# Fuzzy Tricentric Pharmacophore Fingerprints. 2. Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure−Activity Relationships

Fanny Bonachéra and Dragos Horvath*

UMR 8576 CNRS − Unité de Glycobiologie Structurale & Fonctionnelle, Université des Sciences et Technologies de Lille, Bât. C9-59655 Villeneuve d'Ascq CEDEX, France

Received August 30, 2007

Topological fuzzy pharmacophore triplets (2D-FPT), using the number of interposed bonds to measure separation between the atoms representing pharmacophore types, were employed to establish and validate quantitative structure−activity relationships (QSAR). Thirteen data sets for which state-of-the-art QSAR models were reported in literature were revisited in order to benchmark 2D-FPT biological activity-explaining propensities. Linear and nonlinear QSAR models were constructed for each compound series (following the original author's splitting into training/validation subsets) with three different 2D-FPT versions, using the genetic algorithm-driven Stochastic QSAR sampler (SQS) to pick relevant triplets and fit their coefficients. 2D-FPT QSARs are computationally cheap, interpretable, and perform well in benchmarking. In a majority of cases (10/13), default 2D-FPT models validated better than or as well as the best among those reported, including 3D overlay-dependent approaches. Most of the analogues series, either unaffected by protonation equilibria or unambiguously adopting expected protonation states, were equally well described by rule- or p$K_a$-based pharmacophore flagging. Thermolysin inhibitors represent a notable exception: p$K_a$-based flagging boosts model quality, although—surprisingly—not due to proteolytic equilibrium effects. The optimal degree of 2D-FPT fuzziness is compound set dependent. This work further confirmed the higher robustness of nonlinear over linear SQS models. In spite of the wealth of studied sets, benchmarking is nevertheless flawed by low intraset diversity: a whole series of thereby caused artifacts were evidenced, implicitly raising questions about the way QSAR studies are conducted nowadays. An in-depth investigation of thrombin inhibition models revealed that some of the selected triplets make sense (one of these stands for a topological pharmacophore covering the $P_1$ and $P_2$ binding pockets). Nevertheless, equations were either unable to predict the activity of the structurally different ligands or tended to indiscriminately predict any compound outside the training family to be active. 2D-FPT QSARs do however not depend on any common scaffold required for molecule superimposition and may in principle be trained on hand of diverse sets, which is a must in order to obtain widely applicable models. Adding (assumed) inactives of various families for training enabled discovery of models that specifically recognize the structurally different actives.

## 1. INTRODUCTION

The recent development of topological fuzzy pharmacophore fingerprints[1] 2D-FPT, shown to display excellent neighborhood behavior,[2] naturally raised the question of their potential applications in quantitative structure−activity relationships[3−5] (QSARs), empirical mathematical models returning an estimate of the molecular activity as a function of structural descriptors. Relationships between activity and pharmacophore feature distribution descriptors have been intensely studied in chemoinformatics, either in terms of (a) QSAR model buildup or (b) binding pharmacophore[6−8] elucidation attempts. There is however no fundamental distinction between (a) and (b)—selecting and weighing specific elements of the vector describing the overall pharmacophore pattern of a molecule, as in (a), may in principle allow the backtracking of the important, activity-enhancing variables to the actual pharmacophore features in the molecules and thus translate a QSAR model into a pharmacophore hypothesis in the sense of (b). With molecular[9] or pharmacophore field[10] maps of the space surrounding the studied ligands, the space zones corresponding to the relevant field terms may be readily assimilated to the hypothesized binding site regions involved in interactions. This very tempting and straightforward interpretation of CoMFA[9] models has largely contributed to the success of the approach, albeit authors sometimes tend to forget that a statistically valid correlation is not enough evidence for a cause-to-effect relationship between correlating magnitudes.[11] Molecular field maps do however require the construction and alignment of one or several conformer(s) for each compound. Computer-effective overlay-independent descriptors of the pharmacophore patterns typically rely on autocorrellograms[12,13] and pair density distributions.[14] These 'encrypt' the pharmacophore/field pattern information, providing a less straightforward link between descriptors and structural elements in the molecules, but are nevertheless successful in QSAR.[15,16] Pharmacophore triplets[17] or quadruplets[18] provide an even more detailed description, but large size and strong conformer-dependence of 3D triplets dissuaded scientists to use these otherwise than in similarity searches, until recently.[19]

* Corresponding author phone: +333.20.43.49.97; fax: +333.20.43.65.55; e-mail: dragos.horvath@univ-lille1.fr, d.horvath@wanadoo.fr.

10.1021/ci7003237 CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/07/2008

Coding for population levels of specified (setup-dependent) pharmacophore triplets, with topological distances used as a metric of the relative positions of the atoms representing pharmacophore features[20] (hydrophobicity, aromaticity, hydrogen bond donors/acceptors, cations, anions), 2D-FPT contain in principle all the chemical information required to elucidate a binding pharmacophore. They should be able to model compound recognition by an active site, when fed into a QSAR building tool (descriptor selection and weighing procedure) set to detect (i) pharmacophore triplets selectively populated in actives, which are supposed to enhance binding, and (ii) triplets selectively populated in inactives, which are supposed to block it. However, the topological nature of 2D-FTP, only implicitly and imprecisely accounting for the actual 3D interfeature distances perceived by the receptor, is a further obstacle in the way of straightforward mechanistic interpretation. The assumption that triplets (i) actually stand for ligand atoms favorably interacting with the site, while (ii) include atoms clashing with the site, may be far-fetched.

This notwithstanding, 2D-FPT models may still convey more physicochemically meaningful information than equations based on abstract topological indices. Success of 2D-FPT descriptors in QSAR would be good news, because they are computationally cheap (no 3D structure generation and alignment required). This benchmarking study revisits 13 ligand/inhibitor sets concerning various receptors, from various literature sources[13,21−24] where various QSARs, including high-end, 3D overlay-dependent CoMFA models, were proposed. Linear models were generated for all these sets, using the following:

1. Three differently parametrized versions of 2D-FPT, differing in terms of the ‘grid mesh’ size (the step used to enumerate edge lengths of the considered basis triplets), the minimal and maximal considered edge lengths, etc. These include the ‘default’ (D) and the ‘optimal’ (O) setups reported in the original paper, plus a ‘coarse’ version (C).

2. Two variants based on the default 2D-FPT scheme but using a rule-based pharmacophore feature assignment procedure instead of the one based on calculated p$K_a$ values for ionizable groups (D-R, for ‘R’ule-based) and, respectively, abandoning the fuzzy mapping of molecular triplets onto basis triplets in favor of strict matching (D-S, for ‘S’trict matching).

Using the stochastic QSAR sampler (SQS),[25] a set of relevant QSAR equations (having cross-validation scores within the upper end of the spectrum generated at model training) was kept and confronted with validation compounds, for each compound set and 2D-FPT version. SQS-generated relevant model sets typically feature thousands of independent equations selected due to their high cross-validation scores, but literature studies only present a few individual models. Therefore, the benchmarking only reports the statistical criteria of the best validating model from each representative set. As the initial splitting schemes into training/validation sets have been scrupulously followed here, the root-mean-squared prediction errors with respect to validation set compounds (not excluding any outliers) are directly comparable, irrespective of the nature of the reference models (regression, PLS, neural network) and their original calibration procedures (stepwise, stochastic, PLS, etc.).

This study continues with benchmarking the relative performance of different 2D-FPT versions against each other, in order to shed some light on the question of how to optimally generate QSAR-proficient 2D-FPT. Both the impact of explicit modeling of proteolytic equilibria vs rule-based pharmacophore feature assignment and the influence of fuzzy triplet mapping were assessed. Duplicate SQS runs were performed with both default 2D-FPT, the rule-based version D-R, and the nonfuzzy version D-S, in order to generate extended representative model sets, allowing a comparison of average validation propensities according to a previously described approach.[25] Comparison of QSARs based on triplet versions D, O, and C relies on the validation statistics of best validating models, like in literature model benchmarking.

Next, a comparison of best validating linear vs nonlinear SQS models is undertaken in order to further confirm the previously observed trend[25] of improving validation propensities when allowing for preset nonlinear transformations to enter the models.

Eventually, thrombin inhibition models are challenged to predict the affinity of chemically different, cocrystallized thrombin ligands, including two amidine[26,27] derivatives and a pyrazinone[28] adopting a different binding mode. The groups seen to directly interact with the thrombin site will be matched against the atom triplets corresponding to selected 2D-FPT elements.

This paper is structured as follows: in Methods, a brief revisiting of 2D-FPT and of the SQS model building methodology will continue with an outline of the statistical criteria used for benchmarking. An introduction of the employed data sets, followed by the details on the assessment of structural interpretability of 2D-FPT models, complete this section. Results and Discussions will first address the various benchmarking aspects: comparison with literature models, comparative assessment of pharmacophore flagging strategies, of the use of fuzzy logic for triplet generation and of the nonlinearity policy for SQS model buildup. The next addressed key point will be the extrapolability of the trained QSAR models to compounds of different topology. The section will close with the discussion of selected triplets as ‘topological pharmacophores’ matched against the actual pharmacophore points in the structures of cocrystallized ligands. The Conclusions paragraph, concerning the usefulness and interpretability of 2D-FPT triplets as QSAR descriptors, will be extended to a general discussion about the limitations of QSAR buildup and benchmarking caused by restricted training/validation set diversity, in light of the artifacts and chemically meaningless terms seen to enter some of the nonetheless well validating QSARs.

## 2. METHODS

**2.1. 2D-FPT Buildup.** 2D-FPT buildup has been described in detail elsewhere.[1] A basis set of reference pharmacophore triplets is chosen, enumerating all possible combinations of pharmacophore features (Hp-hydrophobic, Ar-aromatic, HA-hydrogen bond acceptor, HD-donor, PC-positive charge, NC-negative charge) of the corners, times all the considered integer edge lengths obeying triangle inequalities, within a finite range [$E_{\min}$, $E_{\max}$] and sampled by an $E_{\text{step}}$ controlling the graininess of 2D-FPT. Next, all triplets of features

**Table 1.** Parameters Controlling 2D-FPT Buildup − Three Considered Setups: D − Default Setup and O − Optimal Setup (Maximizing NB) from Previous Work[1] and C − Coarse Setup[29]

| parameter | description | D | O | C |
|---|---|---|---|---|
| $E_{\mathrm{min}}$ | minimal edge length of basis triangles (number of bonds between two pharmacophore types) | 2 | 4 | 5 |
| $E_{\mathrm{max}}$ | maximal triangle edge length of basis triangles | 12 | 15 | 15 |
| $E_{\mathrm{step}}$ | edge length increment for enumeration of basis triangles | 2 | 2 | 3 |
| $E$ | edge length excess parameter: in a molecule, triplets with edge length > $E_{\mathrm{max}}$+e are ignored | 0 | 2 | 2 |
| $\Delta$ | maximal edge length discrepancy tolerated when attempting to overlay a molecular triplet atop of a basis triangle | 2 | 2 | 3 |
| $\rho_{\mathrm{Hp}} = \rho_{\mathrm{Ar}}$ | Gaussian fuzziness parameter for apolar (hydrophobic and aromatic) types | 0.6 | 0.9 | 0.7 |
| $\rho_{\mathrm{PC}} = \rho_{\mathrm{NC}}$ | Gaussian fuzziness parameter for charged (positive and negative charge) types | 0.6 | 0.8 | 0.3 |
| $\rho_{\mathrm{HA}} = \rho_{\mathrm{HD}}$ | Gaussian fuzziness parameter for polar (hydrogen bond donor and acceptor) types | 0.6 | 0.7 | 0.2 |
| $L$ | aromatic-hydrophobic interchangeability level | 0.6 | 0.5 | 0.7 |
| | number of basis triplets at given setup | 4494 | 7155 | 2625 |

represented in a molecule are analyzed, following a protonation state-dependent pharmacophore typing of the atoms, using shortest-path topological interatomic distances as actual edge lengths. Molecular triplets are then mapped onto basis triplets, using fuzzy logic (each molecular triplet may contribute to the population levels of several similar basis triplets, by increments directly related to their degree of similarity). Total population levels of basis triplets form a sparse vector, the 2D-FPT descriptor, with nonzero elements corresponding to the basis triangles that are either present per se or are represented by similar triplets in the molecule. Table 1 reports the specific setups used to generate the 2D-FPT versions used in this paper, where ‘D’ and ‘O’ correspond to the two setups already discussed in the original publication (therein called FPT-1 and FPT-2, respectively; labels ‘D’ and ‘O’ recall that the former is a default setup, while the latter was shown to optimize NB). An additional ‘C’oarse fingerprint ‘C’ has been considered here, using a larger $E_{\mathrm{step}}$ of 3 and thus relying upon a significantly smaller basis triplet set, while still preserving excellent NB.[29] Two additional variants of the default 2D-FPT were also considered: D-R using a ‘R’ule-based pharmacophore feature assignment strategy rather than the one based on predicted p$K_a$ values for ionizable groups, and D-S, the ‘S’trict fingerprint mapping molecular triplets strictly onto the identical basis triplets or ignoring them as no such triplet is listed within the reference basis set. Both had their NB tested in the original 2D-FPT publication.

