

Reliability of Comparative Molecular Field Analysis Models: Effects of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospholipase A₂ Inhibitors[†]

Angel R. Ortiz,^{*,‡,§} Manuel Pastor,^{||,§} Albert Palomer,[⊥] Gabriele Cruciani,^{||} Federico Gago,[§] and Rebecca C. Wade^{*,‡}

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, Departamento de Fisiología y Farmacología, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain, Laboratorios Menarini, SA, R & D Department, Alfonso XII 587, 08912 Badalona, Spain, and Department of Chemistry, University of Perugia, Via Elce di Sotto, 8, 06100 Perugia, Italy

Received February 26, 1996[⊗]

The effects of data pretreatment, data scaling, and variable selection on three-dimensional quantitative structure–activity relationships derived by comparative molecular field analysis (CoMFA) using the GRID energy function were studied in detail for a set of inhibitors of the human synovial fluid phospholipase A₂ (HSF-PLA₂). The quality of the models was evaluated for predictive power and ability to map the receptor binding site by (a) comparison of predicted and experimental activities using cross-validation and external validation sets and (b) comparison of the regions selected in space in the CoMFA models with a crystal structure of a HSF-PLA₂–inhibitor complex, with optimized comparative binding energy analysis (COMBINE) models (Ortiz *et al.*, 1995) and with structure–activity relationships derived previously for different sets of compounds. It is found that (1) data scaling and dielectric modeling strongly influence CoMFA results. Unscaled data and a uniform dielectric constant of 4 are well suited to GRID-CoMFA studies for the present compound set. (2) The GOLPE and Q²-GRS variable selection methods select variables in roughly the same regions in Cartesian space, but they produce different models in chemometric space and differ in their sensitivity to data scaling and pretreatment and their tendency to overfitting. (3) CoMFA models are consistent with COMBINE models in that they identify approximately the same intermolecular interactions as relevant for activity. Our study provides support for the qualitative receptor-mapping properties of CoMFA models and for the validity of variable selection when applied with care and also provides guidelines for how to evaluate the quality of CoMFA models.

Introduction

One important aim of drug design is to correlate the three-dimensional (3D) structure of drug molecules with their biological activities, i.e., to derive 3D-QSARs.^{1,2} The goal is to be able to design and predict the biological efficacy of new molecules prior to synthesis. At present, one of the most frequently used tools for this task is comparative molecular field analysis (CoMFA).³ It entails the superposition of a set of compounds whose activities have been measured, the computation of interaction energy fields for probes on a grid around each compound, and partial least squares (PLS) statistical analysis to correlate fields with activity and to detect regions around the molecules where there are interactions that have an important impact on activity. It does not require information about the 3D structure of the receptor. However, even when the 3D structure of the receptor is available, the derivation of structure–

activity relationships for a family of ligands may aid lead optimization. In this context, we have recently reported⁴ a new approach for the prediction of ligand binding affinities based on the analysis of the relationship between the binding affinity of a set of known ligands and selected interaction energies with the receptor. This method, referred to as comparative binding energy analysis (COMBINE analysis), makes use of the same statistical tools as CoMFA, but energetic information derived from the structures of ligand–receptor complexes is used to produce better regression models. An advantage of COMBINE analysis compared to molecular mechanics calculations alone is that (like CoMFA analysis) it weights pairwise interaction energies between the individual atoms of the receptor and ligand in order to improve the correlation with binding affinity. This weighting procedure can serve to filter out some of the inaccuracies of potential energy functions and errors in modeling.

The purpose of this paper is to assess the sensitivity of 3D-QSAR models to different data treatments and statistical analysis protocols. Sensitivity to statistical analysis protocols arises because, in 3D-QSAR methods, there is a very large number of quantitative descriptors of molecular interactions from which only a small number of informative descriptors, or variables, that correlate with activity must be detected. Selection of the most informative variables and elimination of background noise is a general problem, common to any 3D-QSAR method like COMBINE or CoMFA. The PLS

[†] Abbreviations: partial least squares (PLS), principal components analysis (PCA), fractional factorial designs (FFD), comparative molecular field analysis (CoMFA), comparative binding energy analysis (COMBINE), human synovial fluid phospholipase A₂ (HSF-PLA₂), three-dimensional (3D), quantitative structure–activity relationship (QSAR), Q²-guided region selection (Q²-GRS).

* To whom correspondence should be addressed (e-mail, wade@embl-heidelberg.de; fax, +49 6221 387 517).

[‡] European Molecular Biology Laboratory.

[§] Universidad de Alcalá.

[⊥] Laboratorios Menarini, SA.

^{||} University of Perugia.

[⊗] Abstract published in *Advance ACS Abstracts*, March 1, 1997.

regression method is able to handle the large matrices of variables used in 3D-QSAR, but when the signal-to-noise ratio is too low, the PLS method can fail to unveil the variables correlated with the activity.⁵

In most CoMFA studies, the structure of the receptor is unknown, and therefore, there is no well-defined way to assign grid spacing and grid box dimensions around the molecules. As a consequence, a large number of ligand-probe interactions must be considered, most of which are irrelevant for explaining the activity. In fact, it has been shown that keeping irrelevant variables in the model can have detrimental effects on predictive ability.⁶ Thus, a method that would successfully select only those variables which have the most significant effect on the biological activity would be of considerable interest. In principle, there are two different strategies that can be pursued in order to increase the signal-to-noise ratio: one is manipulation of the relative scaling or weighting of the interactions, and the other is statistical selection of the relevant variables.

The scaling of variables increases the weight of certain interactions that may therefore dominate the regression model. If these interactions are important for activity, this is a way to improve the signal. Several scaling procedures have been suggested.⁶ A popular one is "autoscaling", i.e., rescaling all independent variables to unit variance, thus giving each variable the same opportunity to influence the PLS regression model. Another method is "block-scaling", in which the data matrix is divided into consistent groups of variables (for example, van der Waals and electrostatic) and each block is given unit variance, with the constraint of maintaining the relative weights of variables inside the block. However, it can be argued that these matrix modifications result in a balance of the interactions which has no physical interpretation. An alternative, physically based approach to weight interactions is to apply a dielectric model to the electrostatic interactions. A dielectric model chosen to properly reproduce the dielectric properties of the receptor binding site can give physically based interaction weights. Unfortunately, there is not a consensus about the best way to reproduce the dielectric properties of protein binding sites, although several *ad-hoc* dielectric models have been proposed (see below).

Several approaches to variable selection based on statistical tools have been put forward. The GOLPE method⁷ essentially works by evaluating the effects of individual variables on the predictive ability of the model. To calculate these effects efficiently, the GOLPE method builds and cross-validates a set of PLS models following a fractional factorial design (FFD) scheme. An alternative approach is the cross-validated Q^2 -guided region selection (Q^2 -GRS) method.⁸ In this method, the grid box is divided into smaller cubic regions which are used to build independent PLS models. The regions producing better models are then "pasted" together to create a "composite" region from which the final regression model is generated.

Data pretreatment and selection of the most informative variables are key steps in any 3D-QSAR study and have therefore received increasing attention in the field.⁷⁻¹² However, the effects on model quality of changing different parameters in the CoMFA methodology are not necessarily independent, and to date, this interdependency has not been studied systematically.

In this paper, different protocols for data scaling and variable selection in CoMFA analysis are assessed by developing regression models for all possible protocol combinations. The quality of the models has been carefully evaluated. A difficulty in model evaluation is that a consensus is lacking on how to quantify the quality of a model. In our opinion, the two main aspects of the model which should be considered are (1) its predictive ability, which can be assessed by internal and external validation, and (2) its ability to mimic the key ligand-receptor interactions in the receptor binding site. We have assessed predictive ability both by internal validation (cross-validation) and by external validation for several prediction sets. We have quantified predictive ability with the SDEP and Q^2 parameters which are defined by:

$$SDEP = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N}} = \sqrt{\frac{PRESS}{N}} \quad (1)$$

$$Q^2 = 1 - \left[\frac{\sum (Y - \hat{Y})^2}{\sum (Y - \langle Y \rangle)^2} \right] \quad (2)$$

where Y is experimental activity, \hat{Y} is predicted activity, $\langle Y \rangle$ is the average experimental activity, and N is the number of molecules. The pharmacophoric patterns obtained in the different models were assessed by comparison to the known crystal structure of a ligand-receptor complex, previous structure-activity studies, and "optimal" COMBINE regression models. The contributions of different residues in the receptor to activity cannot generally be deduced from graphical and geometric analysis of a structure of the ligand-receptor complex alone. Since COMBINE analysis assigns a contribution to the biological activity to each residue in the receptor, it is possible to use COMBINE models as a reference by which to judge the pharmacophoric mapping properties of the CoMFA models in a more objective way. Intramolecular interactions, which are not considered in a CoMFA model, can contribute to COMBINE models.⁴ However, if they are excluded from the COMBINE analysis, predictive models are still obtained; Q^2 is typically about 0.1 lower, but the models have similar or greater predictive ability than the best CoMFA models. The fact that COMBINE models including only intermolecular terms give good predictions indicates that reasonable receptor maps should be achievable with CoMFA.

