J Comput Aided Mol Des (2016) 30:209–217
DOI 10.1007/s10822-015-9893-9

Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign

Gregory Sliwoski[1,2] · Jeffrey Mendenhall[1] · Jens Meiler[1]

Received: 14 September 2015 / Accepted: 23 December 2015 / Published online: 31 December 2015

Abstract Quantitative structure–activity relationship (QSAR) is a branch of computer aided drug discovery that relates chemical structures to biological activity. Two well established and related QSAR descriptors are two- and three-dimensional autocorrelation (2DA and 3DA). These descriptors encode the relative position of atoms or atom properties by calculating the separation between atom pairs in terms of number of bonds (2DA) or Euclidean distance (3DA). The sums of all values computed for a given small molecule are collected in a histogram. Atom properties can be added with a coefficient that is the product of atom properties for each pair. This procedure can lead to information loss when signed atom properties are considered such as partial charge. For example, the product of two positive charges is indistinguishable from the product of two equivalent negative charges. In this paper, we present variations of 2DA and 3DA called 2DA_Sign and 3DA_Sign that avoid information loss by splitting unique sign pairs into individual histograms. We evaluate these variations with models trained on nine datasets spanning a range of drug target classes. Both 2DA_Sign and 3DA_Sign significantly increase model performance across all datasets when compared with traditional 2DA and 3DA. Lastly, we find that limiting 3DA_Sign to maximum atom pair distances of 6 Å instead of 12 Å further increases model performance, suggesting that conformational flexibility may hinder performance with longer 3DA descriptors. Consistent with this finding, limiting the number of bonds in 2DA_Sign from 11 to 5 fails to improve performance.

Keywords Quantitative structure activity relationship · Descriptor · 2D autocorrelation · 3D autocorrelation · Artificial neural network · Virtual high-throughput screening

Abbreviations

2DA	2D autocorrelation
3DA	3D autocorrelation
ANN	Artificial neural network
BCL	BioChemical library
CADD	Computer aided drug discovery
GPCR	G-protein coupled receptor
HTS	High-throughput screen
LB-CADD	Ligand-based CADD
logAUC	Area under the logarithmic ROC curve
LOO	Leave-one-out
QSAR	Quantitative structure–activity relationship
RDF	Radial distribution function
ROC	Receiver operating characteristic
VDW	Van der Waals

Electronic supplementary material The online version of this article (doi:10.1007/s10822-015-9893-9) contains supplementary material, which is available to authorized users.

✉ Jens Meiler
jens@meilerlab.org;
http://www.meilerlab.org

Gregory Sliwoski
gregory.r.sliwoski@vanderbilt.edu

[1] Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology, Institute for Chemical Biology, Vanderbilt University, 7330 Stevenson Center, Station B 351822, Nashville, TN 37235, USA

[2] Institute of Biochemistry, Leipzig University, Brüderstraße 34, 04103 Leipzig, Germany

Introduction

Computer aided drug discovery (CADD) is a multi-faceted approach that implements computational tools into the drug discovery pipeline [1]. CADD can reduce the time and resources required for the development of novel therapeutics. Scientifically, CADD can also provide insights into the complex interaction between small molecule and a biological target protein. Ligand-based CADD (LB-CADD) is one approach that focuses on analyzing the collective chemical properties of a set of active and inactive compounds without leveraging explicit knowledge of the target protein structure. One fundamental principle of LB-CADD is quantitative structure–activity relationship (QSAR) modeling. The goal of QSAR modeling is to define the relation between chemical structure and biological activity in a quantitative way so that the activity of new molecules can be predicted to prioritize acquisition or synthesis. In general, QSAR can be separated into two major components: a quantitative description of molecular structure (descriptor) and a mathematical model that uses these multidimensional descriptors as input to predict activity. Both components come in a variety of flavors and strategies that vary in performance depending on the specific project. Machine learning techniques are the most commonly applied non-linear mathematical QSAR models [2]. For this study, we use Artificial Neural Networks (ANN) as implemented in BCL::ChemInfo [3] to generate our mathematical models across all conditions.

Descriptors of chemical structure are typically computed as a combination of atomic properties (mass, volume, surface area, partial charge, electro-negativity, polarizability, etc.) that are processed with a translation and rotation invariant geometric function to describe the distribution of these properties in the molecular structure. Descriptors can be grouped into five categories, depending on the 'dimensionality' of the small molecule description required: (1D) Descriptors that can be derived from the molecular formula such as molecular weight by summing up all atom masses or total charge by summing up nominal charges. (2D) Descriptors that depend on constitution such as the number of hydrogen bond donors/acceptors, number of ring systems, topological surface area, and some approximations of volume and surface area. A topological index, for example, encodes which atoms a chemically bonded [4]. (2.5D) Configuration-dependent descriptors that encode, for example, the relation of stereo-centers within a topological index [5]. (3D) Conformation-dependent descriptors including Radial Distribution Functions (RDF) [6] and 3-Dimensional Autocorrelation (3DA) [7] that encode aforementioned atomic properties in a three-dimensional fingerprint. (4D) Descriptors that take conformational flexibility into account such as those derived from low energy conformational ensembles [8].