**2.2. The Stochastic QSAR Sampler (SQS).** SQS is based on a hybrid parallelized genetic algorithm-driven engine for selecting both relevant descriptors and their optimal nonlinear transformation rules to enter a model. It uses randomized leave-1/3-out cross-validation to let the in silico Darwinian selection process pick the most robust models. SQS has been shown able to typically retrieve thousands of not overfitted QSAR equations, which successfully passed the subsequent external validation tests.[25] It may, upon request, exclusively mine for linear equations or try to select the most appropriate among a set of predefined nonlinear transformation functions to be used in conjunction with any given descriptor. Linear models have been generated for benchmarking purposes against literature equations, most of which were linear as well. Nonlinear QSARs were also systematically built for all the data sets, in quest of equations potentially outperforming the linear models. SQS proceeds by successively running a Model Builder (MB) with varying operational control parameters which are tuned on-the-fly to maximize MB sampling performance. After each MB run, a set of most relevant models found up-to-date are extracted, using error pattern similarity to decide which models are redundant. At the end, these sets of locally most representative equations are merged, and all models having their leave-1/3-out cross-validated correlation coefficient within a window of 0.2 units at the top end enter the final ‘representative’ model pool of that simulation.

For each of the 13 compound sets, SQS mining for linear models was systematically performed with each of the 5 considered 2D-FPT versions. Linear simulations with D, D-R, and D-S descriptors have been duplicated, and representative model pools were merged in order to allow the estimation of the average validation correlation coefficient shifts attributable to switching from one descriptor variant to the other, according to a previously outlined formalism[25] (also see below). Eventually, a second round of SQS simulations mining for fully nonlinear models was also carried out for all compound sets using all descriptor versions, leading to a total of 5 × 13 (first round, linear) + 3 × 13 (linear model generation duplicates: D, D-R, and D-S descriptors only) + 5 × 13 (nonlinear) = 169 different SQS runs using various Linux and IRIX workstations of the laboratory. This effort lead to a total of 236 852 individual QSAR equations, members of the respective representative sets, all compound series confounded.

**2.3. Statistical Criteria Used for Benchmarking.** This work only reports statistical criteria with respect to the external validation sets used in literature and taken over as such in the present work. Training set and cross-validation criteria are either uninteresting (some general information concerning training set $R^2_{\mathrm{T}}$ values of the selected best validating models will be given in the Results section) or not directly comparable to literature values (cross-validation schemes differ from author to author). The key benchmarking criterion used here is the root-mean-squared prediction error RMSPE of a model $\mu$ with respect to the $N_{\mathrm{VS}}$ molecules $m$ of the external validation sets (VS), where their predicted activities $Y^{\mu}(m)$ are directly compared to experimental values $A(m)$:

$$\mathrm{RMSPE}(\mu) = \sqrt{\frac{\sum_{m \in VS} [Y^{\mu}(m) - A(m)]^2}{N_{\mathrm{VS}}}} \qquad (1)$$

This prediction error might, if desired, be compared to the

**Table 2.** List of the 13 Considered Data Sets[a]

| ID | symbol | description and references | D | D-S | D-R | O | C | training set size | validation set size | no. of inactives |
|---|---|---|---|---|---|---|---|---|---|---|
| ACE | + | angiotensin converting enzyme inhibitors[21] | 3062 | 1498 | 3421 | 3948 | 1683 | 106 | 38 | - |
| AChE | × | acetylcholinesterase inhibitors[21] | 1535 | 808 | 1619 | 1884 | 761 | 74 | 37 | - |
| AT1 | * | angiotensin type-1 receptor activators[22] | 1971 | 1131 | 1900 | 2490 | 948 | 122 | 122 | - |
| AT2 | □ | angiotensin type-2 receptor activators[22] | 1971 | 1131 | 1900 | 2490 | 948 | 122 | 122 | - |
| Art | ■ | artemisinin analogues[23] | 1492 | 803 | 1685 | 1697 | 692 | 142 | 37 | - |
| BZR | ○ | benzodiazepine receptor inhibitors[21] | 1491 | 674 | 1955 | 1195 | 482 | 98 | 49 | 16 |
| Cox2 | ● | cyclooxygenase-2 inhibitors[21] | 1479 | 653 | 1645 | 1518 | 627 | 188 | 94 | 40 |
| DhfR | △ | dihydrofolate reductase inhibitors[21] | 2380 | 1447 | 1807 | 2705 | 880 | 237 | 124 | 36 |
| GPB | ▲ | glycogen Phosphorylase B inhibitors[21] | 1403 | 664 | 1462 | 1144 | 427 | 44 | 22 | - |
| FXa | ▽ | factor Xa inhibitors[13] | 3642 | 2487 | 3639 | 5822 | 2333 | 290 | 145 | - |
| Ster | ▼ | original CoMFA steroids data set[24] | 907 | 382 | 907 | 849 | 362 | 21 | 10 | - |
| Ther | ◇ | thermolysin inhibitors[21] | 2942 | 1693 | 3016 | 3718 | 1497 | 51 | 25 | - |
| Thr | ♦ | thrombin inhibitors[21] | 2790 | 1498 | 2950 | 3475 | 1508 | 59 | 29 | - |

[a] Featuring their ID used in this work, associated symbols used in plots such as Figure 2, a brief description, and referencing plus the total number of pharmacophore triplets populated in at least one of the molecules, depending on the fingerprint version as defined in Table 1.

variance of the experimental property witnessed within the validation set, to calculate the validation set correlation coefficient $R^2_V$:

$$R^2_{\mathrm{V}}(\mu) = \max\left(0{,}1 - \frac{\sum_{m\in \mathrm{VS}} [Y^{\mu}(m) - A(m)]^2}{\sum_{m\in \mathrm{VS}} [\langle A\rangle_{\mathrm{VS}} - A(m)]^2}\right) \qquad (2)$$

As the denominator in eq 2 simply serves to provide an order of magnitude to serve as a reference for the sum of squared residuals, its actual choice may vary from author to author: some use the average over learning set molecules $<A>_{TS}$ rather than the one over validation set compounds (which should make no difference if the VS contains a representative sample of the entire data set—but this is not always the case with the herein adopted compound series and splitting schemes). Also, certain authors report $R^2_V$ values after linearly refitting predicted to experimental values. This amounts to accepting a model if its predictions obey an arbitrary linear relationship to the experiment ($A \approx \alpha Y + \beta$), rather than expecting predictions to equal actual values. Therefore, validation correlation coefficients as reported here may, unlike the average prediction error, *not be comparable to literature values.*

$R^2_V$ is truncated at 0, signaling that any model with prediction errors larger than the ones of a 'null' model will simply count as failing to validate. This only affects benchmarks monitoring average validation propensities over the representative sets of SQS models $\mu$, $<R^2_V(\mu)>_\mu$. Sets of SQS models including few equations that fail to validate with strongly negative $R^2_V$ untruncated values should not be overtly penalized with respect to sets of equations with many but unspectacular validation failures.[25] The benchmarking studies, monitoring the impact of p$K_a$-dependent pharmacophore flagging (or fuzzy mapping) on the average validation propensities of models $<R^2_V(\mu)>_\mu$ rely on a 'minimal guaranteed shift' criterion $\delta_R$ (or $\delta_S$, respectively), which expresses the drift of averages obtained with different descriptor versions ($<R^2_V(\mu)>_{(\mathrm{D-based}\ \mu)}$ vs $<R^2_V(\mu)>_{(\mathrm{D-R-based}\ \mu)}$ and $<R^2_V(\mu)>_{(\mathrm{D-S-based}\ \mu)}$, respectively), corrected by the amount of drift that may affect these averages due to imperfect SQS sampling[25]—see eq 8 in that publication. The larger $\delta_R$ (or $\delta_S$, respectively), the more significant the advantage of using p$K_a$-dependent pharmacophore flagging (or fuzzy triplet mapping, respectively).

Some literature studies also provided classification scores with respect to external subsets—either percentages of inactives correctly classified[21] as such or the percentage of correctly classified molecules[13] (actives as actives and inactives as inactives, respectively). In either case, these criteria were recalculated following the original author's procedures—please refer to cited papers for details.

**2.4. Experimental Data.** Table 2 shows the considered data sets, with their IDs in the present work, the references to the publications[13,21,22,23,24] reporting the previous QSAR studies, the numbers of populated triplets for each 2D-FPT version, and the set sizes. Please refer to the original publications and the Supporting Information for compound set sizes and training/validation set definitions. Except for the artemisinin[23] set, where the explained variable is a global score of antimalarial activity, all the considered studies refer to in vitro binding tests, the explained variables being in all cases dose-dependent indicators of inhibitory potency ($IC_{50}$, $K_i$) on a logarithmic scale. No metabolism-related or pharmacokinetic properties were included, as pharmacophore descriptors are primarily aimed at explaining the affinity of reversible noncovalent site/ligand interactions, while fragment descriptors are better suited to capture reactivity-related properties. Also, although 2D-FPT contain all the chemical information needed to estimate physicochemical properties such as the lipophilic character and derived indices (LogP/LogD, solubility, permeability, etc.) they may be too fine-grained in this respect. Global descriptors such as the total polar surface area may be more useful to predict LogP, rather than allowing for all the possible triplets including polar features to enter a very long and therefore statistically less robust QSAR equation. However, 2D-FPT may prove very helpful to pinpoint specific effects (such as the impact of intramolecular hydrogen bond formation on LogP) in completion to overall polarity indices—the study of possible synergies of 2D-FPT with other categories of descriptors is beyond the purpose of this work.

**2.5. Extrapolability and Structural Interpretation of 2D-FPT-Based Models.** All the representative thrombin (Thr) inhibition models using default 2D-FPT descriptors were challenged to predict the affinity of two chemically

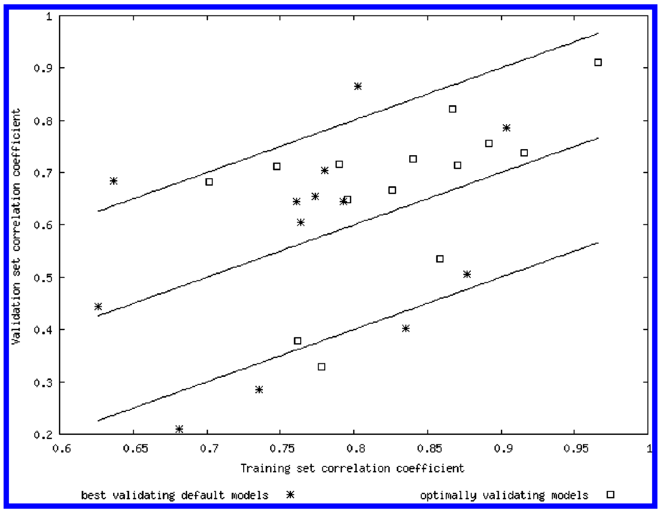


**Figure 1.** Comparative plot of training set (TS; $R^2{}_T$ on X) vs validation set (VS; $R^2{}_V$ on Y) correlation coefficients, for the globally optimal (row 1, Table 3) and default linear QSAR models (row 3, Table 3) of highest VS correlation coefficient, in context of grid lines $R^2{}_V = R^2{}_T$, $R^2{}_V = R^2{}_T - 0.2$, and $R^2{}_V = R^2{}_T - 0.4$, respectively.

different cocrystallized amidine/guanidine derivatives (PDB[30] codes 1BHX and 1D4P), plus a recently published compound,[28] of a radically different chemical class (2BXT). Since none of the Thr-trained equations passed the test—not even with respect to the amidine/guanidine derivatives—a novel series of models was refitted, after enrichment of the Thr training/validation series with presumed Thr inactives taken from 11 other compound sets (excluding FXa). The key pharmacophore triplets of these equations were traced back to their source atoms in the ligands, in order to check whether these include actual ligand-site anchoring points.

## 3. RESULTS AND DISCUSSIONS

**3.1. 2D-FPT-Based Models Compare Favorably with Respect to Published QSARs of Higher Cost/Complexity.** Prior to focusing on predictive success of validation sets—a necessary but not sufficient condition for any QSAR equation meant to serve for actual virtual screening of compound databases—Figure 1 provides, for each compound set, a concise outlook of the relationship between training and validation correlation coefficients, for the linear default and respectively optimal (top $R^2{}_V$) models. It is no surprise that no $R^2{}_V$ -$R^2{}_T$ correlation can be seen across multiple compound sets: while $R^2{}_T$ values may to some extent relate to cross-validation scores (results not shown), correlations between training and validation scores are, even within a family of models based on a same compound set, rather rare (the "Kubinyi paradox"[31]). The reason for showing Figure 1 is to confirm that none of these models, selected due to their high $R^2{}_V$ values, fail to apply to training set compounds. In principle, such models—artifacts 'explaining' the validation set by pure chance but unable to properly accommodate all the training examples—could be visited by the SQS procedure during its random walk in QSAR problem space. However, these are unlikely to enter the set of representative models regrouping only the most successful (training set) cross-validators—indeed, no situation with $R^2{}_V \gg R^2{}_T$ could be evidenced. Reversely, few models have $R^2{}_V \ll R^2{}_T$: the question whether these are 'overfitting' artifacts or whether some validation set compounds fall outside the applicability range granted by the training set will not directly addressed here. However, equivalently large $R^2{}_V$-$R^2{}_T$ gaps are reported in the literature for the concerned compound sets: Cox2, GPB, and Ster (with the thrombin inhibitor set, Thr, only the default linear model displays large training-validation discrepancies).