The data set used in this study is the same as used before in the development of the COMBINE formalism.⁴ It consists of a series of 26 inhibitors (see Figure 1 and Table 1) of the human synovial fluid phospholipase A₂ (HSF-PLA₂; Figure 2).¹³ The three-dimensional structure of the HSF-PLA₂ has been solved by X-ray crystallography both in its native form^{14,15} and in a complex with the transition state analogue LM1220 (see Table 1).¹⁵ The PLA₂ inhibitors used for the derivation of the 3D-QSAR are quite large, with two long hydrocarbon tails present in all inhibitors (see Table 1). Most of them are charged, and they do not all have the same net charge. This means that in order to detect all important electrostatic interactions, a rather large grid box must be used for the CoMFA studies. On the other hand, structural variations are confined to a few places in the molecules, mainly in the transition state analogue group and in the *sn*-3 position. Thus, many of the grid points

Table 1. Chemical Formulae and Activities of the HSF-PLA₂ Inhibitors^a

| name | XLM | YLM | ZLM | SN1 | SN2 | SN3 | SN4 | SN5 | RLM | GLI | % inhibition |
|--------|--------------------|------------------------------------|--|--|---------------------------------|---|--|--|--|----------|--------------|
| lm1166 | -CH ₂ - | -CONH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -CH ₂ CH ₂ OH | -(CH ₂) ₃ CH ₃ | | | | 65 ± 14 (7) |
| lm1192 | -CH ₂ - | -CONH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | | 6 ± 9 (3) |
| lm1216 | -CH ₂ - | -CONH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -CH ₂ CHNH ₃ COO | -(CH ₂) ₃ CH ₃ | | | | 31 ± 27 (3) |
| lm1220 | -O- | -PO ₂ O- | -OPO ₂ O- | -(CH ₂) ₄ - | -C ₆ H ₁₂ | -CH ₂ CH ₂ NH ₃ | -CH ₃ | -(CH ₂) ₃ CH ₃ | | <i>R</i> | 33 ± 3 (3) |
| lm1228 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | <i>R</i> | 78 ± 12 (6) |
| lm1230 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ OH | -(CH ₂) ₃ CH ₃ | | | | 49 ± 3 (3) |
| lm1240 | -O- | -PO ₂ O- | -OPO ₂ O- | -(CH ₂) ₄ - | -C ₆ H ₁₂ | -CH ₂ CH ₂ N(CH ₃) ₃ | -CH ₃ | -(CH ₂) ₃ CH ₃ | | <i>R</i> | 9 (1) |
| lm1245 | -CH ₂ - | -CONH- | -OPOOCH ₃ CH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -CH ₂ CH ₂ CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 24 ± 36 (3) |
| lm1246 | -CH ₂ - | -CONH- | -OPO ₂ CH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -CH ₂ CH ₂ CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 45 (1) |
| lm1258 | -CH ₂ - | -SO ₂ NH- | -OCH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CF ₃ | -(CH ₂) ₃ CH ₃ | | | | 0 (1) |
| lm1261 | -CH ₂ - | -CONH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₇ H ₁₄ | -(CH ₂) ₃ CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 80 ± 6 (3) |
| lm1265 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ CH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 30 ± 18 (3) |
| lm1277 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ CH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 33 ± 18 (3) |
| lm1283 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | | <i>R</i> | 45 ± 26 (3) |
| lm1284 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | <i>S</i> | 12 ± 10 (3) |
| lm1292 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ OH | -(CH ₂) ₃ CH ₃ | | | <i>R</i> | 44 ± 18 (3) |
| lm1293 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ OH | -(CH ₂) ₃ CH ₃ | | | <i>S</i> | 40 ± 19 (3) |
| lm1298 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ NH ₃ | -(CH ₂) ₃ CH ₃ | | | | 4 ± 20 (3) |
| lm1299 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -(CH ₂) ₃ CH ₃ | -(CH ₂) ₃ CH ₃ | | | | 0 (3) |
| lm1300 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₅ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | | 24 ± 24 (3) |
| lm1304 | -CH ₂ - | -SO ₂ NH- | -OSO ₂ CH ₂ - | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | | 28 ± 5 (3) |
| lm1309 | -CH ₂ - | -CONH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ NH ₃ | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | | 28 ± 28 (3) |
| lm1313 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | | | 36 ± 12 (3) |
| lm1338 | -CH ₂ - | -SO ₂ NH- | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | <i>S</i> | 46 ± 15 (7) |
| lm1339 | -CH ₂ - | -SO ₂ CH ₂ - | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | -CH ₂ C ₆ H ₅ | | 79 ± 1 (3) |
| lm1340 | -CH ₂ - | -SO ₂ CH ₂ - | -OPO ₂ O- | -(CH ₂) ₃ CH ₃ | -C ₆ H ₁₂ | -CH ₂ CH ₂ O- | -(CH ₂) ₃ CH ₃ | | | | 29 ± 3 (3) |

^a See Figure 1 for a schematic diagram of the molecules' structures. The GLI fragment corresponds to the glycerol backbone. Its chirality is specified as follows: *R* indicates that both experiments and modeling were performed with the *R* structure; likewise for *S*. For the remaining compounds, a racemic mixture was used in the experiments, but the *R* form was modeled as the most potent chirally resolved compound had *R* stereochemistry. Enzyme activities were measured as described in ref 13b with enzyme isolated from human synovial fluid and the natural substrate, phosphatidylethanolamine so that the experimental conditions were as relevant to human proinflammatory situations as possible. Inhibitor activities (taken from ref 13) are expressed as percent enzyme inhibition (with standard deviation) at 0.01 mol fraction of inhibitor in the substrate vesicles. The number of activity measurements for each compound is shown in parentheses. XI(50) data available for 10 of these compounds (unpublished) show similar trends to the percent inhibition data with a limited linear correlation ($R = 0.86$). This provides support for the validity of correlating percent inhibition data and receptor binding affinity.

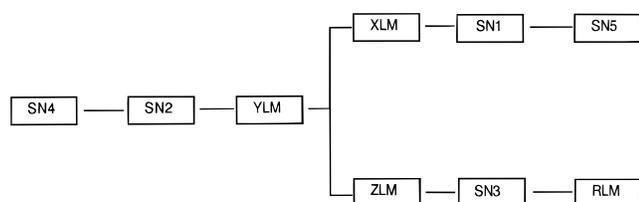


Figure 1. Schematic diagram of the HSF-PLA₂ inhibitors, showing the fragments into which they were divided for the analysis (see also Table 1). The glycerol backbone, corresponding to fragment GLI, is not labeled for clarity.

in the box do not carry any significant information, and therefore the PLA₂ inhibitor set represents a challenging example for any variable selection method.

Methods

COMBINE Analysis. 1. Outline of the Method. A more detailed account of the COMBINE analysis method can be found in ref 4. A summary of the procedure is given here for completeness. The process can be considered as three sequential steps. In the first step, the total binding energy of each ligand is calculated using a molecular mechanics force field. The binding energy is given by the sum of the following terms, expressed on a "residue" basis:

$$\Delta U = \sum_{i=1}^{nl} \sum_{j=1}^{nr} u_{ij}^{dw} + \sum_{i=1}^{nl} \sum_{j=1}^{nr} u_{ij}^{ele} + \sum_{i=1}^{nl} \Delta u_i^{B,L} + \sum_{i=1}^{nl} \Delta u_i^{A,L} + \sum_{i=1}^{nl} \Delta u_i^{T,L} + \sum_{i < j} \Delta u_{ij}^{NB,L} + \sum_{j=1}^{nr} \Delta u_j^{B,R} + \sum_{j=1}^{nr} \Delta u_j^{A,R} + \sum_{j=1}^{nr} \Delta u_j^{T,R} + \sum_{j < i} \Delta u_{ij}^{NB,R} \quad (3)$$

The first two terms on the right hand side of eq 3 describe the intermolecular interaction energies between each residue

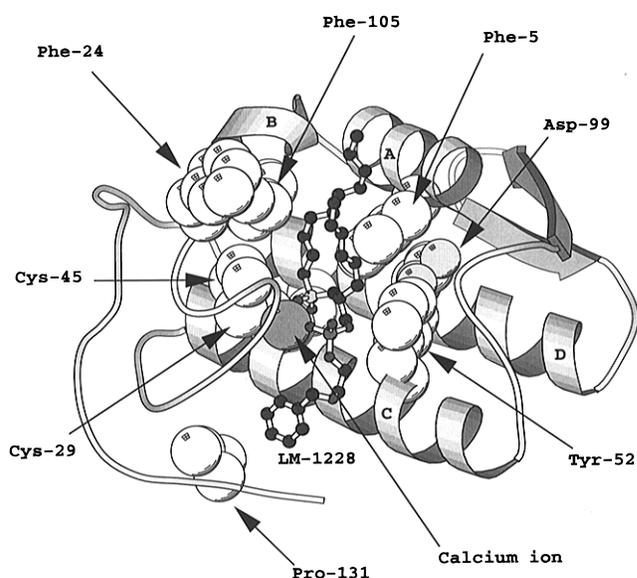


Figure 2. Schematic diagram of HSF-PLA₂ complexed with a representative inhibitor (lm1228). White spheres represent atoms of protein residues lining the binding site that are frequently selected to contribute to regression models in COMBINE analysis (see ref 4 and Table 3 therein). The calcium ion in the active site (shaded sphere) also makes an important contribution to COMBINE models. Other residues whose interactions are selected in COMBINE models are mostly located further from the binding site or are only selected occasionally when multiple models are generated. This diagram was generated with the Molscript program.²⁹

i of the ligand and each residue *j* of the receptor. The next four terms describe changes in the bonded (bond, bond angle, and torsion) and the nonbonded (Lennard–Jones and electrostatic) energies of the ligand upon binding to the receptor, and

the last four terms account for changes in the bonded and nonbonded energies within the receptor upon ligand binding.

In the second step, a matrix is constructed whose columns correspond to each of the energetic terms on the right-hand side of eq 3 and whose rows represent the different compounds studied. A final column containing the inhibitory activities of the compounds is added to the matrix. In the third step, the PLS method⁵ is applied along with a variable selection procedure. Variables of little importance to activity are thus excluded from the QSAR. The final equation for correlating the binding free energy (and hence, inhibitory activity) with a subset of the interactions in the ligand–receptor complex is

$$\Delta G = \sum_{i=1}^n w_i \Delta u_i^{\text{rep}} + C \quad (4)$$

where Δu_i^{rep} are the changes on ligand binding in the energy terms selected using the statistical analysis, w_i are the weights applied to these energy terms, and the constant C accounts for some of the systematic contributions to the binding affinity and statistical errors in the process of fitting eq 4.

2. Model Building and Statistical Analysis. Computer models of all of the complexes between the inhibitors listed in Table 1 and the HSF-PLA₂ were generated using the 3D structure of HSF-PLA₂ cocrystallized with a transition state analogue (LM1220, Table 1) solved at 2.1 Å resolution¹⁵ as described in refs 4 and 16. To obtain conformations for the free inhibitors, each compound was positioned in the nearest local energy minimum conformation to the minimum found in the complex. A detailed account of the modeling procedures can be found in our previous work.^{4,16}

The matrix for the statistical analysis was generated from the energy terms in eq 3 calculated for each inhibitor and was analyzed with the GOLPE 3.0 program.^{7b} There were 26 rows, one for each of the compounds studied. The original matrix had 3310 columns of X -variables corresponding to the energy terms described in eq 3 (constituting the “ X -matrix”) and one Y -variable column corresponding to experimental activity. The X -matrix was then subjected to various pretreatment protocols (see next section) before building PLS models. PLS models were derived with variable selection performed using the GOLPE method,⁷ as described previously.⁴ The final models were tested by cross-validation using 5 random groups and 20 randomizations. This means that objects were assigned to one of five different groups in a random way. Four of these groups were used to build a PLS model, which was then used to predict the activity of the objects of the remaining group. The procedure was repeated five times, each time removing a different group. The whole computation, including the random assignment of the objects to the groups, was repeated 20 times.

