the authors expect that a more strict AD definition might produce reliable results? What does this mean for de-novo structure generation as second step?

- Is it possible to apply such a concept for multi-parameter optimization, e.g. multiple QSAR models combined for predicting compound profiles / selectivity / druglikeness?

- Minor point: Drawings of chemical structures in figure 7 need to be checked.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular modelling, drug design

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.