Scrambling tests are routinely performed by the SQS approach: after termination due to failure to retrieve new fit models, the 10 best performing sets of operational parameters are used to pilot 10 independent attempts to build models on hand of Y-scrambled training set data (each of the 10 attempts relied on a different Y randomization). These attempts go beyond refitting of previously found equations and imply descriptor (re)selection and full-blown cross-validation. Typically, scrambling results met expectations: cross-validated $Q^2$ values were low (below 0.4) in most of the cases. For some compound set/descriptor version combinations, model fitting against unscrambled data failed to reach $Q^2$ values above 0.4—in these situations, there is significant overlap of the $Q^2$ ranges of scrambled and actual models. This is uninteresting—those cases would have been judged to represent QSAR buildup failures anyway, on the sheer basis of their low $Q^2$. Interestingly, some quite high scrambled $Q^2$ of up to 0.6 were obtained for the thrombin and steroids series, irrespectively of the employed descriptor versions. This is a critical warning signal about the extremely low intrinsic diversity of the sets: scrambling lead to swapping of activity values between molecules that are similar enough to 'stand' for others. Nevertheless, the top $Q^2$ scores with proper data exceeded 0.8 in all of these cases—therefore, the representative models discussed here, within a window of 0.2 $Q^2$ units, are all outside the $Q^2$ range covered by scrambling experiments. Under these circumstances, benchmarking may safely be based solely on validation criteria.

Table 3 shows that, with the notable exception of artemisinin analogues, 2D-FPT-based models were found (row 1) to equal or even significantly outperform the best validating published approaches (row 4; relative RMSPE shift in row 9—positive values standing in favor of 2D-FPT). [Albeit the correct classification rate of the FXa linear regression model is slightly lower than reported in literature, the former displays an excellent linear correlation score—which is a more constraining indicator of model quality than a classification rate. The reported GRIND-based discriminant model and the 2D-FPT linear equation are, as far as they can be compared, equipotent predictors.] This proves that 2D-FPT appropriately capture the structural information relative to reversible noncovalent binding to receptor sites and that the SQS methodology successfully mines for properly validating models. It is however inappropriate to claim that 2D-FPT are intrinsically more informative than CoMFA fields, although, for example, FPT-based results do outperform CoMFA even with the rigid and easy-to-align steroid (Ster), when the classical CoMFA drawbacks (uncertainties concerning the relevant conformations and alignment modes, etc.) are of little concern. Observed advantages may alternatively be explained by enhanced model sampling due to the parallelized, computer-intensive SQS approach, which might perhaps have found even better validating approaches if allowed to mine CoMFA field descriptors. This

**Table 3.** Benchmarking with Respect to Literature Results[a]

| | ACE | AChE | AT1 | AT2 | Art | BZR | Cox2 | DhfR | GPB | FXa | Ster | Ther | Thr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.14 | 0.69 | 0.33 | 0.39 | 0.78 | 0.70 | 1.08 | 0.77 | 0.69 | 0.80 | 0.42 | 1.33 | 0.56 |
| | *0.713* | *0.714* | *0.727* | *0.910* | *0.756* | *0.378* | *0.329* | *0.683* | *0.667* | *0.821* | *0.536* | *0.649* | *0.737* |
| | - | - | - | - | - | **81%** | **75%** | **69%** | - | **85%(a)** | - | - | - |
| 2 | D-R | D | C | O | D | D-R | O | D-S | O | D-R | D | C | O |
| | (L) | (N) | (N) | (N) | (N) | (L) | (N) | (N) | (L) | (L) | (N) | (L) | (N) |
| 3 | 1.33 | 0.76 | 0.35 | 0.48 | 0.88 | 0.75 | 1.18 | 0.81 | 0.90 | 0.88 | 0.43 | 1.34 | 0.85 |
| | *0.605* | *0.655* | *0.705* | *0.865* | *0.685* | *0.286* | *0.209* | *0.644* | *0.444* | *0.785* | *0.506* | *0.645* | *0.402* |
| | - | - | - | - | - | **75%** | **68%** | **92%** | - | **84%(a)** | - | - | - |
| 4 | 1.48 | 0.95 | 0.42 | 0.51 | 0.70 | 0.87 | 1.17 | 0.84 | 0.79 | - | 0.69(c) | 1.59 | 0.69 |
| | - | - | - | - | - | **88%** | **70%** | **92%** | - | **88%(a)** | - | - | - |
| 5 | CoMSIA basic | CoMFA | CoMFA | CoMFA | NN(b) | 2.5D | CoMSIA extra | HQSAR | CoMSIA extra | GRIND-PLS | CoMFA | CoMFA | CoMSIA basic |
| 6 | 1.50 | 1.20 | - | - | 0.78(b) | 0.87 | 1.25 | 0.99 | 1.20 | - | - | 2.24 | 0.96 |
| | | | | | | **88%** | **70%** | **81%** | | | | | |
| 7 | **10.1** | **16.8** | **16.7** | **5.9** | **−25.7** | **13.8** | **−0.9** | **3.6** | **−13.9** | - | **33.3** | **15.7** | **−23.2** |
| 8 | **11.3** | **36.7** | - | - | **−12.8** | **13.8** | **5.6** | **18.2** | **25.0** | - | - | **40.2** | **11.4** |
| 9 | **23.0** | **27.4** | **21.4** | **23.5** | **−11.4** | **17.5** | **7.7** | **8.3** | **12.7** | - | **39.1** | **16.4** | **18.8** |

[a] **1** − Validation criteria for the globally optimal, best validating 2D-FPT models: RMSPE (plain text), validation set $R^2_V$ (italics), and percentage of correctly classified inactives (bold) in an additional inactive validation set, except for (a), reporting the overall correct classification rate of both validation set actives and inactives. **2** − 2D-FPT setup and nonlinearity policy (linear, nonlinear) leading to results (1). **3** − Validation criteria, as in 1, of best validating linear model based on default 2D-FPT. **4** − validation criteria (RMSPE, correct classification rates) of most successful models reported in the literature (references in Table 1); RMSPE value (c) not reported as such, was calculated on hand of data reported in Table 2 of that publication.[24] **5** − Methodology leading to models (4). **6** − Validation criteria (as in 4) for literature models of comparable complexity to 2D-FPT equations. Except for artemisinin (b), reporting a linear model based on 2D and 3D descriptors, this row presents 2.5D descriptor-based models.[21] **7, 8, and 9** − relative prediction error decrease of 2D-FPT vs reported models: default 2D-FPT vs best reported, e.g., row 7 = ($RMSPE_4$-$RMSPE_3$)/$RMSPE_4$ (%), default 2D-FPT vs comparable reported (row 8: 3 vs 6) and best 2D-FPT vs best reported (row 9: 1 vs 4), respectively.

notwithstanding, 2D-FPT are clearly able to generate state-of-the-art QSAR models in conjunction with powerful model building procedures. Furthermore, previous results[25] actually showed that, in many cases, valid 2D-FPT models may well be obtained by less aggressive techniques, such as stepwise regression. Computationally effective topological and alignment-independent 2D-FPT have thus significant technical advantages over CoMFA.

Nonlinear approaches occupy the top position of the best validating model in eight out of 13 cases, an observation reinforcing the already reported[25] trend of improving model validation propensity upon allowing SQS to employ predefined nonlinear transformations.

Linear models using the default fingerprint version (row 3) never happened to represent the globally best validating approach. Their performances still equal or exceed the best literature values (row 4, relative shifts in row 7) in ten out of the 13 studied cases. In two situations, default linear models fail to meet literature standards, although better-than-literature 2D-FPT models could be found using other setups and/or nonlinearity policies. In the case of thrombin inhibitors Thr, the top linear model based on the O version performs only slightly better than the default (RMSPE of 0.82 instead of 0.85)—the dramatic drop to the global optimum at 0.56 is a specific consequence of nonlinearity. For glycogen phosphatase B inhibitors GPB, the globally optimal model is linear as well—see the next paragraph for a discussion of D- and O-version GPB equations.

Benchmarking of (row 3) default 2D-FPT equations against literature models of comparable complexity—i.e. linear, overlay-independent models not requesting any buildup of molecular geometries—found these latter (row 6, relative shifts in 8) to be outperformed in eight out of nine cases. Except for the artemisinin analogue series, where row 6 refers to a linear model based on 2D descriptors, other row 6 equations are based on '2.5D' indices. According to the original paper,[21] these models, using a mixture of standard 2D and 3D descriptors, outperform equations solely based on 2D indices. However, since the involved 3D descriptors are whole-molecule indices (such as molecular volume and surface values, replaceable by quick estimators based only on molecular connectivity), the 2.5D models were considered to be acceptable matches to 2D-FPT equations in terms of complexity.

The lesser performance of 2D-FPT with respect to the artemisinin series is actually not surprising. In this case, the monitored activity is an overall, systemic antimalarial potency score, normalized by molecular weight. These compounds are peroxides and act as heme alkylating agents,[32] not in reversibly binding to receptors. Fragment and/or topological descriptors are expected to (and actually do, according to the literature model[23]) better explain such a type of activity. 2D-FPT nevertheless come up with some reasonable models, which does not imply that some kind of pharmacophore recognition is required for the alkylating activity. More likely, the presence of certain pharmacophore triplets may correlate with specific substructures and therefore implicitly relate to reactivity (also see discussions below).

**3.2. 2D-FPT Setup-Dependence of the Validation Performance.** Within this and the following chapters, the validation set correlation score, systematically calculated according to eq 2, is used for comparison—either in terms of top $R^2_V$ of the best validating models of the respective representative sets or in terms of average validation propensity $\langle R^2_V \rangle$ over the representative sets of equations. In principle, due to the stochastic nature of the model builder, it is risky to extrapolate the intrinsic quality of descriptors from validation score differences of single models. However, the analysis of the 39 duplicated SQS simulations for all 13 data sets, performed with D, D-R, and D-S descriptors, respectively, showed that duplicate simulations generate significantly diverging representative sets[25] and different best validating models, which nevertheless have remarkably close $R^2_V$ values. In 19 cases out of the 39 repeats (49%), the top $R^2_V$ value was reproduced within 0.025, and in 25 cases

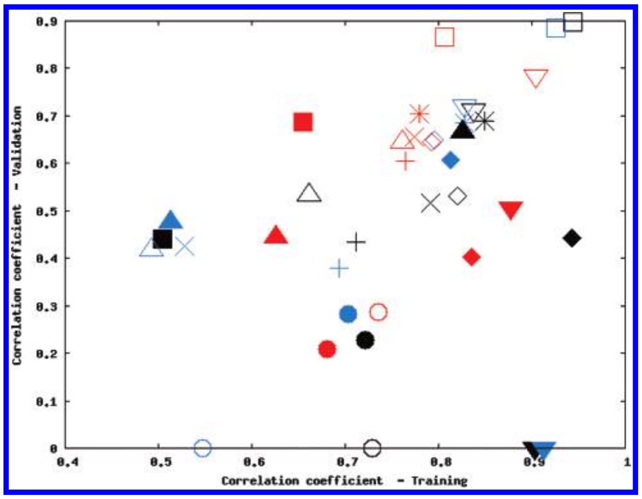


**Figure 2.** Training set (TS; $R^2_T$ on X) vs validation set (VS; $R^2_V$ on Y) correlation coefficients for the best validating linear models obtained for each compound set (see symbol coding in Table 2) and each descriptor setup version (red-D, black-O, blue-C).

(90%) the $R^2_V$ shift did not exceed 0.05. The most irreproducible top $R^2_V$ value, seen to shift by 0.2, concerned the steroid data set in conjunction with nonfuzzy descriptors (D-S). Since repeated simulations virtually never led to top $R^2_V$ value differences above 0.1 units, this may be, in our opinion, taken as the significance threshold (shifts above 0.1, if observed, may be attributed to the differences in chemical information conveyed by each descriptor version).

The monitoring of the descriptor versions found at the basis of globally optimal equations (Table 3, row 1) shows a slight preference for the O setup (winner in 4 cases), followed by D (3), D-R (3), C (2), and finally D-S (1). Figure 2, representing the training and validation set $R^2$ values for every top validating linear model built at a given {compound set, descriptor version} combination, shows that the influence of 2D-FPT setup on resulting QSAR model performance is unfortunately not easy to foresee (compound sets are dot-shape coded, Table 2, while descriptor versions are color coded: D-red, O-black, C-blue).

*3.2.1. Coarse FPT Are the Less Successful in QSAR.* It appears that the coarse, less information-rich, descriptor version C is on the whole less well suited for QSAR modeling: in one case (Art ■) it actually failed to generate any useful models at all, while in two more cases (Bzr ○ and Ster ▼) all of the linear C models failed to validate. In all other cases, the blue mark corresponding to C-based models rarely tops the two others on the Y axis ($R^2_V$)—however, a remarkable exception was observed in the case of thrombin inhibitors (Thr ♦), where the C-based linear model only comes in second to the nonlinear O-based approach (not shown on the plot). No straightforward explanation of this unexpectedly good performance of the C version could be found. None of the C triplets entered either D or O models, and, furthermore, the C version model does not even include any specific triplet featuring the positive charge required for thrombin activity. This is, per se, not surprising, since all the Thr compounds, actives and inactives alike, include at least one protonable group. This training set does not emphasize the fact that the cationic center is important. It is thus likely that the C model exploits some local, family dependent chance correlation between C descriptors and activities. [In QSAR literature, a 'chance' correlation is said to apply within the training set but break down either at the cross-validation stage or, at latest, with respect to validation compounds. If, however, training and validation sets are subsets of the same structural family, 'chance' correlations that persist throughout training, cross-validation, *and* external validation may well exist—and explain the low success rate of QSAR models in actual virtual screening of diverse databases.] D and O models, however, feature at least one triplet involving the positive charge, i.e., implicitly suggesting that a cation flanked by specific groups at specific distances may play a role in binding. Such a conclusion may yet be an overinterpretation: it should not be forgotten that triplets are the only input options in this approach. Therefore, if the actually important element were a pharmacophore *pair*, not a triplet, selecting a triplet involving that pair plus the ubiquitous cation is merely the workaround found by the approach to compensate for a missing explicit pair of descriptors (also see the paragraph dedicated to structural interpretation of Thr models, further below).