CoMFA Studies. A number of CoMFA³ studies using the same data set were carried out. Inhibitors were aligned by superimposing the C α atoms of the protein in all of the modeled complexes, using the LM1220–HSF-PLA₂ complex as a template. The superimposed set of inhibitors is shown in Figure 1 of the Supporting Information. Calculation of steric and electrostatic fields around the aligned set of molecules was performed with the GRID program, version 11,¹⁷ using methyl and proton probes, respectively. The same set of partial atomic charges was used as in the COMBINE analysis and was derived by electrostatic potential fit from semiempirical MNDO calculations. The van der Waals radii were, however, chosen according to the GRID parameter set. A 2 Å grid spacing and a box of 26 × 28 × 22 points were used. Details of the relevant parameters in the matrix generation procedure can be found in Table 1 of the Supporting Information. Three different scaling procedures (variance scaling) were carried out prior to the development of the regression models. Four different dielectric schemes (dielectric scaling) were used to modify the electrostatic interactions. Two different variable selection methods were used. Thus, a total of 3 × 4 × 2 = 24 regression models were derived from the full data set. The scaling procedures are as follows (each scaling and selection procedure will be identified in the rest of the paper by the one-letter code given):

1. Variance Scaling. A. No Scaling (N). The X -matrix, with the energy values as calculated above, was used directly as input to the GOLPE 3.0 program.^{7b} No further modifications were performed in order to derive the regression model. With this procedure, energy variables with higher variance have greater initial weights in the PLS analysis.

B. Autoscaling (A). Each element x_{ij} of the X -matrix, corresponding to an energy term, was autoscaled as follows:

$$x_{ij} = \frac{x_{ij} - \langle x_j \rangle}{\sigma_j} \quad (5)$$

where $\langle x_j \rangle$ is the average value of the variable j in the X -matrix and σ_j is the standard deviation. This procedure assigns all variables unit variance and thus the same initial weight and opportunity to influence the PLS results.

C. Block Scaling (B). The X -matrix is divided into k blocks of related data. In this case, different blocks contain data derived with different probes in the GRID calculations. Each element $x_{ij,k}$ of a block is scaled as follows:

$$x_{ij,k} = \frac{x_{ij,k} \sigma_k}{\sigma_x} \quad (6)$$

where σ_x is the total standard deviation of the X -matrix and σ_k is the standard deviation of the block under consideration. Within each block, the variables are unscaled, and the larger the variance of a descriptor is, the greater its importance in the model. However, the total sum of squares is preserved, and each block has the same importance.

2. Dielectric Scaling. A. Constant Dielectric Model (C). A relative dielectric constant of $\epsilon = 4$ was assigned to the medium surrounding the compounds studied as is common when using the GRID energy function for CoMFA analysis.¹⁷ This value is often assigned to the dielectric constant in the interior of a protein.¹⁸

B. Distance Dependent Dielectric Model (R). A distance dependent relative dielectric constant $\epsilon = r_{ij}$ was also tested. The rationale for using this dielectric model in force field calculations is to implicitly mimic molecular polarization at short separations and the screening effect of charges that are not explicitly modeled at longer separations. This dielectric model is standard in CoMFA calculations using the SYBYL implementation.¹⁹

C. Warshel Dielectric Model (W). This dielectric model was derived by Warshel²⁰ using experimental information about the free energy of electrostatic interactions between charged groups in proteins. Because the aim of CoMFA is to find a significant correlation between molecule–probe interactions and a term related to the binding free energy, this dielectric model may be appropriate for application to the electrostatic field in CoMFA. It has the following dependence on the distance r between the interacting charges:

$$\epsilon(r) = 1 + 60(1 - e^{-0.1r}) \quad (7)$$

D. Hingerty Dielectric Model (H). This dielectric model has been proposed by Hingerty et al.²¹ It is based on a modification of Debye's distance dependent dielectric function, corrected to short distances using the method of image charges for a charge in a cavity immersed in water. This dielectric model has been found to be adequate for modeling electrostatic base-stacking interactions in DNA as it provides energy values that are comparable to those obtained using the finite difference Poisson–Boltzmann equation.²² The following analytical expression is used in order to calculate the macroscopic relative dielectric constant:

$$\epsilon(r) = 78 - 77 \left(\frac{r}{2.5} \right)^2 \frac{e^{r/2.5}}{(e^{r/2.5} - 1)} \quad (8)$$

These models result in very different distance dependencies for electrostatic interaction energies (see Figures 2 and 3 in Supporting Information). The $\epsilon = r_{ij}$ model tends to increase the weight, with respect to the constant dielectric model, of the strongest electrostatic interactions which occur when the

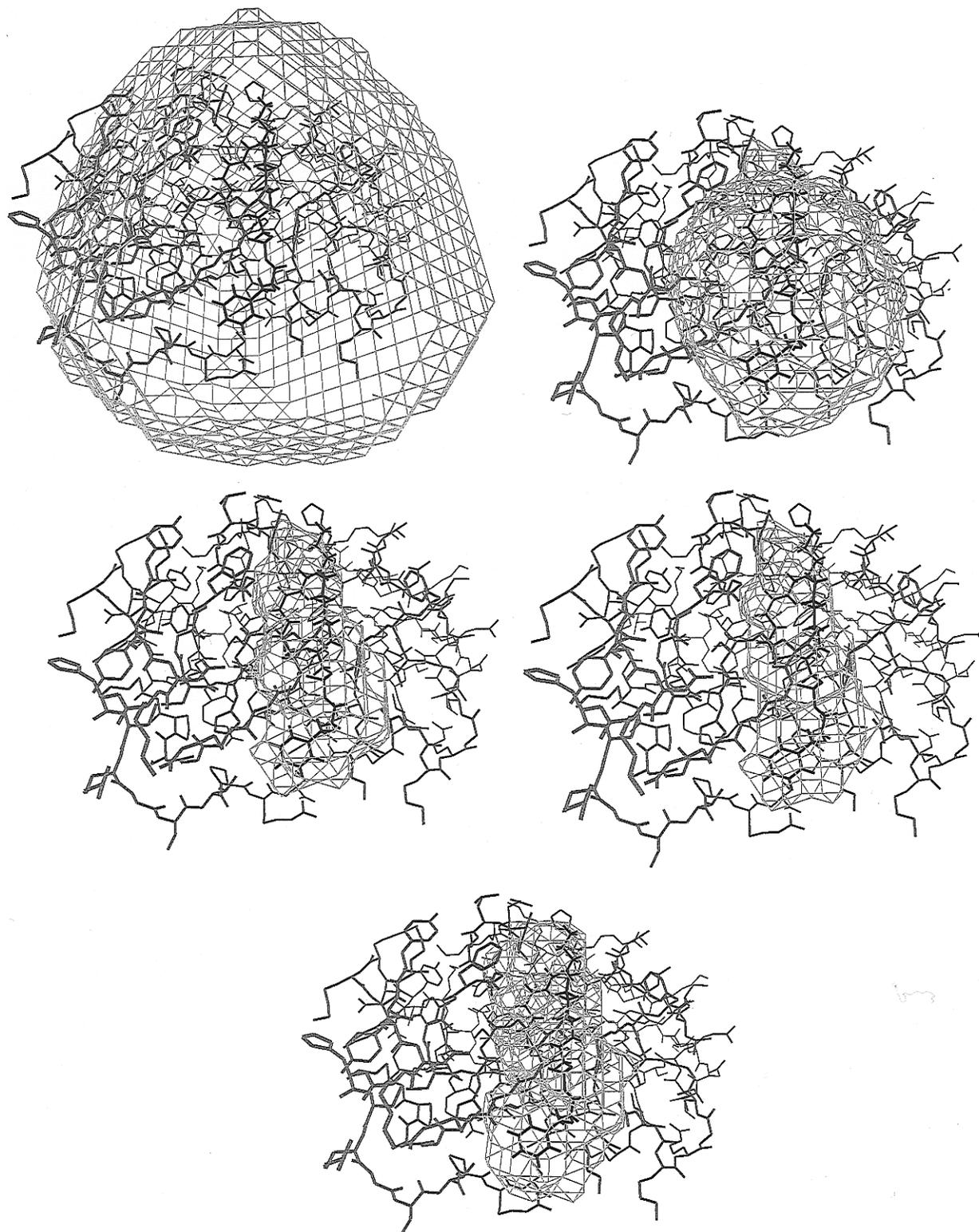


Figure 3. Contour plots of standard deviations in the interaction energy fields for the electrostatic energy with the different dielectric models (top and middle rows) and the steric Lennard-Jones energy (bottom row). Contours are at $2\sigma_{ep}$ and are shown superimposed on the structure of HSF-PLA₂ with inhibitor lm1228 modeled into the active site: (top left) $\epsilon = 4$, (top right) $\epsilon = r_{ij}$, (middle left) Warshel model, and (middle right) Hingerty model.

probe is near the van der Waals surface of the molecules. The Warshel model, on the other hand, strongly damps interactions at all distances. The Hingerty model behaves in a similar way to the constant dielectric model in the neighborhood of the molecules but damps interactions strongly at greater distances, thus tending to suppress long range correlations between the molecular electrostatic interaction fields. The effect of the different models on the standard deviation of the electrostatic potential (σ_{ep}) and its relationship to the steric potential is shown in Figure 3. In this figure, the electrostatic potential

is contoured at $2\sigma_{ep}$. The dielectric model with $\epsilon = 4$ shows, as expected, the highest standard deviation. Virtually all the points in the grid box have values $> 2\sigma_{ep}$. In the case of $\epsilon = r_{ij}$, a sphere of 10 Å radius around the inhibitor's phosphate group encloses the $2\sigma_{ep}$ limit. Both Warshel and Hingerty models behave in a similar way, with strong damping such that the $2\sigma_{ep}$ contour essentially encloses the molecular shape. The linear distance dependent model and the constant dielectric model have similar block weights (see Table 2 of Supporting Information), with electrostatic terms tending to dominate

the variance of the **X**-matrix with weights of about 1.5 compared to 0.8 for the steric terms. In contrast, in both the Warshel and Hingerty models, the electrostatic and steric fields show essentially the same variance (with respective weights of 0.9 and 1.1). Thus, the overall picture of the four matrices is that the distance dependent model tends to focus on similarities between deep minima in the electrostatic potential; the Warshel and Hingerty models focus on those regions near the surface of the molecules, although the Warshel model damps the electrostatic potential in a more homogeneous way, and in the constant dielectric model, the electrostatic potential tends to dominate and long range correlations in the electrostatic potential are important.