A descriptor is considered useful when it provides pertinent information about a compound while adding minimal noise to the overall model. In this respect, the most useful descriptors are the ones with the greatest degree of information density (information used by the model divided by total information). A descriptor that provides no useful information is often ignored by statistical models but can sometimes reduce model performance by overwhelming it with noise [9]. The goal of this paper is to evaluate potential improvements to 2DA and 3DA descriptors [7].

2DA [10] and 3DA [7] descriptors both generate histograms of atom pair distances within a molecule up to a cutoff distance. The major difference between these descriptors that designates their dimensionality is in their representation of interatomic distance. For 2DA, distances are measured in terms of the number of bonds between two connected atoms. 3DA, on the other hand, represents interatomic distance in terms of Euclidian distance typically measured in angstroms. To extend these descriptors beyond the geometric characteristics of a molecule, atom pair distances are weighted by atom properties such as partial charge, electronegativity, etc. The formal definition of 2DA and 3DA is shown in Eq. 1.

$$Autocorrelation\,(r_a, r_b) = \sum_{i}^{n}\sum_{j}^{n}\delta(r_a \leq r_{i,j} < r_b)P_iP_j \quad (1)$$

where r_{ij} is the distance between atoms i and j and n is the total number of atoms in the molecule. P_i and P_j are the atom properties for atoms i and j used to weight the autocorrelation. r_a and r_b define the lower and upper boundaries of each consecutive distance bin.

Weighting 2DA and 3DA with atom properties P_i and P_j allow these descriptors to encode the distribution of specific atom properties within a molecule. These properties may be unsigned in the case of atomic mass or signed in the case of partial charge. However, significant information loss arises when signed atom properties are used to weight 2DA and 3DA due to sign-cancellation. For example, a pair of atoms both with positive partial charges will be encoded the same as a pair with negative partial charges. Therefore, we introduce variations of 2DA and 3DA specifically for heterogeneously signed atom properties called 2DA_Sign and 3DA_Sign, respectively. With 2DA_Sign/3DA_Sign, we separate a single 2DA/3DA histogram into three: negative–negative, positive–positive, and opposite sign property pairs. Comparing 2DA_Sign and 3DA_Sign histograms with their traditional counterparts reveals the different forms of information loss that arise when weighting with signed atom properties. Figure 1a compares a single 2DA weighted with TotalCharge (TotalCharge = $\sigma + \pi$ partial charges) with the three histograms generated for the same molecule's TotalCharge-weighted 2DA_Sign. Figure 1b provides the same illustration for 3DA and 3DA_Sign weighted with

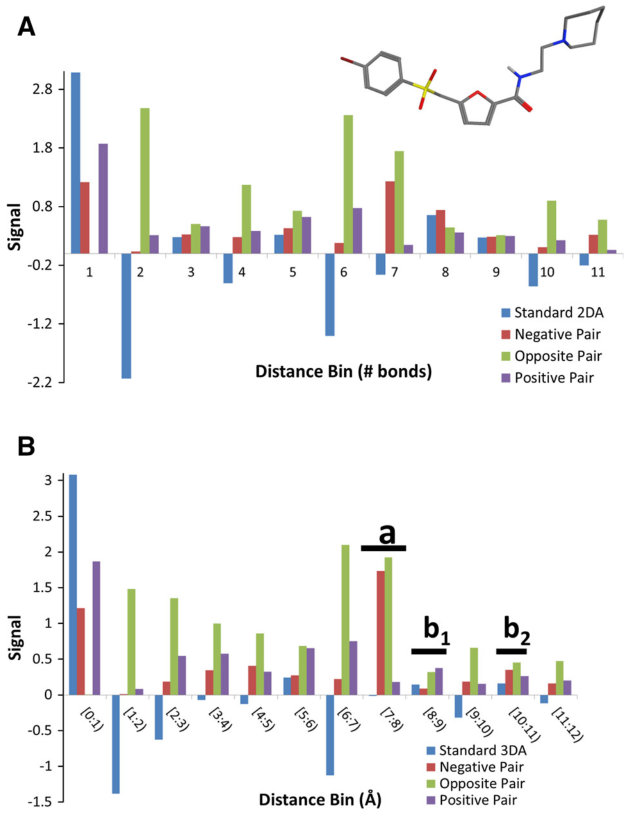

Fig. 1 2DA and 3DA lose information with weighted with signed atom properties. **a** Information loss is revealed when standard 3DA weighted with total atom charge is split into three curves that isolate different sign pairs. 2DA descriptors out to a cutoff distance of 11 bonds are compared for an active compound from screen AID 435034. **b** Information loss is revealed when standard 3DA weighted with total atom charge is split into three curves that isolate different sign pairs. 3DA descriptors out to 12 Å at a resolution of 1.0 Å per bin are compared for the same compound. Sections are highlighted including (*a*) standard 3DA encodes almost no signal for distance bin [7:8), whereas sign pair splitting reveals significant presence of negative sign pairs and opposite sign pairs. *b1* and *b2* standard 3DA encodes equal intensities for bins [8:9) and [10:11), whereas sign pair splitting reveals contribution of negative sign pairs and positive sign pairs are significantly different for these two distance bins