*3.2.2. O-Version Failures and Successes: Why FPT May, within a Structurally Homogeneous Family, Implicitly Behave like Fragment Descriptors.* Like the C fingerprints, the O-based 2D-FPT also failed to generate any properly validating models for the Bzr ○ and Ster ▼ sets.

*Steroid Models.* For the steroid set, the D-version linear model utilizes three triplets, one being favorable for activity: (1) HA4-HA10-Hp12 − two acceptors at 12 bonds apart, e.g., located at both ends of the steroid scaffold and a hydrophobe at 4 bonds from one acceptor and at 10 from the other, e.g., part of the scaffold (in triplet nomenclature,[1] each corner is followed by the length of the opposed edge, in number of bonds). Two other triplets were seen to be most often populated in inactives and decrease predicted activity when present in validation set compounds: (2) Ar4-HA4-HA4 − an equilateral triangle of edge lengths 4 consisting of an aromatic and two acceptors. (3) Ar10-Ar10-HD4 − two aromatic atoms, 4 bonds apart, at the opposite of the steroid scaffold (at 10 bonds) from a hydrogen bond donor.

All these triplets are also part of the O-version basis set, but O-version population levels slightly differ due to different fuzziness and aromatic/hydrophobic equivalence parameters. The levels of Ar10-Ar10-HD4 are particularly low and only come from imperfect mapping of molecular triplets where the aromatic feature is down-weighted because it is actually represented by a hydrophobe. As 2D-FPT are integer value vectors, and given the overall poor match of actual molecular triplets, the actual population level of Ar10-Ar10-HD4 rounds up to either 0 or 1 (out of the 50 arbitrary units standing for a perfect match). Or, with the D setup, fuzziness and aromatic-hydrophobic interchangeability are defined such that the Ar10-Ar10-HD4 population level happens to be correlated with the presence of a hydroxyl group at position 3 of the A ring. All the 3-OH steroids have Ar10-Ar10-HD4 set to 1, and all but one of compounds with Ar10-Ar10-HD4 equaling 1 are 3-OH steroids. 3-OH steroids are inactive—their average activity (alcohols or phenols confounded) is 1.8 log units below the average over the rest of the molecules. When using the O or C setups, however, the privileged relationship between the 3-OH fragment and the particular 2D-FPT element breaks down, with immediate negative impact on the model statistics. The D model,

**Table 4.** Benchmarking of the Impact of Descriptor Fuzziness and p$K_a$-Dependence on the Quality of Resulting QSAR Models[a]

| | best $R^2_V$ | | | $<R^2_V>$ and (variance) | | | guaranteed shift $\delta$ due to | |
|---|---|---|---|---|---|---|---|---|
| set | D | D-S | D-R | D | D-S | D-R | fuzziness | p$K_a$- dependence |
| ACE | 0.605 | 0.526 | **0.713** | 0.260 (0.110) | 0.200 (0.119) | 0.324 (0.131) | +0.049 | −0.043 |
| AChE | 0.655 | 0.627 | 0.658 | 0.051 (0.104) | 0.116 (0.160) | 0.210 (0.178) | 0.000 | **−0.144** |
| AT1 | 0.705 | 0.671 | 0.718 | 0.554 (0.100) | 0.478 (0.115) | 0.489 (0.140) | +0.026 | +0.024 |
| AT2 | 0.867 | 0.873 | 0.884 | 0.744 (0.061) | 0.760 (0.066) | 0.754 (0.071) | −0.006 | −0.002 |
| Art | 0.688 | 0.736 | 0.742 | 0.466 (0.107) | 0.495 (0.156) | 0.549 (0.083) | 0.000 | −0.060 |
| BZR | 0.286 | 0.214 | **0.378** | 0.009 (0.034) | 0.004 (0.021) | 0.023 (0.051) | +0.003 | 0.000 |
| Cox2 | 0.209 | 0.247 | 0.171 | 0.012 (0.029) | 0.029 (0.044) | 0.006 (0.018) | −0.007 | 0.000 |
| DhfR | 0.644 | 0.670 | 0.590 | 0.172 (0.142) | 0.364 (0.166) | 0.173 (0.151) | **−0.139** | 0.000 |
| GPB | 0.444 | 0.498 | 0.426 | 0.033 (0.068) | 0.109 (0.098) | 0.018 (0.065) | −0.029 | 0.000 |
| FXa | 0.785 | 0.819 | 0.841 | 0.639 (0.071) | 0.706 (0.062) | 0.689 (0.070) | −0.055 | 0.037 |
| Ster | 0.506 | 0.345 | 0.457 | 0.003 (0.037) | 0.001 (0.016) | 0.001 (0.022) | 0.000 | 0.000 |
| Ther | **0.645** | 0.623 | 0.439 | 0.321 (0.154) | 0.212 (0.143) | 0.007 (0.039) | **+0.098** | **+0.307** |
| Thr | 0.402 | 0.437 | 0.375 | 0.093 (0.108) | 0.107 (0.142) | 0.050 (0.070) | 0.000 | 0.000 |

[a] In terms of both optimal and average validation propensities ($R^2_V$) of the linear models from the representative SQS sets.

although by any standards much better than reported CoMFA-based QSARs, is yet another example[11] of how QSAR may provide correct predictions based on wrong premises. Since the population level of Ar10-Ar10-HD4 is determined by the −OH group at carbon 3, there is no reason to imply that the aromatic corners of the triplet must be mechanistically involved in modulation of the affinity.

*Benzodiazepine Receptor Inhibitor Models.* Benzodiazepine receptor inhibitor models—including those from literature—all have low validation propensities. In this context, the deceiving behavior of O- and C-based models is not surprising. Since the D-based best validating linear model includes 14 different triplets, tracing the differences between D and O models back to the subjacent descriptors is a difficult task. The fact that several of the triplets entering the D model have edge lengths of two may be the first hint toward a possible explanation: with minimal edge lengths $E_{\min}$ of 4 and 5, respectively, neither the O nor C versions may account for such short-range pharmacophore elements.

*Glycogen Phosphorylase B Models.* GPB offers a counterexample where the O version is the most successful. In this case, all the triplets entering the O model also happen to be members of the D basis set. However, the O fingerprints are less fuzzy than their D counterparts, especially with respect to hydrophobic groups. Attempts to build relevant D models with the triplets entering the O-based top validating equation failed. Furthermore, the best validating linear D-S model, using unfuzzy triplets (see Table 4), performed slightly better than D but worse than the O-based equations. Apparently, the quality of the GPB QSAR models displays a peak at some optimal triplet fuzziness level.

*3.2.3. Influence of p$K_a$-Dependence on QSAR Quality.* Table 4 shows both the optimal $R^2_V$ values and the average $<R^2_V>$ scores over the representative sets of linear models, from duplicate SQS runs using specified 2D-FPT versions. The reported guaranteed shifts are related to the respective (D-S vs D and D-R vs D) average score differences, conservatively corrected by the amount of average shift that might be attributable to $<R^2_V>$ score fluctuations.[25] Positive shift scores suggest the superiority of the default version with respect to rule-based and nonfuzzy approaches, respectively. Both top $R^2_V$ values stand out by more than 0.1 units and guaranteed shifts exceeding 0.1 were highlighted.

Rule-based pharmacophore flagging lead to significantly better top models but not to significantly better average scores for the ACE and Bzr series. It also triggered a significant increase in the average validation propensity of AChE models, without however impacting on the quality of top equations. The only case where switching from D to D-R descriptors is seen to provoke a coherent and very large change of both optimal and average validation scores is the Ther compound set, with a net preference for p$K_a$-dependent D fingerprints. The following paragraphs suggest possible explanations for these observations:

*Angiotensin Converting Enzyme Models.* Within the ACE set, the top validating D-R equation includes five descriptors, compared to four entering the less well performing D top model. The essential difference is the participation of the Ar2-NC2-PC2 triplet in the former but not in the latter. This triplet is populated in α-amino acid moieties (with an actual hydrophobe replacing the aromatic): the D-R model learned that compounds including such moieties are, on the average, more active than others. This hypothesis finds itself confirmed by validation set compounds, of which all three (Figure 3a) that have populated Ar2-NC2-PC2 D-R triplets are nanomolar actives. The D-R Ar2-NC2-PC2 term contributes, for all three, an increment of +2.5 log units which is paramount for correct activity prediction. However, the D-R strategy ignores, by contrast to the p$K_a$-based approach, the existence of populated Ar2-NC2-PC2 in two additional, completely inactive, validation set compounds (Figure 3b). It rightly denies the cation status to the tertiary N atom, erroneously perceived as a quaternary pyridinium by the D flagging scheme, but wrongly ignores protonation of the tertiary amine in the second molecule. This latter is a technical problem that could be fixed by rewriting the default pharmacophore flagging rules precompiled by ChemAxon, which were used as such in this work (the aspect was already mentioned in the previous 2D-FPT paper[1]). Acknowledgment that the Ar2-NC2-PC2 triplet may stem from fragments other than α-amino acid moieties breaks down the correlation between Ar2-NC2-PC2 population levels and activity. Both D and D-R schemes each err once in the flagging of validation set compounds, but the D-R error leads to a happy coincidence, establishing a biased correlation between the pharmacophore triplet and a specific fragment. The number of models exploiting this artifact is however small compared to the set of relevant SQS equations—therefore, average validation propensities were not affected by switching from D to D-R.

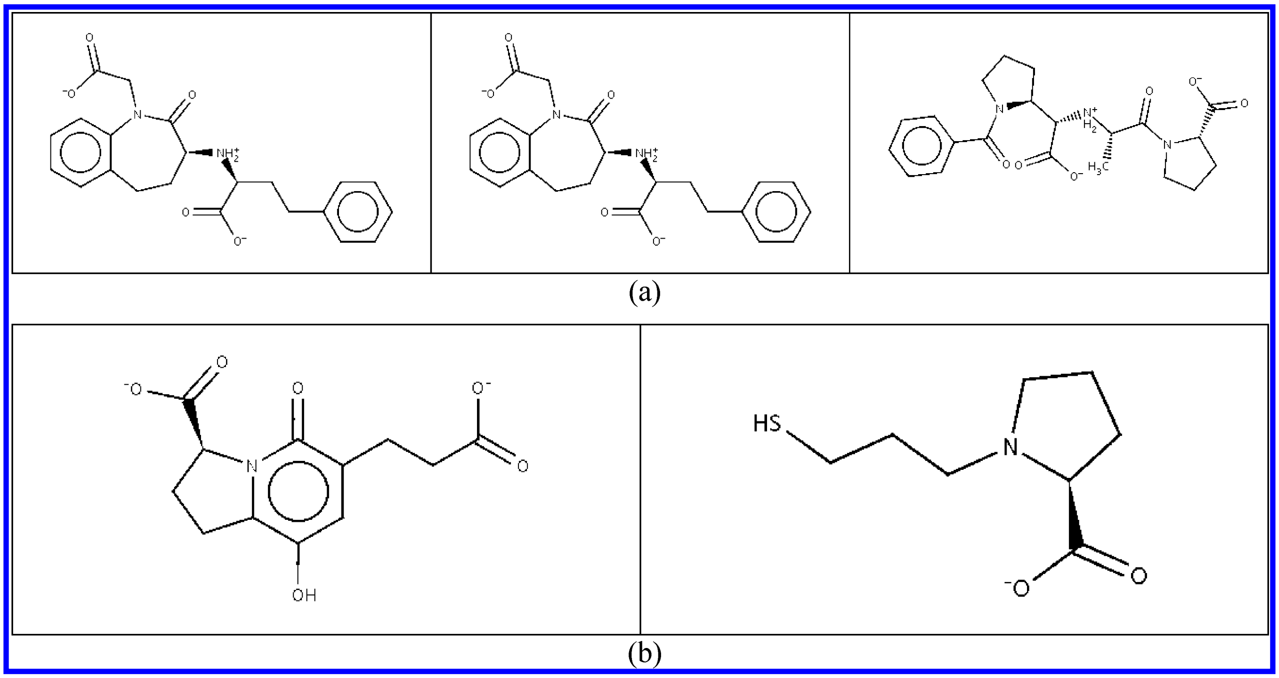


**Figure 3.** ACE compounds featuring the specific Ar2-NC2-PC2 triplet entering the top D-R model: (a) the three validation set compounds populating this triplet according to D-R are nanomolar binders and (b) the D (p$K_a$-sensitive) approach also finds the triplet in these two molecules, where it erroneously assumes a positive charge of the 'quaternary pyridinium' N and it correctly considers tertiary amines to be protonated at pH = 7.4. Both molecules (b) are inactive.