3. Variable Selection. A. GOLPE Method (G). This is the variable selection procedure usually applied within the GOLPE 3.0 program.^{7b} It relies on the validation of a number of reduced models on variable combinations selected according to a FFD strategy.²³ A minimum σ (standard deviation) cutoff of 0.3 was initially applied to all matrices to remove nearly constant variables (trial calculations indicated models were essentially unaffected by reducing the cutoff further below this value). When required by the excessive number of variables, a D-optimal design preselection²⁴ was first carried out in order to filter out redundant variables and retain those variables that tend to correlate with the biological activity while providing the most independent information. For the selection, the partial weights space of a preliminary PLS model was used, and then the D-optimal preselection was made stepwise, by removing 10% of the variables at each step unless this led to overfitting. Overfit was checked by monitoring predictive performance in cross-validation and stopping variable reduction if it had a deleterious effect on the Q^2 value. When this happened or the number of variables dropped below 1500, a FFD variable selection was performed. In these designs, dummy variables were introduced with the ratio of true variables to dummies set to 3:1 and a design combinations to variables ratio of 2:1. The predictive ability of the generated matrices was evaluated by cross-validation, using 5 random groups and 20 randomizations and allowing up to five latent variables. Weights were recalculated after object exclusion. Variables which were determined to be noise were excluded, and variables which were found to be uncertain were retained, using the iterative "fixing and exclusion" method.⁷ This procedure was typically repeated two or three times, and in a final step, uncertain variables were eliminated. The final model was tested by cross-validation using 5 random groups and 20 randomizations.

B. Q^2 -Guided Region Selection (T).⁸ In this method, the initial box was divided into 125 small boxes, and for each box, a PLS analysis was carried out. Only those boxes with a Q^2 value higher than a predefined cutoff were used in the derivation of the final PLS model. The model dimensionality within each of the boxes was determined according to cross-validated predictive ability. The region selection was carried out as implemented in the GOLPE 3.0 program.^{7b} This implementation includes some slight modifications with respect to the original procedure described in ref 8: for each of the small boxes, a step size of 2 instead of 1 Å was used. Moreover, cross-validation within each of the boxes and in the derivation of the final PLS model was carried out using 5 random groups and 20 randomizations rather than leave-one-out cross-validation. Trial calculations indicated that a Q^2 cutoff of 0.2 produced optimal results. In this approach, the number of latent variables chosen for the final model was that yielding the best cross-validated predictive ability.

CoMFA models were first generated for the full 26 compound data set and evaluated by internal cross-validation (as described in the previous two paragraphs). They were then generated for training sets with a reduced number of compounds in order to assess predictive ability on external test sets. Three different test sets were used, and these are described in the next section.

Results and Discussion

Effect of Different Parameters on the Predictive Ability of CoMFA Models. 1. Effect of the Vari-

Table 2. Predictive Performance for the Different **X**-Matrices with Different Variable Selection Procedures^a

| method ^b | LV ^c | SDEC (%) | R^2 | SDEP (%) | Q^2 | X_{sel}^d |
|---------------------|-----------------|----------|-------|----------|-------|-------------|
| N-G-C | 3 | 5.44 | 0.94 | 10.93 | 0.754 | 72 |
| A-G-C | 1 | 11.49 | 0.72 | 13.30 | 0.636 | 56 |
| B-G-C | 3 | 6.22 | 0.92 | 10.92 | 0.754 | 41 |
| N-T-C | 4 | 7.37 | 0.88 | 13.53 | 0.623 | 152 |
| A-T-C | 2 | 15.09 | 0.53 | 21.47 | 0.052 | 499 |
| B-T-C | 3 | 9.23 | 0.82 | 17.03 | 0.404 | 217 |
| N-G-R | 3 | 8.12 | 0.86 | 12.43 | 0.682 | 91 |
| A-G-R | 2 | 9.48 | 0.81 | 13.74 | 0.611 | 130 |
| B-G-R | 3 | 6.68 | 0.90 | 11.24 | 0.740 | 151 |
| N-T-R | 4 | 7.53 | 0.88 | 13.82 | 0.607 | 222 |
| A-T-R | 3 | 12.40 | 0.68 | 19.17 | 0.244 | 118 |
| B-T-R | 3 | 9.96 | 0.79 | 16.36 | 0.449 | 174 |
| N-G-W | 2 | 7.94 | 0.87 | 11.61 | 0.722 | 51 |
| A-G-W | 1 | 11.27 | 0.73 | 20.53 | 0.133 | 134 |
| B-G-W | 2 | 8.02 | 0.86 | 11.49 | 0.728 | 64 |
| N-T-W | 4 | 7.62 | 0.88 | 14.82 | 0.548 | 232 |
| A-T-W | 3 | 12.27 | 0.69 | 18.32 | 0.310 | 156 |
| B-T-W | 3 | 10.86 | 0.75 | 18.82 | 0.272 | 425 |
| N-G-H | 2 | 7.78 | 0.87 | 11.47 | 0.729 | 55 |
| A-G-H | 2 | 8.69 | 0.84 | 11.57 | 0.724 | 95 |
| B-G-H | 2 | 7.97 | 0.86 | 11.83 | 0.712 | 84 |
| N-T-H | 3 | 12.33 | 0.68 | 18.81 | 0.273 | 274 |
| A-T-H | 3 | 12.47 | 0.68 | 18.80 | 0.273 | 118 |
| B-T-H | 3 | 12.46 | 0.68 | 18.91 | 0.265 | 274 |

^a A minimum standard deviation σ cutoff of 0.05 was used to filter out variables of very low variance before statistical analysis.

^b Variable scaling procedure: N, no scaling; A, autoscaling; B, block-scaling. Variable selection method: T, Q^2 -GRS; G, GOLPE. Dielectric model: C, constant ($\epsilon = 4$); R, distance dependent; W, Warshel model; H, Hingerty model. ^c Number of latent variables.

^d Number of selected *X*-variables in the model.

able Selection Procedure on the Predictive Ability. The predictive ability of the regression models obtained by performing a PLS analysis without variable selection on the matrices obtained with different dielectric models and variable scaling procedures is detailed in Table 3 of the Supporting Information. There is no significant predictive ability in any of these matrices as $Q^2 < 0.12$ and SDEP $> 20.6\%$ in all cases. As will be shown, this is due to the very low signal-to-noise ratio in these large matrices.

The predictive ability of the regression models generated after applying variable selection to the different matrices is given in Table 2. The highest Q^2 value (0.754) is obtained with the N-G-C and B-G-C matrices. In general, variable selection with the GOLPE method produces a model with a high Q^2 value (about 0.7). On the other hand, when variable selection is carried out with the Q^2 -GRS method, results are highly dependent on matrix pretreatment and dielectric modeling. Only in three cases does the Q^2 -GRS method produce a predictive model with a Q^2 value above 0.5: with the N-T-C, N-T-R, and N-T-W matrices. For equivalent matrices, the Q^2 value obtained with the Q^2 -GRS variable selection procedure is lower than that obtained with the GOLPE method.

2. Effect of the Energy Cutoff on the Predictive Ability. In order to consider the impact of the maximum energy cutoff on the results, two additional models were derived using an energy cutoff of 5 kcal/mol in the GRID computations. Only Q^2 -GRS variable selection was carried out, with a constant dielectric screening factor of 4.0 for the electrostatic interactions. No-scaling and block-scaling pretreatments were tested. In contrast to the models given in Table 2 for a 30 kcal/mol cutoff, none of the resulting models had significant

predictive ability, and the maximum Q^2 value was 0.26 (see Table 4 of the Supporting Information).

3. Effect of the Grid Spacing on the Predictive Ability. The original Q^2 -GRS method⁸ differs from our implementation in one important aspect: while Cho and Tropsha used two different grid spacings (2 Å for the general regression model and 1 Å for each of the small boxes), we used a 2 Å grid spacing for all stages of the calculation. In order to check the impact of this difference on predictive performance, a comparison of both implementations was performed using the N-T-C matrix. Q^2 values better by about 0.2 are obtained for this matrix when a 2 Å grid spacing is used for the whole procedure (see Table 5 of Supporting Information). This is probably the result of an increased signal-to-noise ratio for the 2 Å spacing grids. In both cases, however, similar boxes are selected for the final model.

4. Effect of Dielectric and Variance Scaling on the Predictive Ability. For both the GOLPE and Q^2 -GRS variable selection methods, the model with the highest predictive ability is found with the unscaled matrices using a relative dielectric constant of $\epsilon = 4$ (see Table 2). With the Q^2 -GRS variable selection method, the ranking of predictivities for the different dielectric models is $C > R > W > H$. This is inversely correlated with the strength of electrostatic damping at distances above 6 Å, which underscores the long range character of electrostatic interactions and suggests that long range correlations of the electrostatic potential are important in structure–activity relationships. Indeed, only the $\epsilon = 4$ and $\epsilon = r_{ij}$ dielectric models produce regressions with good predictive ability when the Q^2 -GRS method is used.

Autoscaling has an undesirable effect on predictive ability in almost all cases. This is consistent with the results of Cruciani and Watson⁸ who analyzed a set of inhibitors of glycogen phosphorylase with the GOLPE variable selection method. Autoscaling overweights variables that have a small influence on predictive ability and do not reflect real structural variations. On the other hand, block-scaling generally improves the predictive ability only slightly when used with the GOLPE method but has a negative effect when used with the Q^2 -GRS method. The fields selected for the best models (i.e., for the N-T-C and N-G-C matrices) are displayed in Figure 4.

5. Analysis with External Validation. So far, we have quantified the predictive quality of the model on the basis of internal validation (cross-validation) only. Even when the internal cross-validation method utilized (random groups cross-validation) is very conservative, it can fail to detect overfitting in the models and thus give overoptimistic indicators of predictive performance. In order to compare the two methods of variable selection, it is of critical importance to confirm that the SDEP and Q^2 parameters derived from internal validation actually reflect the real predictive ability of the models and are not influenced by overfitting. We have done this by external validation.