TotalCharge. Two specific instances of information loss are highlighted in Fig. 1a. In the distance bin [7:8), standard 3DA weighted with TotalCharge contains almost no signal. However, when sign pairs are separated with 3DA_Sign, very strong signals emerge for negative–negative and opposite sign pairs. Because each bin of the histogram represents a sum of atom pairs with similar distances, the positive product of negative–negative and negative product of negative-positive cause their signals to cancel each other. Additionally, standard 3DA contains similar signals at distance bins [8:9) and [10:11). However, when unique sign pairs are split with 3DA_Sign, it becomes clear that these signals represent different distribution of negative–negative and positive–positive sign pairs within these distance bins.

Lastly, by default we use a 2DA that encodes distances up to 11 bonds and 3DA that encodes all atom pair distances up to 12 Å [11]. This distance is sufficient to capture the maximum distance within most small molecules. However, 3D descriptors such as 3DA are computed from a single predicted conformation of each molecule. As interatomic distance increases, the degree of flexibility and rotatable bonds may increase, leading to greater degrees of conformational uncertainty at larger distances. This uncertainty and potential error may interfere with QSAR model training. This issue is 3-dimensional and therefore we test a higher resolution 3DA_Sign variation that is limited to 6.0 Å instead of 12.0 Å. As a comparison, we test a similar variation of 2DA_Sign that is limited to a maximum distance of five bonds instead of 11. Here, no noise is added by conformational flexibility.

To test whether these variations are useful in training QSAR models, we used a generalizable framework for benchmarking the utility of 2DA_Sign and 3DA_Sign [11]. With any novel QSAR descriptor, performance evaluation is both important and challenging. In most cases, a predictive model can disregard information that does not increase performance. However, this is not guaranteed and extra descriptors adding too much noise can decrease performance. Additionally, properties that add noise for one dataset may be useful information for another. One approach is to provide the model with as many descriptors as available and perform iterative steps of descriptor selection where those that fail to significantly improve model performance are discarded. However, with an initial set of n descriptors, there are 2^n possible combinations. Coupled with the importance of cross-validation to avoid over-fitting, this process can quickly become time consuming or even intractable. Additionally, any descriptor selection must be repeated for every target of interest or high-throughput screening (HTS) dataset. Several algorithms have been presented to perform efficient descriptor selection [9]. However, as more descriptors and descriptor variations are developed, it is beneficial to use heuristics to eliminate descriptors unlikely to be beneficial. Therefore, we evaluated our descriptors with a rigorous benchmarking protocol that evaluates model performance across a variety of targets and datasets to identify those that consistently improve model performance.

Results

Developing a standard approach to descriptor benchmarking

The simplest evaluation of a descriptor's utility is through a one-to-one comparison of models trained with and without the descriptor of interest. To keep the total information

Table 1 Nine datasets were used to train models and evaluate model performance across different QSAR descriptor conditions

Pubchem project bioassay ID	Target	Active compounds	Inactive compounds
1798	M1 muscarinic receptor (agonist)	187	61,646
1843	Kir2.1 potassium channel	172	301,321
2258	KCNQ2 potassium channel	213	302,192
2689	Serine threonine kinase 33	172	319,620
435008	Orexin 1 receptor	233	217,925
435034	M1 muscarinic receptor (antagonist)	362	61,394
463087	Cav3 calcium channel	703	100,172
485290	Tyrosyl-DNA phosphodiesterase 1	281	341,084
488997	Choline transporter	252	302,084

PubChem bioassay ID for the overall project is indicated, as well as specific target, total number of confirmed actives, and total inactive compounds

provided to QSAR models in either condition constant, it is best to compare models trained with comparable descriptors or variations. Therefore, performance evaluations were isolated for each descriptor type. 2DA_Sign was compared against 2DA, 3DA_Sign was compared against 3DA, and 3DA/3DA_Sign at 6.0 Å was compared to 3DA/3DA_Sign at 12 Å. To enforce statistical comparability, all ANN parameters and objective functions are kept constant as well as any atom properties used for weighting. This does not always ensure that the total number of descriptors provided to models in both conditions is equal. For example, 3DA_Sign splits different sign pair variants by multiplying a single 3DA histogram into three. To avoid the possibility that 3DA_Sign outperforms 3DA simply because it supplies more descriptors, we decreased the resolution of 3DA_Sign three-fold to keep the total number of properties consistent across conditions. Any increase in model performance, therefore, will not be due to increased input vector length. This strategy is inappropriate for 2DA_Sign evaluation, however, because there is no resolution factor to compensate for the differences in vector size. Therefore, we are forced to evaluate 2DA_Sign with three times as many data-points as 2DA.

Model performance is judged by its ability to predict the activity of compounds it has never seen. Compounds not used for training are evaluated and ranked by their predicted activity. Plotting these predictions as true or false positives generates a receiver operating characteristic (ROC) curve. By computing the area under the curve of a logarithmic x-axis ROC curve, it is possible to score the ratio of true positive predictions to false positive predictions for the high confidence predictions.