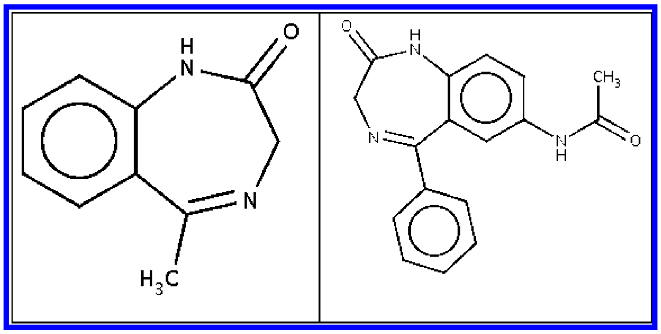


**Figure 4.** Typical Bzr set representatives featuring the imine moiety erroneously taken for an immonium cation by the D-R rule-based flagging strategy.

*Benzodiazepine Receptor Inhibitor Models.* In the case of Bzr inhibitors, the main difference between rule- and p$K_a$-based pharmacophore flagging concerns the imine nitrogen within the 7-membered ring, present in a majority of the compounds. The rule-based approach considers this N to be protonated because it possesses a free electron pair. This is actually not the case at pH = 7.4 (aliphatic imine p$K_a$ values are about 4.0, with ChemAxon predicting values of 3.6 and 3.3 for the phenyl- and diphenylimine moieties in the Bzr compounds from Figure 4). In D fingerprints, this N is ignored, not being basic enough (p$K_a$ cutoff of 5) to be flagged HA. This notwithstanding, imine fragments are nevertheless seen more often in actives than in inactives. Though it is impossible to state whether this relative enrichment is a set-specific accident or whether this fragment is mechanistically needed (electron density effects on conjugated phenyls, conformational constraints guaranteeing proper binding geometries) it makes sense to rely on the imine fragment count to explain activity trends within the set. As a consequence of the flagging error, in the D-R version imines are being assigned a special status (cations are rare features, so that they often stand out as the only 'cation' of the molecule). Therefore, the presence of the imine moiety is straightforwardly expressed by the population levels of specific PC-containing triplets. The better performance of the D-R approach in this case was again a lucky accident.

*Acetylcholine Esterase Models.* Unlike in the two above-mentioned situations witnessing accidental specific improvements of the top-validating D-R models, D-R based AChE models show an improvement of the average validation propensities, a trend not followed by top-validating models. In order to understand this phenomenon, the average prediction errors committed by each of the 1790 D and 2294 representative D-R models, respectively, were monitored for each of the 37 validation set compounds, in search for molecules that were systematically less well predicted by D approaches (Figure 5). It is important to note that in the AChe series the protonation states were explicitly provided in the input files—tertiary amines were protonated, forcing the D-R flagging scheme to recognize them as cations (if plain tertiary amines were input, D-R would have assigned[33] the hydrogen bond donor flag to the tertiary N, while the D flagging scheme is insensitive with respect to the actual protonation status of input compounds). The observed differences between D and D-R models is though not due to the treatment of tertiary amines by the latter. In-depth analysis pinpointed to another—related—flagging artifact of the D-R strategy: in fact, the above-mentioned peculiar flagging of trisubstituted N atoms equally (and erroneously) applies to N-disubsituted amides. Or, all four compounds in Figure 5 happen to belong to this category. The peculiar data set artifact allowing the chemically meaningless flagging of tertiary amides as hydrogen bond donors to translate into more accurate predictions remains obscure.

*Thermolysin Models.* The thermolysin compound set (Ther), the only one to show a consistent amelioration of both top and average validation propensities when using the

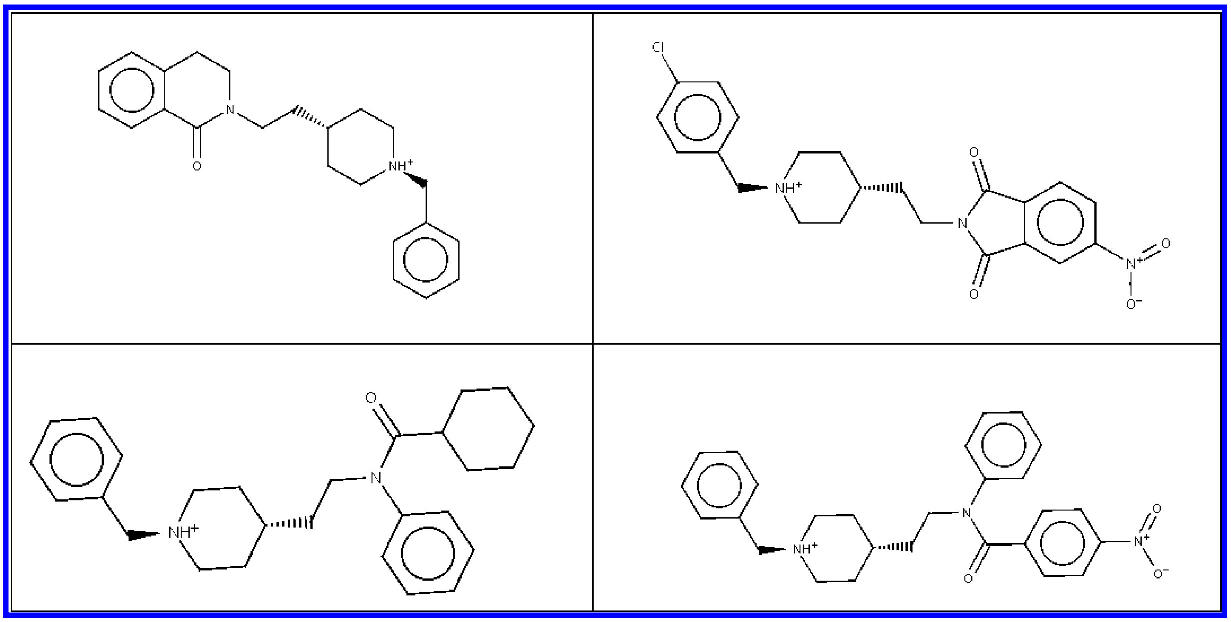


**Figure 5.** AChE validation set inhibitors for which the relevant D-R models provided, on the average, prediction errors smaller by one unit or more compared to the ones committed by D models.

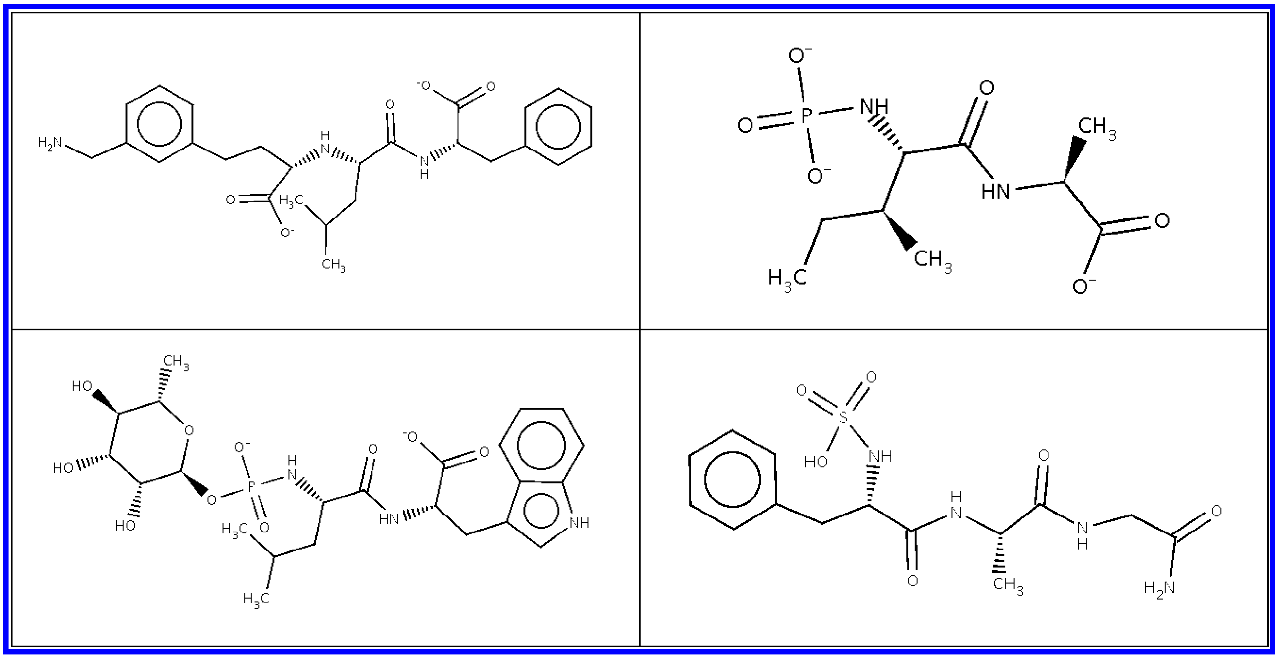


**Figure 6.** Thermolysin (Ther) validation set inhibitors having their activities properly predicted by the top D approach but highly overestimated if the same top D model is used with D-R descriptors.

p$K_a$-dependent flagging scheme, is also a series of outstanding structural diversity. Multiple, both acid and basic ionizable groups—some as atypical as thiophosphates, not perceived as anions by the D-R approach—are often seen in these compounds. Under these circumstances, the clear positive impact of a p$K_a$-dependent approach should not come as a surprise. The top validating D linear model is actually a quite simple equation, involving only four triplets, out of which three have positive coefficients (specifically populated in actives: Ar2-Hp4-Hp4, HA6-HD8-Hp8, and HD4-Hp6-NC4) and one with a negative coefficient (preferentially seen in inactives: Ar10-Ar10-HD6). In order to pinpoint key fingerprint differences upon switching to the D-R version, the validation set activities were also calculated according to the top D model but using D-R population triplet levels. For 10 out of 25 validation compounds, D and D-R population levels were identical, and so were predictions. In 8 cases, however, this led to a significant overestimation of activities (by 1 log unit or more). Figure 6 illustrates four of the concerned examples. For example, in the first represented compound, D-R population levels of both Ar2-Hp4-Hp4 and HD4-Hp6-NC4 were much higher than the corresponding D versions and thus triggered an activity overestimation of ~5 log units. The explanation, however, is not in any way related to different protonation patterns (both D and D-R consider the two carboxylates as deprotonated and the primary and secondary amines as protonated) but to a peculiar difference in flagging strategy. While the D-R approach considers the anionic flag on the negatively charged oxygen, the D strategy assigns it to the carboxylate C, instead of the default hydrophobic flag (oxygens are flagged HDA). The D-R population level of HD4-Hp6-NC4 triplets increases because the contributing atom triplets both have the HD-NC and Hp-NC edges longer by one (the carboxylate $C-O^-$) bond and are therefore a better match for the basis triplet. The Ar2-Hp4-Hp4 levels increase because the D-R version sets additional hydrophobes—the carboxylate C atoms. This is an example of 2D-FPT degeneracy impacting on QSAR propensity: if an atom triplet is responsible for activity, without being unique in the molecule (the others playing no role), then the key triplet will represent only a fraction of the total population level of the matching fingerprint element. This population level may, per se, not discriminate between deletion of the key triplet and deletion of an irrelevant contributor—other triplets, specifically designing the key atoms and their environment, must be taken into account. Setting hydrophobic flags on carboxylates caused ‘drowning’ of the relevant contributions to the Ar2-Hp4-Hp4 population level in noise from the additional, meaningless triplets. Of course, the fitting of specific D-R models, compensating for these flagging differences, lead to equations in which the molecules in Figure 6 are being better predicted than on hand of the D model using D-R fingerprints. This compensation is however incomplete—prediction by the D model with appropriate D fingerprints is still better. Also, the top validating D-R model requires a total of six different triplets, compared to only four for D. Although the D strategy is in this case the clearly better one, its advantages do not stem from capturing any subtle protonation effects but are more likely from the (chemically meaningful) deletion of the hydrophobic character of carboxylate C atoms.

All in all, this work did not produce any clear evidence that p$K_a$-dependent flagging may enhance descriptor performances in QSAR: if such evidence exists, it was unfortunately hidden by noise due to the peculiarities of the compound sets and by the other, unavoidable, flagging scheme differences. All the situations in which the rule-based approach appeared to perform better have been traced down to ‘lucky’ coincidences, where chemically meaningless flags happened to single out specific compound subfamilies, enriched in actives. On the opposite, in the single case where the D version brought clear improvements of both top and average validation propensities, benefits were due to p$K_a$-independent, albeit chemically meaningful, flagging differences. These findings apparently contradict the reported[1] importance of p$K_a$-sensitive flagging in evidencing NB violations (‘activity cliffs’—structurally almost identical compound pairs with nonetheless differing activities). The problem is that differences between D and D-R flagging strategies are not strictly limited to proteolytic equilibria-related effects. In similarity scoring, however, systematic p$K_a$-unrelated flagging differences tend to cancel out: if A and A′ are, for example, two homologous carboxylic acids differing with respect to a single substituent, the dissimilarity score between A and A′ is largely independent of whether $-COO^-$ carbons are both labeled as hydrophobes or both labeled as anions—they are just a common feature of both A and A′. If, however, the differing electronegativities of the varying substituents cause a shift of the $-COOH$ ionization status in A vs A′, the difference is clearly reflected in the dissimilarity score. Things are different for QSAR: ‘noise’ from the allegedly hydrophobic carboxylate carbons happened to accumulate atop of an apparently relevant fingerprint element (Ar2-Hp4-Hp4), decreasing its propensity to enter QSARs and forcing the machine learning process to come up with alternative, less well performing models.