To perform external validation, the initial data set of 26 inhibitors was divided into two groups: a training set of 20 inhibitors and a test set of 6 inhibitors. The molecules in this test set (test set 1) were selected so that they cover the different functional groups present in the inhibitor set and a wide range of activities. The six molecules in test set 1 were lm1293, lm1298, lm1300,

Table 3. Predictive Performance of Models Derived for Different X-Matrices Using the Reduced Training Set of 20 Inhibitors Described in Methods (Test set 1)^a

| method | LV | SDEC (%) | R^2 | SDEP (%) | Q^2 | X_{sel} |
|--------|----|----------|-------|----------|-------|-----------|
| N-G-C | 2 | 10.02 | 0.78 | 14.42 | 0.557 | 51 |
| N-T-C | 3 | 8.26 | 0.85 | 15.23 | 0.506 | 108 |
| N-G-R | 1 | 11.04 | 0.74 | 18.12 | 0.301 | 104 |
| N-T-R | 1 | 15.27 | 0.50 | 19.88 | 0.159 | 310 |
| N-G-W | 2 | 8.79 | 0.83 | 12.70 | 0.656 | 88 |
| N-T-W | 2 | 13.56 | 0.60 | 20.49 | 0.106 | 155 |
| N-G-H | 2 | 9.84 | 0.79 | 13.97 | 0.584 | 68 |
| N-T-H | 5 | 8.67 | 0.83 | 18.63 | 0.261 | 177 |

^a Key as for Table 2.

lm1309, lm1338, and lm1339. New regression models were derived. This was only done for unscaled matrices as the internal validation results showed that they resulted in models with better predictive ability for both variable selection methods than autoscaled or block-scaled matrices. The same four dielectric models were used, and both methods of variable selection were applied, resulting in the $4 \times 2 = 8$ regression models whose predictive ability is summarized in Tables 3 and 4.

As there is a smaller number of inhibitors in the training set, the Q^2 values for internal cross-validation are expected to be smaller than before (see Table 3). With Q^2 -GRS variable selection, internal validation indicates significant predictive ability only for the N-T-C matrix. On the other hand, all the models obtained after GOLPE variable selection have greater predictive ability, and the correlation can be considered significant in at least three cases. It is also remarkable that the Q^2 -GRS method suggests models of different complexity (between one and five latent variables), while the GOLPE method suggests models with two latent variables in all cases except those with the poorest predictive ability.

The external validation results are summarized in Table 4. It appears that, while the internal SDEP always improves (decreases) with variable selection, the same is not always true for the external SDEP. Moreover, there is better agreement between the internal and external SDEP values after variable selection for the models obtained with the Q^2 -GRS method than for those obtained with the GOLPE method.

With both variable selection methods, the best external predictions are obtained when a relative dielectric constant of 4 is used. Figure 5 shows a comparison of the experimental activities of the test set inhibitors with their activities predicted using the N-T-C and N-G-C methodologies. The model obtained after GOLPE variable selection seems to result in a tendency toward predicting the average of the observed activity values. The predictions of the less active compounds are too high, while those of the more active compounds are too low. Predictions carried out with a model derived using COMBINE analysis with the GOLPE variable selection method show a similar tendency, suggesting that the variable selection method is probably at the origin of this problem. On the other hand, the effect of the dielectric treatment shows a similar tendency for both internal and external validation to that found for the models derived from the full data set. The only noticeable difference is the poor performance of the $\epsilon = r_{ij}$ dielectric model compared to its performance with the complete data-set. While the $\epsilon = 4$ model appears

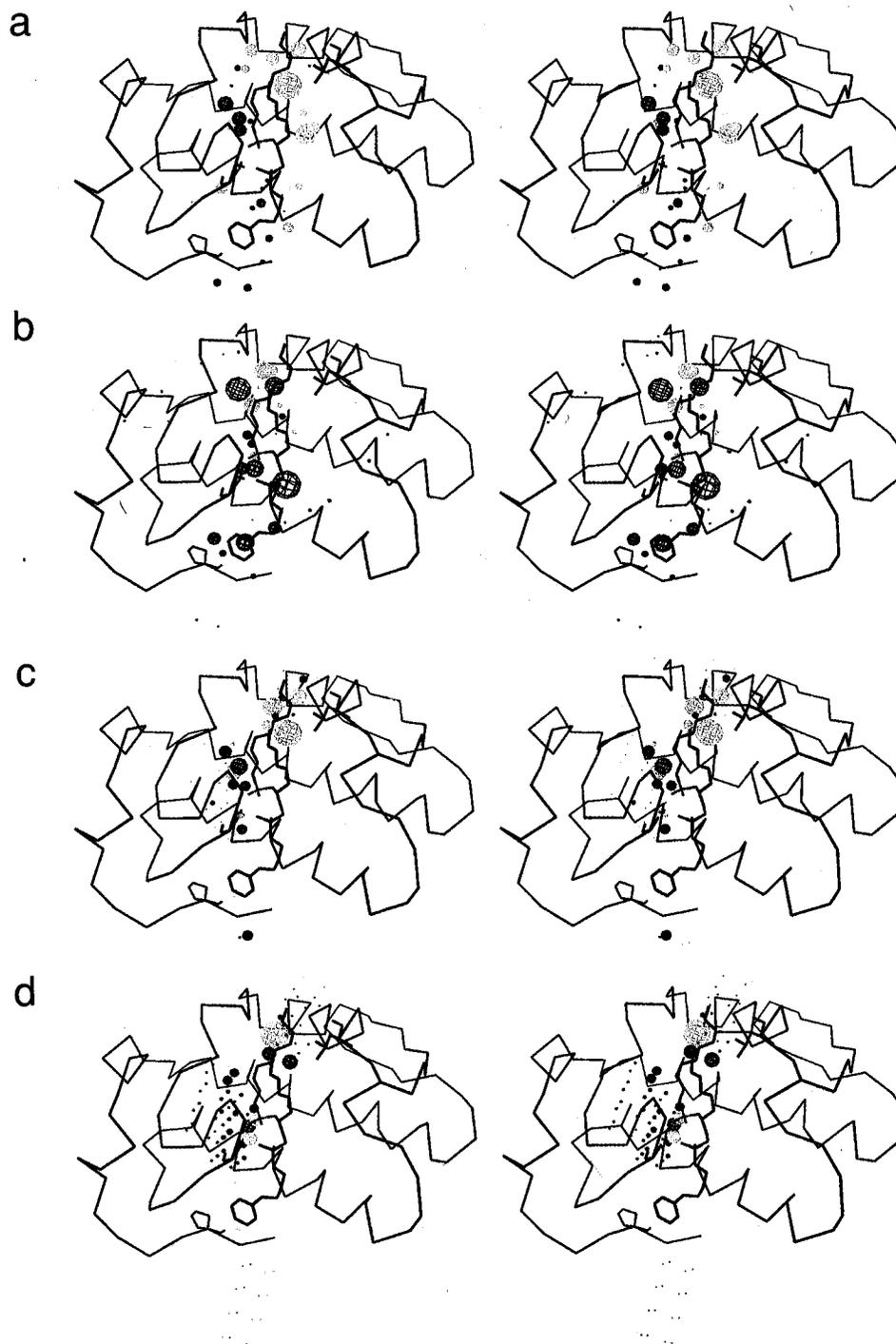


Figure 4. PLS coefficient values at the optimal dimensionality (see Table 2) for the different regression models. Coefficient values are represented by the size and color (yellow, positive; blue, negative) of the spheres superimposed on the structure of HSF-PLA₂ with inhibitor Im1228 modeled into the active site: (a) N-G-C, steric; (b) N-G-C, electrostatic; (c) N-T-C, steric; and (d) N-T-C, electrostatic. The Ca trace of the protein is shown together with the active site calcium ion (represented by a cross) and all non-hydrogen atoms of residues mentioned in the Results and Discussion section.

robust with respect to changes in the training set, the $\epsilon = r_{ij}$ model appears unstable.

These results are based on only one test set. To check for data set dependency in the conclusions, we carried out calculations for two additional test sets (test sets 2 and 3). These two test sets each contain four molecules which were selected using a principal component analysis (PCA) of the original \mathbf{X} -matrix, to fulfill two criteria: (1) they should represent the main clusters of objects present in the score plot, and (2) they should have neighboring objects, which can maintain the structure of the data when they are removed from the training set. An improved pretreatment of the \mathbf{X} -matrix

was also carried out: X -variables that were clustered around two, three, or four values, i.e., that had a clearly discontinuous rather than continuous distribution of values, were removed from the analysis, and the minimum σ cutoff was fixed at a more conservative value of 0.01. In addition, uncertain variables were not removed during variable selection. The results are presented in Table 5.

In general, the new pretreatment and the inclusion of uncertain variables in the final regression model produces slightly better results, but the same trends are observed for the three test sets. The Q^2 -GRS variable selection method gives external SDEP values which are

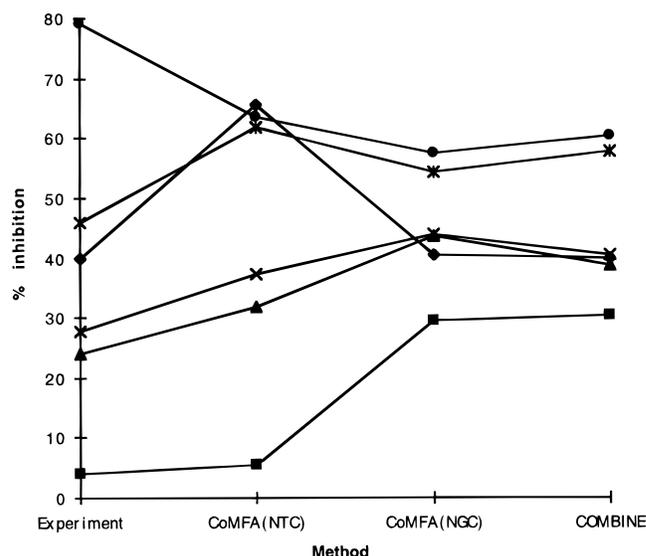
Table 4. Predictive Performance for Models Derived with a Reduced Training Set (as in Table 3) Using Different X-Matrix Pretreatments and Variable Selection Procedures^a