When training and evaluating QSAR model performance, large datasets that cover large chemical spaces are preferred [12]. These datasets often come from high-throughput screening (HTS) projects where active compounds have been verified against a single target. Alternatively, smaller, focused datasets may be used to evaluate novel descriptors using leave-on-out (LOO) cross-validation [13]. However, this method of benchmarking can be misleading and tends to rely heavily on the presence of specific geometries rather than more subtle properties [14]. To apply the most generalizable benchmark possible, we used nine HTS datasets curated from PubChem [11]. These datasets target various proteins including G-protein coupled receptors (GPCRs), kinases, and ion channels. The number of compounds in these datasets range from approximately 61,000 to 344,000. These datasets are detailed in Table 1.

Because each 3D descriptor tested can be weighted with a variety of atom properties, we used nine different atom properties with and without accessible van der Waals (VDW) surface area scaling. Accessible VDW surface area accounts for varying accessibility of different atoms in a molecule arising from overlapping and covered VDW surfaces. Additionally, we provide all models with a standard set of 1D descriptors to achieve a performance baseline that strengthens comparisons. All 1D molecule descriptors and atom properties used for weighting are outlined in Table 2.

2DA_Sign and 3DA_Sign: separating atom properties by sign

The most common method for weighting 2DA and 3DA is with the product of atom properties for each atom pair. For signed properties such as partial charge, information can be lost as the product of two negative values cannot be distinguished from the product of two positive values. To avoid this information loss, we modified the 2DA and 3DA descriptors to allocate atom pairs into one of three histograms depending on the whether the atom properties are both negative, both positive, or opposite. These descriptors are called 2DA_Sign and 3DA_Sign and are designed specifically for signed properties such as partial charge since

Table 2 Properties used to train ANN models are categorized as scalar (one property per molecule) and atom (one property per atom)

Property	Type	Description	Signed
Molecular weight	Molecule	Total weight of molecule	
HBondDonor	Molecule	Total hydrogen bond donors in molecule	
HBondAcceptor	Molecule	Total hydrogen bond acceptors in molecule	
LogP	Molecule	Octanol/water coefficient; solubility	
TotalCharge	Molecule	Total charge of molecule	
NRotBond	Molecule	Number of rotatable bonds	
NAromaticRings	Molecule	Number of aromatic rings	
NRings	Molecule	Number of closed rings	
TopologicalPolarSurfaceArea	Molecule	Total surface area of molecule that is polar	
BondGirth	Molecule	Maximum number of bonds between two toms	
MaxRingSize	Molecule	Number of atoms in largest ring	
MinRingSize	Molecule	Number of atoms in smallest ring	
AromaticAtoms	Molecule	Number of atoms in aromatic rings	
IntersectionAtoms	Molecule	Number of atoms in ring intersections	
AromaticIntersectionAtoms	Molecule	Number of atoms in aromatic ring intersections	
MaxSigmaCharge	Molecule	Maximum sigma charge	
MinSigmaCharge	Molecule	Minimum sigma charge	
TotalSigmaCharge	Molecule	Sum of all sigma charges	
StDevSigmaCharge	Molecule	Standard deviation of all sigma charges	
MaxVcharge	Molecule	Maximum V-charge	
MinVcharge	Molecule	Minimum V-charge	
TotalVcharge	Molecule	Sum of absolute values of all V-charges	
StDevVcharge	Molecule	Standard deviation of all V-charges	
Girth	Molecule	Widest diameter of molecule	
Identity	Atom	Unweighted; 1 for all atoms	
SigmaCharge [15–17]	Atom	Partial charge localized to α-electron system	X
PiCharge [18–20]	Atom	Partial charge localized to π-electron system	X
TotalCharge	Atom	Total partial charge of atom	X
Vcharge [21]	Atom	Partial charge accounting for resonance	X
EffectivePolarizability [22–24]	Atom	Responsiveness of electron density to external field	
IsRingIntersection	Atom	1 if atom is at a non-aromatic ring intersection, 0 otherwise	
IsInAromaticRing	Atom	1 if atom is within aromatic ring, 0 otherwise	
InAromaticRingIntersection	Atom	1 if atom is at an aromatic ring intersection, 0 otherwise	

Molecule properties are used in every condition as a standard baseline of QSAR information and contain general information regarding overall molecular properties. Atom properties are used in every condition to weight the corresponding descriptor (2DA, 3DA, 2DA_Sign, or 3DA_Sign) with and without VDW surface area scaling. Atom properties that are split into unique sign pairs with the 3da_Sign descriptor are indicated as 'signed.' Algorithms used for the implementations of these atom properties are referenced

unsigned properties will solely fill the positive–positive vector. Therefore, when testing the utility of 2DA_Sign and 3DA_Sign, we only apply these new descriptors with signed properties. All unsigned properties are included with standard 2DA or 3DA depending on the condition.