Pinpointing of specific p$K_a$-related effects in QSARs would have been possible if a top common model (or at least models including the same descriptors, with differing coefficients) would have been found for both D and D-R sets. Unfortunately, this was never the case. Trying to use D-R population levels in D models, or vice versa, always leads to significant prediction errors. Optimally validating models do not happen to differ solely because SQS fails to rediscover the same equation when run with the other descriptor set. Top validating models based on one fingerprint version were genuinely incompatible with other descriptors. Moreover, all attempts to fit D-R models with terms entering the top D equations, or vice versa, failed (results not shown). If the same top validating model would hold for both D and D-R series, with only predictions of proteolytic equilibrium-dependent compounds seen to vary in function of the flagging strategy, the direct impact of p$K_a$-dependence could have been monitored. In reality, switching from D to D-R prompts the SQS engine to come up with diverging sets of models, under the combined influence of both the p$K_a$-specific and nonspecific flagging differences that were highlighted above.

**3.2.4. Influence of Fuzzy Mapping on QSAR Quality.** The employment of fuzzy logic at the descriptor build-up stage has no significant impact on QSAR performance—at least not at the level of fuzziness proned by the D version—except for the DhfR inhibitor set. Here, fuzziness actually appears to be detrimental in terms of average validation propensity shifts, though it has no noteworthy impact on the top model quality. The D-S set, with a grid mesh $E_{\text{step}} = 2$, does not capture any information concerning atom triplets separated by an odd number of bonds. Fuzzy logic is mainly a tool to avoid triplets ‘slipping’ through the grid mesh defined by such a rarefied, smaller size triangle basis set and was shown to have a positive effect on the NB of 2D-FPT. However, it does not appear to be essential for QSAR model buildup. This makes sense if recalling that 2D-FPT fingerprints are highly redundant, in the sense that some triplet occurrences are necessarily correlated (if no positive charge is present, then all triplets featuring a PC will simultaneously have population levels of 0, etc.). Nevertheless, note that such interpopulation level correlations must not necessarily be of a linear nature: the more diverse and large the compound sets, the lesser the chance to find an even-edged triplet having its population level linearly correlated to one of the ‘missed’ odd-edge key triplets. If such correlations exist, the population level of the latter may thus implicitly account for one of the ‘missed’ triplets, throughout training and validation sets—very much in the same way in which a given triplet was shown to implicitly monitor the presence or absence of a single functional group. This is apparently the case within the DhfR inhibitor set. Unfortunately, the top validating models cannot reveal any more specific details of the problem, as they have similar validation propensities, and an in-depth analysis of all the relevant (and quite complex) D and D-S models, aimed at understanding the differing average behavior, is too cumbersome to undertake.

On the one hand, fuzziness plays an important role in mimicking the tolerance of certain receptors with respect to varying spacer length between two key groups. There is however no straightforward way to detect such examples throughout the 13 data sets, although it is clear that compound sets with various small substituents around a large central scaffold (such as steroids) are not concerned. On the other hand, too much fuzziness will eventually lead to degenerated fingerprints: as less and less strict edge length matching criteria are imposed, more and more atom triplets—involved in binding or not—will get a chance to contribute an increment to the population level of the given basis triplet. Even with D-S fingerprints, there are chances to find more than one arrangement of three atoms having the required pharmacophore flags and topological distances to match the same basis triplet: if only one of these atom triplets is important for activity, its signal will be buried under the noise from the other fortuitous contributors. Fuzziness only worsens such pitfalls. The impact of fuzziness on QSAR performance is thus different from the impact on neighborhood behavior. In the latter case, considering a pair of close analogues A and A′ with a common scaffold, the fact that in the unfuzzy versions some atom triplets are ignored is not of paramount importance, as the ones slipping between the meshes of the grid will be roughly the same in A and A′. Also, degeneracy due to fuzziness lets contributions from equivalent triplets build up equivalent final population levels, i.e., it has no negative impact on similarity scoring. However, if A and A′ differ in terms of a centrally inserted $-CH_2-$ group between two moieties, then triplets specifically localized within each moiety will appear unchanged in the fingerprint, whereas the mapping of triplets featuring corners from both parts of the molecules may, in the absence of fuzzy logic, vary dramatically as even edge lengths become odd due to $-CH_2-$ insertion and vice versa. The dissimilarity of A and A′ is therefore at risk of being overestimated. In QSAR, however, triplet degeneracy is a serious problem, whereas the issue of varying long-range contributions with spacer length might be circumvented by letting the model simultaneously pick several correlated long-range triplets considering alternative corners adjacent to the actual porters of ligand-site interactions. Among these, some will be populated at odd and others at even spacer lengths—receptor tolerance with respect to varying spacer length may be mimicked without the explicit need for fuzzy logic.

**3.3. Impact of Nonlinearity on QSAR Quality.** As can be seen from Figure 7, the best validating nonlinear models outperform their linear counterparts in a majority of situations. The 59 cases represent compound set/fingerprint version (D, O, C, D-R, D-S) combinations for which both the best linear and the best nonlinear model scored $R^2_V > 0.1$. Out of these, in 36 situations the nonlinear approaches turned out to be more robust validators, sometimes (in 8 cases) by more than 0.1 $R^2_V$ units. The most clear-cut improvements due to nonlinearity (>0.2 $R^2_V$ units) are observed for thrombin models: with D descriptors, nonlinearity allows an improvement of $R^2_V$ from 0.402 to 0.617, with D-R from 0.375 to 0.619, while with O descriptors a jump from 0.442 to 0.737 is observed. At the opposite, only three cases in which the introduction of nonlinearity triggers a decrease by 0.1 $R^2_V$ units could be seen: the most clear-cut is observed for the Ster set and D-R fingerprints (from 0.457 to 0.319).

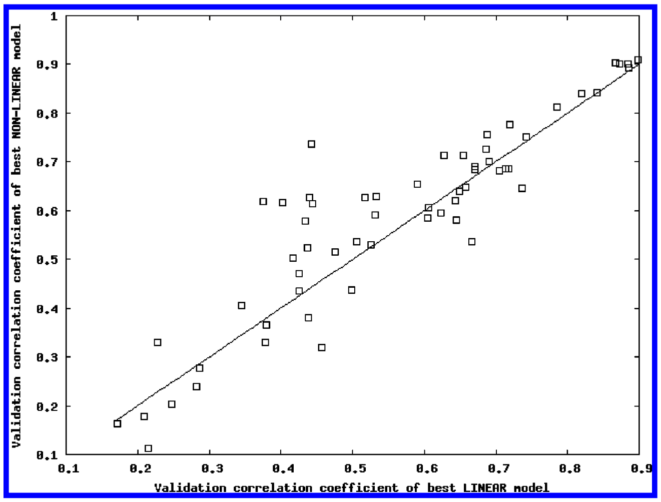
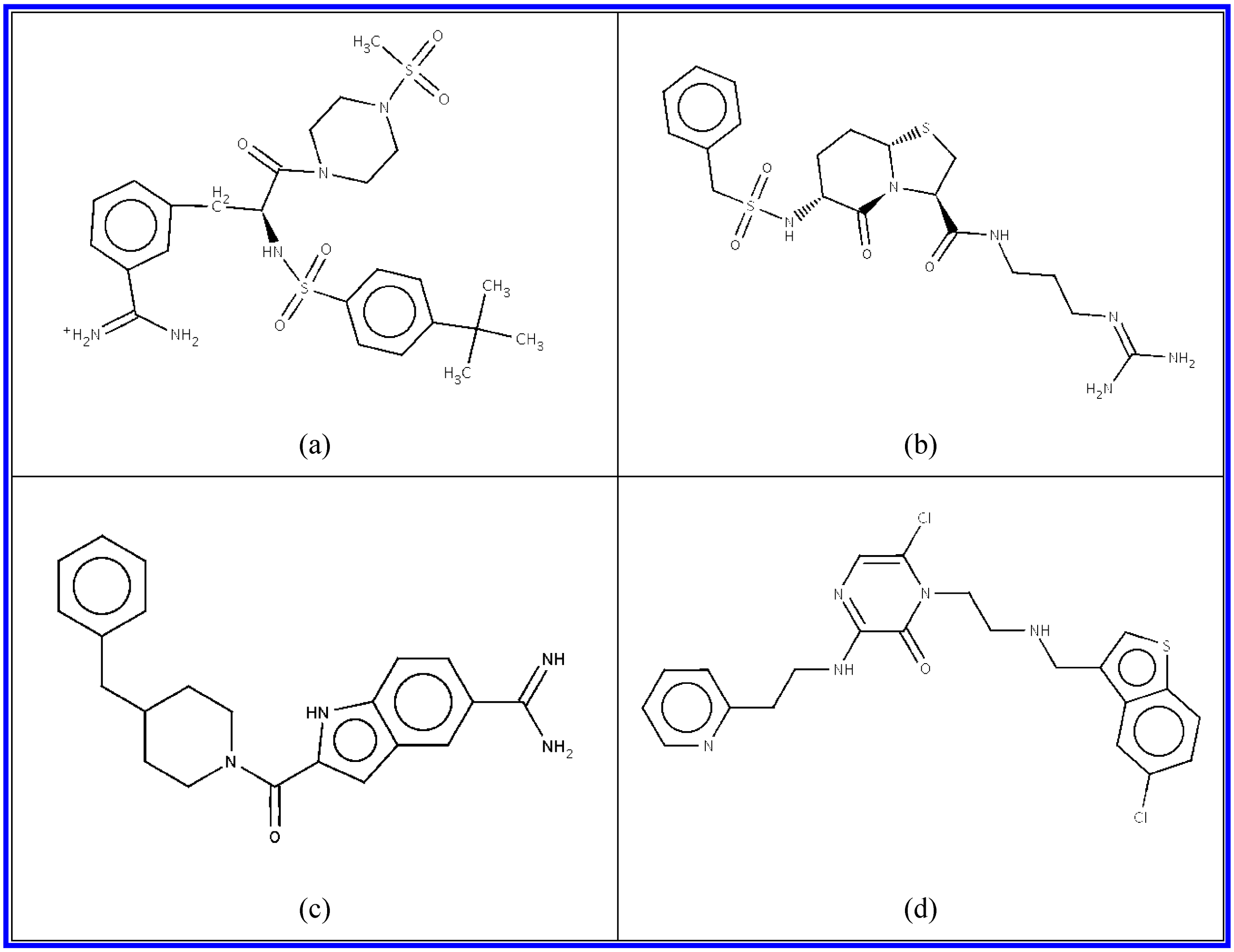


**Figure 7.** Comparative plots of $R^2_V$ values scored, for each of the 13 compound sets, using each of the 5 descriptor versions, by the top validating nonlinear (on Y) and respectively linear (on X) equations (59 out of the 13 × 5 = 65 QSAR problems shown).

Nonlinear models are thus clearly better at extrapolating the knowledge extracted from the learning set to validation set molecules. These results reinforce the similar trend reported in earlier work.[25]

**3.4. Beyond Validation: QSAR Extrapolability to Different Chemotypes.** Successful validation is just a necessary but by no means sufficient guarantee of the actual usefulness of a model in virtual screening of random compound collections. Therefore, an in-depth assessment of QSAR models should, whenever possible, go beyond the simple comparison of $R^2_V$ values. The first attempt to challenge the top validating nonlinear Thr QSAR model (D version) with predicting the activities of chemically different cocrystallized ligands (Figure 8) appeared quite promising at first sight: both compounds (b) and (c)—but not (d)—were predicted active. The latter, however, is known to adopt a different binding mode in the Thr active site—nothing in the training set could have hinted that such molecules may inhibit thrombin. Unfortunately, a closer look at the prediction showed that the high p$K_i$ values predicted for (b) and (c) both stem from a single, very large contribution of the term 11.1×zexp3(HD4-Hp6-PC4). [zexp3(D) = exp[$-3(D-\langle D\rangle)^2/\sigma^2(D)$]—Please refer to Table 1 of the previous publication[25] for more details about the predefined nonlinear transformations in SQS.] Given the standard[25] average <HD4-Hp6-PC4> and variance $\sigma$(HD4-Hp6-PC4) population levels of 1.2 and 6.4, respectively, and knowing that HD4-Hp6-PC4 is not populated in either of the (b) and (c) molecules from Figure 8, the absence of such a triplet contributes 11.1×zexp3(0)=+10 to predicted p$K_i$ values. This makes no sense—according to this model, any molecule without HD4-Hp6-PC4 triplets is a thrombin inhibitor (the considered 12-variable model does not contain any other negative potentially compensating contributions). Indeed, a quick verification confirmed that, according to this model—excellent training and validation statistics notwithstanding—all the compounds from the other sets used in this work should be nanomolar thrombin inhibitors. This is an artifact due to the low diversity of the training/validation set: the HD4-Hp6-PC4 triplet is populated in *all* training and *all* validation molecules, because it stems from the common amidine-phenylalanine moiety: the cation flag is set on the



(a)

(b)

(c)

(d)

**Figure 8.** Part (a) is a typical thrombin inhibitor, featuring the amidine-phenylalanine scaffold characteristic of the Thr set, parts (b)[26] and (c)[27] are chemically different (cocrystallized) amidine/guanidine inhibitors, while part (d)[28] represents a radically new amidine-free class of ligands adopting a different binding mode.

amidine carbon, the donor is the phenylalanine >NH, at 6 bonds from the cation, while the phenyl ring carbon in *para* to the amidine, playing the role of the hydrophobe, is at 4 bonds from both PC and HD (other phenyl carbons also contribute, due to fuzzy mapping). If the phenylalanine carboxylate is coupled to a secondary amine, like in Figure 8(a), there are no other contributions to the population level of HD4-Hp6-PC4. The data set however contains a subset of primary amides: in this case, the HD flag of the CONH group is at 7 bonds from the PC and contributes to the HD4-Hp6-PC4 population. Primary amides have thus significantly higher HD4-Hp6-PC4 levels (i.e., lower zexp3 values), and, furthermore, they are on average significantly less active than secondary amides. Thus, zexp3(HD4-Hp6-PC4) entering the model with a large coefficient makes perfect sense in as far as the family of amidine-phenylalanines is concerned but is faulty outside this restricted applicability domain. Nonlinear models may indeed increase robustness of extrapolation from training to validation set but still do not offer guarantees of actual success in virtual screening. Improved validation set results might come at the price of a restricted applicability range or at least at the price of increased difficulty to properly define the applicability range in the presence of nonlinear terms.