| method | LV | SDEP _e ⁽ⁱ⁾ | SDEP _e ^(vs) | SDEP _i ⁽ⁱ⁾ | SDEP _i ^(vs) |
|--------|----|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
| N-G-C | 1 | 17.31 | 19.44 | 22.96 | 15.92 |
| | 2 | 16.55 | 17.51 | 23.91 | 14.42 |
| | 3 | 16.31 | 16.69 | 25.47 | 15.38 |
| | 4 | 16.51 | 19.07 | 25.17 | 17.30 |
| | 5 | 16.17 | 20.30 | 25.20 | 17.70 |
| N-T-C | 1 | 17.31 | 15.85 | 22.96 | 20.90 |
| | 2 | 16.55 | 14.41 | 23.91 | 17.18 |
| | 3 | 16.31 | 14.68 | 25.47 | 15.23 |
| | 4 | 16.51 | 14.77 | 25.17 | 15.48 |
| | 5 | 16.17 | 15.05 | 25.20 | 15.95 |
| N-G-R | 1 | 18.64 | 20.79 | 21.37 | 18.12 |
| | 2 | 21.38 | 19.66 | 22.94 | 18.34 |
| | 3 | 23.08 | 20.54 | 23.52 | 19.10 |
| | 4 | 21.69 | 20.42 | 23.46 | 19.57 |
| | 5 | 23.26 | 22.39 | 23.46 | 19.49 |
| N-T-R | 1 | 18.64 | 21.46 | 21.37 | 20.13 |
| | 2 | 21.38 | 22.68 | 22.94 | 20.26 |
| | 3 | 23.08 | 24.14 | 23.52 | 20.74 |
| | 4 | 21.69 | 25.08 | 23.46 | 19.06 |
| | 5 | 23.26 | 25.01 | 23.46 | 18.63 |
| N-G-W | 1 | 23.74 | 21.81 | 23.72 | 14.85 |
| | 2 | 23.75 | 20.33 | 23.72 | 12.70 |
| | 3 | 24.57 | 19.91 | 24.54 | 12.78 |
| | 4 | 24.28 | 21.21 | 24.26 | 14.09 |
| | 5 | 24.01 | 20.62 | 24.02 | 14.83 |
| N-T-W | 1 | 23.74 | 21.15 | 23.72 | 23.35 |
| | 2 | 23.75 | 18.66 | 23.72 | 20.49 |
| | 3 | 24.57 | 16.27 | 24.54 | 21.06 |
| | 4 | 24.28 | 16.51 | 24.26 | 22.84 |
| | 5 | 24.01 | 16.92 | 24.02 | 27.30 |
| N-G-H | 1 | 22.12 | 23.09 | 22.67 | 16.28 |
| | 2 | 20.65 | 21.41 | 23.22 | 13.97 |
| | 3 | 21.37 | 23.02 | 24.08 | 14.05 |
| | 4 | 21.96 | 25.62 | 23.56 | 14.08 |
| | 5 | 22.25 | 26.97 | 23.66 | 13.95 |
| N-T-H | 1 | 22.12 | 23.65 | 22.67 | 19.88 |
| | 2 | 20.65 | 24.63 | 23.22 | 21.12 |
| | 3 | 21.37 | 24.45 | 24.08 | 23.38 |
| | 4 | 21.96 | 23.88 | 23.56 | 24.69 |
| | 5 | 22.25 | 24.44 | 23.66 | 25.32 |

^a SDEP values are given for external test set 1 (six molecules) and from internal cross-validation calculations. Key as in Table 2. SDEP values are percentages as follows: SDEP_e⁽ⁱ⁾, external SDEP values obtained using the test set before variable selection; SDEP_e^(vs), external SDEP values obtained using the test set after variable selection; SDEP_i⁽ⁱ⁾, internal SDEP values obtained using the training set before variable selection; SDEP_i^(vs): Internal SDEP values obtained using the training set after variable selection.

in better agreement with the internal ones in the three test sets than the GOLPE method. In some cases, it even gives external SDEP values that are smaller than the corresponding internal SDEPs. In contrast, the external SDEP values from GOLPE variable selection models are always higher than the internal ones.

The external SDEP values for the better models are also largely of the same magnitude as the experimental standard deviations (on average ~15%) for the percent inhibition data. This indicates that models as good as can be expected for this data set have been obtained and, also, that these models are not overfitted. It also suggests that models based on more accurate experimental data, e.g., Xi(50) potencies rather than activities, may result in much better SDEP values. Unfortunately, due to the difficult experimental assay conditions, there were insufficient Xi(50) values measured for this data set for their use in model derivation and testing (see legend to Table 1). Nevertheless, the more approximate percent inhibition data are probably more representative of the data typically available for QSAR derivation

**Figure 5.** Predicted and experimental activities for the six molecules in external test set 1: (◆) lm1293, (■) lm1298, (▲) lm1300, (×) lm1309, (*) lm1338, and (●) lm1339. For the predictions, the optimal model dimensionality as determined by internal cross-validation was used (see Tables 4 and 5).**Table 5.** Predictive Performance of Models Trained and Tested with Different Data Sets^a

| method | test set ^b | LV | R ² | Q ² | SDEP _i ^(vs) | SDEP _e ^(vs) | X _{sel} |
|--------|-----------------------|----|----------------|----------------|-----------------------------------|-----------------------------------|------------------|
| N-G | — | 3 | 0.86 | 0.60 | 13.85 | — | 822 |
| | 1 | 2 | 0.79 | 0.53 | 14.66 | 18.47 | 651 |
| | 2 | 3 | 0.81 | 0.52 | 14.34 | 18.97 | 838 |
| | 3 | 3 | 0.79 | 0.43 | 13.22 | 23.98 | 828 |
| N-T | — | 4 | 0.88 | 0.62 | 13.58 | — | 107 |
| | 1 | 3 | 0.82 | 0.47 | 15.48 | 13.73 | 105 |
| | 2 | 2 | 0.60 | 0.28 | 17.51 | 17.02 | 68 |
| B-G | — | 3 | 0.87 | 0.65 | 13.12 | — | 834 |
| | 1 | 3 | 0.85 | 0.55 | 14.31 | 19.75 | 922 |
| | 2 | 3 | 0.82 | 0.46 | 15.20 | 17.55 | 824 |
| B-T | — | 4 | 0.88 | 0.60 | 13.85 | — | 112 |
| | 1 | 2 | 0.79 | 0.48 | 15.42 | 14.62 | 132 |
| | 2 | 3 | 0.81 | 0.33 | 16.90 | 5.92 | 134 |
| B-T | — | 3 | 0.68 | 0.19 | 15.68 | 21.44 | 94 |

^a Models were derived by applying a minimum σ cutoff of 0.01 to generate the X-matrix and eliminating variables with values clustered in two, three, or four groups rather than adopting a continuous distribution. A constant dielectric of $\epsilon = 4$ was used. Key is as for Tables 2 and 4. ^b Test set definition: —, no test set (i.e., regression model using the full data set); 1, molecules lm1293, lm1298, lm1300, lm1309, lm1338, and lm1339 (test set 1 is the same as the test set for Table 3); 2, molecules lm1298, lm1166, lm1228, and lm1220; 3, molecules lm1258, lm1261, lm1339, and lm1240.

in the pharmaceutical industry. This implies that our results may be of greater significance and practical value than those that would be obtained with more accurate experimental data.

It is worth noting, however, that the Q²-GRS method for the block-scaled model with test set 2 produces a very low external SDEP of 5.92 (predicted and experimental values are nearly in perfect agreement for all molecules in the test set). As also apparent from the performance of the other models described in this work, the Q²-GRS method is quite unstable with respect to the robustness of internal predictive power to input parameters and produces quite different models for different data sets. Depending on the data set, the final model is based on between 60 and 140 X-variables from

Table 6. Effect of Elimination of Uncertain Variables in FFD Selection^a

| method | set ^b | LV | R ² | Q ² | SDEP _i (vs) | SDEP _e (vs) | X _{sel} |
|---------------------------------------|------------------|----|----------------|----------------|------------------------|------------------------|------------------|
| No Elimination of Uncertain Variables | | | | | | | |
| N-G | — | 3 | 0.86 | 0.61 | 13.85 | — | 822 |
| | 2 | 3 | 0.81 | 0.52 | 14.34 | 18.97 | 838 |
| | 3 | 3 | 0.79 | 0.43 | 13.22 | 23.98 | 828 |
| B-G | — | 3 | 0.87 | 0.64 | 13.12 | — | 822 |
| | 2 | 3 | 0.82 | 0.46 | 15.20 | 17.55 | 838 |
| | 3 | 3 | 0.80 | 0.48 | 12.54 | 21.76 | 828 |
| Elimination of Uncertain Variables | | | | | | | |
| N-G | — | 4 | 0.89 | 0.56 | 14.64 | — | 43 |
| | 2 | 2 | 0.79 | 0.55 | 13.87 | 20.90 | 42 |
| | 3 | 3 | 0.84 | 0.59 | 11.14 | 26.32 | 33 |
| B-G | — | 5 | 0.92 | 0.61 | 13.73 | — | 39 |
| | 2 | 3 | 0.84 | 0.46 | 15.17 | 24.64 | 31 |
| | 3 | 2 | 0.76 | 0.46 | 12.84 | 22.96 | 41 |

^a Parameters and key as in Table 5. ^b Test set definition: —, no test set (i.e., regression model using the full data set); 2, molecules lm1298, lm1166, lm1228, and lm1220; 3, molecules lm1258, lm1261, lm1339, and lm1240.

one to three selected sub-boxes and has between two and five latent variables. In contrast, the models derived with the GOLPE method have a similar number of selected *X*-variables (650–920) and latent variables in all the test sets and are thus more stable.

6. Effect of Retaining Uncertain Variables in GOLPE Variable Selection. As stated above, the GOLPE method evaluates the effect of individual variables on the predictive ability of the models. However, some variables might have no clear effect on the predictive ability. These uncertain variables can be either kept in the model or removed. In the original description of the GOLPE method, 7 it was suggested that uncertain variables should be eliminated, but the authors' experience has shown that, in the field of 3D-QSAR, it is better to retain uncertain variables in order to minimize the risk of overfitting. We have tested the impact of elimination of uncertain variables during the selection process, and the results are shown in Table 6. The external SDEP is greater after elimination of uncertain variables for the two external test sets even though the internal SDEP is slightly smaller. Thus, elimination of uncertain variables clearly has a negative effect on predictive ability and increases the risk of overfitting.

It is instructive to consider the evolution of the internal and external SDEP parameters as a function of the number of selected variables. The SDEP_i value drops throughout the variable selection procedure (see Figure 4 of Supporting Information). However the SDEP_e value only decreases by a small amount until elimination of uncertain variables is carried out. It then increases considerably, indicating that overfitting occurs.

7. Chemometric Analysis of the Models. In previous sections, the effect of some of the CoMFA parameters on internal and external indices of predictive ability has been described. However, prior to interpreting these findings to draw conclusions, the PLS models obtained should be analyzed from a chemometric point of view. For this analysis, we have used two CoMFA models, one obtained with GOLPE variable selection and the other obtained with Q²-GRS variable selection (the N-G-C and N-T-C models described in Table 5, respectively).