Models trained with signed properties encoded with 2DA_Sign outperformed models trained with standard 2DA for all properties across all datasets tested. The average performance as measured by the area under the logarithmic ROC curve (logAUC) was 0.335. Compared with the average standard 2DA logAUC of, 0.295, using 2DA_Sign in place of 2DA for signed atom properties resulted in an increase in performance of approximately 13.8 %. Model performance across nine datasets is compared for 2DA and 2DA_Sign in Fig. 2.

As mentioned, 3DA_Sign was encoded with a larger distance step size as 3DA to ensure that the input vector lengths between the two conditions remained constant. Despite the lower resolution, 3DA_Sign improved model performance over standard 3DA in all datasets. Average model performance across nine datasets as measured by logAUC was 0.358 when applying signed properties with

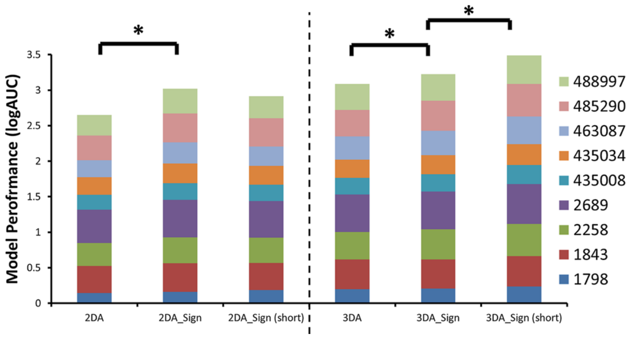

Fig. 2 Model performance is compared across nine datasets for descriptor modifications. Model performance is evaluated as logAUC (area under the logarithmic ROC curve between 0.001 and 0.1) for different QSAR descriptor methods. Different colored datasets are indicated by their Pubchem HTS project assay ID. 2DA_Sign significantly increases model performance (*2DA_sign vs 2DA, paired *t* test $p < 0.0001$, n = 9). Limiting 2DA to 5 bond lengths with 2DA_Sign (short) instead of 11 (2DA_Sign) does not increase performance. 3DA_Sign significantly increases model performance when compared to using standard 3DA with signed properties (*3DA_Sign vs 3DA paired *t* test $p < 0.05$, n = 9). Limiting maximum atom pair distance to 6.0 Å in 3DA_Sign (short) significantly increases model performance when compared to limiting maximum atom pair distance to 12 Å (*3DA_Sign vs 3DA_Sign (short) paired *t* test $p < 0.001$, n = 9)

3DA_Sign (vs 0.343 with 3DA), an average improvement of 4.4 % (paired *t* test $p < 0.05$). Model performance across nine datasets is compared for 3DA and 3DA_Sign in Fig. 2.

Finally, we tested limiting the maximum atom pair distance encoded for 3DA/3DA_Sign to 6.0 Å instead of 12.0. By focusing on the first 6.0 Å at higher resolution, model performance increased significantly from an average performance as measured by logAUC of 0.358–0.381 (6.4 % improvement, paired *t* test $p < 0.001$). Figure 2 compares model performance across nine datasets when encoding atom pair distances up to 12 versus 6.0 Å. When 2DA_Sign is limited to maximum distance of 5 bonds instead of 11, on the other hand, performance is not increased. Instead, there is a non-significant decrease in average performance to logAUC 0.324.

Discussion

This study outlines a general QSAR descriptor benchmarking technique that can be used to evaluate novel descriptors. Three potential QSAR descriptor modifications are evaluated using this generalizable benchmark strategy. Descriptors represent small molecules as vectors of numerical properties that can train ANNs to predict small molecule activity towards a specific target. These descriptors come in a continuously growing range of dimensions and information content. Coupled with the high degree of customization for many descriptors, training models using every available descriptor is not only computationally inefficient, but may introduce noise that hinders model performance. Therefore, an evaluation of a novel descriptor is critical before including it with QSAR model application. This evaluation must also be applied across multiple datasets with different targets. By nature, these biological targets may focus on different property demands, thereby making a broad statement of a descriptor's utility difficult.

The first descriptor tested is a variation of 2DA that is designed for weighting with signed atom properties. Multiplying two negative properties produces the same result as multiplying two equivalent positive properties, leading to misinformation for molecules with two or more atoms with negative properties. Additionally, histogram bins represent a sum of all atom pairs connected by a specific number of bonds. This can lead to additional information loss when opposite signed signals are added. To avoid these problems, we introduce 2DA_Sign to replace standard 2DAs when weighting with signed atom properties. 2DA_Sign generates three histograms of equal length, splitting atom pairs into negative–negative, positive–positive, and opposite signs. Using 2DA_Sign in place of 2DA for signed properties resulted in an average model performance increase of 13.8 % across nine HTS datasets.

Secondly, we tested a variation of 3DA called 3DA_Sign that treats unique sign pairs the same as 2DA_Sign. Because 3DA vector length is controlled by maximum cutoff distance and resolution, it was possible to adjust the resolution of 3DA_Sign to ensure constant input vector length. Encoding signed atom properties with 3DA_Sign in place of 3DA increased model performance across all nine datasets by approximately 4.4 %.