None of properly validating ($R^2_V$>0.4) representative nonlinear D models succeeded to specifically highlight (i.e., predict at submicromolar inhibition levels) (b) and (c) by contrast to randomly chosen inactives. Fortunately, 2D-FPT models are overlay-independent, which allows sets of arbitrarily high diversity to be used for training (training compounds need not have a common core, in order to be superimposable). Therefore, 125 compounds representing a randomly picked 10% of the other 11 data sets (FXa excluded), assumed to be inactive against thrombin (p$K_i$ set to 4.0), were added to the initial Thr series. The resulting 'expanded' (ThrEx, 213 compounds) set was split into 169 training and 44 validation compounds and resubmitted to the SQS-driven nonlinear model buildup with D fingerprints. This time, the representative set of SQS equations featured two properly validating models, being both able to discriminate between Thr inhibitors and randomly picked compounds and to predict that (b) and (c) are submicromolar Thr inhibitors. Out of these two, one furthermore returned an excellent estimation of 50 nM for the affinity of (d) compared to the experimentally[28] reported 3 nM. This 12-variable nonlinear equation ($R^2_T$ = 0.864, RMSPE = 0.73, $R^2_V$ = 0.762) is given below, with *zQ(D:a:v)* denoting the predefined nonlinear functions[25] to be applied to descriptor *D* after its average/variance rescaling (*z*-transformation) with respect to the average value *a* and the variance value *v*, i.e. *zQ(D:a:v)=Q[(D-a)/v]*:

$$pK_i^{pred} = 0.07 \times \mathit{HA8Hp6PC4} - 8.1 \times 10^{-4}\,\mathit{Ar4Ar10HA10} - 0.57 \times \mathit{HP10PC8PC10} - 2.1 \times 10^{-4}(\mathit{HA12Hp12PC12})^2 - 0.16 \times (\mathit{Hp6Hp6PC8})^2 + 0.3 \times \mathrm{zexp}(\mathit{Ar10HP10PC6}{:}2.3{:}13) - 0.45 \times \mathrm{zexp3}(\mathit{Ar6Ar8Hp12}{:}29.3{:}93.4) + 0.5 \times \mathit{zsig}3(\mathit{Ar6HA2Hp6}{:}119.1{:}156.8) - 0.5 \times \mathrm{zexp}(\mathit{Ar8Ar10NC4}{:}3.4{:}18.9) + 0.97 \times \mathrm{zexp3}(\mathit{Ar8HA6Hp6}{:}46.6{:}124.7) + 0.58 \times \mathit{zsig}(\mathit{HA10HA12Hp4}{:}18.8{:}77.9) + 3.46 \times \mathrm{zexp}(\mathit{Ar6HA12NC10}{:}0.2{:}2.0) \quad (3)$$

Equation 3 has the remarkable property to capture contributions not met in inactives but found in both active amidine-phenylalanine derivatives and (b), (c), or (d). It is however not the top validating nonlinear ThrEx model: this latter ($R^2_{\mathrm{T}} = 0.901$, RMSPE = 0.32, $R^2_{\mathrm{V}} = 0.956$) includes Thr-family specific contributions not shared by the structurally different inhibitors. As far as the machine learning process is left to focus only on the differences between actives and inactives within the Thr set, models exploiting all the idiosyncratic correlations due to the peculiar constitution of the data set perform well at training and cross-validation and will be selected. Some of these reveal themselves as meaningless when confronted to the diverse inactives of the extended set. Therefore, refitting with respect to the extended set leaves room for some less family specific, more general models to make it into the representative pool of equations as well.

It is also worth pointing out that out of 1113 distinct models—all of which boast outstanding training and validation criteria ($R^2_{\mathrm{V}}$>0.7)—only two stood up to the challenge of predicting compounds outside the training chemical family. In general, the QSAR problem is considered as solved if one well validating model has been found—what is the use of generating all these equally well performing 'redundant' models? The importance of aggressive QSAR problem space sampling resides in the fact that such 'redundant' models will cease to behave similarly when confronted to external molecules. The lower the informational content of the training set, the lower are the success expectations for any actual virtual screening based on thereon trained models, no matter how training is conducted. With stepwise/deterministic approaches, few equations—most likely all irrelevant—will be built. SQS may well enumerate relevant equations—but it will be impossible to guess which are the ones, unless an external test set can be used for further evaluations. The key advantage of SQS is that external sets may be too small to be useful at training (adding one or two external compounds to a homogeneous family does not help, with no cross-validation being possible) and yet allow for the discarding of most of the many thousands of sampled models, keeping only the ones that were not (yet) proved wrong. Classical QSAR buildup producing few equations is likely to end up with no models at all after confrontation with the external molecules.

**3.5. Structural Interpretation of 2D-FPT Models—Do Topological Pharmacophores Make Sense?** There is to our knowledge no direct experimental evidence of the binding mode of the amidine-phenylalanine derivatives of the Thr set. However, given the binding modes of related compounds and the overlay hypotheses standing at the basis of the original QSAR studies[34] concerning this family, it may be safely assumed that they would occupy all the three known binding pockets of thrombin. For example, compound (a) from Figure 8 would place the benzamidine moiety in $P_1$, the less hydrophobic sulfonylpiperazine substituent in $P_2$, and the t-Bu-phenyl group in $P_3$. Figure 9 illustrates this expected binding mode atop of the experimental bound geometry of compound (c). Clearly, compounds (a) and (c) are topologically different: in the former, the substituents filling the pockets feature a 'star' topology $P_1(-P_2)P_3$ centered on the phenylalanine α carbon, whereas in (c)—as well as in (b)—this arrangement is linear: $P_1-P_2-P_3$. Compounds like (b) or (c) have to adopt a U-shape geometry to close their $P_1$ and $P_3$ moieties up. This is a challenge to 2D-FPT-based models, since (a) and (b,c) do not share any topological triplets spanning all the three pocket-filling moieties: the $P_1-P_3$ topological distance in (b) or (c) is much larger than in (a). However, 3D-distance based common pharmacophore triangles might be found if the proper U-shaped fold is considered—should it be thus concluded that topological pharmacophore-based models prove unable to perform 'lead-hopping' from the star topology of the Thr series to the linear arrangement of the alternative ligands (b) and (c). Obviously not, since eq 3 applies to both of these topologies.

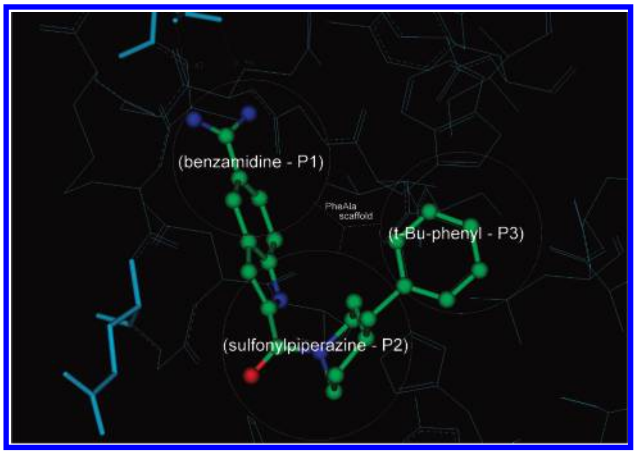


**Figure 9.** Thrombin active site with cocrystallized ligand—Figure 8(c)—and hypothesized binding mode of Thr set amidine-phenylalanine derivative from Figure 8(a).

The high predicted p$K_{\mathrm{i}}$ values for the compounds in Figure 8 mainly stem from three main contributions. The highest one, an increment of +3.4 due to 3.46 × $z$ exp($Ar6HA12NC10$:0.2:2.0) is constant for all four molecules, since none includes any negative charge. The term signals that compounds featuring such a triplet are not likely to be active—a 'lesson' learned from the additional inactives entering the ThrEx set. This makes sense insofar as thrombin clearly prefers cationic compounds. However, a negative charge will perhaps be detrimental even if it is not a part of this peculiar triplet chosen here.

Next, the Gaussian function of Ar8-HA6-Hp6 contributes with 0.6 to 0.9 p$K_{\mathrm{i}}$ units—the largest contribution seen in (a), where the triplet is represented once (fuzzy population level of 63), while the lowest occur if the triplet is not populated—in (c) and (d). Given the large variance of this triplet population level within the set of representative drugs used for 2D-FPT calibration,[1] this term may play an important (activity-detrimental) role only in molecules containing several such triangles.

Insofar, the prediction that compounds (a)−(d) are active was only based on the fact that they are free of unwanted features, seemingly causing an affinity loss. Other contributions are, with one key exception, quite small (less than ± 0.5 p$K_i$ units) and tend to cancel out. The remaining term is of paramount importance, based on a triplet actively favoring activity (HA8-Hp6-PC4). Figure 10 exemplifies the actual occurrences of this triplet in the molecules (recall that there are other atom triplets contributing, besides the highlighted ones—notably the ones including the symmetrically situated C atom in piperazine/cyclohexylamine rings). The triplet highlights two essential elements of the actual thrombin binding pharmacophore: the cation (amidine) interacting with Asp 189 from $P_1$ and the $P_2$ hydrophobic moiety 'sandwiched' between Trp 60 and Tyr 83. As the $P_1$ and $P_2$ binding moieties are topologically close in both (a) and (b/c), this particular triplet ensures model extrapolability from one topological family to the other.

Intriguingly, there is no role in binding directly attributable to the hydrogen accepting carbonyl of the triplet. However, this carbonyl is nevertheless 'important'—not structurally, but chemically, for synthesis reasons. As acylation is a preferred building block coupling reaction, it is not astonishing to see a conserved carbonyl throughout diverse series of compounds that were conceived as timers matching the three thrombin binding pockets. This is a nice example showing that QSARs will never represent absolute training-set independent laws, as training sets will always be biased, be it only for chemical feasibility reasons.

The $P_3$ pocket does not appear to play any important role: according to eq 3, compounds filling in $P_1$ and $P_2$ already score better than micromolar. This is arguably wrong, but, unfortunately, the data set presented to the machine learning tool cannot unambiguously tell whether hydrophobic groups in the $P_3$ pocket are absolutely necessary for activity or not. There are two compounds without a large hydrophobic group bound to the phenylalanine N, in which this group is actually not substituted at all and therefore cationic. The compounds are inactive, but this is too little evidence to make the model learn that hydrophobes in $P_3$ are important. Actually, the unsubstituted inactive compounds happen to be properly predicted, due to a penalty stemming from the square of the Hp6-Hp6-PC8 term. The extra positive charge in N-unsubstituted phenylalanines leads to increased population levels of this negatively weighted triplet, i.e., the model seems to suggest that inactivity is due the additional free charge. As far as the only examples missing a $P_3$ hydrophobe are also the only ones with a protonated phenylalanine N; there is no reason to prefer one explanation over the other.

Although the success of eq 3 appears to be partly due to the 'illusion' that the $P_3$ pocket may be ignored as something filled 'by default' with a hydrophobe in all the examples given, this does not mean that models accommodating various ligand topologies will be impossible to build once that the training set is furnished with enough examples to document the influence of pharmacophore pattern variation in the $P_3$ region. Such an equation may be based on several triplets, each regrouping elements from ($P_1$ and $P_2$), ($P_2$ and $P_3$), and ($P_2$ and $P_3$), respectively, binding moieties: there is no need to enter a triplet having each corner from a different moiety, since such triplets will not be shared thorough topologically different series.