Figure 6 shows the partial weights plot for these two models. In the N-T-C model, a few *X*-variables show a

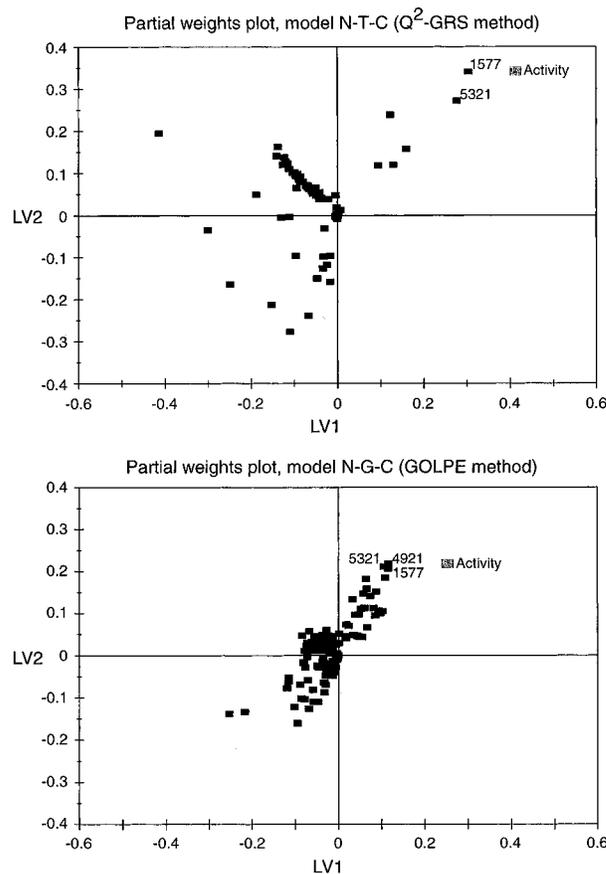


Figure 6. Partial weights plots for the *X*-variables and the activity in the first two latent variables (LV1 versus LV2) for the N-T-C and N-G-C models listed in Table 5 (see text for details).

surprisingly high correlation with the activity, while all the rest seem to correlate very little with activity. Variable 1577 shows the best correlation with activity (see Figure 5 of Supporting Information). While there is little correlation between the value of variable 1577 and the activity for compounds of intermediate activity, most of the less active molecules (lm1298, 4% inhibition; lm1192, 6% inhibition; and lm1240, 9% inhibition) have low interaction energies for variable 1577, and the molecules with the highest activity (lm1261, 80% inhibition; lm1339, 79% inhibition; and lm1228, 74% inhibition) have the maximum positive energy for variable 1577, corresponding to the maximum positive energy cutoff at 30 kcal/mol. Variable 1577 is a van der Waals interaction and is positioned at the grid point within the largest yellow sphere in Figure 4a,c. This grid point is within the van der Waals volume of some of the compounds in the series, and therefore the interaction energy between the methyl probe and these compounds has a large positive value which is truncated to the positive energy cutoff value (5 or 30 kcal/mol). The unusual behavior and high influence on the model of variable 1577 are due to the fact that so many of the molecules take the value of the maximum positive cutoff.

Obviously, variables like 1577 and 5321, which show a good correlation with the activity, will be included in the model by both methods of variable selection, but in the case of the Q²-GRS models, these variables take on a more important role, as is evident from the partial weights plots (Figure 6). This difference arises from the differences in the methods of variable selection: the Q²-

GRS method selects a box of variables only when this set of variables has a predictive ability better than a given cutoff. Thus, the method will fail to include variables in the final model which, even if important for explaining the activity, cannot produce a good correlation without the presence of variables from other boxes. As a consequence, the Q^2 -GRS models contain a smaller number of variables correlated with the activity than the GOLPE models. This effect appears clearly in the loading and partial weights plots, but it is not so easily recognized by only looking at the number of selected variables. It should also be remembered that many of the variables are included in the final model only because they belong to the same box as some important ones. On the other hand, the GOLPE models are much less dependent on the few highly correlated variables because they contain many other variables which also show significant correlations with activity.

The fact that the Q^2 -GRS models rely heavily on variable 1577 helps to explain the findings described in previous sections:

1. Reduction of the maximum positive energy cutoff from 30 to 5 kcal/mol will decrease the predictive ability of Q^2 -GRS models (Table 6) because this transformation will greatly reduce the correlation between activity and variable 1577. In addition, this change will decrease the weight of such variables and their influence in the PLS models.

2. Changes in X -variable scaling (with both statistical and dielectric schemes) are more likely to perturb regression models based on a small number of variables. This partially explains the instability of the Q^2 -GRS models (see Tables 4 and 6). In contrast, GOLPE models are more stable with respect to the number of variables selected and the dimensionality of the models.

3. The sensitivity of the external predictions can be understood from the correlation between activity and variable 1577. The external SDEP values of Q^2 -GRS models depend directly upon how the test set is selected. Test set 2 contains points that can be well described by variable 1577, and therefore, the external SDEP is even lower than the internal SDEP. On the other hand, test set 3 contains compound Im1258, which is an outlier as regards the correlation between variable 1577 and activity (with a maximum value for variable 1577 and a low activity). As a consequence, the external SDEP value for test set 3 is higher than the internal SDEP and higher than the SDEP obtained for test set 2.

This analysis raises the question of whether variable 1577 represents a "real" effect or whether the correlations only appear by chance. On one hand, as will be discussed later, this variable might represent the hydrophobic interaction of the *sn*-1 chain of the inhibitors with Leu-2 of the protein. On the other hand, as there are no other variables close to 1577 in latent variable space, there is no chemometric support for the significance of this variable, and the effect might be a chance occurrence.

Comparison of COMBINE and CoMFA Models.

One of the potential advantages of the CoMFA 3D-QSAR approach over conventional QSAR is the putative ability of CoMFA to provide results that can be interpreted in terms of the interaction energies involved in the binding process. As a consequence, it is important not only that a CoMFA model has the ability to forecast the biological activity of new molecules but also that the

resulting coefficient contour plots reflect the actual nature of the ligand–receptor interaction. Several authors have compared the fields selected using the CoMFA approach with crystallographically determined ligand–receptor geometries, evaluating the "match" between the selected fields and the position of residues in the ligand binding site graphically.^{9,10,25} However, all of these comparisons suffer from the fact that the magnitude of the contributions of the active site residues to the differences in binding affinity are generally unknown. Thus, the simple superimposition of fields and residues in some region of space is not sufficient to evaluate the receptor-mapping properties of CoMFA. On the other hand, comparison of the energy terms selected using the COMBINE analysis with the selected CoMFA fields can be used to partially overcome this difficulty. COMBINE models were found⁴ to be robust to chance correlation by three methods: external blind cross-validation, substitution of the activity vector by random numbers, and random permutation of the activities among the compounds. They thus provide a reliable basis for evaluation of the CoMFA results. Here, it is shown that the models derived with the two methods share common features.

Figure 2 shows the "typical" intermolecular interactions selected and quantified in a COMBINE analysis of HSF-PLA₂ inhibitors (see ref 4). Qualitatively, the important interactions for activity are as follows. Binding affinity is dominated by electrostatic interactions with the calcium ion located in the active site. Several van der Waals interactions then modulate the affinity of the inhibitors. Some of the residues in the B-helix (top left of Figure 2) and the calcium binding loop form a rigid wall sensitive to the conformation of the *sn*-2 chain. As a result of the sp² geometry of the transition-state analogue group, amide-based inhibitors tend to display poorer interactions with this wall compared to sulfonamide-based inhibitors with sp³ geometry, and this explains part of the differences in activity. On the other side of the binding site, Phe-5 and Tyr-52 form a pocket in which the glycerol moiety of the inhibitor fits and an optimal fit into this pocket increases activity.⁴ Finally, Pro-131 forms van der Waals interactions with the benzyl moiety at the end of the *sn*-3 chain of some of the inhibitors, which tend to increase the activity.

Indirect validation of these structure–activity relationships (SARs) comes from independent studies on HSF-PLA₂ inhibitors. Schevitz et al.²⁶ arrive at similar SARs for a completely different set of compounds which are indole-based inhibitors. According to their results, chelating the calcium ion produces a large increase in the ligand affinity. Moreover, they observe strong steric constraints in those portions of the inhibitors interacting with the cleft formed by the B-helix and the calcium binding loop of the enzyme. Wheeler et al.,²⁷ who studied the substrate specificity of short chain phospholipid analogues to HSF-PLA₂, also arrived at similar SARs to those obtained with COMBINE analysis.

Comparison of the selected regions in the most predictive CoMFA models, namely, N-T-C and N-G-C, with each other and with the COMBINE model indicates a rough correspondence of the important regions and that similar interpretations of binding differences are obtained (compare Figures 2 and 4). In both CoMFA models, a zone of electrostatic field points that increase the activity (blue spheres in Figure 4b,d) is selected near

the position of the calcium ion. This region is less evident in the N-T-C model, in which this interaction tends to spread out over grid points in the same sub-box. This dispersion appears to be a general characteristic of Q^2 -GRS-derived models, which tends to make the "pharmacophoric regions" less interpretable and results in poorer correspondence with the interaction points in the receptor. The "forbidden" region corresponding to the position of the B-helix of the receptor is selected in the steric fields of both CoMFA models (blue spheres in Figure 4a,c). However, the glycerol binding pocket is lost in the N-T-C model, although it is clearly observable (yellow spheres) in the N-G-C model. This effect could be an artifact of the box division methodology in the Q^2 -GRS method. There is no clear correspondence between the models in the region around the benzyl portion of the inhibitors. While COMBINE analysis selects Pro-131, N-G-C models select a set of grid points forming an electrostatic ring around the benzyl ring. In some CoMFA studies, this has been interpreted as evidence of π - π interactions with the receptor.²⁵ However, this interpretation is inconsistent with the ligand-receptor complexes in this case. Also, in the N-T-C model, the regions around the benzyl ring of the inhibitor are located outside the receptor structure and lack a clear physical meaning. There are also selected regions that are difficult to explain in structural terms. Thus, in the N-G-C model, a steric field is selected at the position of Gly-32 in the enzyme (leftmost yellow sphere in Figure 4a). The regression models indicate that filling this region with ligand atoms should improve activity. However, this is hard to rationalize in the light of the complexes, as a serious steric clash between the ligand and the receptor would take place in this region. One possible explanation for such physically unreasonable steric field regions is the difficulty of making predictions from CoMFA models for compounds with substituents that occupy space not sampled by the training set. Indeed, it can be expected that a positive steric region in a CoMFA field, where addition of steric bulk is beneficial to activity, is likely to be associated with a negative steric region a little further away, since the reason that adding some steric bulk is favorable is that it is positioned close to receptor atoms. On the other hand, both CoMFA models select a favorable steric interaction with Leu-2 (largest yellow contour in Figure 4a,c), although a directly corresponding interaction is lacking in COMBINE models, despite several intrareceptor terms being selected in the A-helix.⁴ It was shown in the previous section that this interaction could originate from an artifact due to the energy cutoff used in the calculation of the \mathbf{X} -matrix in the CoMFA studies. However, the interaction with Leu-2 has a clear physical meaning in structural terms, as a result of the hydrophobic interactions between Leu-2 and the *sn*-1 chain of the inhibitor. It is interesting that the interaction with residue 2 in PLA₂ enzymes has been advocated as a factor influencing selectivity for indole-based inhibitors.²⁸