Because of the signed nature of the information provided by 2DA_Sign and 3DA_Sign over traditional autocorrelation, it is possible that targets placing higher demands for charged active compounds may benefit significantly more from these descriptor improvements than targets that require more neutral compounds. Therefore, it was important to examine the charge demands of nine datasets used for evaluation to ensure that they contained active compounds with diverse charge profiles. In Table 3, formal charge populations are listed for all active compounds across all datasets. This reveals a range of charge profiles across active datasets with 435034 containing the lowest percent of neutral compounds (39 %) and 435008 containing the highest (88 %). Additionally, the percent of active compounds with positive formal charges varies significantly from 2 % (485290) to 59 % (435034) and the percent of active compounds with negative formal charges varies from less than 1 % (1843, 463087) to 46 % (485290).

To evaluate whether a higher presence of formal charges in the active compounds allows for a greater performance increase when using 2DA_Sign or 3DA_Sign, the Pearson's correlation coefficient was calculated between the

Table 3 Percent of active compounds with formal charges varies across datasets

Pubchem project bioassay ID	Percent neutral actives	Percent positive actives	Percent negative actives	Percent zwitterion actives
1798	74	19	5	1
1843	55	42	1	2
2258	80	6	11	2
2689	65	13	20	2
435008	88	4	5	3
435034	39	59	1	0
463087	86	13	0	1
485290	47	2	46	5
488997	48	33	6	12

Active compounds across all datasets were analyzed for the presence of formal charges to ensure that target datasets with diverse formal charge preferences were tested due to the nature of the 2DA_Sign and 3DA_Sign descriptors. Active compounds are broken down into neutral, overall positive, overall negative, and zwitterion properties below

presence of formal charges within the active compounds and the performance increase for each dataset. When comparing the percentage of neutral actives within a dataset with the performance increase seen when using the improved autocorrelation descriptor over the traditional one, no significant correlation was found with 2DA_Sign (r = 0.10) or 3DA_Sign (r = 0.18). Additionally, no correlation was found between the performance increases with 2DA_Sign and the percentage of active compounds with overall positive charge (r = −0.24) or overall negative charge (r = 0.12). However, a moderate but significant negative correlation was found between the performance increase with 3DA_Sign and the percentage of actives with overall positive charge (r = −0.67, one-tailed $p < 0.05$). Additionally, a moderate but significant positive correlation was found between performance increase and negative charge (r = 0.67, one-tailed $p < 0.05$). This suggests a potential link between the formal charge demand of active compounds and the improved performance seen with using 3DA_Sign to separate the autocorrelation of signed atom properties. However, the opposite correlations with regards to positive and negative charge presence and the fact that this correlation is not reflected with 2DA_Sign makes it difficult to predict the specific relationship. Additionally, the comparative performance increases seen with both primarily neutral and largely charged active compounds suggests that the improvements seen with these descriptors are independent of specific charge demands of the targets. All plots examining potential relationships between active compound formal charges and performance increases can be found in the supplemental information (supplemental figure S1: percentage neutral actives; supplemental figure S2: percentage positively charged actives; supplemental figure S3: percentage negatively charged actives; supplemental figure S4: percentage zwitterion actives).

Lastly, we tested a maximum atom pair distance limitation of 6.0 Å instead of 12.0. Although 12.0 Å covers the maximum width of many small molecules, encoding longer atom pair distances can provide false information in cases of high molecular flexibility or rotation. A 6.0 Å limitation, on the other hand, focuses more on fragments within the molecule that are relatively invariant with respect to the arbitrary choice of conformer used to represent each molecule. Additionally, shorter distances can be sampled at a higher resolution without increasing input vector size. We found that limiting the maximum atom pair distance to 6.0 Å significantly increases performance across nine datasets by an average of 6.4 %. The fact that we see no performance increase when limiting 2DA_Sign to a distance cutoff of 5 bonds supports the conclusion that the increase in performance of a limited 3DA_Sign is linked to conformational flexibility. However, both results support the notion that 2DA and 3DA descriptors with maximum radius of 5 bonds and 6 Å, respectively, are sufficient to describe molecular structure for QSAR studies.

In conclusion, we present three recommendations for ANN-based QSAR descriptor selection: (1) Encoding signed properties in standard 2DA results in information loss that can significantly decrease model performance. Therefore, it is preferable to split unique sign pairs as with 2DA_Sign. (2) Similar information loss can be seen with standard 3DA and unique sign pairs should be split as with 3DA_Sign. (3) Limiting 3DAs to encode atom pairs up to 6.0 Å instead of 12.0 can significantly improve model performance.