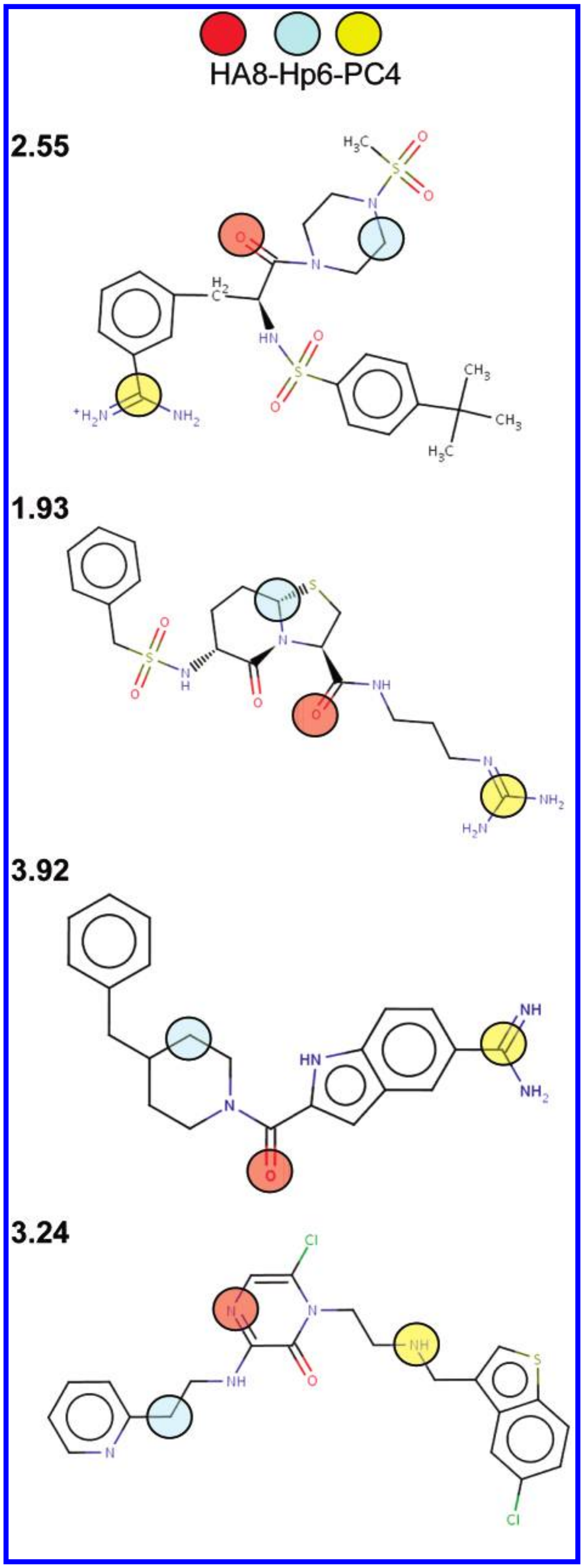


**Figure 10.** Color-coded display of the occurrence of the key Thr affinity modulating triplet HA8-Hp6-PC4 in the four chemically different inhibitors.

The successful prediction of a typical compound from Figure 8(d) is due to the herein present triplet HA8-Hp6-PC4. However, this very same topological pharmacophore

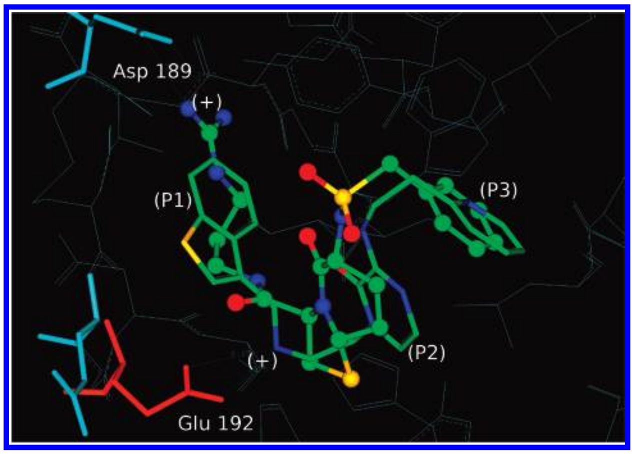


**Figure 11.** Superimposed thrombin active sites with aligned cocrystallized ligands—Figure 8(b),(d). Glu 192 forming the atypical salt bridge with the cation of compound (d) is seen (in red, vs default light blue) to shift its side chain in order to interact.

seen to previously match elements binding to the $P_1$ and $P_2$ binding pickets covers in compound (d) elements seen to go into $P_3$ (the hydrophobic spot of the pyridine linker) and $P_2$, respectively. This is a purely accidental example of ‘inverse’ degeneracy, where the same topological triplet may be accommodated in two different ways in an active site. [This ‘inverse degeneracy’ is antonymic to the previously illustrated ‘classical’ degeneracy of 2D-FPT, where different atom triplets may indistinctively contribute to the population level of the same basis triplet.] The cation now forms a salt bridge with Glu192, which reorients its side chain in order to enter this interaction. While the HA corner of the relevant triplet has no direct binding role in compounds (a)−(c), the pyrazone oxygen, which is an alternative contributor to HA8-Hp6-PC4 (not highlighted in Figure 10 for the sake of simplicity), is actually involved in the interaction with the main chain >NH of Gly 216. All this is however anecdotic: the QSAR model did not foresee that the thrombin active site supports this alternative binding mode. Out of the two equations that correctly extrapolated the activity of (b) and (c), on the one hand, of the family of (a), only eq 3 included triplets also shared by (d). Due to the peculiarities of the thrombin active site, the two distinct binding modes might be explained by the same model. From a practical point of view, using eq 3 in a virtual screening would have triggered a major breakthrough in thrombin inhibitor research, (serendipitously) leading to a completely new family. Unfortunately, there are no deterministic recipes to find such models, if ever they happen to exist.

## 4. CONCLUSIONS

As far as the benchmarking exercise goes, 2D-FPT-based QSARs fare extremely well, outperforming not only 2D and 3D-index-based models but also the elaborate, overlay-based CoMFA approaches. The biological property less well handled by pharmacophore triplet models is, unsurprisingly, the heme alkylating activity of artemisinin analogues—the only studied property not reflecting a reversible noncovalent target inhibition process, conceptually associated with ‘binding pharmacophores’. 2D-FPT are thus information-rich and relevant descriptors of site-ligand recognition processes. The study of optimal 2D-FPT fuzziness highlighted the problem of 2D-FPT degeneracy, which may be of serious concern in descriptor selection-based QSARs (much more so than in similarity scoring), although pharmacophore triplets suffer much less from this problem than pairwise descriptors.

Nevertheless, the ‘topological pharmacophores’ defined by triplets entering 2D-FPT models are not necessarily representatives of ligand-site anchoring points. This work highlighted the very limited scope of the training and validation sets typically used for QSAR buildup and benchmarking, showing many situations where the successful QSAR fitting and validation relied on family specific idiosyncrasies. Another symptom of training set limitations is the generation of models predicting high activity values by default and relying on penalizing terms to reduce the score for the known inactives containing ‘unwanted’ features, or these models predicting high activities for any molecules too small to contain any triplets, be it wanted or unwanted, are thus senseless.

Although p$K_a$-dependent pharmacophore flagging was proven to be more rigorous than the rule-based one, leading to a much better understanding of the molecular similarity principle,[1] in QSAR studies, set-specific artifacts gained the upper hand over p$K_a$-related effects: the best performing flagging scheme was often the one best exploiting some set-specific coincidence.

The broad range of encountered set-specific artifacts (and which surely appeared under different forms with the various descriptors used in the cited literature studies) is a serious incentive to reconsider the actual sense of QSAR buildup, validation, and benchmarking on such limited series. In light of the many examples of chemically flawed equations brilliantly passing ‘external’ validation tests—against new members of the training family, more precisely—the present work suggests that (a) any training set should be completed with a set of diverse (presumed) inactives before QSAR buildup. This is easily feasible with 2D-FPT and other overlay-independent descriptors but problematic with CoMFA and related tools. (b) An additional challenge against topologically different actives should be regularly included in benchmarking. General equations based on chemically meaningful terms may be enumerated upon extensive sampling of the QSAR problem space, among many other successfully validating family specific models. They are likely to perform reasonably well, without being the best in terms of training/validation scores (therefore, deterministic QSAR build-up procedures may not find them). The challenge to predict topologically different actives is needed to highlight them among the many apparently redundant alternative models.

Concerning the interpretability of 2D-FPT models, it must be pointed out that these were excellent tools to highlight training set deficiencies: the chemically interpretable terms responsible for observed artifacts allow a straightforward comprehension of the problem. Whether or not selected triplets match actual binding pharmacophores is mainly a question of training set diversity. 2D-FPT may lead to valuable QSAR models, provided the training set diversity is sufficient to force the learning of key features, not of secondary pharmacophore signatures that serendipitously reflect subsets locally enriched in actives. If this is the case, the applicability range of such models may extend over several chemotypes—and may even go beyond expectations if the targeted active site offers alternative models to accommodate a topological triplet.

The setup files.xml controlling 2D-FPT buildup are available upon request from the author.

**Supporting Information Available:** Thirteen considered data sets plus the compiled extended thrombin inhibitor set ThrEx, for each set, a two-column (SMILES, activity score) <set>.smi.txt file, a list of the molecules entering the validation set <set>.vset.txt, and the activity-descriptor matrices <set>.<descriptor-version>.txt are available, for each descriptor version D, D-R, D-S, O, and C (all files—Unix ASCII). This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

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

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

(3) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Towards generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and related methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.

(4) Milicevic, A.; Nikolic, S.; Trinajstic, N. Toxicity of aliphatic ethers: A comparative study. *Mol. Diversity* **2006**, *10*, 95−99.

(5) Adam, M. Integrating research and development: the emergence of rational drug design in the pharmaceutical industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(6) Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A. M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. Pharmacophore-based design of HIV-1 integrase strand-transfer inhibitors. *J. Med. Chem.* **2005**, *48*, 7084−7088.

(7) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.

(8) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold hopping with molecular field points: identification of a cholecystokinin-2 (CCK2) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-based CCK2 antagonists. *J. Med. Chem.* **2005**, *48*, 6790−6802.

(9) Cramer, R. D., III; Patterson, D. E.; Bunce, J. E. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(10) Horvath, D. ComPharm − Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design. In *QSPR /QSAR Studies by Molecular Descriptors*; Diudea, M., Eds.; Nova Science Publishers, Inc.: New York, New York State, 2001; pp 395−439.

(11) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Chemoinformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; pp 117−137.

(12) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687−698.

(13) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687−2694.

(14) Fechner, U.; Paetz, J.; Schneider, G. Comparison of Three Holographic Fingerprint Descriptors and their Binary Counterparts. *QSAR Comb. Sci.* **2005**, *24*, 961−967.

(15) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of d(1) dopaminergic antagonists for database mining. *J. Med. Chem.* **2005**, *48*, 7322−7332.

(16) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *J. Med. Chem.* **2005**, *48*, 6563−6574.

(17) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(18) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1998**, *38*, 144−150.

(19) Sciabola, S.; Morao, I.; de Groot, M. J. Pharmacophoric Fingerprint Method (TOPP) for 3D-QSAR Modeling: Application to CYP2D6 Metabolic Stability. *J. Chem. Inf. Model.* **2007**, *47*, 76−84.

(20) Güner, O. F. *Pharmacophore Perception, Use and Development in Drug Design, IUL Biotechnologies Series;* Güner, O. F., Eds.; International University Line: La Jolla, CA. 2000.

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

(22) Skold, C.; Karlen, A. Development of CoMFA models of affinity and selectivity to angiotensin II type-1 and type-2 receptors. *J. Mol. Graphics Modell.* **2007**, *26*, 145−153.

(23) Guha, R.; Jurs, P. C. Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.

(24) Coats, E. A. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discovery Des.* **1998**, *12−14*, 199−213.

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

(26) Wagner, J.; Kallen, J.; Ehrhardt, C.; Evenou, J. P.; Wagner, D. Rational design, synthesis and X-ray structure of two selective noncovalent thrombin inhibitors. *J. Med. Chem.* **1998**, *41*, 3664−3674.

(27) Chirgadze, N. Y.; Sall, D. J.; Klimkowski, V. J.; Clawson, D. K.; Briggs, S. L.; Hermann, R.; Smith, G. F.; GiffordMoore, D. S.; Wery, J. P. The crystal structure of human alpha-thrombin complexed with LY178550, a nonpeptidyl, active site-directed inhibitor. *Prot. Sci.* **1997**, *6*, 1412−1417.

(28) Bulat, S.; Bosio, S.; Papadopoulos, M. A.; Cerezo-Galvez, S.; Grabowski, E.; Rosenbaum, C.; Matassa, V. G.; Ott, I.; Metz, G.; Schamberger, J.; Sekul, R.; Feurer, A. Design and Discovery of Novel, Potent Pyrazinone-Based Thrombin Inhibitors with a Solubilizing Amino P1−P2-Linker. *Lett. Drug Des. Discovery* **2006**, *3*, 289−292.

(29) Horvath, D. et. al. − unpublished work.

(30) RCSB Protein Data Bank. http://www.rcsb.org/pdb/ (accessed Oct 08, 2007).

(31) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(32) Cazelles J.; Robert A.; Meunier B. Alkylation of heme by artemisinin, an antimalarial drug. *C. R. Acad. Sci., Ser. IIc: Chim.* **2001**, *4*, 85−89.

(33) Chemaxon − Pmapper user guide. http://www.chemaxon.com/jchem/doc/user/PMapper.html#config (accessed Oct 10, 2007). Also check the pharma-frag.xml configuration file in the JChem distribution.

(34) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.

CI7003237