Conclusions

While the present study shows that predictive 3D-QSARs can be obtained for a particularly challenging data set (noisy but typical of many data sets studied in medicinal chemistry projects), it highlights some of the difficulties that may arise in achieving sound, reliable

3D-QSAR regression models. In summary, the main findings are as follows:

1. Parameters such as dielectric constant, maximum energy cutoff, grid-spacing, and variable scaling procedure have a strong influence on CoMFA results. In the present case, in which the GRID energy function¹⁷ was used, a uniform dielectric with a relative dielectric constant of 4 seems to be the best suited dielectric model of those tested. This value of the dielectric constant is often assigned to the protein interior¹⁸ in continuum electrostatics calculations, and its use here implies that long range correlations of the electrostatic potential make an important contribution to structure-activity correlations. A grid-spacing of 2 Å shows better performance than a 1 Å spacing, probably as a result of an increased signal-to-noise ratio.⁵ Autoscaling has a negative effect on predictive ability. Block-scaling improves the predictive ability slightly when used together with GOLPE variable selection but has a negative effect when used with the Q^2 -GRS method.

2. Variable selection with either the Q^2 -GRS⁸ or GOLPE⁷ variable selection method only slightly improves the external predictive ability of the models. However, provided that care is taken in order to avoid overfitting, variable selection enables predictive models to be detected and their quality assessed. More specifically, internal and external validation indices like the SDEP parameter (eq 2) tend to converge after variable selection with greater convergence being observed for the Q^2 -GRS method. Importantly, both methods of variable selection tend to select variables in roughly the same regions of space leading to similar physical interpretations. Most of these regions appear plausible in the context of the 3D structure of the receptor. The regions selected are also reasonably consistent with those selected in the COMBINE analysis and those found in independent structure-activity studies by other authors.

3. The Q^2 -GRS variable selection method is very sensitive to scaling procedures and maximum energy cutoff values, while the GOLPE method seems more robust. GOLPE models are also easier to interpret than Q^2 -GRS models. However, the GOLPE method is more prone to producing overfitted models, a risk that is higher if uncertain variables are eliminated.

4. Validation of CoMFA models is always difficult. External validation provides a stringent test, but the results must be analyzed carefully. Any index of predictive ability should be interpreted in relation to the underlying structure of the data to avoid misleading conclusions. In particular, PLS models should be examined with the aid of plots of variables in principal component space.

Acknowledgment. We thank Professor Paul B. Sigler for kindly providing us with the coordinates of the HSF-PLA₂-inhibitor complex, Dr. P. J. Kraulis for the MOLSCRIPT program, and Dr. P. J. Goodford for provision of the GRID program. A.R.O. is the recipient of a predoctoral fellowship from the Comunidad Autónoma de Madrid, Spain. This work was supported in part by the Spanish CICYT (SAF projects 94/0630 and 96/231).

Supporting Information Available: Tables 1-5 and Figures 1-5 giving details of the methods and results as referred to in this paper (8 pages). Ordering information is given on any current masthead page.

References

- (1) (a) Jolles, G.; Wooldridge, K. R. H., Eds. *Drug Design: Fact or Fantasy?* Academic Press: London, 1984. (b) Dean, P. M. *Molecular Foundations of Drug-Receptor Interactions*; Cambridge University Press: Cambridge, 1987. (c) Hansch, C., Sammes, P. G., Taylor, J. B., Eds. *Comprehensive Medicinal Chemistry. Vol. IV: Quantitative Drug Design*; Pergamon Press: New York, 1990.
- (2) Kubinyi, H., Ed. *3D-QSAR in Drug Design. Theory, Methods and Applications*; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993.
- (3) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (4) (a) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by Comparative Binding Energy Analysis: Application to Human Synovial Fluid Phospholipase A₂ Inhibitors. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; J. R. Prous Science Publishers: Barcelona, 1995; pp 439–443. (b) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.
- (5) Cramer, R. D., III. Partial Least Squares (PLS): Its strengths and limitations. *Perspect. Drug Discovery Des.* **1993**, *1*, 269–278.
- (6) (a) Cruciani, G.; Clementi, S.; Baroni, M. Variable selection in PLS analysis. In *3D-QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp 551–564. (b) Wold, S.; Johansson, E.; Cocchi, M. PLS - Partial least-squares projections to latent structures. In *3D-QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp 523–550.
- (7) (a) Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20. (b) GOLPE 3.0, Multivariate Infometric Analysis, Perugia, Italy, 1995.
- (8) Cho, S.-J.; Tropsha, A. Cross-validated R²-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060–1066.
- (9) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589–2601.
- (10) Greco, G.; Novellino, E.; Pellecchia, M.; Silipo, C.; Vittoria, A. Effects of variable selection on CoMFA coefficient contour maps in a set of triazines inhibiting DHFR. *J. Comput. Aid. Mol. Des.* **1994**, *8*, 97–112.
- (11) Davis, A. M.; Gensmentel, N. P.; Johansson, E.; Marriott, D. P. The use of the GRID program in the 3D-QSAR analysis of a series of calcium-channel agonists. *J. Med. Chem.* **1994**, *37*, 963–972.
- (12) (a) Kubinyi, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294. (b) Kubinyi, H. Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. *Quant. Struct. Act. Relat.* **1994**, *13*, 393–401.
- (13) (a) Garcia, M. L. Manuscript in preparation. (b) Cabre, F.; Carabaza, A.; Garcla, A. M.; Gomez, M.; Garcla, L.; Mauleon, D.; Carganico, G. Presentation at the VIIIth International Congress on Prostaglandins and Related Compounds, Montreal, Canada, 1992. (c) Carganico, G.; Mauleon, D.; Garcla, M. L. Spanish Patent WO-94/28004, 1994.
- (14) Wery, J.-P.; Schevitz, R. W.; Clawson, D. K.; Bobbitt, J. L.; Dow, E. R.; Gamboa, G.; Goodson, T.; Hermann, R. B.; Kramer, R. M.; McClure, D. B.; Mihelich, E. D.; Putname, J. E.; Sharp, J. D.; Stark, D. H.; Teater, C.; Warrick, M. W.; Jones, N. D. Structure of a recombinant human rheumatoid arthritic synovial fluid phospholipase A₂ at 2.2 Å resolution. *Nature* **1991**, *352*, 79–82.
- (15) Scott, D. L.; White, S. P.; Browning, J. L.; Rosa, J. J.; Gelb, M. H.; Sigler, P. B. Structures of free and inhibited human secretory phospholipase A₂ from inflammatory exudate. *Science* **1991**, *254*, 1007–1010.
- (16) Pisabarro, M. T.; Ortiz, A. R.; Palomer, A.; Cabre, F.; Garcla, M. L.; Wade, R. C.; Gago, F.; Mauleon, D.; Carganico, G. Rational modification of human synovial fluid phospholipase A₂ inhibitors. *J. Med. Chem.* **1994**, *37*, 337–341.
- (17) (a) GRIN, GRID and GRAB (version 11), 1994, Molecular Discovery Ltd., West Way House, Elms Parade, Oxford OX2 9LL, England. (b) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849–857. (c) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083–1094. (d) Wade, R. C.; Clark, K. J.; Goodford, P. J. Further development of hydrogen-bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 140–147. (e) Wade, R. C.; Goodford, P. J. Further development of hydrogen-bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 148–156.
- (18) Sharp, K. A.; Honig, B. Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301.
- (19) Tripos Associates, 1699 S. Hanley Rd, St. Louis, MO.
- (20) Warshel, A.; Russell, S. T. Calculations of electrostatic interactions in biological systems and in solution. *Quart. Rev. Biophys.* **1984**, *17*, 283–422.
- (21) Hingerty, B. E.; Ritchie, R. H.; Ferrell, T. L.; Turner, J. E. Dielectric effect in biopolymers: the theory of ionic saturation revisited. *Biopolymers* **1985**, *24*, 427–439.
- (22) Friedman, R.; Honig, B. The electrostatic contribution to DNA base-stacking interactions. *Biopolymers* **1992**, *32*, 145.
- (23) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley and Sons: New York, 1978; Chapter 12.
- (24) Mitchell, T. J. An algorithm for the construction of "D-optimal" experimental designs. *Technometrics* **1974**, *16*, 203.
- (25) De Priest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372–5384.
- (26) Schevitz, R. W.; Bach, N. J.; Carlson, N. Y.; Chirgadze, D. K.; Clawson, R. D.; Dillard, S. E.; Draheim, L. W.; Jones, N. D.; Mihelich, E. D.; Olkowski, J. L.; Snyder, D. W.; Sommers, C.; Wery, J. P. Structure-based design of the first potent and selective inhibitor of human non-pancreatic secretory phospholipase A₂. *Nature Struct. Biol.* **1995**, *2*, 458–465.
- (27) Wheeler, T. N.; Blanchard, S. G.; Andrews, R. C.; Fang, F.; Gray-Nunez, Y.; Harris, C. O.; Lambert, M. H.; Mehrotra, M. M.; Parks, D. J.; Ray, J. A.; Smalley, T. L., Jr., Substrate specificity in short-chain analogs at the active site of human synovial fluid phospholipase A₂. *J. Med. Chem.* **1994**, *37*, 4118–4129.
- (28) Verlinde, C. L. M. J.; Dijkstra, B. W. Drug or tool, design or serendipity? *Nature Struct. Biol.* **1995**, *2*, 429–432.
- (29) Kraulis, P. J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **1991**, *24*, 946–950.

JM9601617