Methods

HTS dataset curation

Nine datasets were used to evaluate descriptor performance. Specific details regarding all curation steps have been previously described [11]. However, relevant details for all datasets have been summarized:

Datasets were selected for high-throughput screening assays that focused on a single well-defined biological target protein. Active compound sets contain only those hits that were verified in confirmatory assays and did not show cross-activity with other targets tested against. Additionally, only those sets with at least 150 confirmed active compounds were used and the final collection was designed to encompass a variety of pharmaceutically relevant target protein classes. Specifically, for dataset 1798, positive allosteric modulators of M1 were identified in the primary calcium flux assay AID 626. Only actives compounds that were verified in confirmatory assay AID 1741 were kept for the dataset and those with cross activity with M4 (AID 1488) were removed. The final ratio of active to inactive compounds is 1:329. For 1843, inhibitors of the inward-rectifying potassium ion channel Kir2.1 were identified in the thallium flux assay AID 1672. Only active compounds verified in confirmatory assays AID 2032 and AID 463252 were kept and compounds showing non-specific activity in additional verification assays were removed. The final active to inactive ratio is 1:1751. For 2258, potentiators of the KCNQ2 potassium channel were identified in the primary thallium flux assay AID 2239. Only active compounds that were verified in confirmatory assay AID 2287 were kept and false positives or compounds showing non-specific effects in additional assays were removed. The final active to inactive ratio is 1:1418. For 2689, inhibitors of the serine/threonine kinase 33 were identified in the primary screen AID 2661. Only active compounds that were verified in confirmatory screen 2821 were kept and non-selective compounds identified in additional assays were removed. The final active to inactive ratio is 1:1858. For 435008, antagonists of the orexin 1 receptor were identified in three primary screens AID 485270, AID 463079, and AID 434989. Only active compounds verified in confirmatory screens AID 504701 and AID 504699 were kept for the active dataset. The final active to inactive ratio is 1:935. For 435034, negative allosteric modulators of the M1 receptor were identified in the primary calcium flux assay AID 628. Only active compounds verified in confirmatory assay AID 677 were kept and compounds that showed cross activity with the M4 receptor in assay AID 860 were removed. The final active to inactive ratio is 1:169. For 463087, inhibitors of the T-type calcium channel Cav3 were identified in the primary calcium flux assay AID 449839. Only actives that were verified in confirmatory screens were kept as the active compound dataset. The final active inactive ratio is 1:142. For 485290, inhibitors of the tyrosyl-DNA phosphodiesterase 1 were identified in the primary screen 485290. Only active compounds verified in the confirmatory assay AID 489007 were kept as the active compound dataset. The final active to inactive ratio is 1:1213. For 488997, inhibitors of the choline transporter were identified in the primary screen AID 488975. Only active compounds verified in all three confirmatory assays AID 493221, AID 504840, and AID 588401 were kept and compounds showing non-specific activity in additional assays were removed. The final active to inactive ratio is 1:1198. Inactive compounds for all datasets represent those identified in the primary screens.

Generation of numerical descriptors for QSAR model creation

Numerical descriptors and QSAR models were generated and evaluation over nine HTS datasets detailed in Tables 1, 2 and 3. The curation of these datasets has been previous outlined [11]. 3D conformations of all small molecules were generated using the CORINA [25] software package.

The BioChemical Library (BCL) software was used to generate all molecular descriptors tested in this study. All descriptors and atom properties used to weight 2DA and 3DA descriptors are described in Table 2. When weighting autocorrelation descriptors, all atom properties are represented with and without accessible surface area scaling.

Standard length 2DA and 2DA_Sign descriptors contained a cutoff distance of 11 bonds (12 values). Shortened 2DA_Sign descriptors contained a cutoff distance of five bonds (6 values). 3DA descriptors tested at a 12.0 Å cutoff were calculated for a step size of 0.167 Å (72 total values). 3DA descriptors tested at a 12.0 Å cutoff were calculated for a step size of 0.084 Å (72 total values). 3DA_Sign descriptors tested at 12.0 Å cutoff were calculated for a step size of 0.5 Å (3 × 24 = 72 total values). 3DA_Sign descriptors tested at 6.0 Å cutoff were calculated for a step size of 0.25 Å (3 × 24 = 72 total values).

Artificial neural network model architecture and training

All ANN models were trained using back propagation and a sigmoid transfer function with a simple weight update of $\eta = 0.05$ and $\alpha = 0.5$, a hidden layer of 32 neurons, 0.1 visible neuron dropout, and 0.5 hidden neuron dropout. Each dataset was divided into two sets of compounds: compounds used to train the model (training) and compounds kept hidden from the model during training to evaluate predictability after training has completed (independent). Five-fold cross-validation was used where 20 individual ANN models were trained for each HTS dataset by rotating which compounds appeared in the training and independent sets. Final active or inactive prediction for each independent compound was taken as a consensus across models for which that compound appeared in the independent set. The objective function used during training was the area under the logarithmic receiver operating characteristic (ROC) curve [26, 27] (logAUC [28]) between false positive rates of 0.001 and 0.1.

ANN model performance evaluation

All models were evaluated with the same objective function used for training. ROC curves with a logarithmic x-axis were generated for consensus predictions sorted by predicted activity and the area under the curve as calculated for the range of 0.001–0.1 (the top 10 % of predicted compound activities). For all statistical comparisons, two-tailed paired *t* tests were performed between descriptor conditions across the nine HTS datasets.

Figures and artwork

All graphs were generated with Microsoft Excel 2007. Molecule structures were generated with Molecular Operating Environment (MOE, Chemical Computing Group).

Acknowledgments Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842, R01 DK097376, R01 HL122010, R01 GM073151, U19 AI117905) and NSF (CHE 1305874).

References

1. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr (2014) Computational methods in drug discovery. Pharmacol Rev 66(1):334–395. doi:10.1124/pr.112.007336
2. Salt DW, Yildiz N, Livingstone DJ, Tinsley CJ (1992) The use of artificial neural networks in QSAR. Pestic Sci 36(2):161–170. doi:10.1002/ps.2780360212
3. Butkiewicz M, Lowe EW, Meiler J (2012) Bcl::ChemInfo—qualitative analysis of machine learning models for activation of HSD involved in Alzheimer's Disease. In: Computational intelligence in bioinformatics and computational biology (CIBCB), 2012 IEEE symposium on, 9–12 May 2012, pp 329–334. doi:10.1109/cibcb.2012.6217248
4. Trinajstić N (1992) Chemical graph theory. In: Mathematical chemistry series, 2nd edn. CRC Press, Boca Raton
5. Balaban AT (1998) Topological and stereochemical molecular descriptors for databases useful in QSAR, similarity/dissimilarity and drug design. SAR QSAR Environ Res 8(1–2):1–21. doi:10.1080/10629369808033259
6. Hemmer MC, Steinhauer V, Gasteiger J (1999) Deriving the 3D structure of organic molecules from their infrared spectra. Vib Spectrosc 19(1):151–164. doi:10.1016/S0924-2031(99)00014-4
7. Broto P, Moreau G, Vandycke C (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies. Perception of molecules: topological structure and 3-dimensional structure. Eur J Med Chem 19(1):61–65
8. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc 119(43):10509–10524. doi:10.1021/ja9718937
9. Shahlaei M (2013) Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chem Rev 113(10):8093–8103. doi:10.1021/cr3004339
10. Moreau G, Broto P (1980) The auto-correlation of a topological-structure—a new molecular descriptor. Nouv J Chim 4(6):359–360
11. Butkiewicz M, Lowe EW Jr, Mueller R, Mendenhall JL, Teixeira PL, Weaver CD, Meiler J (2013) Benchmarking ligand-based virtual high-throughput screening with the PubChem database. Molecules 18(1):735–756. doi:10.3390/molecules18010735
12. Kubinyi H, Folkers G, Martin YC (1998) 3D QSAR in drug design. Qdsar, vol 2. Kluwer, Dordrecht
13. Kiralj R, Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J Braz Chem Soc 20:770–787
14. Manchester J, Czermiński R (2009) CAUTION: popular "Benchmark" data sets do not distinguish the merits of 3D QSAR methods. J Chem Inf Model 49(6):1449–1454. doi:10.1021/ci9000508
15. Gasteiger J, Marsili M (1978) A new model for calculating atomic charges in molecules. Tetrahedron Lett 19(34):3181–3184. doi:10.1016/S0040-4039(01)94977-9
16. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36(22):3219–3228. doi:10.1016/0040-4020(80)80168-2
17. Guillen MD, Gasteiger J (1983) Extension of the method of iterative partial equalization of orbital electronegativity to small ring systems. Tetrahedron 39(8):1331–1335. doi:10.1016/S0040-4020(01)91901-5
18. Bauerschmidt S, Gasteiger J (1997) Overcoming the limitations of a connection table description: a universal representation of chemical species. J Chem Inf Comput Sci 37(4):705–714
19. Streitwieser A (1961) Molecular orbital theory for organic chemists. Wiley, New York
20. Gasteiger J, Saller H (1985) Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. Angew Chem Int Ed Engl 24(8):687–689. doi:10.1002/anie.198506871
21. Gilson MK, Gilson HS, Potter MJ (2003) Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. J Chem Inf Comput Sci 43(6):1982–1997
22. Gasteiger J, Hutchings MG (1983) New empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. Tetrahedron Lett 24(25):2537–2540
23. Gasteiger J, Hutchings MG (1984) Quantitative models of gas-phase proton-transfer reactions involving alcohols, ethers, and their thio analogs. Correlation analyses based on residual electronegativity and effective polarizability. J Am Chem Soc 106(22):6489–6495. doi:10.1021/ja00334a006
24. Miller KJ (1990) Additivity methods in molecular polarizability. J Am Chem Soc 112(23):8533–8542. doi:10.1021/ja00179a044
25. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem Rev 93(7):2567–2581. doi:10.1021/cr00023a012
26. Cleves AE, Jain AN (2006) Robust ligand-based modeling of the biological targets of known drugs. J Med Chem 49(10):2921–2938. doi:10.1021/Jm051139t
27. Hristozov DP, Oprea TI, Gasteiger J (2007) Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. J Comput Aided Mol Des 21(10–11):617–640. doi:10.1007/s10822-007-9145-8
28. Clark RD, Webster-Clark DJ (2008) Managing bias in ROC curves. J Comput Aided Mol Des 22(3–4):141–146. doi:10.1007/s10822-008-9181-z